

E

Earthquake Clusters over Multi-dimensional Space, Visualization of

DAVID A. YUEN¹, WITOLD DZWINEL²,
YEHUDA BEN-ZION³, BEN KADLEC⁴

¹ Dept. of Geology and Geophysics,
University of Minnesota, Minneapolis, USA

² Dept. of Computer Science, AGH University
of Sci. and Technol., Kraków, Poland

³ Department of Earth Sciences,
University of Southern California, Los Angeles, USA

⁴ Department of Computer Science,
University of Colorado, Boulder, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Earthquakes Clustering](#)

[Multidimensional Feature Space](#)

[The Detection and Visualization of Clusters
in Multi-Dimensional Feature Space](#)

[Description of the Data](#)

[Earthquake Visualization by Using Clustering
in Feature Space](#)

[Remote Problem Solving Environment \(PSE\)
for Analyzing Earthquake Clusters](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

Glossary

Grid Virtual metacomputer, which uses a network of geographically distributed local networks, computers and computational resources and services. *Grid Computing* focuses on distributed computing technologies,

which are not in the traditional dedicated clusters. *Data Grids* – represent controlled sharing and management of large amounts of distributed data.

Problem solving environment (PSE) A specialized computer software for solving one class of problems. They use the language of the respective field and often employ modern graphical user interfaces. The goal is to make the software easy to use for specialists in fields other than computer science. PSEs are available for generic problems like data visualization or large systems of equations and for narrow fields of science or engineering.

Global seismographic network (GSN) The goal of the GSN is to deploy permanent seismic recording stations uniformly over the earth's surface. The GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second, and to provide for high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is also the goal of the GSN to provide for real-time access to its data via Internet or satellite. Over 75% of the over 128 GSN stations meet this goal as of 2003.

WEB-IS A software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet through the interaction between client and server.

Scientific visualization is branch of computer graphics and user interface design that are dealing with presenting data to users, by means of patterns and images. The goal of scientific visualization is to improve understanding of the data being presented.

Interactive visualization is a branch of graphic visualization that studies how humans interact with computers to create graphic illustrations of information and how this process can be made more efficient. **Remote-visualization** – the tools for interactive visualization of high-resolution images on remote client machine, rendered and preprocessed on the server.

OpenGL A standard specification defining a cross-language cross-platform API for writing applications that produce 2D and 3D computer graphics.

Sumatra-Andaman earthquake An undersea earthquake that occurred at 00:58:53 UTC (07:58:53 local time) December 26, 2004, with an epicenter off the west coast of Sumatra, Indonesia. The earthquake triggered a series of devastating tsunamis along the coasts of most landmasses bordering the Indian Ocean, killing large numbers of people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Sri Lanka, India, and Thailand.

Earthquake catalog Data set consisting of earthquake hypocenters, origin times, and magnitudes. Additional information may include phase and amplitude readings, as well as first-motion mechanisms and moment tensors.

Pattern recognition The methods, algorithms and tools to analyze data based on either statistical information or on a priori knowledge extracted from the patterns. The patterns for classification are groups of observations, measurements, objects, defining feature vectors in an appropriate multidimensional feature space.

Data mining Algorithms, tools, methods and systems used in extraction of knowledge hidden in a large amount of data.

Features denoted f_i or F_j (i, j – feature indices) – a set of variables which carry discriminating and characterizing information about the objects under consideration. The features can represent raw measurements (data) f_i or can be generated in a non-linear way from the data F_j (features).

Feature space The multidimensional space in which the F_k vectors are defined. Data and feature vectors represent vectors in respective spaces.

Feature vector A collection of features ordered in some meaningful way into multi-dimensional feature vectors F_l (F_l where l – feature vector index) that represents the signature of the object to be identified represented by the generated features F_l .

Feature extraction The procedure of mapping source feature space into output feature space of lower dimensionality, retaining the minimal value of error cost function.

Multidimensional scaling The nonlinear procedure of feature extraction, which minimizes the value of the “stress” being the function of differences of all the distances between feature vectors in the source space and corresponding distances in the resulting space of lower dimensionality.

Data space The multi-dimensional space in which the data vectors f_k exist.

Data vector A collection of features ordered in some meaningful way into multi-dimensional vectors f_k (f_k, k – data vector index) and $f_k = [m_k, z_k, x_k, t_k]$ where m_k is the magnitude and x_k, z_k, t_k – its epicentral coordinates, depth and the time of occurrence, respectively.

Cluster Isolated set of feature (or data) vectors in data and feature spaces.

Clustering The computational procedure extracting clusters in multidimensional feature spaces.

Agglomerative (hierarchical) clustering algorithm The clustering algorithm in which at the start the feature vectors represent separate clusters and the larger clusters are built-up in a hierarchical way. The procedure repeats the process of gluing-up the closest clusters up to the stage when a desired number of clusters is achieved.

k-Means clustering Non-hierarchical clustering algorithm in which the randomly generated centers of clusters are improved iteratively.

Multi-resolutional clustering analysis Due to clustering a hierarchy of clusters can be obtained. The analysis of the results of clustering in various resolution levels allows for extraction of knowledge hidden in both local (small clusters) and global (large clusters) similarity of multidimensional feature vectors.

N-body solver The algorithm exploiting the concept of time evolution of an ensemble of mutually interacting particles.

Non-hierarchical clustering algorithm The clustering algorithm in which the clusters are searched for by using global optimization algorithms. The most representative algorithms of this type is **k-means** procedure.

Definition of the Subject

Earthquakes have a direct societal relevance because of their tremendous impact on human community [59]. The genesis of earthquakes is an unsolved problem in the earth sciences, because of the still unknown underlying physical mechanisms. Unlike the weather, which can be predicted for several days in advance by numerically integrating non-linear partial differential equations on massively parallel systems, earthquake forecasting remains an elusive goal, because of the lack of direct observations and the fact that the governing equations are still unknown. Instead one must employ statistical approaches (e. g., [61,72,82]) and data-assimilation techniques (e. g., [6,53,81]). The nature of the spatio-temporal evolution of earthquakes has

to be assessed from the observed seismicity and geodetic measurements. Problems of this nature can be analyzed by recognizing non-linear patterns hidden in the vast amount of seemingly unrelated information. With the proliferation of large-scale computations, data mining [77], which is a time-honored and well-understood process, has come into its own for extracting useful patterns from large incoherent data sets found in diverse fields, such as astronomy, medical imaging, combinatorial chemistry, bio-informatics, seismology, remote sensing and stock markets [75]. Recent advances in information technology, high performance computing, and satellite imagery have led to the availability of extremely large data sets, exceeding Terabytes at each turn, that are coming regularly to physical scientists who need to analyze them quickly. These data sets are non-trivial to analyze without the use of new computer science algorithms that find solutions with a minimal computing complexity. With the imminent arrival of petascale computing by 2011 in USA, we can expect some breakthrough results from clustering analysis. Indeed, clustering has become a widely successful approach for revealing features and patterns in the data-mining process. We describe the method of using clustering as a tool for analyzing complex seismic data sets and the visualization techniques necessary for interpreting the results. Petascale computing will also spur visualization techniques, which are sorely needed to understand the vast amounts of data compressed in many different kinds of spaces, with spatial, temporal and other types of dimensions [78]. Examples of clusters abound in nature include stars in galaxies, hubs in airline routes and centers of various human relationships [5]. Clustering comes from multi-scale, nonlinear interactions due to the rock rheology and earthquakes.

Introduction

Earthquake clustering is automatically implicated by the classical Gutenberg–Richter relationship [40], which specifies the frequency of earthquakes between some small magnitude cutoff and a certain large magnitude around 8 [39]. This empirical finding with a broad magnitude range implies that the largest seismic events are surrounded by a large number of smaller events. This clustering may have both spatial and temporal dependences. One of the goals of earthquake clustering studies is to find these special points in a high-dimensional space related to the nature of the dimensional space associated with earthquake dynamics [24,25]. One major goal of this chapter is to introduce the reader to the notion of searching for clustering points in dimensional spaces higher than the

3D physical space we are used to. This concept is crucial to our understanding of the clustering points of earthquakes in these higher-dimensional spaces, which may enable progress in forecasting earthquakes. Information in seismicity data sets can be both relevant and irrelevant from the point of view of deterministic earthquake dynamics. It can be also “entangled” and impossible to be interpreted with normal human perception. The role of data mining is to have a mathematically rigorous algorithm for extracting relevant information from this deluge of data, and make it understandable. Clustering techniques, which are commonly used today in many fields, ranging from biology (e. g., [26]) to astrophysics, allows us to produce specially crafted data models that can be employed for predicting the nature of future events. In more complex cases, these special data models can work in concert with formal mathematical and physical paradigms to give us deeper physical insight.

The concept of clustering has been used for many years in pattern recognition [2,50,78]. The clustering can use more (e. g. [54]) or less mathematically rigorous principles (e. g. [33]). Nowadays clustering and other feature extraction algorithms are recognized as important tools for revealing coherent features in the earth sciences [32,65,66,67], bioinformatics [51] and in data mining [37,43,44,57]. Depending on the data structures and goals of classification, different clustering schemes must be applied [36,55].

In this chapter we emphasize the role of clustering in the understanding of earthquake dynamics and the way to visualize and interpret the computed results from clustering. All the seismic events occurring over a certain region during a given time period can be viewed as a single cluster of correlated events. The strength of mutual correlations between events, such as correlations in spatial and time positions along with magnitude, cause this single cluster to have very complex internal structure. The correlations – the measures of similarity between events – divide the global cluster into variety of small clusters of multi-scale nature, i. e., small clusters may consist of a cascade of smaller ones. Coming down the scale we record clusters of more and more tightly correlated events. Exploring the nature of events belonging to a single cluster, we can extract common features they possess. Having more information about events belonging to the same cluster we can derive hidden dependences between them. Moreover, we can anticipate the type of an unknown event belonging to a certain cluster from the character of the other events of this cluster.

In the following sections we describe the idea of clustering and the new idea of higher dimensions associated with data sets. We also demonstrate the results of cluster-

ing analysis of both synthetic and real data. Long synthetic data were derived by using a model for a segmented strike-slip fault zone in a 3D elastic half-space [7]. The real data represent short time (5 years interval) seismic activities of the Changbaishan volcano (the north-east frontier of the North China craton) and the Japanese Archipelago. Lastly, we also highlight the role of visualization of clusters as an important tool for understanding this type of new data arrangement, and we describe the role played by remote visualization environment specially devised for visualization of earthquake clusters.

Earthquakes Clustering

Statistical Laws as Elementary Building Bricks of Earthquake Models

The earthquake prediction problem is of fundamental importance to society and also geosciences. Progress in this field is hampered, mainly because many important dynamic variables – such as stress – are not accessible for direct observations. Moreover, instrumental observations of seismicity are possible only for a fraction of a single large earthquake cycle. Overcoming these difficulties will require combining analyzes of model and observed data by using knowledge extraction instruments. The fundamental process of knowledge extraction is finding dependences between data and/or between model parameters. They can be revealed as patterns (clusters) in time, spatial and feature (parameter) space domains. The most elementary dependences can be expressed in the form of semi-empirical functional laws.

There are a few basic statistical laws which represent the basis for earthquake models development. The frequency-size statistics of regular tectonic earthquakes (excluding swarms and deep focus earthquakes) follow the Gutenberg–Richter relation [39,80,84]:

$$\log N(M) = a - bM \quad (1)$$

where N is the number of events with magnitude larger than M and a , b are constants giving, respectively, the overall seismicity rate and relative rates of events in different magnitude ranges. Observed b -values of regional seismicity typically fall in the range 0.7–1.3.

Aftershock decay rates are usually be described by the Omori–Utsu law [71,79]:

$$\Delta N / \Delta t = K(t + c)^{-p} \quad (2)$$

where N is the cumulative number of events, t is the time after the mainshock, and K , c , and p are empirical constants. The epidemic-type aftershock-sequences (ETAS)

model combines the Omori–Utsu law with the Gutenberg–Richter frequency-magnitude relation for a history-dependent occurrence rate of a point process in the form (e. g., [61])

$$\lambda(t|H_t) = \mu + \sum_{t_i < t} \frac{K_0 \exp[\alpha(M_i - M_c)]}{(t - t_i + c)^p} \quad (3)$$

where α is a constant background rate, M_i is the magnitude of earthquake at time t_i , M_c is a lower magnitude cut-off, H_t denotes the history, and the productivity factor $K_0 \exp[\alpha(M_i - M_c)]$ gives the number of events triggered by a parent earthquake with magnitude M_i . The ETAS model is used widely in analysis of seismic data, owing to its built-in clustering associated with the incorporation of the Gutenberg–Richter and Omori–Utsu laws. Examples of recent applications can be found in [45,62,68].

These results can be used to derive additional properties such as average recurrence times (e. g., [4,18,19,20,68,89]). It is usually defined as the number of years between occurrences of an earthquake of a given magnitude in a particular area. For example, the probability of a devastating earthquake striking the greater San Francisco Bay Region over the following 25 years (2007–2031) is 0.62 [68]. Corral [18,19,20] proposed the existence of a universal scaling law for the probability density function $H(\tau)$ of recurrence times (or interevent times) τ between earthquakes in a given region:

$$H(\tau) \cong \lambda \times f(\lambda\tau) \quad (4)$$

The function $f(x)$ appears to be similar for many different seismic regions, which suggests some universal properties. The average rate λ represent the region specific constant, whose reciprocal is the only relevant characteristic time for the recurrence times. Molchan [58] showed that under general conditions, the only universal distribution of inter-event times in a stationary point process is exponential. Hainzl et al. [42] and Saichev and Sornette [68] discussed relations between statistics of interevent times, the ETAS model of triggered seismicity, and the Corral [18,19] distribution of Eq. (4).

In the context of earthquake prediction it is important to analyze earthquake cycles with repeating sequences of events such as foreshocks, mainshocks and aftershocks (e. g., [9,74,80]). Apart from qualitative tendencies reflected by statistical laws, the earthquakes exhibit various types of more subtle spatio-temporal clustering, i. e., grouping of events of the same type both in time and in spatial coordinates. The recognition of these patterns followed by the analysis of the reasons of their appearance may lead to the development of improved prediction algorithms.

In the following section we present a closer look of clustering as a knowledge extraction technique and a possible way of its application to earthquake data analysis.

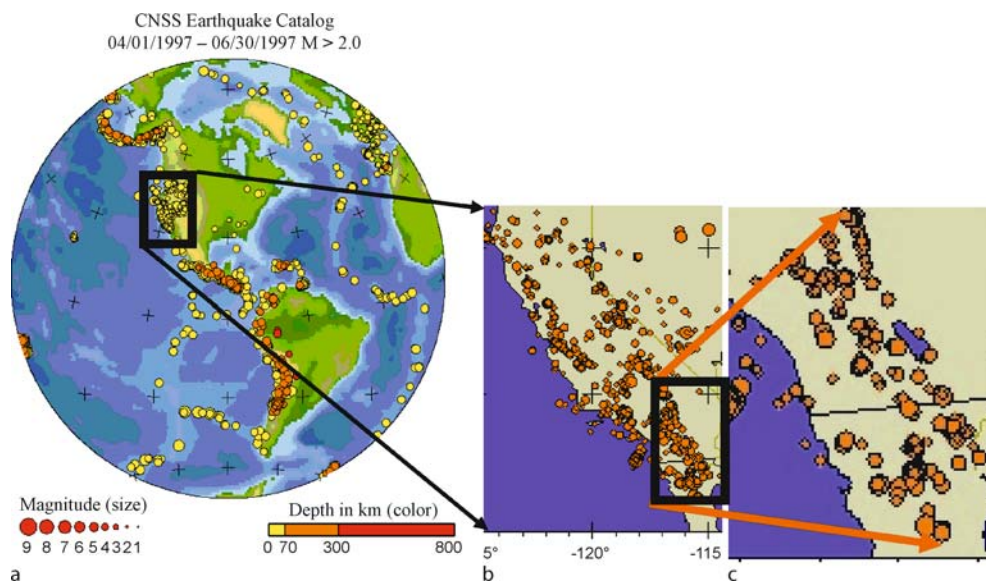
Basic Concepts of Clustering

Clustering analysis is a mathematical concept whose main useful role is to extract the most similar (or dissimilar) separated sets of objects according to a given similarity (or dissimilarity) measure [2]. Clustering is one of the most fundamental processes generated by nature. For example, people gathering in groups, tribes, demonstrations, parties, cities, produce clusters. Similarly, towns and cities are clusters of buildings while galaxies are clusters of stars. The local computer networks and bacterial colonies are also clusters. The objects forming clusters can be the clusters of smaller objects, which in turn, are clusters of even smaller and smaller building bricks. The complexity of cluster structure reflects the complexity of the real world. The clusters of various shapes, densities and sizes, with additional attributes as colors, transparency etc. built up patterns, which are the fingerprints of all multi-scale processes and phenomena. The clusters are the primitives of the patterns.

The same notion of clustering concerns geographical locations and other properties of earthquakes. In Fig. 1a we present a spatial distribution of earthquake

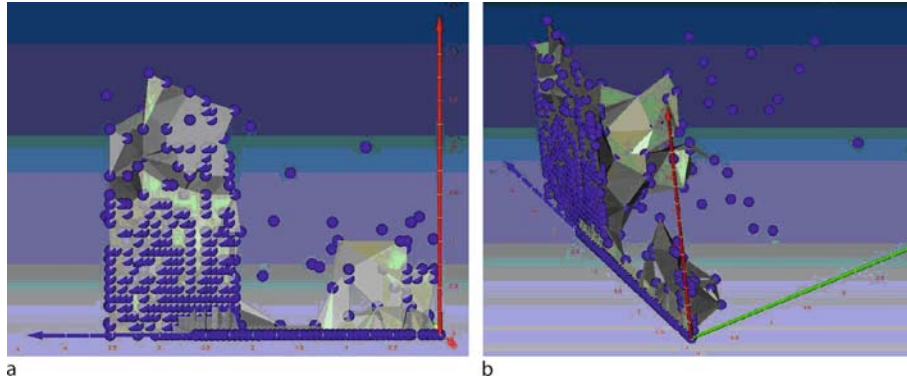
epicenters in the western hemisphere of the Earth (data from <http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). One can see with the naked eye that their distribution is far from being uniform. We observe both elongated and oblate structures – the earthquake clusters – separated at this resolution by large holes of seismically quiescent area.

Properties of the clusters result from properties of the generating processes. The shape and structure of clusters are visual representation of information on these processes. Therefore, detection of clusters and their analysis is the first step for knowledge extraction from this information. For example, as shown in Fig. 1a, the earthquake clusters on Earth are located in geologically active regions, mainly, on the edges of colliding tectonic plates. The distribution and shape of the earthquake clusters follow the borders between the plates. In Fig. 1b we show the large earthquake cluster from Fig. 1a located at the US western coast. One can distinguish here many smaller clusters of different density separated by geologically inactive area. A similar pattern (see Fig. 1c) is observed by zooming-in one of denser clusters from Fig. 1b. This multi-resolutional and self-similar system is characteristic for many critical phenomena ► [Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of \[3,9,16,74\]](#). The worldwide fault network has a fractal structure (or multifractal) [22,27,79]. Wavelet-based multi-fractal anal-



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 1

Multiscale character of the earthquake clusters. The epicenters of earthquakes of various depth and magnitude are displayed. The data come from the CNSS Earthquake Catalog (<http://quake.geo.berkeley.edu/cnss/maps/cnss-map.html>). a the western hemisphere, b the US western coast c California and Nevada



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 2

Seismic activity of the Changbaishan volcano during 5 years time span from 07.1999 to 05.2004 (the north-east frontier of the North China craton) [47]. The plates represent the seismic events in 3-D feature space attributed by eruption time (*blue axis*), magnitude (*red axis*) and distance to the epicenter (*green axis*) coordinates. The clusters are rendered using the *wrap point* technique (the Amira visualization package www.amiravis.com). Two different positions of coordinates are shown. The *large cluster* representing the earthquake swarm is preceded by the *small precursory cluster* of seismic activity and quiescent time period

ysis [27] shows clearly several distinct scaling domains in earthquake catalogs revealing rich self-similar multi-scale structure. However, the spatial structure of earthquake clusters alone is inadequate to formulate plausible hypotheses about earthquake dynamics. More information is required.

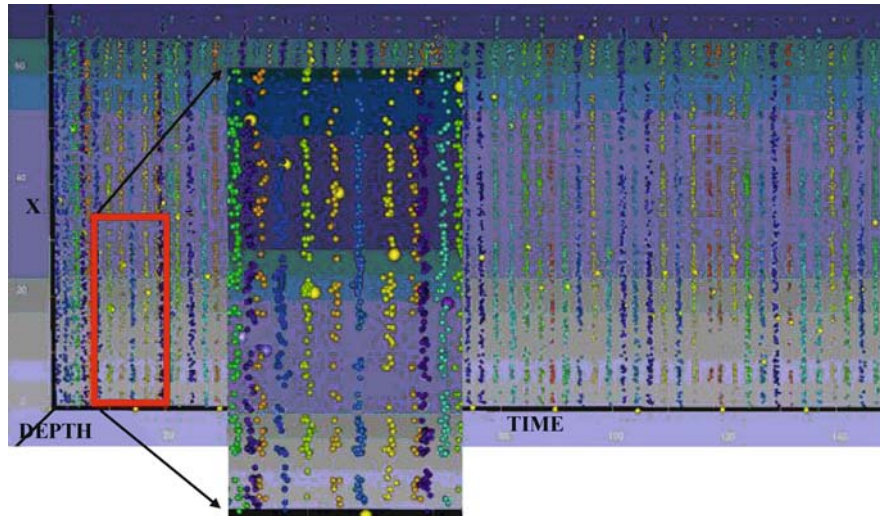
As shown in Fig. 1, besides the geographical location, earthquakes have additional features such as the time and depth of occurrence and the amount of energy released (proportional to $10^{\alpha m}$ with $\alpha \sim 1.5$ and m the magnitude). These attributes can be used as additional coordinates of, so called, feature space (e.g., [78]). In Fig. 2 we display the earthquake clusters representing the seismic activity nearby the Changbaishan volcano in an abstract 3-D feature space. Apart from geographical location – represented by the distance from the epicenter – other coordinates (features) are employed: the time of occurrence and the magnitude of the earthquake. As shown in Fig. 2, the large cluster of seismic activity is preceded by the small precursory cluster and low activity region. The larger cluster is characterized by the seismic events from broader interval of magnitudes and with satellite earthquakes more distant from the epicenter than in the preceding smaller cluster.

The dynamics of the volcanic earthquakes covers only a period of 5 years. The time is too short to conclude about the long-time earthquake dynamics. To obtain data covering much longer time period we used synthetic data generated by numerical simulations of seismicity on a heterogeneous fault governed by 3-D elastic dislocation theory, power-law creep and boundary conditions corresponding to the central San Andreas Fault [7,28,29]. In Fig. 3 we

represent seismic activity during 150 years. This period contains $M_f \sim 1 - 3 \times 10^4$ events (represented in Fig. 3 by colored dots) in the magnitude interval [3.3–6.8]. Unlike in the Changbaishan case, the seismic events have one more feature – the earthquake depth. Thus the feature space has now four dimensions. In Fig. 3 we display the data distribution in time-depth-position 3-D space. The fourth dimension – the magnitude – is displayed in Fig. 3 by the size of dot. To make the situation clearer only the large earthquakes with magnitudes $m > 6$ (large dots) and the smallest ones $m < 4$ (small dots) are distinguished in Fig. 3. As shown in Fig. 3 and in [24], the synthetic seismic events with magnitudes $m < 4$ produce stripe-like clusters in the data space. They precede large earthquakes ($m > 6$) and are separated in time by the regions of mixed type of events (i. e., with $4 < m < 5$).

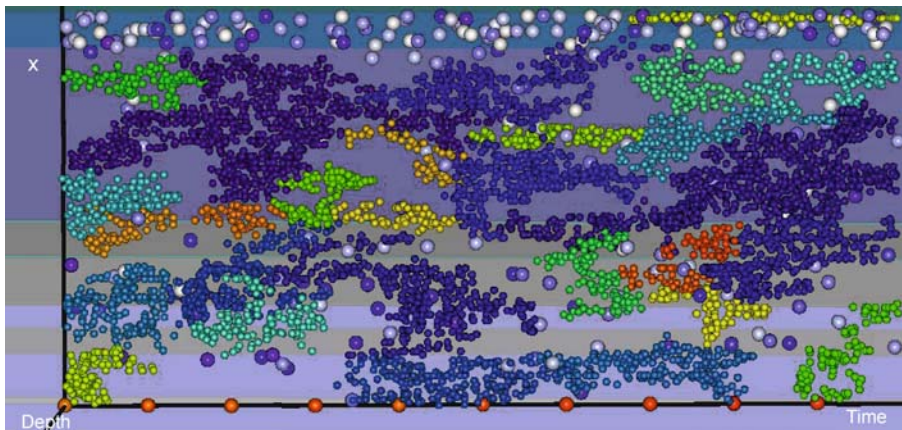
Another system of earthquake clusters are shown in Fig. 4. The synthetic data ($M_f \sim 10^5$ events) corresponding to the seismic activity during 1500 years were generated by the same model [7] for similar geological and boundary conditions. Only medium size events with $4.5 < m < 6$ were taken for clustering. In addition to the local strip like clusters of smaller events ($m < 4$) detected for 150-years data, one can observe in Fig. 4a distinct spatio-temporal patchwork structure of clusters of medium sized events ($4.5 < m < 6$). These clusters follow spatio-temporal changes in strength-stress properties of the fault in the region simulated.

In summary, we can highlight very fundamental properties of earthquakes, multi-resolutional clusters are built up by the earthquake epicenters. The clustering is a dynamical process involving many spatio-temporal scales.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 3

The plot reconstructing seismic activity during 150 years from synthetic data [7] (horizontal distance – X, depth – z; visualized by using the Amira visualization package [1]). Large events (with magnitude $m > 6$) are shown as distinctly larger dots on the background of the lowest magnitude events ($m < 4$). There are visualized patches of low magnitude events preceding larger events [24]. The separate clusters are marked in colors



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 4

The plot reconstructing seismic activity during 1500 years from synthetic data [7]. The largest clusters obtained for events with magnitudes $4.5 < m < 6$. Large events ($m > 6$) are shown as distinctly larger plates. The separate clusters are marked in colors

The dynamic nature of earthquake clusters in a very long time horizon is obvious because everyone can expect that tectonic plates will change dynamically the geo-mechanical properties of the Earth crust. In a long time period covering thousands years, the patterns from Fig. 1 will evolve following the changes in the fault network. More mysterious is the character of earthquake dynamics in spatio-temporal scales allowing for making realistic predictions. We show that in the medium-time period lasting more than a hundred of years the seismic events

may produce periodic system of clusters in approximately equal time intervals with increasing and decreasing seismic activity. The large earthquakes, preceded by the quiescent time periods, appear. The short-time dynamics reveal additionally, that the earthquake swarms are signaled by the smaller precursory cluster of seismic activity. The earthquake attributes such as the magnitude, and the epicenter depth, allow for better interpretation of emerging clusters and exploration of hypotheses space. Therefore, for studying various aspects of earthquake dynamics, in-

cluding their prediction, we have to analyze the cluster structures in multi-dimensional feature space, to be sure that none of important information will be lost or neglected.

Multidimensional Feature Space

In Fig. 1 every point i representing one out of M_f earthquakes has two dimensions – the geographical coordinates $\mathbf{x}_i = [x_1, x_2]$ of the epicenter. The point can be treated as 2-D vector \mathbf{f}_i in the feature space where $\mathbf{f}_i = \mathbf{x}_i$. Assuming additional coordinates, at the highest level of resolution, a single seismic event i can be represented as a five-dimensional data vector $\mathbf{f}_i = [m_i, z_i, \mathbf{x}_i, t_i]$ where m_i is the magnitude and \mathbf{x}_i, z_i, t_i – its epicentral coordinates, depth and the time of occurrence, respectively. The spatio-temporal clusters can be extracted by 3-D visualization similar as those of Figs. 3, 4 distinguishing extra dimension by the size of dots and colors. Only clusters in the three spatially visualized dimensions can be extracted, while the other attributes associated with the earthquake characteristics are used for discriminating among the different types of clusters.

As shown in Figs. 3, 4 and in [24], at the lowest resolution level we can analyze the data locally by looking for clusters with similar events. However, considering a single event on a given area as a feature vector [78] cannot be a good approach from a generalization point of view. The number of events is usually large. There are many noisy background events, which destroy the relevant clusters or produce artificial ones. Moreover, the clustering of raw data neglects the important statistical information, which concerns the entire inspected area. An alternative approach exists in which the entire seismic area can be described as a multidimensional feature vector evolving in time. In the following these features will represent descriptors a_k (seismicity parameters) corresponding to different statistical properties of all the events measured in a given time interval. The number of descriptors N defines the dimensionality of the feature vector $\mathbf{F}_i = [a_1, a_2, \dots, a_N]$, $i = 1, 2, \dots, M$. The vector represents not a single seismic event but it corresponds to seismic situation on the whole controlled area in the subsequent time interval indexed by i . The number of feature vectors M is equal to the number of time intervals in which the descriptors are computed. The index i is a discrete equivalent of time. We expect that the features vectors representing different moments of time also have the tendency to produce clusters in the abstract N -dimensional feature space. Monitoring changes of these time-series in abstract N -dimensional space may be used as a proxy for

the evolution of stress and a large earthquake cycle on a heterogeneous fault [9].

To explain this approach better, let us assume that we have to analyze the client behaviors in a hypermarket. We can watch every client separately assuming that it can be defined as a feature vector consisting of only two coordinates: the time he entered the shop, money spent. Then we can try to find clusters emerging with time during a shopping day. This cannot be easy due to both a large number of feature vectors (clients) producing statistical noise and lack of correlations between them. Another approach consists in treating as a feature vector not a single client but every subsequent time interval $t_i = i \times \Delta T$ ($i = 0, 1, 2, \dots, M_F$; $t_i < t_e$; $M_F = (t_e - t_b)/\Delta T$ and t_b – beginning of the working day and t_e – closing time). Let the coordinates of the subsequent feature vector define the following descriptors averaged in Δt : the number of people inside the shop (crowding), the flow, items bought, money spent per person. We note that now the number of feature vectors will be substantially smaller than in the previous approach but the dimensionality of feature space is larger. Let us assume that as a result of clustering we extract two distinct clusters. The first one consists of feature vectors (time intervals) from between 10.00–11.00 and 13.00–14.00. The cluster is characterized by very small values of the first three descriptors (crowding, flow, numbers of items sold) and relatively large expenses. The second cluster consists of time intervals from between 8.00–9.00 and 16.00–18.00 with all descriptors large. We could conclude that the first cluster consists of time intervals from shopping hours that are the favorite for wealthy retired people from the rich village in the neighborhood, while the second cluster is associated with the rush in shopping just before before the beginning and after the end of working hours.

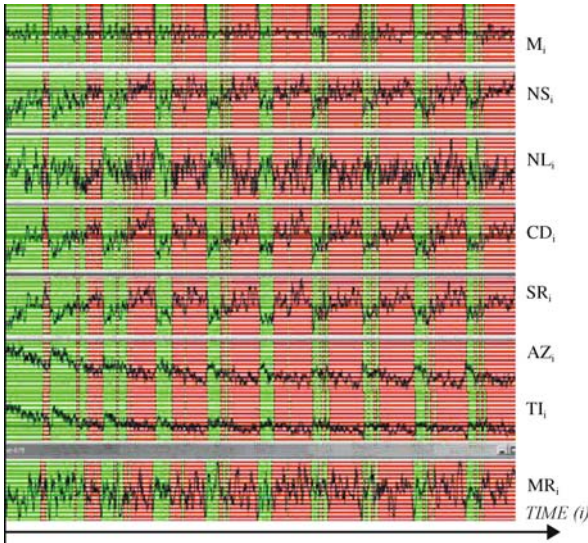
In the same way, the clusters of feature vectors (time intervals) consisting of seismicity parameters should reflect the similarity between seismic activities in various time intervals. As we show before the large seismic events are preceded by precursory events, reflected by an abnormal seismic activity in the whole area. We suppose that these moments of time are similar in the context of the set of seismicity parameters selected. Thus the feature vectors corresponding to precursory events should belong to the same cluster. The idea of predictive system based on clustering consists in detecting the clusters of former foreshocks and signal if the current feature vector – which represents current seismic situation over the area – is or is not the member of this clusters.

The seismicity parameters are computed as time and space averages in a given time and space intervals within

Earthquake Clusters over Multi-dimensional Space, Visualization of, Table 1

Definition of seismicity parameters

NS	Degree of spatial non-randomness at short distances. The differences between distributions of event distances and distances between randomly distributed points.
NL	Degree of spatial non-randomness at short distances.
CD	Spatial correlation dimension calculated on the basis of correlation integrals and on interevent distances.
SR	Degree of spatial repetitiveness represents the tendency of events with similar magnitudes to have nearly the same locations of hypocenters.
AZ	Average depth of the earthquake occurrence.
TI	Inverse of seismicity rate – time interval in which a given (constant) number of events occurs.
MR	Ratio of the numbers of events falling into two different magnitude ranges = $M_f(m \geq M_0)/M_f(m < M_0)$.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 5

The exemplary set of seismicity parameters $\{M, NS, NL, CD, SR, AZ, TI, MR\}$ in time (i – subsequent number of the feature vector) for a file from the 1500-years synthetic data catalog (from [25]). The green and red strips show the time moments belonging to the two different clusters. The green cluster corresponds to the time intervals of lower while the red cluster of higher seismic activities. The time series represent about $M_f = 10^3$ feature vectors F_i .

a sliding time window with a length ΔT and time step dt . The values of a_k represents one of the following seismicity parameters: NS, NL, CD, SR, AZ, TI, MR. The value of dt was assumed to be equal to the average time difference between two recorded consecutive events while ΔT is equal to about 1/10 of the average time distance between

two successive large events ($m > 6$ or $m > 5$). By increasing the values of dt and ΔT one can obtain smoother time series due to better statistics. On the other hand, poorer prediction characteristics can be expected then. We define the seismicity parameters as shown in Table 1 [28,29].

The seismicity parameters produces seven time series and create the abstract 7-dimensional feature space of time events $F_i = (NS_i, NL_i, CD_i, SR_i, AZ_i, TI_i, MR_i)$ where i are discretized values of time $t = t_b + i\Delta T$. In Fig. 5 we display an example set of seismicity parameters (with average magnitude M) for synthetic data [25]. The precise location of the clusters and the visualization of the clustering results are significant challenges in clustering over multi-dimensional space. In the following section we present briefly the basics of clustering and algorithms needed in this venture.

The Detection and Visualization of Clusters in Multi-Dimensional Feature Space

Our main challenge is to devise a clustering scheme which can divide the M feature vectors x_i $i = 1, 2, \dots, M$ into k separate groups (clusters). More formally, assuming that $X = \{x_i\}_{i=1, \dots, M}$ and $x \in \mathbb{R}^N$; $x_i = \{x_{i1}, x_{i2}, \dots, x_{iN}\}$ we define as an k -clustering of X , i. e., the partition of X into k clusters C_1, \dots, C_k provided three conditions are met:

- $C_i \neq \emptyset$, $i = 1, \dots, k$ – the clusters are non empty sets,
- $\cup_{i=1, \dots, k} C_i = X$ – the sum of elements inside clusters is equal to the total number of feature vectors,
- $C_i \cap C_j = \emptyset$, $i \neq j$, $j = 1, \dots, k$ – each feature vector belongs to only one cluster.

The computational problem with clustering is that the number of possible clustering of M vectors into k groups is given by the Stirling numbers (very large numbers) of the second kind:

$$S(M, k) = \frac{1}{k!} \sum_{i=1}^k (-1)^{k-i} \binom{k}{i} \cdot i^M. \quad (5)$$

Some values of $S(N, k)$ are: $S(15, 3) \approx 2 \times 10^6$, $S(20, 4) \approx 45 \times 10^9$, $S(25, 8) \approx 7 \times 10^{17}$, $S(100, 5) \approx 2 \times 10^{68}$. Knowing that the value of N in typical clustering problems can be 10^2 to 10^9 and more we see that the clustering problem is intrinsically hard and exhaustive search – looking through all possible clusterings – cannot be considered. The special clustering schemes based on the proximity measures between feature vectors have to be exploited. The basic steps must be followed in order to develop a clustering task are the following:

1. **Feature selection** – Features must be properly selected to encode as much information as possible. Parsimony

and minimum redundancy among the features is a major goal.

2. *Proximity measure* – This is the measure how “similar” (or “dissimilar”) two features vectors are.
3. *Clustering criterion*, which depends on the interpretation of the term “sensible”, depending on the type of clusters expected in the data set e. g., oblate, elongated, “bridged”, circular etc.
4. *Clustering algorithms*. Choose a specific algorithmic scheme that unravels the clustering structure of the data set.
5. *Validation and interpretation of results* are the final processes of clustering.

There are two principal types of clustering algorithms: non-hierarchical and agglomerative schemes [2,50,78].

Clustering Techniques

The non-hierarchical clustering algorithms are used mainly for extracting compact clusters by using global knowledge about the data structure. The well known *k-means* based schemes [78], consist in finding the global minimum of the following goal function:

$$J(w, z) = \sum_j \sum_{i \in C_j} |x_i - z_j|^2, \quad (6)$$

where: z_j is the position of the center of mass of the cluster j , while x_i are the feature vectors closest to z_j . To find a global minimum of function $J()$, one repeats many times the clustering procedures for different initial conditions [48]. Each new initial configuration is constructed in a special way from the previous results by using the methods from [48,87]. The cluster structure with the lowest $J(w, z)$ minimum is selected.

Agglomerative clustering schemes consist in the subsequent merging of smaller clusters into the larger clusters, basing on proximity and clustering criteria. Depending on the definition of these criteria, there exist many agglomerative schemes such as: average link, complete link, centroid, median, minimum variance and nearest neighbor algorithm. The hierarchical schemes are very fast for extracting localized clusters with non-spherical shapes. The proper choice of proximity and clustering criteria depend on many aspects such as dimensionality of data. For example, a smart clustering criterion based on linked-list scheme for finding neighbors used for molecules clustering is completely worthless for clustering N -dimensional data for which it has extremely high computational complexity. All of agglomerative algorithms suffer from the problem of not having properly defined control parameters, which can be matched for the data of interest and

hence can be regarded as invariants for other similar data structures.

Majority of the classical clustering algorithms require knowledge on the number of clusters. However this number is usually unknown a priori. Furthermore, these methods do not perform well in the presence of heavy noise or outliers. Recently, new methods have been proposed that can: deal with noisy data, discover non-spherical clusters and allow for automatic assessment of number of clusters. Some important examples are the Chameleon [55], DBSCAN [70] and CURE [38] algorithms. Unfortunately, these methods are suited only for low dimensional data and are rather inefficient limiting their use for data mining of large-scale sets. For clustering of large data sets of multidimensional data other approaches are in great demand. In the innovative work by Frey and Dueck [33] the authors use the concept of “affinity propagation,” which takes as input measures of similarity between pairs of data points. Real-valued messages are exchanged between data points until a high-quality set of exemplars and corresponding clusters gradually emerges. Affinity propagation promises to find clusters with much lower error than other methods, and it can do this in less than one-hundredth the amount of time.

Clustering schemes do not produce univocal results. For low dimensional 2-3-D spaces human eye can decide whether the clustering result is optimal or not. However, it becomes hopeless for higher dimensions. There exist many techniques for visualization multidimensional clusters. One of them is the multi-dimensional scaling (MDS) (see overview of mapping techniques in [73]) – the most powerful non-linear mapping technique. This method allows for visualization of the multidimensional data in 2-D or 3-D and for interactive extraction of clusters.

Multidimensional Scaling

Multi-dimensional scaling or MDS is mathematically a non-linear transformation of N -dimensional data onto n -dimensional space, where $n \geq N$ [23,73,78]. The MDS algorithm bases on the “stress function” criterion. The goal is to maintain all the distances between points $R_i \in \omega \subset \mathbb{R}^N$ in the Euclidean 3-D (or 2-D) space with a minimum error. The “stress function” can be written as follows:

$$E(\omega, \omega') = \sum_{j < i} s_{i,j}^{w,mi} \cdot (s_{i,j} - s'_{i,j})^{mi} = \min$$

$$\text{where: } s'_{i,j} = (y_i - y_j) \cdot (y_i - y_j), i, j = 1, \dots, M, \quad (7)$$

and $D_{i,j}$ – is a squared distance between points \mathbf{R}_i , $\mathbf{R}_j \in \omega \subset \mathbb{R}^N$ and $\mathbf{r}_i, \mathbf{r}_j \in \omega' \subset \mathbb{E}^3$ – coordinates of the respective points in 3-D Euclidean space. The values of w and m_i are the parameters of transformation.

The result of mapping depends on the quality of the minimum obtained for the “stress function”. Usually the dimensionality of the “stress function” domain is very high and is equal to $N \cdot M$, i.e., thousands, in the smallest and billions in large problems. For more than $M = 10^3$ feature vectors, the high dimensionality of source space and data complexity may cause the resulting low dimensional patterns to be completely illegible. The application of standard numerical algorithms for finding global minimum of this multimodal, non-linear and complex criterion becomes hopeless. Therefore, for visualization of $M > 10^3$ multidimensional data samples, more reliable minimization techniques extracting global minimum of the “stress function” are required. In [23] we proposed N-body solver by ODE’s as a heuristic means. The algorithm is as follows:

1. The initial configuration of M interacting “particles” is generated in \mathbb{E}^3 ,
2. Every “particle” corresponds to the respective N -dimensional point from \mathbb{R}^N ,
3. The “particles” interact with each other with $\Phi_{i,j}$ particle-particle potential:

$$V_{i,j} = \frac{1}{4} \cdot k \left(r_{i,j}^2 - a_{i,j}^2 \right)^2 \quad (8)$$

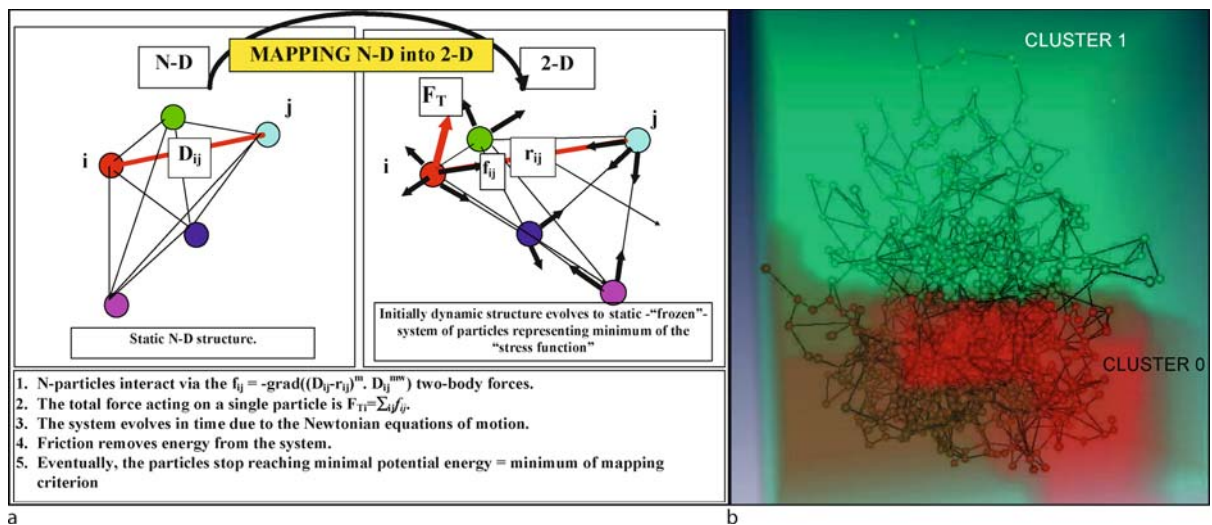
(k – is the stiffness factor) and the energy produced is dissipated by the friction force proportional to the velocity of the particles.

4. The system of particles evolves according to the Newtonian equations of motion.

In this way the interactions between each pair of particles are described by various spring like potentials, dependent on the separation distance between particles r_{ij} and the distance D_{ij} between respective multidimensional points in \mathbb{R}^N . If the distance between particles i and j in the output 2(3)-D space is smaller than the distance between respective i and j feature vectors in the source N -D space these points repel one another. Otherwise, i.e. the distance is larger, the particles attract one another. By using the *leap-frog* numerical scheme for time-integration [41] the following formula for velocities and positions of “particles” can be derived from the Newtonian equations:

$$\begin{aligned} \mathbf{v}_i^{n+1/2} &= \frac{(1-\varphi)}{(1+\varphi)} \cdot \mathbf{v}_i^{n-1/2} \\ &+ \frac{\alpha \Delta t}{(1+\varphi)} \left\{ \sum_{j=1}^K (r_{i,j}^n - a_{i,j}^2) \mathbf{r}_{i,j}^n + \frac{g}{\alpha} \mathbf{i}_z \right\} \quad (9) \\ \mathbf{r}_i^{n+1} &= \mathbf{r}_i^n + \mathbf{v}_i^{n+1/2} \cdot \Delta t \\ \alpha &= \frac{k}{m}, \quad \varphi = \frac{\lambda}{2m} \cdot \Delta t, \end{aligned}$$

where \mathbf{v}_i^n , — the particle i , n – the time-step number, $m = 1$ – particle mass.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 6

a The conceptual diagram of MDS transformation, b The clusters from Fig. 6 mapped by using multidimensional scaling into 3D space for synthetic seismic data catalog A covering 1500 years

As it is common in molecular dynamics [41], the system of “particles” evolves in time until the global (or close to the global) minimum of Eq. (8) (the total potential energy of the particle system) is gained. Two free parameters, λ and k , have to be fit to obtain the stable state, where the final positions of frozen “particles” reflect the result of N -D to 3-D mapping. The conceptual scheme of MDS exploiting N -body solver is shown in Fig. 6A. In Fig. 6B we present the feature vectors shown in Fig. 5, using the 7-dimensional feature space which has been transformed by using the MDS procedure and mapping onto the 3-D space. Take a look on the movies (Movie 1 and 2 in Supplementary Materials), which shows how rotation in the 3-D space can help in cluster recognition.

Description of the Data

Natural Datasets

We analyze the observed and synthetic earthquake catalogs for three time intervals of 5, 150 and 1500 years re-

spectively. The observed data (Fig. 7) represents seismic activities of the Japanese islands collected by the Japan Meteorological Agency (JMA).

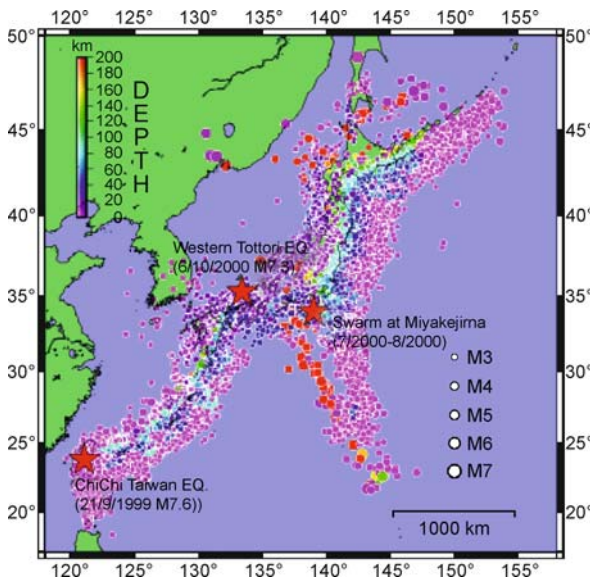
The JMA Catalog consists of 915,829 events detected in Japan Islands between 1923 and January 31, 2003. The original catalog includes also events with magnitudes less than 1.0. The lowest magnitudes were determined by using a detection level, estimated from the Gutenberg–Richter frequency-size distribution. We have assumed that the cutoff magnitude of earthquake is equal to 3 ($m > 3$). We do not use any cutoff depth of hypocenter events. The seismic events shown in Fig. 7, were recorded during the 5 years time interval from October 1, 1997 to January 31, 2003. The data set processed consists of $M = 42\,370$ seismic events with magnitudes m , position in space (latitude X , longitude Y , depth z) and occurrence time t .

To analyze the seismic activity in longer time periods, we use data from synthetic catalogs generated by numerical earthquake models [7].

Physical Model of Earthquake Dynamics

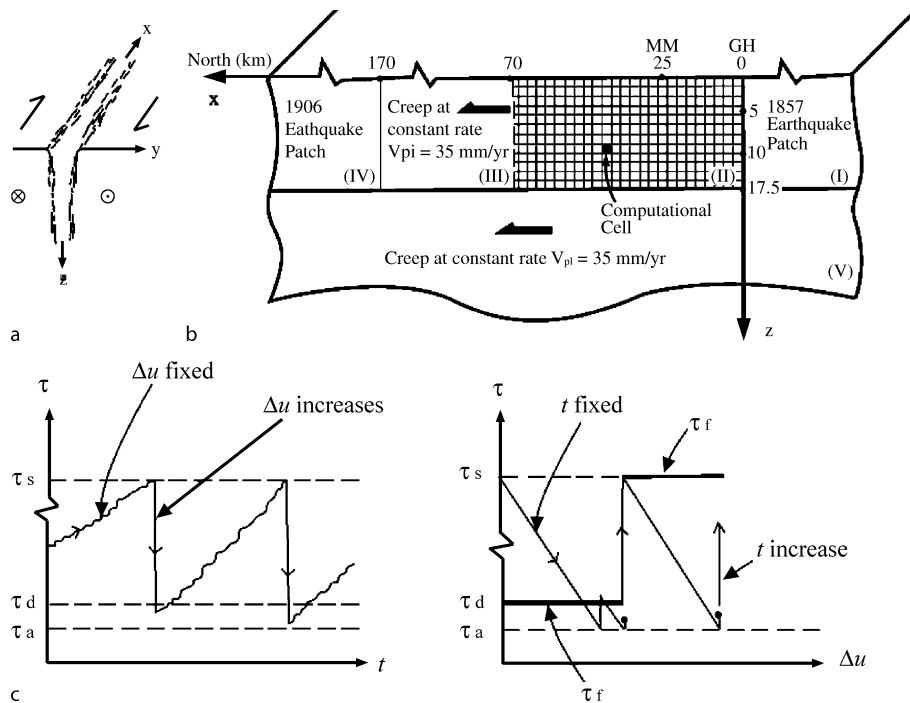
The synthetic catalogs are generated by the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space, based on earlier developments of Ben-Zion and Rice [10,11]. The model attempts to account for statistical properties of earthquake ruptures on long and narrow fault zones with bends, offsets, etc (Fig. 8a), represented by a cellular structure in a 2D plane with discrete cells and spatial variations of frictional parameters (Fig. 8b). The model contains a computational grid (region II of Fig. 8b) where evolving stress and seismicity are generated in response to ongoing loading imposed as slip boundary conditions on the other fault regions. Regions III and V creep at constant plate velocity of 35 mm/yr, while regions I and IV follow staircase slip histories with recurrence times of 150 yr. The stress transfer due to the imposed boundary conditions and failing grid cells is calculated by using a discretized form of a boundary integral equation and employing the static solution for dislocations in a 3D elastic half-space [10,63].

Deformation at each computational cell is the sum of slip contributions from brittle and creep processes. The brittle process (Fig. 8c) is governed by distributions of static friction τ_s , dynamic friction τ_d , and arrest stress τ_a . The static friction characterizes the brittle strength of a cell until its initial failure in a given model earthquake. When stress τ at a cell reaches the static friction, the strength drops to the dynamic friction for the remaining duration of the event. The stress at a failing cell drops to the arrest level τ_a , which may be lower than τ_d to ac-



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 7

Seismic activities around the Japanese Archipelago with a time period of 5 years. We use the hypocentral data provided by the Japan Meteorological Agency (JMA). The magnitude of the earthquakes (JMA magnitude) and their depths are represented by differences of the radius of the circle and colors, respectively. The red stars symbolize large events such as: Chi-Chi Taiwan earthquake (21/9/1999 M7.6 latitude 23.8 longitude 121.1), Swarm at Miyakejima (7/2000-8/2000 latitude 34.0 longitude 139.0), Western Tottori earthquake (6/10/2000 M7.3 latitude 35.3 longitude 133.4) (from [25])



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 8

The schematics of the model of Ben-Zion [7] for a segmented strike-slip fault zone in a 3D elastic half-space

commodate dynamic overshoot, producing local slip governed by dislocation theory [17,63]. The static friction, dynamic friction, and arrest stress are connected via a dynamic overshoot coefficient $D = (\tau_s - \tau_a)/(\tau_s - \tau_d)$. If the stress transfer from failing regions increases the stress at other cells to their static or dynamic strength thresholds, as appropriate, these cells fail and the event grows. When the stress at all cells is below the brittle failure thresholds, the model earthquake ends and the strength at all failing cells recovers back to τ_s . The creep process is governed by a power-law dependence of creep-velocity on the local stress and space-dependent coefficients that increase exponentially with depth and with distance from the southern edge of the computational grid. The chosen parameters produce an overall “pine-tree” stress-depth profile with a “brittle-ductile” transition at a depth of about 12.5 km, and variable stress-along-strike profiles with a gradual “brittle-creep” transition near the boundary between regions II and III (see Ben-Zion [7] for additional details). The model generates many realistic features of seismicity compatible with observations, including frequency-size and temporal event statistics, hypocenter distribution with depth and along strike, intermittent criticality, accelerated seismic release, scaling of source time functions and more (e. g., [9,29,56,88]).

Synthetic Catalogs

Synthetic data generated by computational models can comprise many events covering large spatial areas and extremely long time spans. Moreover, the synthetic data retain the statistical reliability of the results. The data are free of measurement errors, which occur in estimating earthquake magnitudes and hypocentral locations, and do not suffer from incomplete recording of small events, which exist in natural catalogs. These are significant advantages for our study, which attempts to illustrate clearly the performance of clustering analysis and visualization techniques.

In Sect. “Description of the Data” we analyze synthetic catalogs generated by two model realizations (A and M) of Ben-Zion [7]. The catalogs contain the time, location and magnitude of earthquakes calculated by the model for 150 and 1500 years. Extensive numerical simulations with several different classes of models, summarized by Ben-Zion [8] and Zöller et al. [9], suggest that the degree of disorder in fault heterogeneities is a tuning parameter of the earthquake dynamics. Catalog A is generated by a model realization tailored to the Parkfield section of the San Andreas fault. Catalog M is generated by a realization of a more-disordered system like the San Jac-

into fault or the Eastern California Shear Zone in Southern California. In both data sets the time interval covers all events ($M \sim 1 - 3 \times 10^4$) that have occurred in the last 150 years of simulated fault activity. These simulations were repeated for ten times larger time scale i. e. 1500 year interval (the number of events $M \sim 10^5$) covering hundreds of large earthquakes ($m > 6$) and correspondingly wider time window.

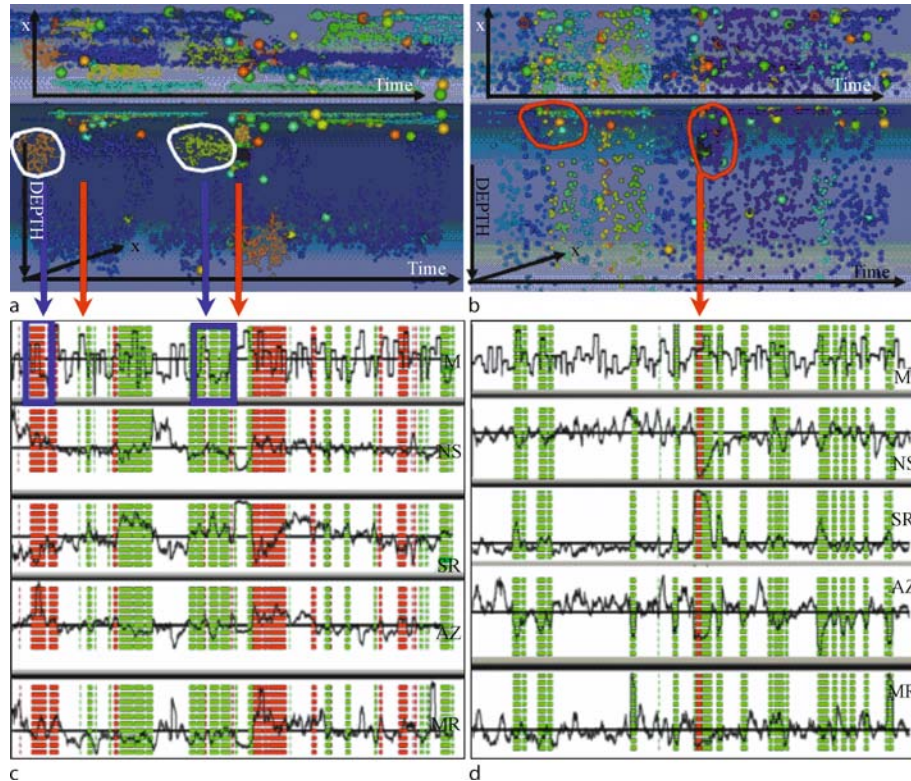
The seismicity parameters were obtained by averaging the data using a sliding time window of constant width ΔT and shift dt . We employ $\Delta T = 10$ days and $dt = 2$ days for the Japanese data, $\Delta T = 10$ months and $dt = 2$ months for the 150-years synthetic data and $\Delta T = 30$ months and $dt = 6$ months for the data covering 1500 years time period. Each parameter in the clustering was normalized with respect to the standard deviation.

Earthquake Visualization by Using Clustering in Feature Space

Short-Time Period

Results of clustering of the observed Japanese seismic catalogs (see Fig. 7) both in raw data and in feature spaces are shown in Fig. 9. At the data resolution level a single seismic event i can be represented as a multi-dimensional data vector $f_i = [m_i, z_i, X_i, Y_i, t_i]$ where: m_i is the magnitude, X_i – the latitude, Y_i – the longitude, z_i and t_i – the depth and the time of occurrence, respectively. The seismic events are visualized with the Amira package in Fig. 9a,b as irregular clouds of colored dots with (z, x, t) coordinates.

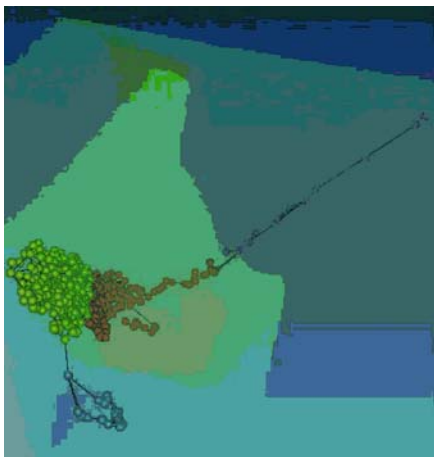
In accordance with the Gutenberg–Richter relationship, we find that the number of events from various



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 9

Real seismic data [49] analyzed by using clustering in both the data a,b and the feature (c,d) spaces. In panels a and b one can see the results of clustering in the data space (from two different perspectives, X-Time and Depth-Time) for small ($3 < m < 4$) and medium magnitude ($4 < m < 6$) events, respectively, represented by small dots. The different colors of the dots denote different clusters. Large events are visualized by the larger spheres. Their colors show the difference in magnitudes m (red – the largest, green – the smallest). The clusters in panels a,b encircled in red display the places with the largest seismic activity, while those in white represent the clusters of small precursory events. The red, white and green stripes in panel c and d representing 4 (out of 7) seismic parameters and maximum magnitude M show the clusters of similar time events for situations corresponding to panels a and b, respectively. The Amira visualization package was used (<http://www.amiravis.com>)

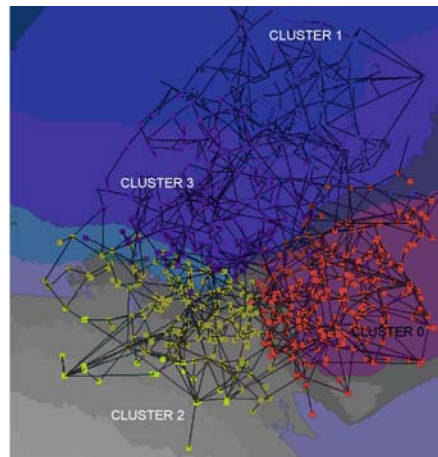
ranges of magnitudes differs considerably, and divide the entire set of data onto three subsets. The first one comprises the small, the second medium and the last one represents the largest earthquakes displayed in Fig. 9a,b as big dots. The deepest earthquakes $z > 150$ km are not displayed in the Fig. 9. The various shades represent the magnitudes of earthquakes from $m = 6$ (green) to $m = 7$ (red). In Fig. 9 we present the clustering results in both the data f_i and the feature F_i spaces. We look for clusters of similar seismic events (data space) and time events (feature space). The dots (data vectors), belonging to the same clusters, have the same color. The Fig. 9a,b is very rich in cluster-like forms, some of them hard to interpret. Correspondence of the cluster structure of data $f_j (j = 1, \dots, M_f)$ with the clusters of averaged events $F_i (i = 1, \dots, M_F)$ in the feature space can reveal interesting information. As one can see from the panel C, only three clusters are obtained in the feature space consisting of small data events ($3 < m < 4$). The green cluster corresponds to two relatively large time intervals of small events preceding Miyakejima earthquake and many smaller post shock periods. The time events F_i from this cluster represent averaged data events f_j , mainly shallow (AZ) of high degree of spatial repetitiveness SR and small diversity of magnitudes (MR). The red cluster consists of deeper events of smaller repetitiveness, and more diversified in magnitudes. The larger time interval of this type of behavior is recognized just after Miyakejima shock. The white cluster is not interesting in this scale of small events and includes all other events (including the earthquake swarm).



In panel D we display the seismicity parameters, which form three clusters of time events obtained for seismic events of larger magnitude $4 < m < 6$. Clusters of these events have different structure than in the previous case. They are parallel to X-depth plane. The borders between clusters roughly correspond to the borders of successive showers of the earthquakes. The red cluster comprises only the earthquakes corresponding to the Miyakejima swarm encircled in red in Fig. 9b. As we can see by the MDS visualization displayed in Fig. 10a, this cluster is made up from a needle of time events sprouting away from the two remaining and oval clusters. The green cluster in Fig. 9d represents the deep events, diversified in magnitude of high repetitiveness and rather high degree of spatial non-randomness at short distances (NS). As shown in Fig. 9, these time events represent mainly the post-swarm series of shocks. The white cluster, as before, includes all the other events.

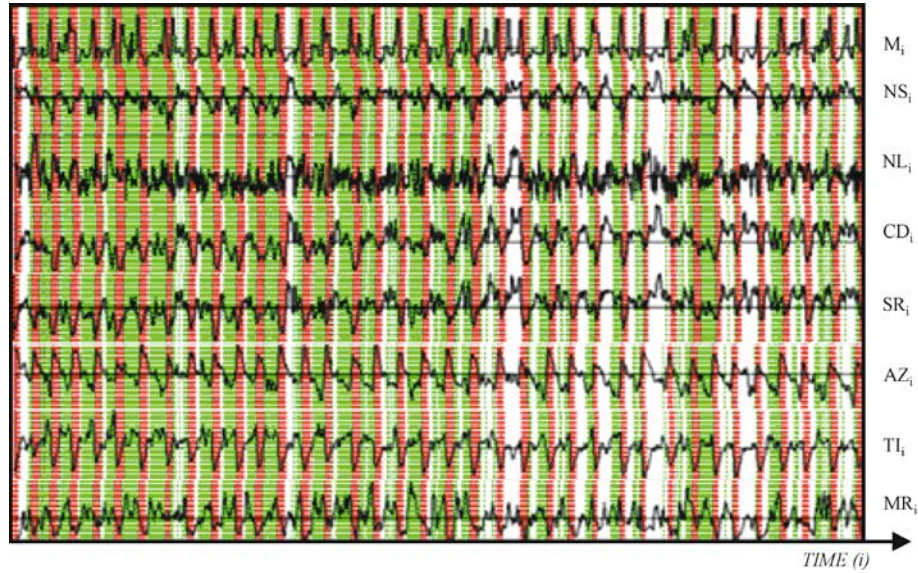
Time Period of 150 Years

In Fig. 11 we display the time series of seismicity parameters computed for the complete synthetic data catalog A. These time series follow the situation from Fig. 3 where dots represent separate data events. The green, red and white strips in Fig. 11 separate 3 clusters of similar time events represented by 7-dimensional feature vectors. In Fig. 10b these clusters are visualized due to the MDS transformation of 7-dimensional feature space into 3-D space. In Fig. 10b each dot represents a 7-dimensional feature vector mapped into 3-D space by MDS transformation.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 10

The clusters from feature space mapped into 3D space for a realistic short-time interval seismic data. The small blue cluster at the bottom represents the events at the end of the time interval, which are averaged within a shrinking time window. b The synthetic seismic data catalog A covering 150 years



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 11

The seismicity parameters $\{M, NS, NL, CD, SR, AZ, TI, MR\}$ in time for synthetic data catalog A (from [25])

From the top panel of Fig. 11 displaying the largest events M in the sliding time window, we may conclude that the white (blue in Fig. 10b) and red clusters from Figs. 10, 11b comprise time events, which correspond to the aftershock effects. The white cluster represents the net aftershock events, while the red one includes the earthquake effects averaged in sliding time window. Conversely, the green cluster (yellow in Fig. 10b) contains the time events preceding the earthquakes.

The selectivity in time of the seismicity parameters depends on the width ΔT and shift dt of the sliding time window. Due to space and time averaging, it is impossible to correlate precisely the appearance of an earthquake with the rest of the seismicity parameters when two earthquakes are too close to each other. Therefore, the sequence of green-red-white cluster events can be broken (Fig. 11) into time domains with many large earthquakes. As shown in Fig. 11 the occurrence of the largest events correlates well with the minima of NS , CD , SR , TI , and maxima of AZ , MR parameters. This means that the occurrence of large earthquakes is preceded by increasing spatial diffusion of events and increasing seismicity rate. Moreover, the results confirm the some findings from the real data in a shorter time-scale:

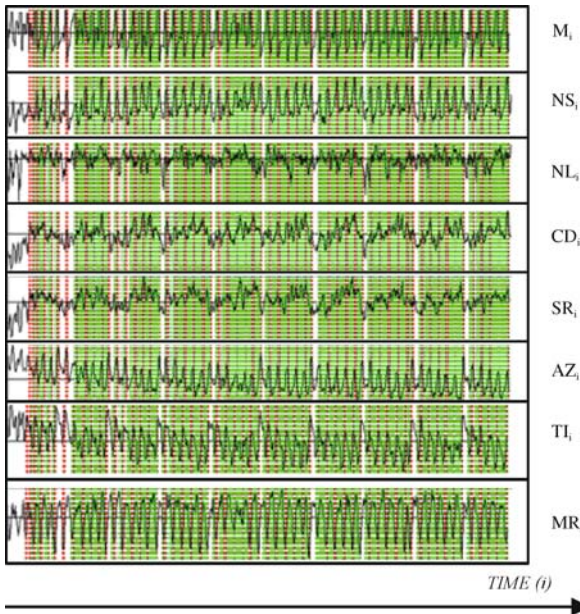
1. The events preceding large earthquakes are shallow and have small magnitudes. They have also higher degree of spatial repetitiveness than events from different clusters.

2. The earthquakes accompanying and following the mainshock are rather large in magnitude, deep, have high seismicity rates and low spatial correlation dimension (this drops off rapidly at the onset of large events),

The analysis of synthetic data shows clearly that the clusters in the feature space reflect well the periodicity of increasing and decreasing seismic activity in a given area. For this scale, however, the fine scale characteristics of precursory and after-shock effects become fuzzy.

Time Period of 1500 Years

In Figs. 4, 5, 6b and Figs. 12, 13 we visualize the feature vectors for data covering 1500 year period for two models: the A model with a Parkfield-type asperity and the M model with multi-size-heterogeneities. In Fig. 4 and Fig. 12 one can recognize two types of clusters with different sizes. The larger cluster comprises feature vectors forming approximately 150-year long periodic time intervals, which are represented by red strips in Fig. 4 and by green strips in Fig. 12. The second cluster consists of feature vectors from periodic gaps colored in green in Fig. 4 and in white in Fig. 12. This anomalous cluster corresponds to the periodic changes in the character of seismic activities. The third cluster (see Fig. 13), marked in red for M type of data in Fig. 12, consists of periodic and short time intervals representing rapid bursts of seismic activity within every 150-year interval.



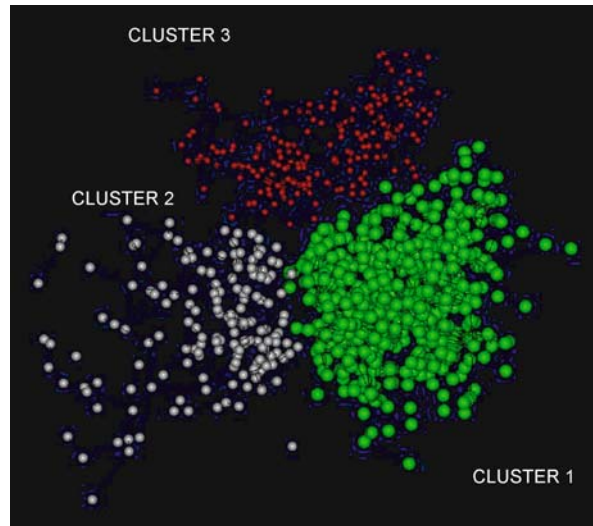
Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 12

The seismicity parameters with time for synthetic (catalog M) for seismic data representing time interval of 1500 years. The red and green strips depict the events belonging to red and green clusters from Fig. 7b, respectively

In both A and M models the gaps between 150-year long intervals are correlated with decrease of: the correlation dimension (CD), degree of spatial repetitiveness (SR) and seismicity rate. These gaps are preceded by large earthquakes. The simulations used for generating the datasets incorporate imposed large earthquakes on regions (I) and (IV) of Fig. 8b that bound the computational grid (region II), as staircase boundary conditions with a step at every 150 years. The analysis detected the effects of these boundary conditions on the seismicity that is calculated in region II.

There are also evident differences between the A and M data in the time intervals belonging to the second cluster. For A environment the gaps between 150-year intervals are greater and the secondary periodicity within them is less clear. Moreover, within time intervals from the second cluster, the degree of spatial non-randomness decreases at long distances (NL) for M model while for A data it decreases at short distances (NS). In addition, the average depth of earthquakes (AZ) is then clearly larger for A model, while for M data it remains on the average level.

In sum, by means of analyzing earthquake clusters in feature space over long time-scale, we can investigate important characteristics of seismic activity such as:



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 13

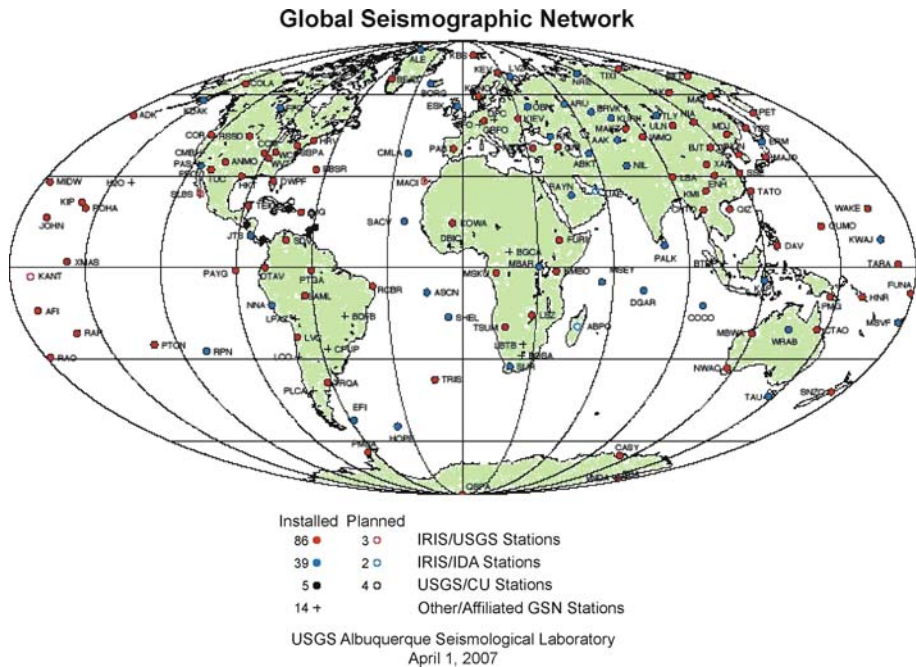
The clusters from Fig. 12 mapped by using multidimensional scaling from a 7-dimensional feature space into 3D space for synthetic seismic data catalog M covering 1500 years

1. The occurrence of hierarchical time-periodicity in seismic activity caused by increase of short-time correlations and their destruction, respectively. Correlations can be broken both due to short-wave and long-wave resonances of the Poincare type (e.g. during largest earthquakes) [Sornette, 2004].
2. The dependence of seismic activities on the ambient rheological and geological properties of the environment, which strongly modify the cluster structure of the feature vectors.

Remote Problem Solving Environment (PSE) for Analyzing Earthquake Clusters

Need for Remote Visualization and Analysis

We need fast access to large databases in order to forecast earthquakes by observation of similarities between thousands and millions of seismic events by visualization of earthquake clusters. The largest earthquake catalogs comprise TBytes of data. Taking into account also the data from tsunami earthquakes and micro-earthquakes in mines, the total amount of data collected by seismic centers spread all over the world is humongous. Moreover, knowledge extraction of earthquake precursors may demand exploration of cross-correlation relationships among many different catalogs. Therefore, both



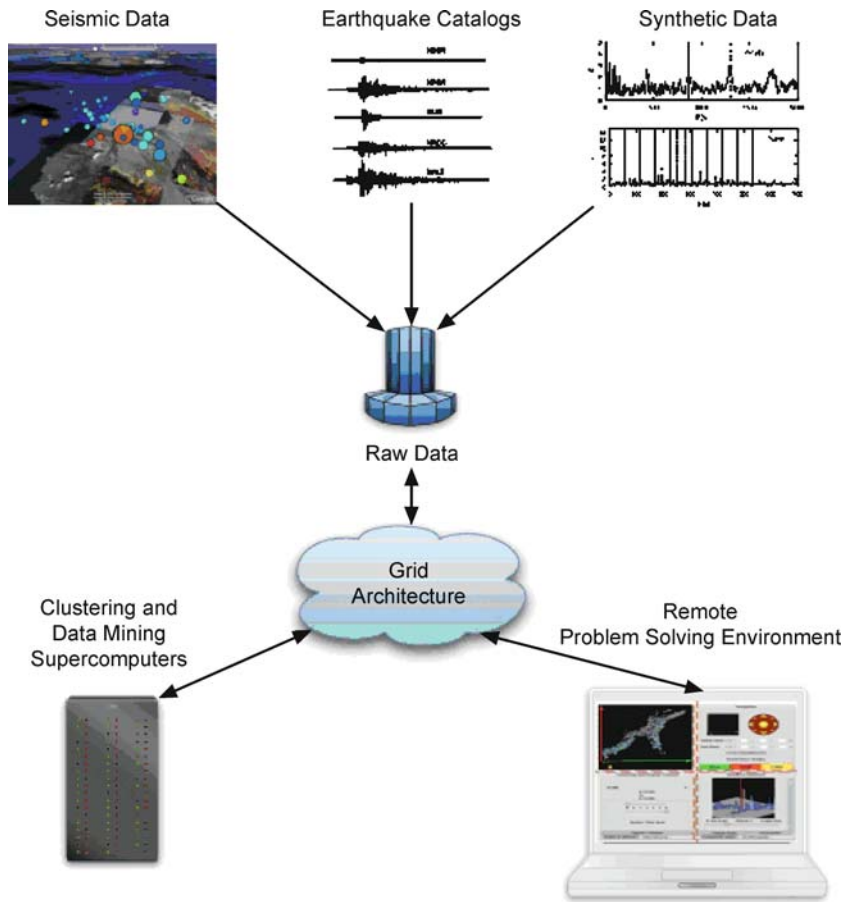
Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 14
Worldwide distribution of earthquake seismographic stations (© USGS)

fast communication between data centers and large disk spaces are sorely needed.

As shown in Fig. 14, earthquake seismograph stations, which collect earthquake data from regions with high seismic activity, are distributed worldwide. Therefore, the unprocessed data needs to be stored and then transferred to a dedicated remote server for data processing. After processing, the results must be returned to data acquisition centers and/or other clients. Broadband access to remote facilities dedicated specifically to pattern recognition and visualization allows for scrutinizing local data catalogs by using peer-to-peer connections of data acquisition centers to data preprocessing servers. Clients in the network can automatically compare various types of earthquake catalogs, data measured in distant geological regions, and the results from various theoretical and computational models. By comparing data accessible in the network we have a chance to eliminate the environmental factors and to extract the resultant earthquake precursory effects.

Integration of a variety of hardware, operating systems, and their proper configuration results in many communication problems between data centers. Efficient, reliable, and secure integration of distributed data and software resources, such as pattern recognition and visualization packages, is possible only within the GRID paradigm

of computing [13,31]. The GRID mode of computing has flourished rapidly in recent years and has facilitated collaboration and accessibility to many types of resources, such as large data sets, visualization servers and computing engines. Scientific teams have developed easy-to-use, flexible, generic and modular middleware, enabling today's applications to make innovative use of global computing resources. Remote access tools were also produced to visualize huge datasets and monitor performance and data analysis properties, effectively steering the data processing procedures interactively [21]. The TeraGrid project (<http://www.teragrid.org>) is a successful high-performance implementation of such a GRID infrastructure and is being used as an integrated, persistent computational resource at universities and laboratories across the USA. The TeraGrid development impacts also the earthquake science. The National Science Foundation has awarded the Southern California Earthquake Center 15 million service units of computer processing time on supercomputers nationwide [Grid Today, August 2007]. These computational resources will be used for simulating thousands of possible earthquakes scenarios in Southern California, including the largest breaks on the San Andreas fault (www.scec.org/cybershake). SCEC will be able to simulate the most disastrous earthquakes ($M > 7$), such as events that could produce Katrina-scale disasters.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 15
 Data acquisition, storage, processing, and remote problem solving environments

We discuss the idea of an integrated problem-solving environment (PSE) intended for the analysis of earthquake clusters for the prediction of earthquakes. A simplified scheme for data acquisition and visualization of earthquake clusters is displayed in Fig. 15. This system promotes portability, dynamic results on-demand, and collaboration among researchers separated by long distances by using a client server paradigm. This is provided through a lightweight front-end interface for users to run locally while the a remote server takes care of intensive processing tasks on large databases, off-screen rendering, and data visualization.

Grid Environment

In general, large datasets and high-performance computing resources are distributed across the world. When collaboration and sharing of resources are required, a computational GRID infrastructure needs to be in place to con-

nect these servers (see, e.g., [15]). There must exist protocols available to allow clients to tap into these resources and harness their power. The computational grid can be seen as a distributed system of “clients”, which consists of either “users” or “resources” and proxies. A GRID can be implemented using an event brokering system designed to run on a large network of brokering nodes. Individually, these brokering nodes are competent servers, but when connected to the brokering system, they are able to share the weight of client requests in a powerful and efficient manner. Examples of this include GRID Resource Brokering [30] and NaradaBrokering.

These GRID architectures are well suited to the functionality of a PSE for earthquake cluster analysis and as an integrated computational environment for data exchange and common ventures. The seismic data centers from the networking point of view represent a complex hierarchical cluster structure. They are located geographically in the regions of high seismic activity within heavily popu-

lated areas of economic importance. Therefore, the seismic data centers create distant superclusters of various “density” of computational resources corresponding to the size and importance of the regions. These superclusters are sparse in the sense of computational resources devoted for earthquake detection and data acquisition. However, these same structures contain important computational, scientific and visualization facilities with strong interest in the analysis of earthquake data and earthquake modeling. The efficient interconnection of these sites is of principal interest. Due to the “small world network” structure of GRID architectures it is possible to select the most efficient routing schemes, considerably shortening the average communication path length between brokers. GRID architectures are appropriate to link the clients, both users and resources, together. Construction of efficient and user friendly Problem Solving Environments requires integration of data analysis and visualization software within the GRID environment, in such a way that it can be easily accessed via the Internet. We created an integrated data interrogation toolkit to act as a PSE for visualization and clustering of seismic data, which we call WEB-IS.

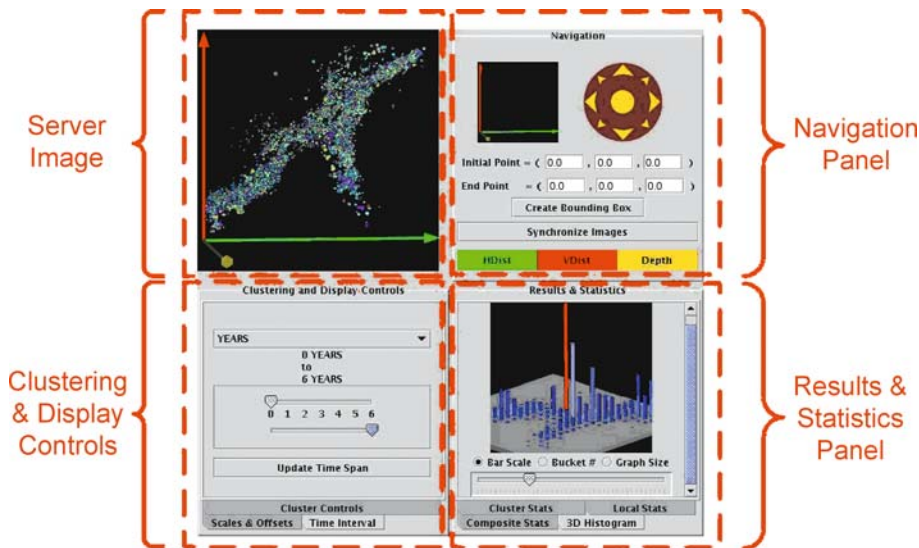
Example of Remote PSE

WEB-IS is a software tool that allows remote, interactive visualization and analysis of large-scale 3-D earthquake clusters over the Internet [85] through the interaction between client and server. WEB-IS acts as a PSE through a web portal used to solve problems by visualizing and ana-

lyzing geophysical datasets, without requiring a full understanding of the underlying details in software, hardware and communication [34,52]. As shown in Fig. 16, the primary goal of WEB-IS in the geosciences is to provide middleware that sits between the modeling, data analysis tools and the display systems that local or remote users access. In the case of large and physically distributed datasets, it is necessary to perform some preprocessing and then transmit a subset of the data to one or more processes or visualization servers to display. The details of where and how the data migrates should be transparent to the user. WEB-IS makes available to the end users the capability of interactively exploring their data, even though they may not have the necessary resources such as sufficient software, hardware or datasets at their local sites. This method of visualization allows users to navigate through their rendered 3-D data and analyze for statistics or apply earthquake cluster analysis. To the client, the process of accessing and manipulating the data appears simple and robust, while the middleware takes care of the network communication, security and data preparation.

Complete realization of an earthquake clustering PSE consists of:

- 1. Data analysis tools to implement earthquake clustering techniques;
- 2. High performance visualization techniques using OpenGL or Amira;
- 3. The Grid environment;
- 4. Integration toolkit, such as WEB-IS.



Earthquake Clusters over Multi-dimensional Space, Visualization of, Figure 16
WEB-IS is an example of a remote earthquake clustering PSE

These exist and can work both independently and coupled in a single special purpose system. This system can be developed creating the backbone of the sophisticated computational data acquisition environment, which can be devised specifically for earthquake clustering or for general needs of the geophysical community. Equipped with only PDAs or laptops, and working on location in unreachable desert terrains with remote data acquisition centers or perhaps just analyzing data in one of the many computation facilities located around the globe, geophysicists will be enabled unlimited access to data resources spread all over the world.

We see the principal goal of our work in contributing to the construction of a global warning system, which can be used for prediction of catastrophes such as various types of earthquakes along the circum Pacific belt, where there is a great concentration of people. For example, similar methodology can be used for tsunami earthquake alerting. Theoretical models of faulting and seismic wave propagation used for the computation of radiated seismic energy from broad-band records at teleseismic distances [14] can be adapted to the real-time situation when neither the depth nor the focal geometry of the source is known accurately. The distance-dependent approximation was used in [60]. By analyzing some singular geophysical parameters such as the energy-to moment ratio H [60] for regular earthquakes, the results obtained from the theoretical models agree well with values computed from available source parameters (e.g., as published by the National Earthquake Information Center). It appears however that the so called “tsunami earthquakes” – characterized by the significant deficiency of moment release at high frequencies – yield the values of H considerably different the regular earthquakes. Thus H value can be used as a suitable criterion for discriminating various types of earthquakes in a short duration of time, like an hour. However, this hypothesis holds only for a few cases. For, so called, “tsunamigenic earthquakes” this difference is not so clear. Moreover, the value of the moment computed on the base of long-period seismic waves can be underestimated. For example, analysis of the longest period normal modes of the Earth, $0S_2$ and $0S_3$, excited by the December 26, 2004 Sumatra earthquake [76], yields an earthquake moment of $1.3 \cdot 10^{30}$ dyn-cm, approximately three times larger than the $4 \cdot 10^{29}$ dyn-cm measured from long-period surface waves. Therefore, instead of a single-value discrimination we recommend using more parameters (dimensions) for detecting tsunami earthquakes. As shown in [64] and [83], one could employ other T-phase characteristics such as its duration, seismic moment, and spectral strength or even similar features associated with the

S-phase. We believe that the lack of success in predicting earthquakes still comes from the lack of communications between researchers and difficulties in free and fast access to the various types of data. Therefore, we hope that globalization of computation, data acquisition and visualization resources, together with fast access through a scale-free network, will provide a triumphant solution to this problem.

Future Directions

In this chapter we endeavor to bring across the basic concept of clustering and its role in earthquake forecasting. Indeed we find that the clustering of seismic activities reflects both the similarity among them and their correlation properties. As discussed in, e.g., Ben-Zion et al. [9], Saichev and Sornette [68] and Zöller et al. ▶ [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#), there exists an evolutionary process or memory between successive earthquakes, which impact the distribution of the inter-event times. We believe that by means of earthquake clustering we can capture the essence of this predictive information [27]. Therefore, in order to carry out real-time earthquake forecasting for short-time scales, it is necessary to derive a thorough understanding of all families of earthquake clusters produced over an earthquake-prone region.

We stress here that in obtaining this type of information one must first be able to detect the precise location of the significant clusters, by filtering out simultaneously the noise and the outliers. While the existence of spatial-temporal clusters is important, they do not reveal the subtle information hidden behind the relations among the data events, such as: spatial-temporal correlation dimensions, correspondence between the numbers of small and large magnitude events, degree of spatial randomness, repetitiveness at different distances and other factors. The features – “descriptors” or seismicity parameters – constructed from the empirical knowledge of the researcher should be largely independent and should represent aptly distinctive features, which are useful for the purpose of pattern recognition. Unlike single events described only by spatio-temporal features (and magnitude), the N -dimensional feature vectors can represent better the dynamics of the seismically active area in different moments of time. By following the basic rules of learning theory, we may be able to arrive at the number N and quality of features, which can assure the generalization power of the data and allow us to construct reliable data-models or classifiers.

We have shown that clustering, as a well-honed tool in data mining and pattern recognition, represents the clas-

sifier without the teacher, which means that the nature of the clustering is unknown and its exact background must be guessed at from expert knowledge and analysis of the cluster properties. Clustering is a process based on a priori knowledge extraction for constructing the hypothesis space needed for reliable classifiers that can be taught and used for forecasting [25]. However, the quality of these data models depends strongly on the quality of hypothesis space constructed. Consequently, it depends on the quality of clusters extraction. The major problem comes from the lack of a universal clustering scheme, thus making the clustering process somewhat subjective. In this case we must visualize the multidimensional feature space. Visual confirmation gives one a confidence concerning the validity of the clusters and we can then adjust for the optimal clustering procedures by removing the noise and outliers. Among the major goals of earthquake clustering, we can include the following salient points:

- classification of the chaotic properties of seismicity patterns [35], for example to recognize the three main groups of shocks: foreshocks, mainshocks and aftershocks or to remove the temporary clustering to estimate the background seismicity;
- understanding the correlations between observed properties of earthquakes in different domains (e.g., space, time, number, size);
- understanding the relations between various physical parameters of the models and properties of the generated earthquakes;
- investigating the multi-scale nature of the cluster structure and reconstructing the important and hidden information associated with the stress characteristics.

Classification of type of shocks seems to be an unresolved problem because there are no observable differences between foreshocks, main shocks and aftershocks [68]. Each earthquake is able of triggering other earthquakes according to the basic laws from [46,69]. Despite this difficulty, as shown in [9], it is possible to construct some sort of stochastic classifiers based on theoretical footing. The method proposed here closely related to the epidemic-type aftershock sequence (ETAS) model [61]. It is important that the principal characteristics of ETAS-based models correspond to experimental verifications, i.e., they treat all earthquakes on the same footing and there is not distinction between foreshocks, main shocks and aftershocks. The key points of the method are the probabilities of one event being triggered by a previous event (e.g., [82]). Making use of these probabilities, we can reconstruct the functions associated with the characteristics of earthquake

clusters to test a number of plausible hypotheses about the earthquake clustering phenomena.

As shown above by our results on seismicity clustering for the three different time epochs, clustering can be truly regarded as a coarse-graining procedure. We can see details from the smaller scales are erased, thereby exposing the general trends associated with the long correlation length. For large data bases covering long time intervals we can unveil the shorter timescale characteristics by removing the background events, using successive clustering. Eventually, we can build up the strong classifiers. In the case where the long-time data catalogs are missing, we can employ the stochastic classifiers advocated Ben-Zion et al. [9] for prior thresholding of the background data or what is sometimes called “fuzzification” [86]. By this procedure we can construct the hypothesis space for data models by clustering (or fuzzy clustering) procedures.

The results discussed in this paper contribute to the development of improved software infrastructure for analysis of seismicity. A combined clustering analysis of observed and synthetic data, aided by state-of-the-art visualization of multidimensional clusters, undoubtedly lead to improved earthquake forecasting algorithms with shorter time windows of increased probability of large seismic events.

Acknowledgments

This research was supported by NSF ITR and Math-Geo grants. WD acknowledges support from the Polish Committee for Scientific Research (KBN) Grant No. 3T11C05926. YBZ acknowledges support from the NSF, USGS and SCEC.

Bibliography

Primary Literature

1. Amira visualization package. <http://www.amiravis.com>
2. Andenberg MR (1973) Clusters analysis for applications. Academic Press, New York
3. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomena. *J Geophys Res* 94(B11):15635–15637
4. Bak P, Christensen K, Danon L, Scanlon T (2002) Unified scaling law for earthquakes. *Phys Rev Lett* 88:178501
5. Barabasi AL, Jeong H, Neda Z, Ravasz E, Schubert A, Vicsek T (2002) Evolution of the social network of scientific collaborations. *Physica A* 311(3–4):590–614
6. Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press, Cambridge, pp 346
7. Ben-Zion Y (1996) Stress, slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
8. Ben-Zion Y (2001) Dynamic rupture in recent models of earthquake faults. *J Mech Phys Solids* 49:2209–2244

9. Ben-Zion Y (2003) Appendix 2, key formulas in earthquake seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part B. Academic Press, pp 1857–1875
10. Ben-Zion Y, Rice JR (1993) Earthquake failure sequences along a cellular fault zone in a three-dimensional elastic solid containing asperity and nonasperity regions. *J Geophys Res* 98:14109–14131
11. Ben-Zion Y, Rice JR (1995) Slip patterns and earthquake populations along different classes of faults in elastic solids. *J Geophys Res* 100:12959–12983
12. Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. *J Geophys Res* 108(B6):2307–27
13. Berman F, Fox GC, Hey AJG (2003) Grid computing – making the global infrastructure a reality. Wiley Series in Communications Networking and Distributed Systems, pp 1007
14. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91:2095–2112
15. Bollig EF, Lyness PA, Nacar MA, da Silveira PR, Erlebacher G, Pierce M, Yuen DA (2007) VLAB: Web services, portlets, and workflows for enabling cyber infrastructure in computational mineral physics. *J Phys Earth Planet Inter* 163:333–346
16. Chen C-C, Rundle JB, Li H-C, Holliday JR, Turcotte DL, Tiampo KF (2006) Critical point theory of earthquakes: Observations of correlated and cooperative behavior on earthquake fault systems. *Geophys Res Lett* L18302
17. Chinnery M (1963) The stress changes that accompany strike-slip faulting. *Bull Seismol Soc Am* 53:921–932
18. Corral A (2005) Mixing of rescaled data and Bayesian inference for earthquake recurrence times. *Nonlinear Process Geophys* 12:89–100
19. Corral A (2005) Renormalization-group transformations and correlations of seismicity. *Phys Rev Lett* 95:028501
20. Corral A, Christensen K (2006) Comment on earthquakes descaled: On waiting time distributions and scaling laws. *Phys Rev Lett* 96:109801
21. da Silva CRS, da Silveira PRC, Karki B, Wentzcovitch RM, Jensen PA, Bollig EF, Pierce M, Erlebacher G, Yuen DA (2007) Virtual laboratory for planetary materials: System service architecture overview. *Phys Earth Planet Inter* 163:323–332
22. Davy P, Sornette A, Sornette D (1990) Some consequences of a proposed fractal nature of continental faulting. *Nature* 348:56–58
23. Dzwinel W, Blasiak J (1999) Method of particles in visual clustering of multi-dimensional and large data sets. *Future Gener Comput Syst* 15:365–379
24. Dzwinel W, Yuen DA, Kaneko Y, Boryczko K, Ben-Zion Y (2003) Multi-resolution clustering analysis and 3-D visualization of multitudinous synthetic earthquakes. *Vis Geosci* 8:12–25
25. Dzwinel W, Yuen DA, Boryczko K, Ben-Zion Y, Yoshioka S, Ito T (2005) Nonlinear multidimensional scaling and visualization of earthquake clusters over space, time and feature space. *Nonlinear Process Geophys* 12:117–128
26. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 95:14863–14868
27. Enescu B, Ito K, Struzik ZR (2006) Wavelet-based multiscale analysis of real and simulated time-series of earthquakes. *Geophys J Int* 164:63–74
28. Eneva M, Ben-Zion Y (1997) Techniques and parameters to analyze seismicity patterns associated with large earthquakes. *J Geophys Res* 102(B8):785–795
29. Eneva M, Ben-Zion Y (1997) Application of pattern recognition techniques to earthquake catalogs generated by models of segmented fault systems in three-dimensional elastic solids. *J Geophys Res* 102:24513–24528
30. Ferreira L (2002) Introduction to grid computing with Globus IBM Redbook series. IBM Corporation <http://ibm.com/redbooks>
31. Foster I, Kesselman C (eds) (1998) Building a computational grid: state-of-the art and future directions in high-performance distributed computing. Morgan-Kaufmann, San Francisco
32. Freed AM, Lin J (2001) Delayed triggering of the 1999 Hector Mine earthquake by viscoelastic stress transfer. *Nature* 411:180–183
33. Frey BJ, Dueck D (2007) Clustering by Passing Messages Between Data Points, *Science* 315(5814):972–976
34. Garbow ZA, Erlebacher G, Yuen DA, Sevre EO, Nagle AR, Kaneko Y (2002) Web-based interrogation of large-scale geophysical datasets and clustering analysis of many earthquake events from desktop and handheld devices. American Geophysical Union Fall Meeting, Abstract
35. Goltz C (1997) Fractal and chaotic properties of earthquakes. In: Goltz C (ed) Lecture notes in earth sciences, vol. 77. Springer, Berlin, p 3–164
36. Gowda CK, Krishna G (1978) Agglomerative clustering using the concept of nearest neighborhood. *Pattern Recognit* 10:105
37. Grossman RL, Karnath Ch, Kegelmeyer P, Kumar V, Namburu RR (2001) Data mining for scientific and engineering applications. Kluwer, Dordrecht
38. Guha S, Rastogi R, Shim K (1998) CURE: An efficient algorithm for large databases. In: Proceedings of SIGMOD '98, Seattle, June 1998, pp 73–84
39. Gutenberg B (1942) Earthquake magnitude, intensity, energy and acceleration. *Bull Seismol Soc Am* 32:163–191
40. Gutenberg B, Richter CF (1954) Seismicity of the earth and associated phenomena. Princeton University Press, Princeton
41. Haile PM (1992) Molecular Dynamics Simulation. Wiley, New York
42. Hainzl S, Scherbaum F, Beauval C (2006) Estimating background activity based on interevent-time distribution. *Bull Seismol Soc Am* 96:313–320. doi:10.1785/0120050053
43. Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press, Cambridge
44. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: Data mining, inference and prediction. Springer, New York, pp 533
45. Helmstetter A, Sornette D, Grasso J-R (2003) Mainshocks are aftershocks of conditional foreshocks: How do foreshock statistical properties emerge from aftershock laws. *J Geophys Res* 108:2046
46. Helmstetter A, Kagan Y, Jackson D (2005) Importance of small earthquakes for stress transfers and earthquake triggering. *J Geophys Res* 110:B05S08

47. Hong H, Kadlec DJ, Yuen DA, Zheng Y, Zhang H, Liu G, Dzwinel W (2004) Fast timescale phenomena at Changbaisan volcano as inferred from recent seismic activity. *Eos Trans AGU Fall Meet.* 85(47) <http://www.agu.org>
48. Ismail MA, Kamel MS (1989) Multi-dimensional data clustering utilizing hybrid search strategies. *Pattern Recognit* 22(1):77–89
49. Ito T, Yoshioka S (2002) A dike intrusion model in and around Miyakejima, Niiijima and Kozushima. *Tectonophysics* 359:171–187
50. Jajuga K, Sokolowski A, Hermann H (eds) (2002) Classification, clustering and data analysis. Springer, Berlin, pp 497
51. Jones NC, Pevzner P (2004) An introduction to bioinformatics algorithms. MIT Press, Cambridge
52. Kadlec BJ, Yang XL, Wang Y, Bollig EF, Garbow ZA, Yuen DA, Erlebacher G (2003) WEB-IS (Integrated System): An overall view. *Eos Trans AGU* 84(46), Fall Meet. Suppl., Abstract NG11A-0163
53. Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, pp 341
54. Karypis G, Han E, Kumar V (1999) Chameleon: A hierarchical clustering algorithms using dynamic modeling. *IEEE Computer* 32(8):68–75
55. Karypis G, Aggarwal R, Kumar V, Shekhar S (1999) Multi-level hypergraph partitioning: applications in VLSI domain. *IEEE Trans on Very Large Scale Systems Integration (VLSI)* 7(1):69–79
56. Mehta AP, Dahmen KA, Ben-Zion Y (2006) Universal mean moment rate profiles of earthquake ruptures. *Phys Rev E* 73:056104
57. Mitra S, Acharya T (2003) Data mining: multimedia, soft computing and bioinformatics. Wiley, New Jersey
58. Molchan GM (2005) Interevent time distribution of seismicity: A theoretical approach. *Pure Appl Geophys* 162:1135–1150
59. National Research Council (2003) Living on an active earth, perspectives on earthquake sciences. The National Academies Press, Washington DC
60. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: the $S/M0$ discriminant for tsunami earthquakes. *J Geophys Res* 103(B11):23885–23898
61. Ogata Y (1999) Seismicity analysis through point-process modeling: A review. *Pure Appl Geophys* 155:471–507
62. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. *Tectonophysics* 413:13–23
63. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. *Bull Seismol Soc Am* 82:1018–1040
64. Okal EA, Alasset P-J, Hyvernaud O, Schindele F (2003) The deficient T waves of tsunami earthquakes. *Geophys J Int* 152:416–432
65. Rundle JB, Gross S, Klein W, Ferguson C, Turcotte DL (1997) The statistical mechanics of earthquakes. *Tectonophysics* 277:147–164
66. Rundle JB, Klein W, Tiampo K, Gross S (2000) Linear pattern dynamics in nonlinear threshold systems. *Phys Rev E* 61(3):2418–2143
67. Rundle JB, Turcotte DL, Klein W (eds) (2000) GeoComplexity and the physics of earthquakes. American Geophysical Union, Washington, pp 284
68. Saichev A, Sornette D (2007) Theory of earthquake recurrence times. *J Geophys Res* 112(B4):1–26
69. Saichev A, Helmstetter A, Sornette D (2005) Power law distributions of offspring and generation numbers in branching models of earthquake triggering. *Pure Appl Geophys* 162:1113–1134
70. Sander J, Ester M, Krieger H (1998) Density based clustering in spatial databases, The algorithm DBSCAN and its applications. *Data Min Knowl Discov* 2(2):169–194
71. Shcherbakov R, Turcotte DL (2004) A modified form of Bath's law. *Bull Seismol Soc Am* 94:1968–1975
72. Shcherbakov R, Turcotte DL, Rundle JB (2005) Aftershock statistics. *Pure Appl Geophys* 162(6–7):1051–1076
73. Siedlecki W, Siedlecka K, Sklanski J (1988) An overview of mapping for exploratory pattern analysis. *Pattern Recognit* 21(5):411–430
74. Sornette D (2006) Critical phenomena in natural sciences. Springer Series in Synergetics, Berlin, pp 528
75. Sornette D, Johansen A, Bauchaud J-P (1996) Stock market crashes, precursors and replicas. *J Phys Int Finance* 5:167–175
76. Stein S, Okal E (2004) Ultra-long period seismic moment of the great December 26, 2004 Sumatra earthquake and implications for the slip process. <http://www.earth.nwu.edu/people/emile/research/Sumatra.pdf>
77. Tan P-N, Steinbach M, Kumar V (2005) Introduction to data mining. Addison Wesley, Boston, pp 769
78. Theodoris S, Koutroumbas K (1998) Pattern recognition. Academic Press, San Diego
79. Turcotte DL (1997) Fractals and chaos in geology and geophysics, 2nd Edn. Cambridge University Press, New York
80. Utsu T (2002) Statistical Features of Seismicity. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International Handbook of Earthquake and Engineering Seismology, Part A. Academic Press, pp 719–732
81. Van Aalsburg J, Grant LB, Yakovlev G, Rundle PB, Rundle JB, Turcotte DL, Donnellan A (2007) A feasibility study of data assimilation in numerical simulations of earthquake fault systems. *Phys Earth Planet Inter* 163:149–162
82. Vere-Jones D (1976) A branching model for crack propagation. *Pure Appl Geophys* 114(4):711–726
83. Walker DA, McCreery CS, Hiyoshi Y (1992) T-phase spectra, seismic moment and tsunamigenesis. *Bull Seismol Soc Am* 82:1275–1305
84. Wesnousky SG (1994) The Gutenberg-Richter or characteristic earthquake distribution, which is it? *Bull Seismol Soc Amer* 84:1940–1959
85. Yuen DA, Garbow ZA, Erlebacher G (2004) Remote data analysis, Visualization and Problem Solving Environment (PSE) based on wavelet analysis in the geosciences. *Vis Geosci* 8:83–92. doi:10.1007/x10069-003-0012-z
86. Zadeh LA (1996) Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh, World Scientific Series In Advances In Fuzzy Systems. World Scientific Publishing, River Edge, pp 826
87. Zhang Q, Boyle R (1991) A new clustering algorithm with multiple runs of iterative procedures. *Pattern Recognit* 24(9):835–848
88. Zöller G, Hainzl S, Ben-Zion Y, Holschneider M (2006) Earthquake activity related to seismic cycles in a model for a heterogeneous strike-slip fault. *Tectonophysics* 423:137–145. doi:10.1016/j.tecto.2006.03.007
89. Zöller G, Ben-Zion Y, Holschneider M (2007) Estimating recurrence times and seismic hazard of large earthquakes on an

individual fault. *Geophys J Int* 170:1300–1310. doi:10.1111/j.1365-246X.2007.03480.x

Books and Reviews

- Ertoz L, Steinbach M, Kumar V (2003) Finding clusters of different size, shapes and densities in noisy, high-dimensional data. Army High Performance Center, technical report, April 2003
- Yuen DA, Kadlec BJ, Bollig EF, Dzwinel W, Garbow ZA, da Silva C (2005) Clustering and visualization of earthquake data in a grid environment, vol 10/1. *Vis Geosci* <http://www.springerlink.com/content/n60423820556/>

Earthquake Damage: Detection and Early Warning in Man-Made Structures

MARIA I. TODOROVSKA
Department of Civil Engineering,
University of Southern California, Los Angeles, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Literature Review
Damage and Damage-Sensitive Features
Structural Models and Identification
Examples
Future Directions
Bibliography

Glossary

Structural health monitoring Is the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. It is often used interchangeably with *structural damage detection*.

Inter-story drift Is the ratio between the relative horizontal displacements at two levels of the structure and the distance between them. It is important to distinguish between drift resulting from *deformation* of the structure, which is directly related to damage, and drift resulting from the deformation of the soil and rocking of the structure as a rigid body. It is also important to estimate reliably the drift due to permanent displacement (its “DC” component), which cannot be done reliably using data from vibrational sensors unless six degrees of freedom of motion (three translations and three rotations) are recorded.

Soil-structure interaction (SSI) Is a process occurring during vibration of structures founded on flexible soil, in which the structure and soil interact, and their motions are modified. *Kinematic interaction* refers to the effects of scattering and diffraction of the incident seismic waves from the soil excavation for the foundation. *Dynamic interaction* refers to the effects caused by the inertia forces of the structure and foundation, which lead to deformation of the soil, and results in modification of the *resonant frequencies* and *damping* of the response of the structure, foundation and soil acting as a system.

Resonant frequencies of vibration Of a structure on flexible soil are those of the *soil-structure system*, and the energy of the vibrational response is concentrated around these frequencies. They depend on the stiffness of the building and that of the soil. *Fixed-base frequencies of vibration* are the resonant frequencies of the structure on rigid soil, and depend only on the stiffness of the structure. Loss of stiffness of the structure due to damage results in reduction of the fixed-base frequencies, and indirectly of the system frequencies. Monitoring changes in the fixed-base frequencies is most reliable because it eliminates the effects of the soil, which can exhibit (recoverable) nonlinear behavior during strong shaking.

Definition of the Subject

Structural health monitoring and structural damage detection refers to the process of determining and tracking the structural integrity and assessing the nature of damage in a structure. Being able to detect the principal components of damage in structures as they occur during an earthquake or soon after the earthquake, or the absence of it, before physical inspection is possible, is an important and challenging problem. Considering the challenges faced and the potential benefits for safety and for minimizing disruption of productivity, structural health monitoring has the elements of a *grand challenge* problem in civil engineering [12].

Structural damage can be described by the following five attributes: existence, location, type, extent, and prognosis for the remaining useful life. Structural damage is a complex state, which can occur on different time scales, suddenly during some catastrophic event such as earthquake or explosion, or gradually over the life of the structure, due to deterioration of the structural materials by aging, service, and exposure to environmental influences. This article is concerned primarily with identification of the most significant components in the space of complex

patterns of damage caused by earthquakes. Damage in structures also can be described on different spatial scales, e. g. from small defects and localized damage in a component, to global state of damage of the structural system. Hence the damage detection methods are classified as *local* and *global*. The local methods are those for nondestructive testing (NDT) of materials, which can determine the location of the damage in a structural component. They involve use of actuators (radiating ultrasonic waves into the structural element), and require access to the element. The global methods assess the overall state of damage of the structural system (as it reflects on its overall performance during an extreme event). The focus of this review is on the global methods, and intermediate scale methods, which can point to the part of the structure that has been damaged.

Structural damage detection and early warning involve: (1) *recording* some sensory data, (2) *identification* of some structural parameter(s) sensitive to damage (e. g. natural frequencies of vibration, or wave travel times), some characteristic of response (e. g. levels of inter-story drift) that can be correlated with damage, or some other patterns (e. g. abrupt changes in the response detected as novelties), (3) *comparison* of the result of the identification with some knowledge base of correlation of such patterns with levels of damage, and (4) *decision making* (e. g. whether to evacuate or continue occupancy). Because of various uncertainties, the answer can be only expressed probabilistically, and the decision will also depend on the nature of the use of the structure and level of tolerance of the user.

The earliest and most wide-spread methods of structural damage detection are those based on data from vibrational sensors. In fact, the hope to eventually be able to detect hidden damage has been one of the motivations for the development and deployment of seismic sensors in structures. The first strong motion recordings in a building are those during the $M = 5.4$ Southern California earthquake of October 2, 1933, obtained in the Hollywood Storage Building, the instrumented structure in US with the longest history of recording earthquakes [63]. The earliest identification methods consisted of estimation of the building resonant frequencies and damping, from energy distributions of small amplitude ambient noise and forced vibration tests [3], as well as from earthquake records [63]. These studies identified the resonant frequencies and damping of the soil-structure system, which depend on the properties of the soil, and can change significantly even when there is no damage. Detailed system identification studies from full-scale test vibration data that separate the effects of the soil-structure

interaction appeared in the 1970s, following theoretical developments that helped understanding the phenomenon of soil-structure interaction [13,32,33,34,71]. Thirty years later, such studies are still rare, due to a combination of factors, one of which is the inadequate coverage of this topic in the graduate curricula, and the other is the emphasis of earthquake engineering research on laboratory experimentation and numerical simulations, rather than on the full-scale testing of structures [63].

Despite the progress made to date in instrumentation of structures as well as in development of theoretical methods, structural health monitoring systems are deployed in structures only on an experimental basis. The main obstacles to the routine practical deployment of such systems are: (1) the high cost of sensors and monitoring systems, which limits the number of structures that are instrumented and the detail of the measurements (spatial resolution, e. g.), (2) the low sensitivity and robustness of the methods, and ability to discriminate between changes in the damage sensitive feature caused by damage from changes caused by other factors (e. g. age, level of excitation, and weather), and (3) the paucity of data recorded in damaged structures necessary to calibrate the health monitoring methods. Consequently, the main challenges for future research are: (1) to design low cost but high performance sensors and monitoring systems, making it possible to densely instrument many structures, (2) to develop methods that are robust and sensitive enough to detect also light damage (in particular one that is not visible), and (3) to build a knowledge base that can help reliably relate observed patterns in the data with actual observations of damage.

Recently, structural identification and health monitoring of buildings by detecting changes in wave travel time through the structure has received revived attention and has proven to be very promising [20,24,25,37,39,41,42,53,54,65]. Exploratory applications to data from damaged buildings [53,54] showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gives results consistent with the spatial distribution and degree of the observed damage).

Introduction

This volume would not be complete without addressing the catastrophic consequences of earthquakes, and damage in soil-structure systems, which is a complex, multidimensional, and highly interrelated set of phenomena.

Since the early days, the mathematical formulation of practical earthquake engineering problems has been dominated by *linear* differential equations [58], which *cannot* lead to chaos. Nevertheless, cost and the increasing needs of society have pushed the design into the nonlinear regimes of large deformations increasing the possibility of encountering chaotic dynamic phenomena in structural response, and have increased the complexity of the possible damage outcomes. However, working with parameters that produce chaotic output reduces the ability to predict the outcome. The chaotic behavior of nonlinear systems does not completely exclude the possibility to predict the response, but introduces an upper bound (prediction horizons) [30]. Then the remaining question is over what time-scales can the predictions still be reliable. Also, the prediction of response requires a realistic physical *model*, while the practical outcome of most work in engineering remains *empirical*. Consequently, there is a conflict in the classical engineering description of the world. This conflict is in part due to the assumption that nature is moving forward, according to a deterministic law, and in part due to the fact that engineers model the world based on incomplete data, and thus working with unverifiable representation. This leads to the question what models are good for. The problem is further aggravated by the fact that the art of dynamical modeling tends to be neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [2]. In the following review of structural health monitoring in earthquake engineering, it is accepted that there is a modeling problem, and the success of a method is gauged by the degree to which its predictions match the observed outcomes.

During the last several decades, stochastic processes have been used to help analyze the irregular behavior of deterministic systems with too many variables to be described in detail. Stochastic processes have been used also to model the deterministic response of structures to earthquake and wind forces, and as an approximate description of deterministic systems sensitive to their initial conditions. In some analyses, random noise is added to the model to account for the differences between the behaviors of model and prototype. This noise represents no more than lack of knowledge of the system structure or inadequacy of the identification procedure [23].

Following a damaging earthquake, buildings, bridges, dams and other structures are physically inspected for damage, and their safety is assessed. To assess the safety of buildings, the city departments of public safety (or their equivalents) dispatch inspectors to the field to “walk through” each building and write a report on the observed damage and safety concerns to its occupants. On the ba-

sis of such assessments, a color tag can be assigned to the building: (1) green if the structure is safe, (2) yellow if it has been damaged and needs to be evacuated, but is safe for the occupants to return to retrieve their belongings, and (3) red if it has been damaged to a degree that it is unsafe for the occupants to return to the structure [1]. When the affected area is relatively large, such inspection takes time (several weeks or longer), and the tagging is often first preliminary, to be revised at a later time after a preliminary inspection of all buildings has been completed. Such walk-in inspections can detect only damage that is visible, and there is always considerable subjectivity in the assessments. The major problem with such inspections is however the timeliness, as aftershocks following the earthquakes can further damage a structure that has survived the main event but is weakened, and endanger the occupants. Another problem is the loss of function of a structure that may be safe, until a more detailed inspection and assessment is possible. This is particularly important for critical facilities, such as hospitals, as well as for major businesses, such as banks, for which interruption of work can cause major financial losses. Without a doubt, the ability to detect damage in structures early, as it occurs or soon after the earthquake, using some structural health monitoring system, and assess the state of safety of the structure before physical inspection is possible, can benefit society immensely. Ideally, based on instrumental data, such systems would be able to detect also hidden damage that is not visible to the naked eye. There would be benefit even when the damage is obvious, if that information is available immediately after the earthquake. To be effective, however, such systems must be sensitive enough to detect at least the significant damage, and also be accurate enough, to avoid false alarms and unnecessary and costly service interruption.

The objective of this article is to review the basic principles on which such systems operate, and to present some illustrative examples of several robust methods applied to full-scale buildings. This is followed by a discussion of remaining critical issues and directions for future research, in the view of the author.

Literature Review

Earthquake Damage Detection in Structural Health Monitoring Research

Earthquake damage detection in civil structures, such as buildings and bridges, is closely related to structural health monitoring of structures such as light aerospace structures, rotating machinery and offshore platforms, for example, that are of concern to other disciplines. A review

of recent developments in this broader field, as applied to civil and mechanical systems, can be found in Chang et al. [7] and Liu et al. [31]. The earliest, and still the most popular methods for civil structures are those that use data from vibrational sensors, and detect changes in the vibrational characteristics of the structure – frequencies of vibration and mode shapes. Detailed reviews of vibrational methods in the general area of structural health monitoring can be found in a report by Doebling et al. [11], its shorter version as a journal paper [10], and a follow up report by Sohn et al. [40]. Another recent review of the vibrational methods can be found in Carden and Fanning [4].

These detailed reviews conclude that the currently available vibrational methods can determine if the structure has been damaged, but cannot indicate precisely the location of the damage, and are therefore referred to as *global*. Most vibrational methods monitor changes in the *modal* properties of the structures (modal frequencies and mode shapes). The stated difficulties associated with these methods include: (1) the presence of other factors than damage that produce similar effects on the monitored parameters not easy to isolate (e.g. the effects of soil-structure interaction on the measured frequencies of vibration, as well as environmental influences such as temperature and rain; [8,46,47]); (2) the redundancy of the civil engineering structures, which results in low sensitivity of the method (i.e. small change of the overall stiffness and consequently of the measured frequencies) when the damage is localized; and (3) dependence on detailed prior analytical models and/or prior test data for the detection and location of damage (supervised learning), which may not be readily available for a structure, may be outdated, and even when available represent only an idealization of the real structure [7,11]. Further critical issues identified are (4) the scarcity of objective comparisons of different procedures applied to a common data set, and (5) the number and location of sensors (techniques to be seriously considered for implementation in the field should demonstrate that they can perform well for small numbers of measurements). Finally, Doebling et al. [11] conclude that “while sufficient evidence exists to promote the use of measured vibration data for the detection of damage in structures, using both forced-response testing and long-term monitoring of ambient signals, the research needs to be more focused on the specific applications and industries that would benefit from this technology... Additionally, research should be focused more on testing of real structures in their operating environment, rather than laboratory tests of representative structures.”

In the follow up review, Sohn et al. [40] mention as outstanding problems: The reliance on analytical models

to obtain the structural parameters from the data, not only in methods involving direct inversion, but also in those that use neural networks; and that the damage sensitive features are also sensitive to changes of the environmental and operational conditions of the structures. They mention as one of the most significant improvements since the previous review [11] the signal processing methods that do not rely on detailed analytic models, such as novelty/outlier analysis, statistical process control charts, and simple hypothesis testing (unsupervised learning), shown to be very effective to identify the onset of damage growth, and the presence of damage but not the damage type. In this article, one such method – based on detection of novelties using wavelets – is reviewed and illustrated. Another significant advancement is the availability of more affordable MEMS sensors, as well as fiber optics, and piezoceramic sensors, and of wireless data communication technology.

In structural health monitoring literature, the vibrational methods are referred to as *global*, due to the relatively small number of sensors typically installed in structures, and can detect only significant damage [11,40]. The cost of seismic monitoring systems is still high, and trade-offs have to be made between the detail of the instrumentation of a particular structure and the number of structures that are instrumented. The truly *local* methods are those for nondestructive testing (NDT) of materials, which can detect the location of cracks or some other defects in a structural member. These methods typically use: (1) ultrasonic waves, which are attenuated quickly along the wave path, (2) need an actuator to create such waves, and (3) require direct access to the structural member, usually not readily available. Consequently, they are used to detect the location of the damage in a particular structural member, known or suspected to have been damaged, but are too costly and impractical for structural health monitoring of an entire structure [7]. To make a difference for society, structural health monitoring and early warning systems have to be reasonably priced so that they can be installed in many structures.

Earthquake Damage Detection in Earthquake Engineering Research

In the earthquake engineering research, earthquake damage detection emerges from system identification studies of full-scale structures (typically involving identification of their frequencies of vibration and damping) from ambient and forced vibration test data, or earthquake records. Consequently, it is *data driven*, in contrast to the structural health monitoring research, which focuses on methodolo-

gies, validated mostly on “clean” numerically simulated data, and sometimes on laboratory data or small amplitude full-scale data. In the US, the earliest system identification studies from full-scale data follow the first deployment of strong motion instruments in structures [3], and continue through the 1960s [9,19,70]. More sophisticated studies from the system identification point-of-view using earthquake response data appear in the 1970s, following the San Fernando, California earthquake of 1971, which produced strong motion records in many buildings in the Los Angeles metropolitan area [66,67,68,69]. A significant finding of these studies is that the building frequencies of actual structures vary significantly as a function of the level of the response. The variation is such that the fundamental frequency decreases during the largest shaking, but recovers afterwards during the remaining smaller amplitude shaking, or during subsequent shaking from aftershocks or small amplitude tests. The recovery may be partial or complete, and a large reduction of frequency of vibration during the earthquake is not always associated with visible damage. This is an important fact, as the decrease of the fundamental frequency of vibration is used as one of the global indicators of damage in structural health monitoring research, and also because many sophisticated structural identification methods are based on the assumption of stationarity and time invariance of the response.

Further, system identification studies of structures using earthquake records, considering the effects of the interaction of the structural vibrations with the vibration of the surrounding soil, appear in the 1960 and 1970s. The most detailed such full-scale studies are probably those of the Millikan library in Pasadena [33,34,71]. Understanding and consideration of the effects of soil-structure interaction in system identification and health monitoring of structures is of *fundamental importance* for the development of reliable methodologies, as this phenomenon is an integral part of the seismic response, and affects the estimation of both the frequencies of vibration and the inter-story drift, both used to infer about the state of damage. Nevertheless, these effects are typically ignored in structural health monitoring research. A detailed literature review on full-scale studies of soil-structure interaction can be found in Trifunac et al. [63], and a discussion of critical issues in recording and interpreting earthquake response of full-scale structures can be found in Trifunac and Todorovska [60,61].

Damage and Damage-Sensitive Features

The damage of a structure can be described by the following five states: (1) no damage, (2) repairable (light and

moderate) damage, (3) irreparable damage, (4) extreme damage, and (5) collapse [14].

Damage is associated with large deformations of the structural elements (usually expressed via the inter-story drift), which cause yielding of the structural steel or steel reinforcement and cracking of the structural concrete. Also, damage causes changes of the structural vibrational characteristics (frequencies of vibration), and wave propagation characteristics (wave velocities/travel times). This section presents the rationale for damage detection algorithms based on monitoring such changes. The concepts are illustrated on a simple soil-structure interaction model.

Structural Models and Identification

Structure as an Oscillator

From an elementary vibrational viewpoint, a structure responds to earthquake shaking as an oscillator characterized by its frequencies of vibration. The fixed-base frequencies are those of free vibration of the structure on *rigid* ground. They are the eigenvalues of a boundary value problem, and the associated eigenfunctions are referred to as mode shapes in structural engineering. The fixed-base frequencies depend only on the properties of the structure, i. e. on the structural stiffness and mass, while their dependence on the structural damping is small for most structures, which are lightly damped. In the linear range, the response of an n -degree-of-freedom system to earthquake shaking is a superposition of the modal responses. The contribution of the fundamental mode is usually the largest, and in engineering design structures are often represented by an equivalent single degree-of-freedom oscillator. For a single degree-of-freedom oscillator, the natural frequency of vibration is

$$\omega_1 = \sqrt{k/m}, \quad (1)$$

where k is its stiffness and m its mass. The frequency of such an oscillator is affected little by typical fluctuations of the mass due to variations in the life load of the structure, and is mostly affected by changes in the stiffness. Damage would cause loss of stiffness, and consequently reduction of the fixed-base frequency of vibration. If $\omega_{1,\text{ref}}$ is a reference frequency corresponding to reference stiffness k_{ref} , then for the damaged structure

$$(\omega_1/\omega_{1,\text{ref}})^2 = k/k_{\text{ref}}. \quad (2)$$

As the fixed-base frequency depends on the *overall* stiffness of the structure, it is by definition a *global* property, and would not change much due to localized damage of civil structures, which are designed to be highly redundant. One advantage of detecting damage by monitor-

ing changes in fixed-base frequency of vibration is that, in the ideal case when the ground is practically rigid (as compared to the structure), and the excitation is relatively broadband, the fixed-base frequency can be determined using only one sensor, on the roof, as the frequency of the peak of the Fourier transform of the roof response. The availability of recorded response at ground level would produce a more accurate estimate, as a transfer-function can be computed between the roof and ground level response motion. Changes in the frequency versus time can be estimated from the Fourier transform in moving windows in time.

Buildings are founded on soil, which is flexible and deforms under the action of forces from the incident waves and from the vibrating structure. Even if rigid, a structure founded on soft soil will vibrate, with the soil acting as a spring. The soil adds both *flexibility* and *dissipation mechanism* to the vibrations of the structure and soil, which act as a coupled system. Two sources of dissipation are (1) scattering of the incident waves from the foundation and (2) radiation of energy into the soil (through vibration of the foundation, which acts as a source of waves radiated in semi-infinite medium). The third source of dissipation is in the structure, and includes a distribution of frictional sources, and hysteretic damping during nonlinear response. The soil-structure system has its own resonant frequencies and “damping”, which is a combination of the contributions from the structure and from the soil. The fundamental frequency of the system is always lower than the fundamental fixed-base frequency of the structure, but the associated system damping can be larger or smaller than the damping of the structure alone, depending on the radiation damping and relative stiffness of the structure with respect to the soil.

In conclusion, the difficulties with Fourier-type analyses for identification of the building frequencies are that these give the resonant frequencies and equivalent damping of the *system*, which depend on the soil, and that they are global properties. Also, there is no knowledge base of changes in such frequencies (for different types of structures and different types of soils) related to different degrees of damage.

Structure as a Wave Guide

Alternatively, the seismic response can be represented as a superposition of waves that propagate through the structure, reflect from its exterior and interior boundaries and interfere [22,39,41,48,49,50,56,57]. Loss of stiffness due to local damage would cause delays in the wave propagation through the damaged part, which could be detected using

seismic response data recorded on each side of the damaged area, along the wave path. A change in wave travel time would depend *only on the changes of the physical properties between the sensors*. Hence, the wave methods are more sensitive to local damage than the modal methods, and should be able to point out the location of damage with a relatively small number of sensors. Additionally, the local changes in travel time are not sensitive to the effects of soil-structure interaction (as demonstrated in [44,45]), which is a major obstacle for the modal methods based on detecting changes in the structural frequencies.

The basic principles of the method are as follows. It is based on D’Alambert’s solution of the wave equation, and representation of the structural response as a superposition of waves traveling through the structure. In contrast, the modal methods are based on representation in the Fourier domain, as superposition of modes of vibration.

The wave travel time between two points

$$\tau = d/V_s, \quad (3)$$

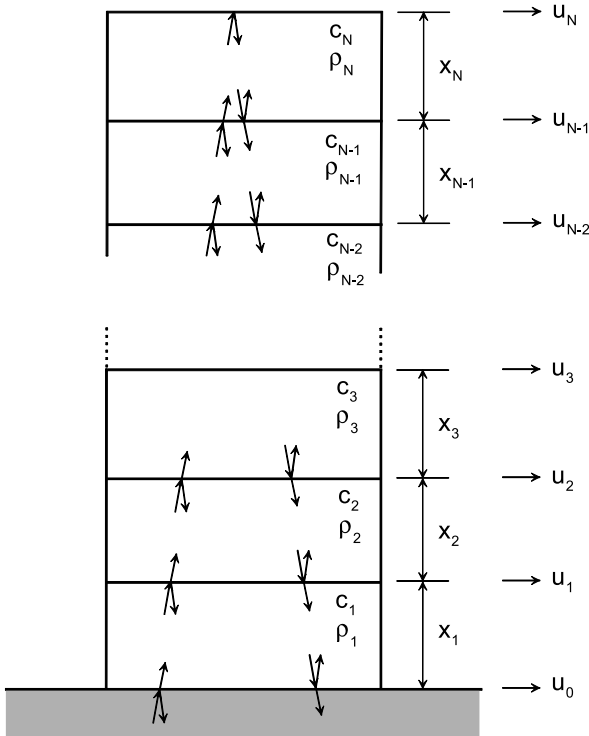
where d is the distance traveled and V_s is the equivalent shear wave velocity in the part of the building between the two sensors. The latter is related to the rigidity via

$$V_s = \sqrt{\mu/\rho}, \quad (4)$$

where μ is the shear modulus and ρ is the density. Hence, reduction of rigidity due to damage will produce a reduction of the equivalent shear wave velocity, which will produce an increase in the pulse travel time, relative to the travel time for the undamaged state. Let μ_{ref} be reference rigidity, and $V_{s,\text{ref}}$ and τ_{ref} be the corresponding shear wave velocity and wave travel time. Then their changes are related as follows

$$\frac{\tau}{\tau_{\text{ref}}} = \frac{1}{V_s/V_{s,\text{ref}}} = \frac{1}{\sqrt{\mu/\mu_{\text{ref}}}}. \quad (5)$$

Global changes can also be detected by monitoring the *total* wave travel time from the base to the roof of a building. Let τ_{tot} be the travel time of seismic waves from the point of fixity (ground level) to the roof. Then the building fundamental fixed-base frequency $f_1 = 1/(4\tau_{\text{tot}})$ assuming that the building as a whole deforms like a shear beam. Based on this relation, f_1 can be estimated using data from only two horizontal sensors. While the goodness of this approximation of f_1 may vary from one building to another, the changes in $f_1 = 1/(4\tau_{\text{tot}})$ will still depend *only* on changes in the building itself, and not on changes in the soil, and monitoring of changes in such an estimate of f_1 can be used as a global indicator of damage in a building [44,45].



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 1
Layered building model

Figure 1 shows a conceptual model for the analysis, in which the building is a horizontally layered medium, with the interfaces between layers at the floor slabs. For vertically incident waves, or/and a narrow building, the layered medium will be traversed by waves propagating upward and downward. Let the excitation be a pulse. At each interface, an incident pulse will be split into a reflected pulse, and a transmitted pulse, and at the roof total reflection will occur. The transmission and reflection coefficients will depend on the impedance contrast between the layers, in particular on the shear wave velocities, which will change due to loss of stiffness caused by damage. Because of reflections and material damping, an incident wave pulse will attenuate as it propagates through the structure, and will be also modified in a dispersive medium. This is schematically illustrated in Fig. 1. The total wave motion propagating upward in a layer will be a superposition of all the pulses, those from direct incidence and those from different generations of reflections. The same applies for the pulses propagating downward. The downward propagating pulses that are reflected back into the building, from the interface with the soil, will interfere with the newly incident pulses just transmitted into the building. Eventu-

ally, constructive interference will occur, and the standing waves will be formed, which are the fixed-base modes of vibration of the building.

The wave travel times can be detected by tracing the propagation of a pulse. Such a pulse can be created by signal processing of recorded earthquake response data, i. e. by deconvolution of the recorded response, which results in the *system impulse response functions*. These can be obtained by computing the transfer-functions between the motion at a particular level and the reference motion, and then computing inverse Fourier transform. The location of the virtual source would coincide with the location of the sensor that recorded the reference motion. Let $u_{\text{ref}}(t)$ be the reference motion, $u_i(t)$ be the motion at level i . Then the impulse response at that level, $h_i(t)$, can be computed as

$$h_i(t) = \text{FT}^{-1} \left\{ \frac{\hat{u}_i(\omega) \bar{\hat{u}}_{\text{ref}}(\omega)}{|\hat{u}_{\text{ref}}(\omega)|^2 + \varepsilon} \right\}, \quad (6)$$

where the hat symbol indicates Fourier transform, the bar indicates complex conjugate, and ε is a regularization parameter, used to avoid dividing by a very small number [39]. At the reference level, the transfer-function is unity, and its inverse is a Dirac delta function.

Proof-of-concept applications to two buildings damaged by earthquakes, and to an analytical model of a building-foundation-soil system showed that the method (1) is robust when applied to damaging levels of earthquake response data, (2) is not sensitive to the effects of soil-structure interaction, and (3) is local in nature (i. e. gave results consistent with the spatial distribution and degree of the observed damage) [44,45,53,54]. The damaged buildings are the former Imperial County Services Building – a 6-story RC structure in El Centro, California, damaged by the 1979 Imperial County earthquake and later demolished [52,54], and the 7-story RC building in Van Nuys, damaged by both the 1971 San Fernando and the 1994 Northridge earthquakes [53,62]. Another application is to a building in Banja Luka in former Yugoslavia, using records of 20 earthquakes, one of which led to levels of response that might have caused structural damage, but no damage was reported following a detailed inspection [65]. This study was aimed at learning about the *threshold change* in the building fixed-base frequency, estimated from wave travel time, associated with damage.

While this method is local, its spatial resolution is limited by the number of sensors. A minimum of two sensors (at the base and at the roof) are required to determine if the structure has been damaged, and additional sensors at the intermediate floors would help point out the part of

the structure that has been damaged. For example, one additional sensor between these two would help identify if the damage has been in the part of the structure above or beyond that sensor.

There have been only a few publications in the literature on wave propagation methods for structural health monitoring and damage detection in civil structures other than the NDT methods [20,35,37,41,53,54,64]. Similar wave travel time analyses (using deconvolution or the NIOM method) of buildings that have not been damaged include Kawakami and Oyunchimeg [24,25], Snieder and Şafak [39], Kohler et al. [26], and Todorovska [44,45]. These studies show that the wave travel times reflect well the characteristics of the buildings studied. A recent review can be found in [53,54].

In conclusion, the advantages of this wave method are its local nature achieved with a relatively small number of sensors, its insensitivity to the effects of soil-structure interaction, and the ability to estimate the structural fixed-base frequency using data from only two sensors (one at the base and one at the roof), which will extend the usability of old data. An outstanding issue to its implementation is the lack of a knowledge base relating changes in wave travel times (and fixed-base frequency) with different levels of damage for different types of structures.

Inter-Story Drift

Structural damage of a building under seismic loads occurs primarily due to large *lateral* deformations of its columns and shear walls, as they are by design much stiffer in the vertical direction to carry the static gravity loads. A measure of the lateral deformations is the inter-story drift. The inter-story drift is also a good indicator of the damage to the architectural (nonstructural) components (partition walls, facade, windows, etc.), which can be costly. As the value of the structure is only about 10–25% of the total construction cost of a building, the damage to the nonstructural components represents a significant portion of the total repair cost following an earthquake. For these reasons, the inter-story drift is one of the performance parameters considered in design. It is important to note that the structural and nonstructural damage are related only to the drift caused by *deformation* of the structure, and not by the drift caused by *rigid body motion*.

The level of structural damage (to a particular element and to the structure as a whole) associated with a particular level of inter-story drift varies depending on the type of structure, height and ductility, among other factors, and is still *not a completely resolved issue in structural engi-*

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Table 1

Drift ratios (in %) associated with various damage levels (based on [14])

State of damage	Ductile MRF	Nonductile MRF
No damage	< 0.2	< 0.1
Repairable damage		
Light	0.4	0.2
Moderate	< 1.0	< 0.5
Irreparable damage	> 1.0	> 0.5
Severe damage, life safe, partial collapse	1.8	0.8
Collapse	> 3.0	> 1.0

neering [14]. To illustrate this correlation, Table 1 shows some values of drift associated with different levels of damage (simplified from [14]) for ductile and nonductile moment resisting frames (MRF), and based on experimental data, field observations and measurements and theoretical analyses. (Ductile are those structures that can undergo large nonlinear deformations before failure as opposed to the nonductile ones, which experience quick brittle failure soon after exceeding the linear range of response). It can be seen from Table 1 that, roughly, inter-story drift > 1% for ductile and > 0.5% for nonductile moment resisting frames causes damage beyond repair, and drift > 3% and > 1% for the same type of frames is significant for life safety.

Drift-based assessment of the state of damage of a building following an earthquake would require: (1) measurement of the drift during the earthquake shaking, and (2) knowledge base of values of drift associated with different states of damage for the particular structure. The accuracy of the assessment would depend on the accuracy of both the measurements and knowledge base, as discussed in the following.

The drift is commonly estimated from the difference of displacements obtained by double integration of recorded velocities in the structure [28]. While in the past these calculations were performed by specialists, after the data had been manually collected, at present, such calculations can be done in near real time either using telemetry or at the site by “client” software supplied by the instrument manufacturer. Such estimates of drift however are limited by: (1) the inability to estimate reliably the *static* component of the drift associated with permanent deformations (i. e. the drift at $\omega \rightarrow 0$), which is not negligible for structures experiencing large deformations in the nonlinear range of response, when damage occurs, and (2) the inability to separate the drift due to deformation of the structure

(which is directly related to damage) from the drift due to rigid body rocking because of inadequate instrumentation.

The inability to estimate reliably the static part of the displacement (and drift) is due to the fact that the traditional (translational) sensors are sensitive also to rotational motions of their support [16,59], which produce errors in the recorded translations and the integrated displacements mimicking permanent displacement [16]. This problem can be solved, by deploying sensors recording all six components of motion (three translations and three rotations) and performing appropriate instrument correction. Such future deployments and their assessment are of interest to and have been advocated by the International Working Group on Rotational Seismology [29].

The *dynamic* (at $\omega > 0$) drift due to deformation of the structure only is not simple to estimate, especially for structures on soft soil, with significant rocking response of their foundation. The rocking motions of the foundation are due to the wave nature of the incident seismic waves, and also due to feedback forces from the structure acting on the soil. The foundation rocking results in relative horizontal displacement between two floors and is not related to damage. Such excessive relative displacements, can affect the stability of the structure, which may collapse before yielding occurs in its members, but that is out of the scope of this article. The *average dynamic* floor rocking can be calculated from the difference of vertical motions recorded by two sensors on that floor, assuming the floor slab is rigid, but such sensor configurations are not routinely installed even in recent denser deployments in buildings. If the building foundation is fairly rigid, the rigid body rocking of the structure can be estimated from two vertical sensors at foundation level. Unfortunately, even such data is lacking for most of the significant earthquake records in buildings, and even in recent dense deployments (e. g. in [6]). It is noted that vertical sensors are also less sensitive to rotation of their support and to cross-axis motion [15,43].

It should be noted here that permanent displacements can be measured directly using GPS (Global Positioning System), and there have been such deployments in long period structures [5]. While GPS measurements are not contaminated by rotation, they are limited by the fact that what is measured are only the roof *absolute* displacements, which makes it impossible to separate the displacement due to deformation of the structure from the rigid body horizontal translation and rocking. The other two limitations in the presently available systems are the small sampling rate (10–20 Hz) and the limited resolution of GPS for civilian applications (± 1 cm horizontally and ± 2 cm vertically; [5]).

Damage estimation algorithms based on published damage versus drift relationships (e. g. in [1]) started to be implemented by manufacturers of strong motion instruments in structural seismic monitoring systems but there is no data yet of their performance. Despite errors in the assessment resulting from the mentioned difficulties, such algorithms are robust when applied to earthquake data and can be useful within a suite of methods.

Matrices like the one in Table 1 [14] can serve as a knowledge base in assessing the class of damage state for a given maximum drift reached. Such matrices are associated with scatter, due to the variability from one structure to another within the same class. Another source of scatter is the source of the data. Because of the limited amount of full-scale earthquake response data, information for such relationships is complemented by laboratory data (e. g. pushover tests). While the drift in the former is the total drift, which includes the drift due to rigid body motion, the drift in the latter is only due to deformation of the structural elements.

In conclusion, outstanding issues in measuring the drifts are: (1) separation of the drift due to deformation of the structure only, and (2) estimation of the static component of the drift. It may be possible to resolve these issues by deploying six degrees-of-freedom sensors. An outstanding issue in the knowledge base is more accurate drift versus damage state relations for specific buildings.

System Identification Considering the Effects of Soil-Structure Interaction – Example

As mentioned earlier, both for frequency-based identification and for damage assessment based on drift, the effects of soil-structure interaction have a significant effect on the reliability of the estimation. This section presents a simple soil-structure interaction model, in which the building is represented as a shear beam. It illustrates the different contributions to the inter-story drift, the difference between fixed-base and apparent building frequencies and their relationship, and the relationship between the model fixed-base frequencies and wave travel times. More detailed analysis can be found in [44,45].

The model is shown in Fig. 2. It consists of a shear beam of height H and fundamental fixed-base frequency of vibration f_1 , representing the building, and a rigid foundation of width $2a$ embedded in elastic half-space. The excitation, in general, is an incident wave (plane P and SV or a Rayleigh wave). The motion on the surface of the half-space in the absence of any structures and excavations, acting as scatterers, is commonly referred to as “free-field.” The effective motion at the base of the

due to translation of the base only, and $u_\varphi(\xi)$, which is due to rotation of the base only, where

$$u_\Delta(\xi) = \Delta \frac{\cos k_S(H - \xi)}{\cos k_S H} \quad (10)$$

$$u_\varphi(\xi) = \frac{\varphi}{k_S} \frac{\sin k_S \xi}{\cos k_S H}, \quad (11)$$

where $k_S = \omega/V_S$ and $V_S = \sqrt{\mu_b/\rho_b}$ is the shear wave velocity in the building. Equations (10) and (11), reflecting the interference conditions in the building, imply fundamental fixed-base frequency of the structure $f_1 = V_S/(4H)$ and overtones at $f_n = (2n - 1) V_S/(4H)$, $n > 1$. If τ is the time it takes for a wave to propagate from the base (at $\xi = 0$) to the top (at $\xi = H$), the interference conditions in the shear beam imply

$$f_1 = 1/(4\tau). \quad (12)$$

Let us now consider the frequencies of vibration. If the building did not deform, the foundation and the building would oscillate freely as a rigid body with frequency f_{RB} such that

$$\frac{1}{f_{RB}^2} = \frac{1}{f_H^2} + \frac{1}{f_R^2}, \quad (13)$$

where f_H and f_R , referred to as the horizontal and rocking foundation frequency, depend on the stiffness of the foundation and on the system mass [33]. If the building is flexible and would freely vibrate on a fixed base with fundamental frequency f_1 , on flexible soil it would freely vibrate with fundamental frequency f_{sys} , which is the soil-structure system frequency, and is a result of the coupling between the vibration of the building and the vibrations of the foundation. The following relationship holds approximately

$$\frac{1}{f_{sys}^2} = \frac{1}{f_{RB}^2} + \frac{1}{f_1^2}. \quad (14)$$

This relationship implies that $f_{sys} < \min(f_1, f_{RB})$, i. e. f_{sys} is always lower than both f_1 and f_{RB} , and that if f_1 and f_{RB} differ significantly, then f_{sys} would be closer to the smaller one of them. How much f_1 would differ from f_{sys} would depend on the *relative* stiffness of the soil compared to the building. The energy of the response of vibrating systems is concentrated around their resonant frequencies, which are measured from the frequency of the peaks of the corresponding transfer-functions. The energy of the building roof response (absolute and relative) will be concentrated around $f = f_{sys}$.

Of interest is how to estimate the relevant quantities from recorded response during an earthquake. If the building foundation is fairly rigid, and there are at least two appropriately located vertical sensors to compute the foundation rocking φ (average value), then $u^{rel}(\xi)$ can be computed. To measure f_{sys} , the driving motion Δ_{inp} is also needed, so that the transfer function between the building response and Δ_{inp} can be computed. Motion from a nearby free-field site can be used for that purpose, but such sites are often not available, and truly free-field sites practically do not exist in urban areas. Also, for most instrumented buildings, the foundation rocking cannot be estimated because of the lack of two vertical sensors even under the ideal conditions that the foundation behaves as rigid.

Consequently, in reality, for most instrumented buildings, the true relative roof displacement cannot be estimated from the recorded data, but only the *apparent* relative displacement

$$\begin{aligned} u_{app}^{rel}(H) &= u(H) - \Delta \\ &= u^{rel}(H) + \varphi H \end{aligned} \quad (15)$$

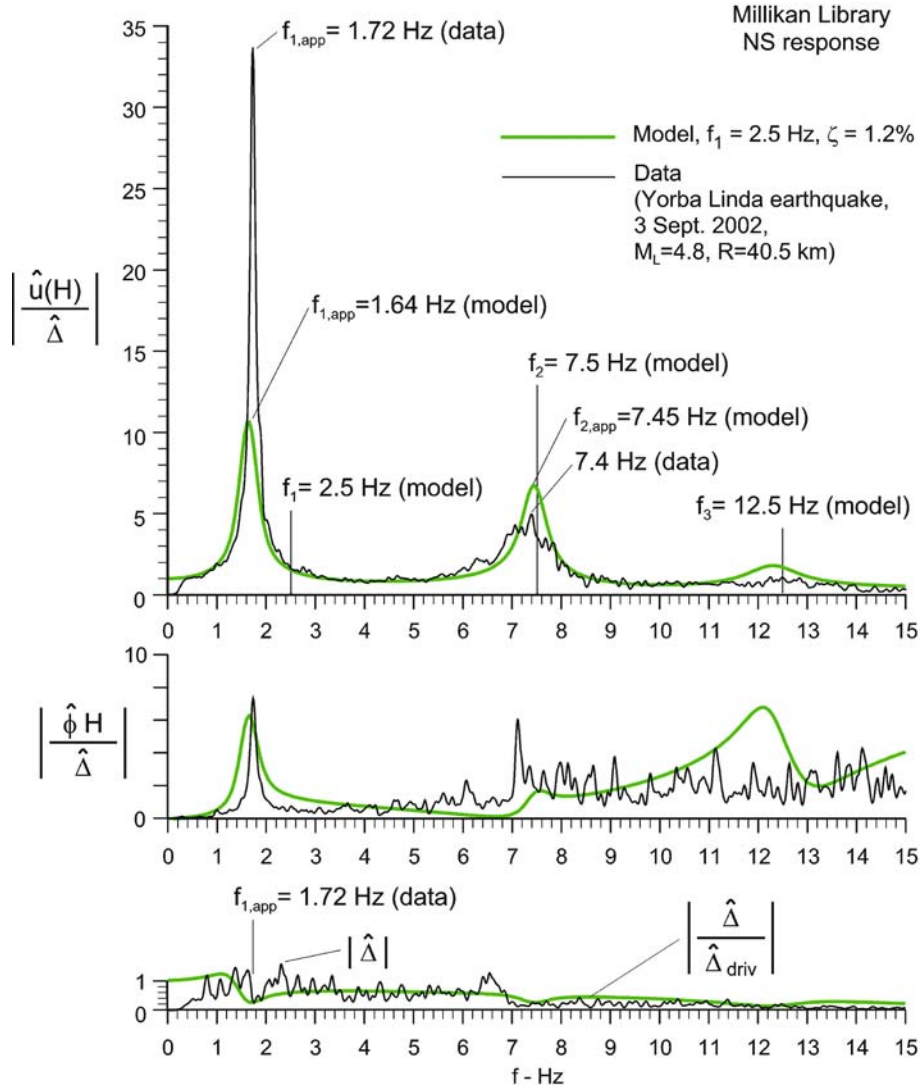
which includes the contribution of the roof displacement due to rigid body rotation, and only the transfer-function $|u_{app}^{rel}(H)/\Delta|$ can be computed, the peak of which gives the *apparent* building frequency f_{app} , which is different from both the fixed base frequency and the system frequency.

What is of interest for structural health monitoring is that the energy of the roof response will be concentrated around $f = f_{sys}$, not around $f = f_1$. It is also significant that the damage will depend on $u^{rel}(\xi)$, while what is usually measured is $u^{rel}(\xi) + \varphi H$.

Figure 3 (redrawn from [45]) shows a comparison of model and measured transfer-functions for a model of the NS response of Millikan library. The model has $f_1 = 2.5$ Hz, height $H = 44$ m, shear wave velocity in the soil 300 m/s, and Poisson ratio 0.333, while the data are from the Yorba Linda earthquake of 2002. Figure 4 shows the corresponding impulse response function for a virtual source at the ground floor. It can be seen that there is a very good qualitative agreement despite the model simplicity and roughly chosen parameters.

Novelty Detection in the Recorded Response

Novelty detection is used in data mining to detect unusual events in data. The unusual events are *outliers* deviating from the *trend*. Within the framework of multi-resolution analysis, the trends and novelties are determined by



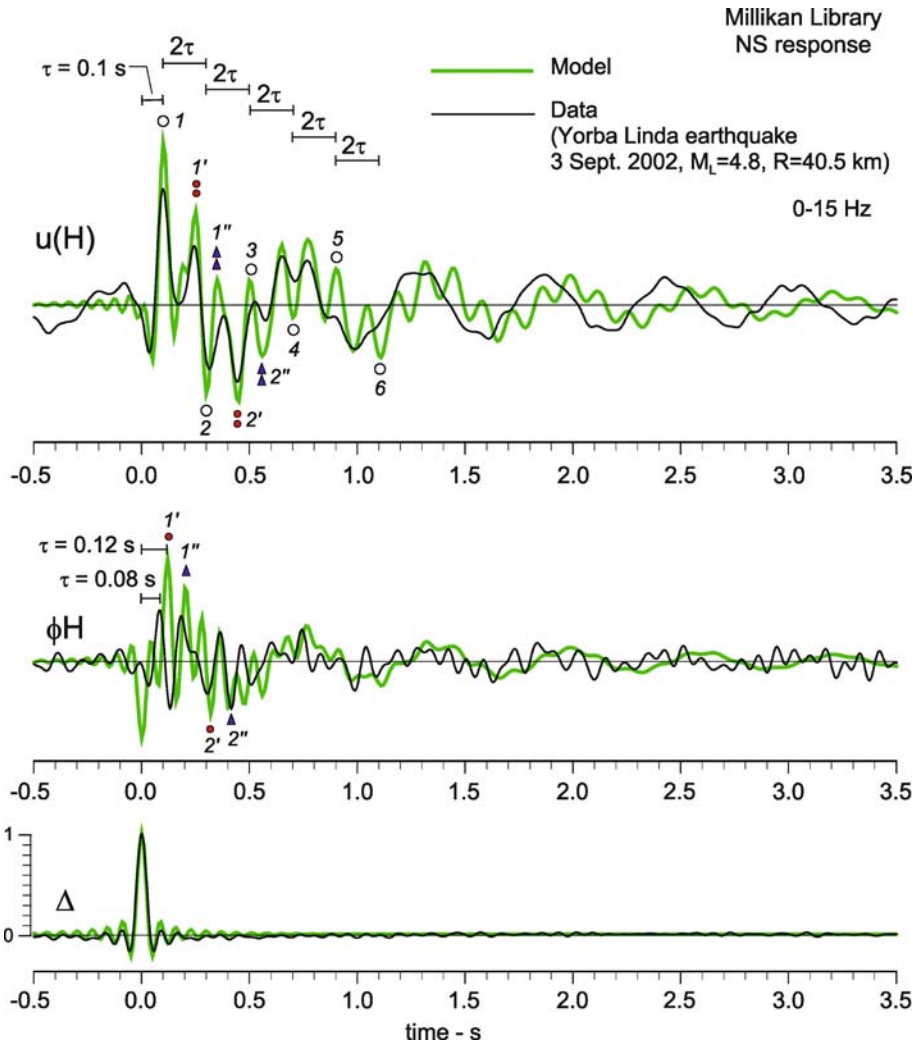
Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 3

Model (thick line) versus Yorba Linda, 2002, earthquake (thin line) NS response: transfer-functions of roof response (top), and base rocking response (middle) with respect to horizontal response of ground level. The plot in the bottom shows the model horizontal response at ground level for unit driving motion (thick line), and the Fourier spectrum of the earthquake response at ground level (thin line) on a relative scale

splitting the signal in two subbands, one smooth (low frequency) and the other one containing the detail (high frequency). By consecutively splitting the smooth subband, trends and detail are obtained at different resolution levels. If J is the last level, then there will be J detail subbands D_i , $i = 1, \dots, J$ and one smooth subband S_J . The last smooth subband can be expanded in a basis of scaling functions $\varphi_{J,k}(t)$, and each of the detail subbands – in a basis of wavelet functions $\psi_{j,k}(t)$, leading to the representation of a discrete time signal $s[n]$, $n = 1, \dots, N$

$$\begin{aligned}
 s[n] &= \sum_{j=1}^J D_j[n] + S_J[n] \\
 &= \sum_{j=1}^J \sum_{k=1}^{N/2^j} d_{j,k} \psi_{j,k}[n] + \sum_{k=1}^{N/2^J} s_{J,k} \varphi_{J,k}[n]. \quad (16)
 \end{aligned}$$

The coefficients of the expansion, $d_{j,k}$ and $s_{J,k}$, can be computed using the fast wavelet transform. The pyramid algorithm on which it is based [36] is shown in Fig. 5.



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 4

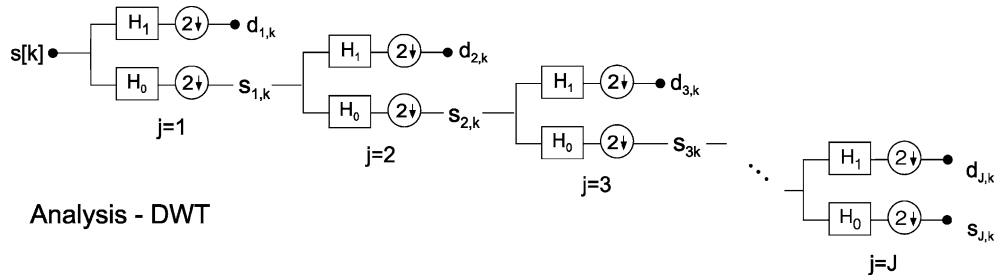
Model (*thick line*) versus Yorba Linda, 2002, earthquake (*thin line*) NS response: impulse responses of roof (*top*), and base rocking (*middle*) to input impulse at ground level (*bottom*), i.e. roof horizontal motion, foundation rocking and ground level horizontal motions deconvolved with the resultant horizontal motion at ground level

The wavelet functions $\psi_{j,k}[n]$, where j is a level and k is the time shift, are localized both in frequency and in time, and each wavelet is a projection of the signal onto the corresponding tile of the phase plane. For a wavelet basis that is orthonormal, the square of a wavelet coefficient represents the energy of the signal in the corresponding tile of the phase plane.

The damage detection method is based on the assumption that, when damage occurs and there is a sudden loss of stiffness, *there will be some abrupt change in the response that would produce novelties*. These would be seen as spikes in the time series of the square of the detail co-

efficients (e.g. $d_{1,k}^2$, $k = 1, \dots, N/2$ for the highest detail coefficients) plotted versus the central time of the corresponding wavelet. These spikes indicate high frequency energy in the response. For data with Nyquist frequency 25 Hz, the novelties can be best seen in the highest detail subband (12.5 to 25 Hz), which is away from the frequency of the first few modes of typical buildings, where the response is amplified by the structure.

Applications to numerically simulated response of simple models with postulated damage [17,18] have shown that this method *can point out very precisely the time of damage*, but the changes are detectable only if the spikes



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 5
The pyramid algorithm for the fast wavelet transform



a



b



c

Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 6
Imperial County Services (ICS) building: a view (towards north); b photographs of damage: columns F1 and F2 at the ground floor; and c column F1

in the wavelet coefficients are above the noise. Further, the magnitude of the novelties is larger if the sensor is closer to the location of the damaged member, and may be difficult to detect if the sensor is far from the location of damage. There have been only few applications to earthquake response records in buildings. Rezai et al. [38] and Hou et al. [18] have shown that there *are* novelties (spikes) in earthquake records of damaged buildings, but have not discussed and extracted other possible causes. Todorovska and Trifunac [51,55] presented a detailed analysis of the correspondence between the spatial distribution and amplitudes of the detected novelties and the observed damage for the Imperial County Services building (see illustrations in Sect. “Examples”), and also analyzed the “noise.” Their study shows that: (1) the spatial distribution and magnitudes of the novelties were generally consistent with the spatial distribution and degree of the observed damage, (2) the timing of those suggesting major damage agreed with the time of significant drops in frequency and of large inter-story drifts, and (3) were much larger in the transverse response, in which the building was stiffer.

In summary, the method of novelties is very effective in determining the time of occurrence of damage, and can reveal the spatial distribution and degree of damage if there is sufficiently dense instrumentation. Unresolved issues are how to distinguish novelties that are not caused

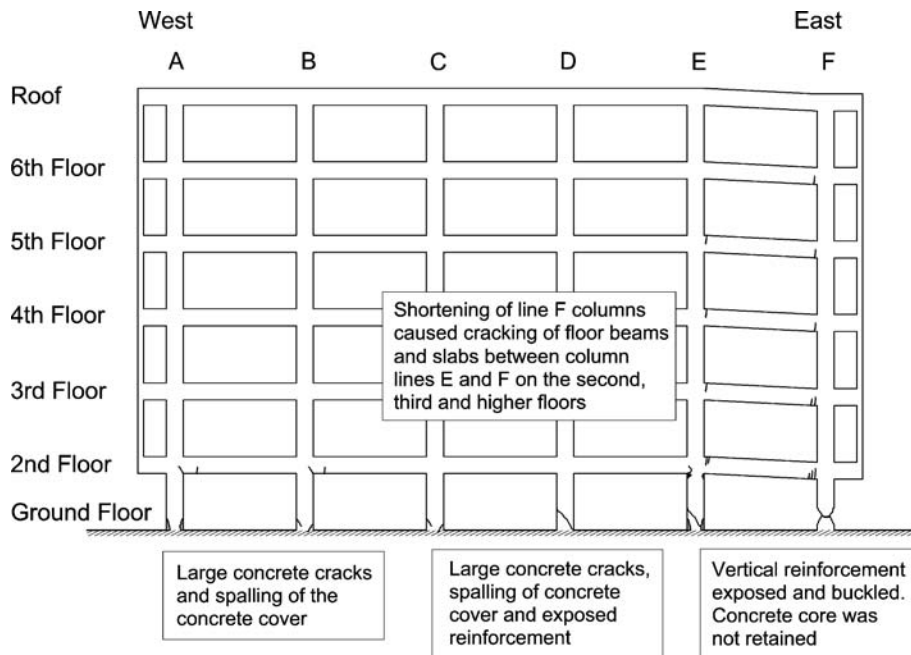
by damage, and small novelties due to larger damage far from the sensor from those due to small damage close to the sensor.

Examples

In this section, the methods previously described are illustrated for the former Imperial County Services (ICS) building – a rare example of an instrumented building damaged by an earthquake, for which description of damage and the strong motion data are available. The building is first described and the strong motion data of the Imperial Valley earthquake, which severely damaged the building.

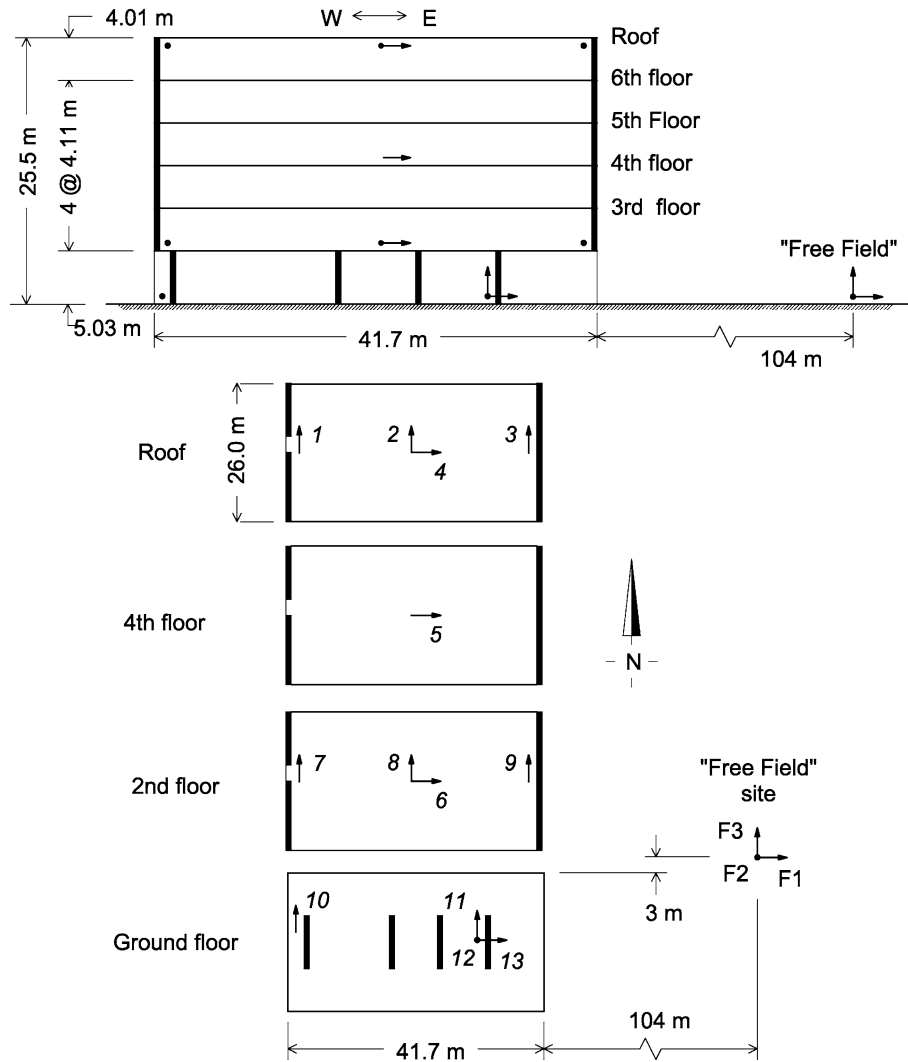
The ICS building was a 6-story reinforced concrete structure located in El Centro, California (Fig. 6a). It was designed in compliance with the 1967 Uniform Building Code, and its construction was completed in 1969. It had plan dimensions 41.70×26.02 m, height 25.48 m, and pile foundation. Up to depth of 9 m, the underlying soil consisted of soft to medium-stiff damp sandy clay with organic materials, with inter-layers of medium dense moist sand, and beneath 9 m it consisted of stiff, moist sandy clay and silty clay [27].

The building was severely damaged by the Imperial Valley earthquake of October 15, 1979 ($M = 6.6$), and was



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 7

ICS building: schematic representation of the damage following the 1979 Imperial Valley earthquake (reproduced from [27])

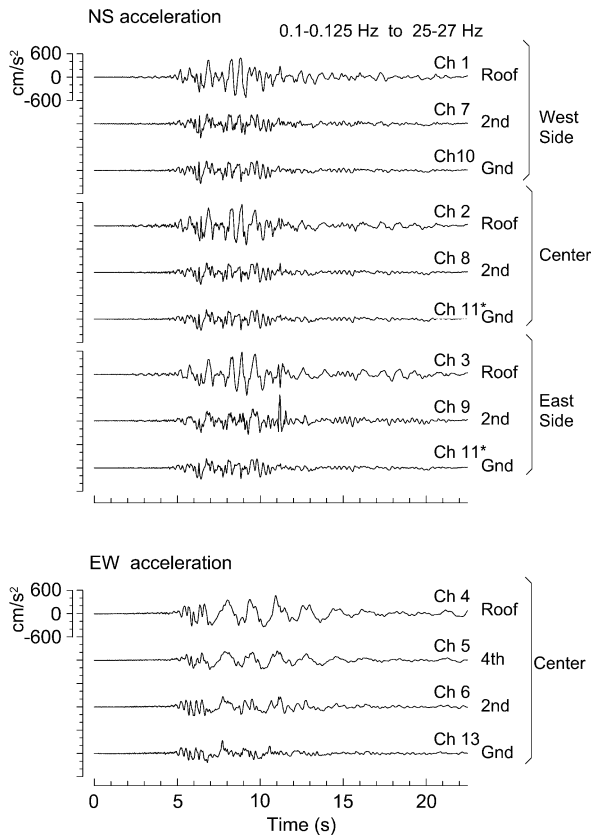


Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 8
ICS building: layout of the seismic monitoring array

later demolished (Fig. 6b,c). Figure 7 shows a schematic representation of the observed damage. The major failure occurred in the columns of frame F (at the east end of the building) at the ground floor. The vertical reinforcement was exposed and buckled, and the core concrete could not be contained, resulting in sudden failure and shortening of the columns subjected to excessive axial loads. This in turn caused an incipient vertical fall of the eastern end of the building, causing cracking of the floor beams and slabs near column line F on the second, third and higher floors. Columns in lines A, B, D, and E also suffered damage. Columns in frames A and E did not suffer as extensive damage as shortening and buckling of the reinforcement

in line F at the east side, but large concrete cracks and exposed reinforcement could be seen near the base. In the columns in interior frames B through E, visible cracks and spalling of the concrete cover were also observed [27].

The building was instrumented by a 16-channel seismic monitoring array (installed by the California Geological Survey, formerly the California Division of Mines and Geology) consisting of a 13-channel structural array of force balance accelerometers (FBA-1), with a central analog recording system, and a tri-axial SMA-1 accelerometer in the "free field," approximately 104 m east from the northeast corner of the building (Fig. 8). Figure 9 shows the accelerations (corrected) during the Impe-



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 9

ICS building: accelerations (NS and EW components) recorded during the 1979 Imperial Valley earthquake

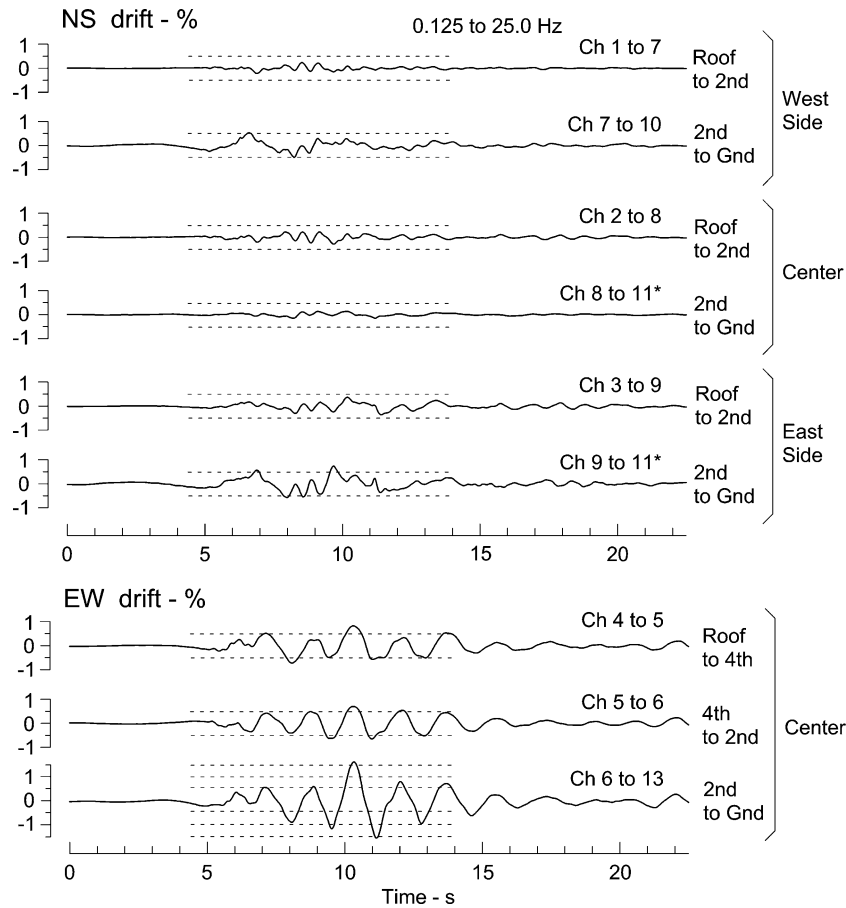
rial Valley earthquake. The peak accelerations at the roof and ground floor were 571 cm/s^2 and 339 cm/s^2 in the NS direction and 461 cm/s^2 and 331 cm/s^2 in the EW direction.

Figure 10 shows the NS (top) and EW (bottom) inter-story drifts computed from band-pass filtered displacements (between 0.1–0.125 Hz and 25–27 Hz) (redrawn from [52]). Hence, they represent only a limited view of the actual drifts – through a tapered window in the frequency domain, and a combination of the drift due to *rigid body rocking* (one of the effects of soil-structure interaction) and drift due to relative *deformation* of the building. The horizontal lines show 0.5%, 1%, and 1.5% drift levels. The plotted drifts suggest: (1) “soft” first story in both NS and EW directions, (2) larger flexibility in the EW direction, and (3) significant torsional response, probably amplified by the wave passage, and by the asymmetric distribution of stiffness in the NS direction at the soft first story (see Fig. 8). It can be seen that during the most severe shaking, the inter-story drifts exceeded 0.5% for NS

and 1.5% for EW motions, consistent with irreparable to severe damage (Table 1).

Figure 11 shows results of time frequency analysis (using Gabor transform) for the EW response (redrawn from [52]). Parts a and b show the ground floor accelerations, and the roof relative displacements (at the center of the building), both included as background information. Part c shows the skeleton (the thicker line), which is a smoothed estimate of the amplitude envelope of the estimated signal, which is the relative roof response near the first system frequency. The thin line is the actual amplitude envelope (that for the broad-band signal), determined by Hilbert transform. This plot is included to help monitor rapid changes in the amplitude of the signal and artifacts in the estimate of instantaneous frequency caused by violations of the asymptoticity condition. Part d shows the Fourier spectra of the relative roof displacement (the solid line), and of the ground floor acceleration (the dashed line, on a relative scale), both included as background information. Part e shows the variations of the system frequency as a function of amplitude of response (estimated from the ridge and skeleton of the Gabor transform), with the arrows indicating the direction of increasing time. Part f shows the variations of the system frequency versus time, estimated from the ridge of the Gabor transform. The missing segments and the dashed lines in parts e and f correspond to time intervals where the estimates cannot be obtained or are not believed to be reliable, due to rapid variations of the envelope of the amplitude, and/or very weak “signal.” The rectangle in part f with sides $2\sigma_t = 1.42 \text{ s}$ and $2\sigma_\nu = 0.22 \text{ Hz}$ illustrates the theoretical uncertainty of the estimates due to the finite resolution of the Gabor transform. In practice, the uncertainty is larger due to violations of the asymptoticity assumption. Finally, the numbered open dots (occurring at different times in parts b, c, e, and f correspond to some characteristic points in time associated with changes in amplitude or frequency, as well as a few other points in-between. It can be seen that the EW frequency dropped rapidly from $\nu \approx 0.88 \text{ Hz}$ at $t \approx 3.5 \text{ s}$ to $\nu \approx 0.67 \text{ Hz}$ at $t \approx 7 \text{ s}$ ($\Delta\nu \approx 0.21 \text{ Hz} > \sigma_\nu$; $\Delta\nu/\nu \approx 24\%$), and then continued to drop gradually to $\nu \approx 0.53 \text{ Hz}$ at $t \approx 17 \text{ s}$ ($\Delta\nu \approx 0.14 \text{ Hz} \approx \sigma_\nu$; $\Delta\nu/\nu \approx 20.9\%$).

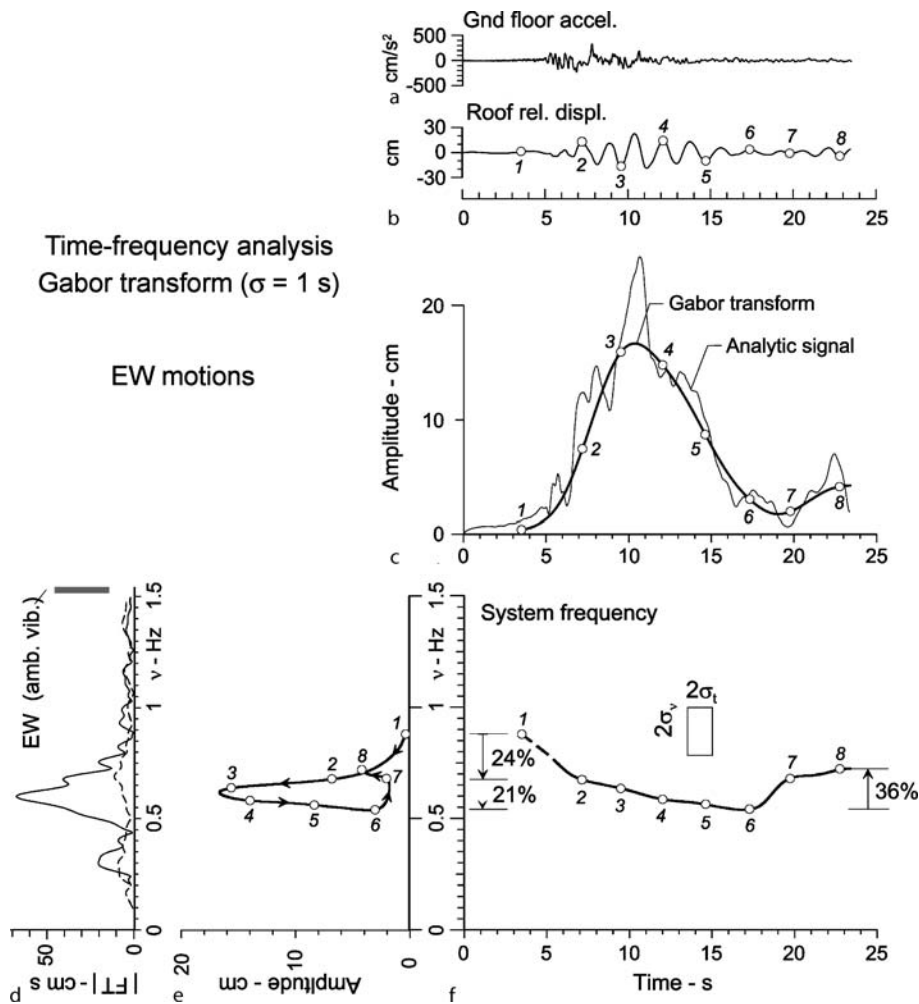
Figure 12 shows results of impulse response analysis for the EW response (redrawn from [53]). The different types of lines correspond to different time intervals of the recorded motion, before, during, and after the major damage occurred: $t < 7 \text{ s}$, $7 < t < 13 \text{ s}$, and $t > 13 \text{ s}$ (based on novelty analysis, discussed below). The plots on the left correspond to an input impulse at the ground floor, and those on the right – to an input impulse at the top. The



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 10
ICS building: inter-story drifts during the Imperial Valley earthquake

latter plots show two waves propagating downwards, one acausal (in negative time, representing the wave going up) and one causal (in positive time). The delays in the pulse arrival during the second and third time interval are obvious, and are consistent with the occurrence of damage, as determined using other methods. The wave travel times suggest, for EW motions initial wave velocities of 201 m/s through the first floor, 183 m/s between the 2nd and 4th floors, and 111 m/s between the 4th floor and roof. The velocity of an equivalent uniform shear beam is 142 m/s. Figure 13 shows the corresponding reduction of stiffness. It can be seen that, for EW motions, the reduction was the largest in the first story (80% during the second time window), but was also large in the upper stories (72% between the 2nd and 4th floors, and 60% between the 4th floor and roof). This is consistent with the spatial distribution of the observed damage (Fig. 6), which was the largest in the first story.

Figure 14 shows the results of novelty analysis, for the EW accelerations (part a) and for the NS accelerations at the east side of the building, where the most severe damage occurred (part b) (redrawn from [55]). The inter-story drifts (in %) between the corresponding stories are also shown, by a solid line for NS and by a dashed line for EW motions. Selected novelties are identified by letters. Novelties T1–T3 are believed to be caused by damage, and are seen in all channels. Novelties G1–G3 and g1–g4 originate in the ground motion, and L1–L6 are possibly caused by local damage close to the sensor, or by other causes. By far the largest novelty is T3, which has amplitude more than an order of magnitude larger than all other novelties in the NS acceleration at the 2nd floor at the east side of the building, where the most severe damage (failure of the first story columns of frame F) occurred. The timing of T3 suggests that the collapse of the columns of the first story occurred at about 11.2 s after trigger. The other two large



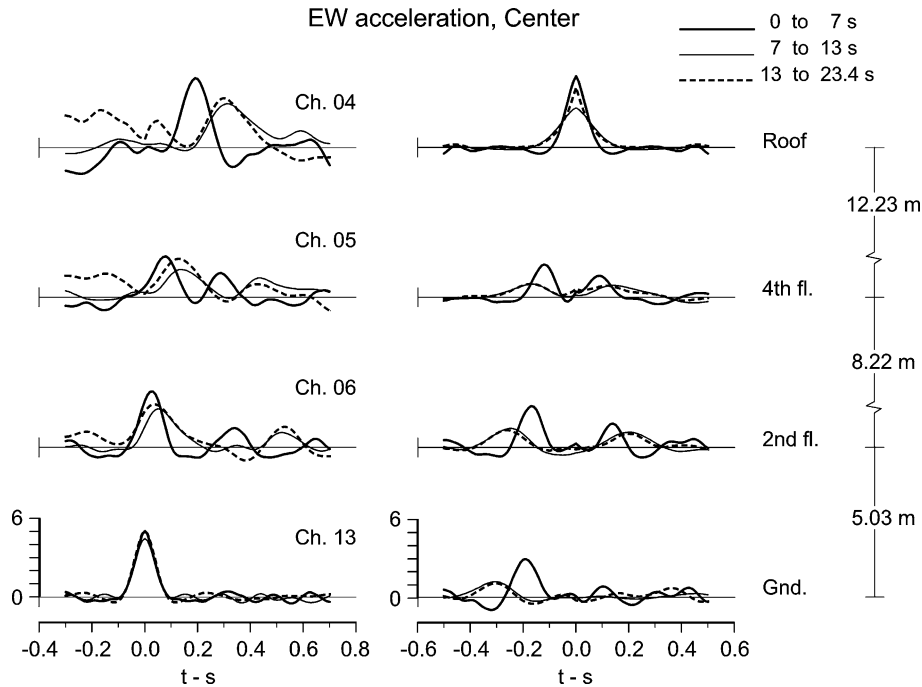
Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 11
ICS building: time frequency analysis for EW response

novelties consistent with the observed damage, T1 and T2, occurring at about 8.2 s and 9.2 s after trigger, indicating damage that weakened the structure, before the collapse of the first story columns.

Figure 15 (redrawn from [54]) shows a comparison of different values of frequency for EW motions: f_1 from wave travel times (the gray line), system frequency f_{sys} estimated from time-frequency analysis (the red line; [52], and f_1 using ETABS models [27]. T1, T2, and T3 mark the times of occurrence of major damage, as indicated by novelties in the response [51,51]. It can be seen that f_1 from wave travel times is consistent with the results of other independent studies.

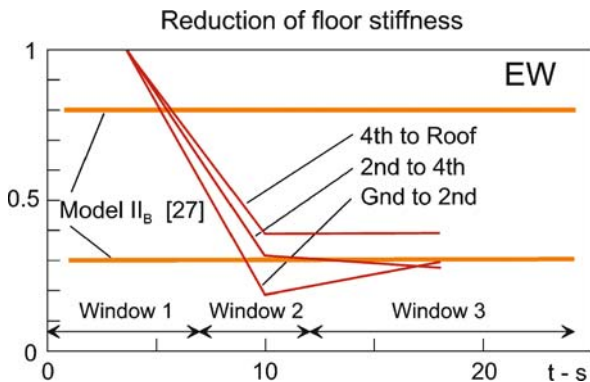
Finally, Fig. 16 shows results for another building, the Van Nuys 7-story hotel, which has been damaged by

earthquakes [53]. It shows a comparison of fixed-base frequency f_1 during 11 earthquakes estimated from wave travel times, and system frequency f_{sys} during the same earthquakes estimated by time frequency analysis (Gabor transform), as well as estimates of f_{sys} during ambient vibration tests. The analysis shows that, during the San Fernando earthquake, f_1 decreased by about 40% (relative to its value within the first 5 s from trigger), which corresponds to a decrease in the global rigidity of about 63%. During the Northridge earthquake, f_1 decreased by about 22% (relative to its value within the first 3 s from trigger), which corresponds to a decrease in the global rigidity of about 40%. The analysis also showed that, although f_{sys} was always smaller than f_1 , their difference varied, contrary to what one could expect from a linear soil-structure



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 12

ICS building: Impulse response analysis and wave travel times for EW response, and for a virtual source at the ground floor (left) and at the roof (right)



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 13

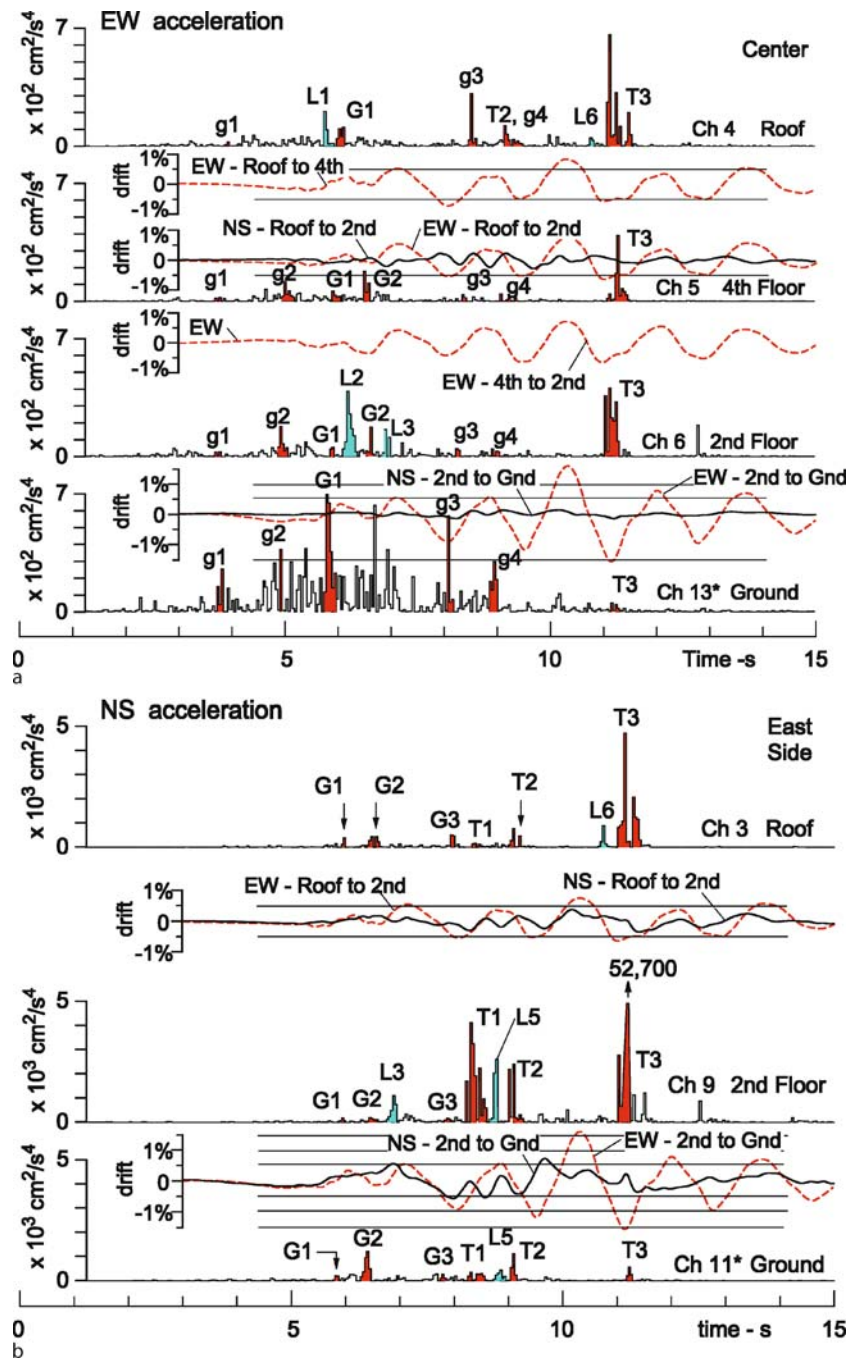
ICS building: reduction of floor stiffness versus time [54]

interaction model. It also showed that while f_{sys} was significantly lower during the Landers and Big Bear earthquakes, compared to the previous earthquakes, f_1 did not change much, which is consistent with the fact that these earthquakes (which occurred about 200 km away from the building) did not cause any damage. The study concluded that monitoring changes in f_{sys} can lead to false alarms about the occurrence of damage, and that f_1 , as estimated

from wave travel times by the proposed method, is a much more reliable estimator of damage.

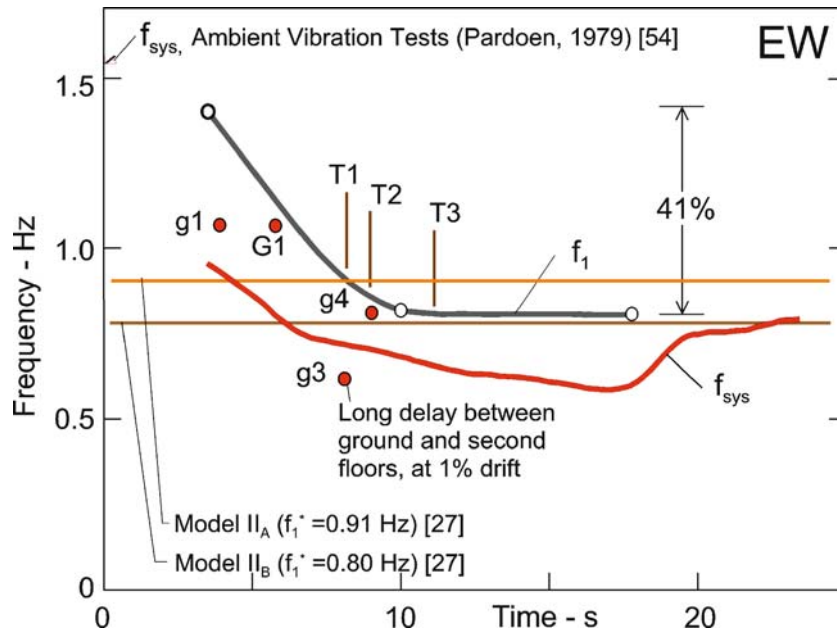
Future Directions

A successful system for earthquake damage detection and early warning would involve applications of technologies in fields other than structural mechanics and engineering, such as sensing, data communication, signal processing, artificial intelligence, and decision analysis. The end of the 20th and the beginning of the 21st centuries have been marked by a revolution in the development and affordability of the technologies in these other fields. Much research in structural health monitoring for civil structures has been directed towards *adaptation* of these technologies to civil structures. The remaining challenge is to develop a system that is robust, redundant and well calibrated, which will neither miss significant damage nor produce many false alarms. Achieving this would require focusing the efforts and resources to further develop those methodologies that are robust when applied to real structures and data, and to calibrate them using documented full-scale data. Further enhancement of the spatial resolution of such methods would benefit from inexpensive and reliable new sensors.



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 14

a ICS building: novelties analysis of the EW accelerations at the center of the building. b Same as Fig. 13a but for the NS accelerations at the east end of the building



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 15

ICS building: comparison of results for EW motions from different methods. System frequency f_{sys} from time frequency analysis, fixed base frequency f_1 from wave travel time analysis, and times of occurrence of novelties, T1, T2, and T3

All this will have to be accomplished by continuously expanding our experience in dealing with the complexities of metastable damage states of engineering structures, which will gradually become more feasible with the formulation of realistic physical models. Nevertheless, the practical outcome of most approaches in engineering will probably remain empirical. Also, the art of dynamical modeling will have to be further developed, especially for the assessment of the damaged states of engineering structures that are highly nonlinear and chaotic. In the end, in structural health monitoring, and in design of earthquake resistant structures, the fact that some modeling problems will remain will have to be accepted. However, considerable progress will be achieved if the success is gauged by the degree to which the predictions match observations in the full-scale structures, contributing towards safety and minimization of disruption and productivity of society in seismically active regions.

Bibliography

1. Applied Technology Council (1989) Procedures for post-earthquake safety evaluation of buildings. Report ATC-20. Redwood City
2. Beltrami E (1987) Mathematics for Dynamic Modeling. Wiley, New York
3. Carder DS (1936) Vibration observations. In: Earthquake Investigations in California 1934–1935. US Dept. of Commerce, Coast and Geological Survey, Special Publication No 201. Washington DC, pp 49–106
4. Carden EP, Fanning P (2004) Vibration Based Condition Monitoring: a Review. Struct Health Monit 3(4):355–377. doi:10.1177/1475921704047500
5. Celebi M, Sanli A (2002) GPS in pioneering dynamic monitoring of long-period structures. Earthq Spectr 18(1):47–61
6. Celebi M, Sanli A, Sinclair M, Gallant S, Radulescu D (2004) Real-time seismic monitoring needs of a building owner—and the solution: a cooperative effort. Earthq Spectr 20(2):333–346
7. Chang PC, Flatau A, Liu SC (2003) Review paper: health monitoring of civil infrastructure. Struct Health Monit 2(3):257–267
8. Clinton JF, Bradford SK, Heaton TH, Favela J (2006) The observed wander of the natural frequencies in a structure. Bull Seism Soc Am 96(1):237–57
9. Crawford R, Ward HS (1968) Determination of the natural periods of building. Bull Seism Soc Am 54(6A):1743–1756
10. Doebling SW, Farrar CR, Prime MB (1998) A summary review of vibration-based damage identification methods. Shock Vib Dig 30(2):91–105. doi:10.1177/058310249803000201
11. Doebling SW, Farrar CR, Prime MB, Shevitz DW (1996) Damage identification and health monitoring of structural and mechanical systems from changes in their vibration characteristics: a literature review. Report LA-13070-MS. Los Alamos National Laboratory, Los Alamos
12. Farrar CR, Worden K (2007) An introduction to structural health monitoring. Phil Trans R Soc A 365:303–315. doi:10.1098/rsta.2006.1928

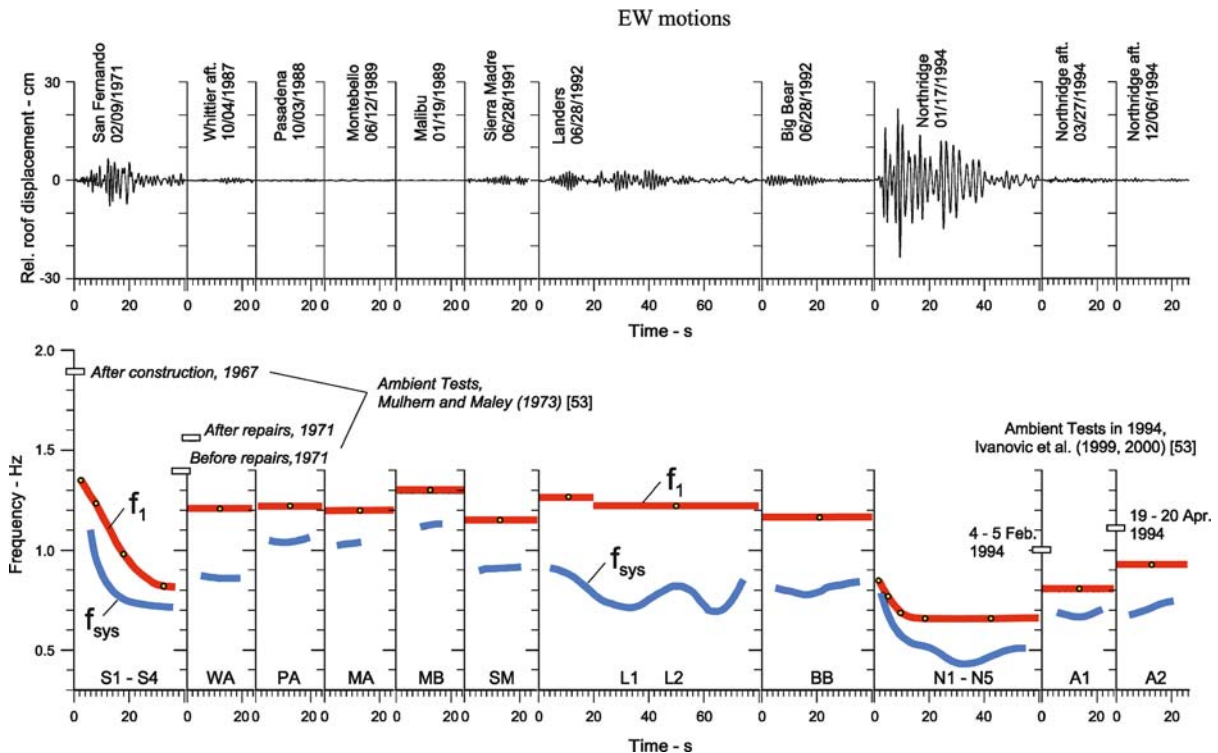
Van Nuys Hotel

Data: 11 earthquakes and 5 ambient vibration tests in 24 years

fsys – form time-freq. energy distribution, **f1** – from wave travel times.

f1 decrease: 1971 San Fernando - by ~40%. 1994 Northridge - by ~22%.

Difference between **f1** and **fsys** is not constant (see Landers and Big Bear).



Earthquake Damage: Detection and Early Warning in Man-Made Structures, Figure 16

Variations of f_1 and f_{sys} in the Van Nuys building during the 11 earthquakes, between February of 1971 and December of 1994. Measured values of f_{sys} during five ambient vibration tests: (1) in 1967, following construction, (2) in 1971, after San Fernando earthquake and before repairs, (3) in 1971 after the repairs, (4) in January of 1994, 18 days after the Northridge earthquake, and (5) in April of 1994, after the building was restrained by wooden braces

13. Foutch DA, Luco JE, Trifunac MD, Udwadia FE (1975) Full-scale three-dimensional tests of structural deformations during forced excitation of a nine-story reinforced concrete building. Proc. of the US National Conference on Earthquake Engineering. Ann Arbor, pp 206–215
14. Ghobarah A (2004) On drift limits associated with different damage levels. Proc. of the International Workshop on Performance-Based Design, 28 June–1 July 2004, Bled, Slovenia, pp 4321–332
15. Graizer VM (1991) Inertial seismometry methods, Izvestiya. Earth Phys Akad Nauk SSSR 27(1):51–61
16. Graizer VM (2005) Effect of tilt on strong motion data processing. Soil Dyn Earthq Eng 25:197–204
17. Hera A, Hou Z (2004) Application of wavelet approach for ASCE structural health monitoring benchmark studies. J Eng Mech ASCE 130(1):96–104
18. Hou Z, Noori M, Amand R (2000) Wavelet-based approach for structural damage detection. J Eng Mech ASCE 126(7): 677–683
19. Hudson DE (1970) Dynamic tests of full scale structures. In: Wiegel RL (ed) Earthquake Engineering. Prentice Hall, pp 127–149
20. Ivanović SS, Trifunac MD, Todorovska MI (2001) On identification of damage in structures via wave travel times. In: Erdik M, Celebi M, Mihailov V, Apaydin N (eds) Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumenta-

- tion for Civil Engineering Structures, 2–5 June, 1999. Kluwer, Istanbul, pp 447–468
21. Kalkan E, Graizer V (2007) Multi-component ground motion response spectra for coupled horizontal, vertical, angular accelerations and tilt. *Indian J Earthq Technol* (special issue on Response Spectra) 44(1):259–284
 22. Kanai K (1965) Some new problems of seismic vibrations of a structure. *Proc. of the Third World Conf. Earthquake Eng*, 22 January, 1 February, 1965. Auckland and Wellington, New Zealand, pp II-260–II-275
 23. Kapitaniak T (1991) *Chaotic Oscillations in Mechanical Systems*. Manchester Univ. Press, Manchester
 24. Kawakami H, Oyunchimeg M (2003) Normalized input-output minimization analysis of wave propagation in buildings. *Eng Struct* 25(11):1429–1442
 25. Kawakami H, Oyunchimeg M (2004) Wave propagation modeling analysis of earthquake records for buildings. *J Asian Archit Build Eng* 3(1):33–40
 26. Kohler MD, Heaton T, Bradford SC (2007) Propagating waves in the steel, moment-frame Factor building recorded during earthquakes. *Bull Seism Soc Am* 97(4):1334–1345
 27. Kojić S, Trifunac MD, Anderson JC (1984) A post earthquake response analysis of the Imperial County Services building in El Centro. Report CE 84-02. University of Southern California, Department of Civil Engineering, Los Angeles
 28. Lee VW, Trifunac MD (1990) Automatic digitization and processing of accelerograms using PC. Dept. of Civil Eng. Report CE 90-03. Univ. Southern California, Los Angeles
 29. Lee WHK, Celebi M, Todorovska MI, Diggles MF (eds) (2007) *Rotational Seismology and Engineering Applications*. Online Proceedings for the First International Workshop, September 18–19 September, Menlo Park. US Geological Survey, Open-File Report 2007-1144 <http://pubs.usgs.gov/of/2007/1144>
 30. Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) *Nonlinear Dynamics and Predictability of Geophysical Phenomena*. Geophysical Monograph 83, IUGG, vol 18 pp 1–5
 31. Liu SC, Tomizuka M, Ulsoy G (2006) Strategic issues in sensors and smart structures. *Struct Control Health Monit* 13:946–957
 32. Luco JE, Trifunac MD, Udwadia FE (1975) An experimental study of ground deformations caused by soil-structure interaction. *Proc. US National Conf. on Earthq. Eng. Ann Arbor, MI*, pp 136–145
 33. Luco JE, Trifunac MD, Wong HL (1987) On the apparent change in the dynamic behavior of a nine-story reinforced concrete building. *Bull Seism Soc Am* 77(6):1961–1983
 34. Luco JE, Trifunac MD, Wong HL (1988) Isolation of soil-structure interaction effects by full-scale forced vibration tests. *Earthq Eng Struct Dyn* 16:1–21
 35. Ma J, Pines DJ (2003) Damage detection in a building structure model under seismic excitation using dereverberated wave machines. *Eng Struct* 25:385–396
 36. Mallat SG (1989) Multiresolution approximations and wavelet orthonormal bases of $L_2(\mathbb{R})$. *Trans Am Math Soc* 315:69–87
 37. Oyunchimeg M, Kawakami H (2003) A new method for propagation analysis of earthquake waves in damaged buildings: Evolutionary Normalized Input-Output Minimization (NIOM). *J Asian Archit Build Eng* 2(1):9–16
 38. Rezaei M, Rahmatian P, Ventura C (1998) Seismic data analysis of a seven-storey building using frequency response function and wavelet transform. *Proc. of the NEHRP Conference and Workshop on Research on the Northridge, California Earthquake*, 17 January, 1994. CUREe, Oakland, pp 421–428
 39. Snieder R, Şafak E (2006) Extracting the building response using interferometry: theory and applications to the Millikan Library in Pasadena, California. *Bull Seism Soc Am* 96(2):586–598
 40. Sohn H, Farrar CR, Hemez FM, Shunk DD, Stinemates DW, Nadler BR (2003) *A Review of Structural Health Monitoring Literature: 1996–2001*, Report LA-13976-MS. Los Alamos National Laboratory
 41. Şafak E (1998) Detection of seismic damage in multi-story buildings by using wave propagation analysis. *Proc. of the Sixth US National Conf. on Earthquake Eng. EERI, Oakland*, Paper No 171, pp 12
 42. Şafak E (1999) Wave propagation formulation of seismic response of multi-story buildings. *J Struct Eng ASCE* 125(4):426–437
 43. Todorovska MI (1998) Cross-axis sensitivity of accelerographs with pendulum like transducers: mathematical model and the inverse problem. *Earthq Eng Struct Dyn* 27:1031–1051
 44. Todorovska MI (2009) Seismic interferometry of a soil-structure interaction model with coupled horizontal and rocking response. *Bull Seism Soc Am* 99-2A (in press)
 45. Todorovska MI (2008) Soil-structure system identification of Millikan Library north-south response during four earthquakes (1970–2002): what caused the observed wandering of the system frequencies? *Bull Seism Soc Am* 99-2A (in press)
 46. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil-structure system frequency: examples based on poroelasticity and a comparison with full-scale measurements. *Soil Dyn Earthq Eng* 26(6–7):708–717
 47. Todorovska MI, Al Rjoub Y (2008) Environmental effects on measured structural frequencies – model prediction of short term shift during heavy rainfall and comparison with full-scale observations. *Struct Control Health Monit* (in press) [doi:10.1002/stc.260](https://doi.org/10.1002/stc.260)
 48. Todorovska MI, Lee VW (1989) Seismic waves in buildings with shear walls or central core. *J Eng Mech ASCE* 115(12):2669–2686
 49. Todorovska MI, Trifunac MD (1989) Antiplane earthquake waves in long structures. *J Eng Mech ASCE* 115(12):2687–2708
 50. Todorovska MI, Trifunac MD (1990) A note on the propagation of earthquake waves in buildings with soft first floor. *J Eng Mech ASCE* 116(4):892–900
 51. Todorovska MI, Trifunac MD (2005) Structural Health Monitoring by Detection of Abrupt Changes in Response Using Wavelets: Application to a 6-story RC Building Damaged by an Earthquake. *Proc. of the 37th Joint Panel Meeting on Wind and Seismic Effects*, 16–21 May, 2005. Tsukuba, Japan. US Japan Natural Resources Program (UJNR), pp 20
 52. Todorovska MI, Trifunac MD (2007) Earthquake damage detection in the Imperial County Services Building I: the data and time-frequency analysis. *Soil Dyn Earthq Eng* 27(6):564–576
 53. Todorovska MI, Trifunac MD (2008) Impulse response analysis of the Van Nuys 7-storey hotel during 11 earthquakes and earthquake damage detection. *Struct Control Health Monit* 15(1):90–116. [doi:10.1002/stc.208](https://doi.org/10.1002/stc.208)
 54. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building III: analysis

- of wave travel times via impulse response functions. *Soil Dyn Earthq Eng* 21(5):387–404. doi:10.1016/j.soildyn.2007.07.001
55. Todorovska MI, Trifunac MD (2008) Earthquake damage detection in the Imperial County Services Building II: analysis of novelties via wavelets. *Struct Control Health Monit* (submitted for publication)
 56. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part I: theoretical models. *Soil Dyn Earthq Eng* 21(3):211–223
 57. Todorovska MI, Trifunac MD, Ivanović SS (2001) Wave propagation in a seven-story reinforced concrete building, Part II: observed wave numbers. *Soil Dyn Earthq Eng* 21(3):225–236
 58. Trifunac MD (2007) Early History of the Response Spectrum Method, Dept. of Civil Engineering, Report CE 07-01. Univ. Southern California, Los Angeles, California
 59. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. *Soil Dyn Earthq Eng* 21(4):275–286
 60. Trifunac MD, Todorovska MI (2001) Evolution of accelerographs, data processing, strong motion arrays and amplitude and spatial resolution in recording strong earthquake motion. *Soil Dyn Earthq Eng* 21(6):537–555
 61. Trifunac MD, Todorovska MI (2001) Recording and interpreting earthquake response of full-scale structures: In Erdik M, Celebi M, Mihailov V, Apaydin N (eds) *Proc. of the NATO Advanced Research Workshop on Strong-Motion Instrumentation for Civil Engineering Structures*, 2–5 June, 1999. Kluwer, Istanbul, p 24
 62. Trifunac MD, Ivanovic SS, Todorovska MI (1999) Experimental evidence for flexibility of a building foundation supported by concrete friction piles. *Soil Dyn Earthq Eng* 18(3):169–187
 63. Trifunac MD, Todorovska MI, Hao TY (2001) Full-scale experimental studies of soil-structure interaction – a review. *Proc. of the 2nd US Japan Workshop on Soil-Structure Interaction*, 6–8 March, 2001. Tsukuba City, Japan, pp 52
 64. Trifunac MD, Ivanović SS, Todorovska MI (2003). Wave propagation in a seven-story reinforced concrete building, Part III: damage detection via changes in wave numbers. *Soil Dyn Earthq Eng* 23(1):65–75
 65. Trifunac MD, Todorovska MI, Manić MI, Bulajić BĐ (2008) Variability of the fixed-base and soil-structure system frequencies of a building – the case of Borik-2 building. *Struct Control Health Monit* (in press). doi:10.1002/stc.277
 66. Udawadia FE, Jerath N (1980) Time variations of structural properties during strong ground shaking. *J Eng Mech Div ASCE* 106(EM1):111–121
 67. Udawadia FE, Marmarelis PZ (1976) The identification of building structural systems I. The linear case. *Bull Seism Soc Am* 66(1):125–151
 68. Udawadia FE, Marmarelis PZ (1976) The identification of building structural systems II. The nonlinear case. *Bull Seism Soc Am* 66(1):153–171
 69. Udawadia FE, Trifunac MD (1974) Time and amplitude dependent response of structures. *Earthq Eng Struct Dyn* 2:359–378
 70. Ward HS, Crawford R (1966) Wind induced vibrations and building modes. *Bull Seism Soc Am* 56(4):793–813
 71. Wong HL, Trifunac MD, Luco JE (1988) A comparison of soil-structure interaction calculations with results of full-scale forced vibration tests. *Soil Dyn Earthq Eng* 7(1):22–31

Earthquake Early Warning System in Southern Italy

ALDO ZOLLO¹, GIOVANNI IANNACONE²,
VINCENZO CONVERTITO², LUCA ELIA²,
IUNIO IERVOLINO³, MARIA LANCIERI²,
ANTHONY LOMAX⁴, CLAUDIO MARTINO¹,
CLAUDIO SATRIANO¹, EMANUEL WEBER²,
PAOLO GASPARINI¹

¹ Dipartimento di Scienze Fisiche, Università di Napoli “Federico II” (RISSC-Lab), Napoli, Italy

² Osservatorio Vesuviano, Istituto Nazionale di Geofisica e Vulcanologia (RISSC-Lab), Napoli, Italy

³ Dipartimento di Ingegneria Strutturale, Università di Napoli “Federico II”, Napoli, Italy

⁴ Alomax Scientific, Mouans-Sartoux, France

Article Outline

Glossary

Definition of the Subject

Introduction

Earthquake Potential and Seismic Risk
in the Campania Region

Seismic Network Architecture and Components

Real-Time Data Transmission System

Network Management and Data Archiving

Real-Time Earthquake Location
and Magnitude Estimation

Real-Time Hazard Analysis
for Earthquake Early Warning

Future Directions

Bibliography

Glossary

Data transmission system A multi-component device aimed at the transmission of seismic signals over a distance, also denoted as a telecommunication system. Each data transmission system consists of two basic elements: a transmitter that takes information and converts it to an electromagnetic signal and a receiver that receives the signal and converts it back into usable information.

Modern telecommunication systems are two-way and a single device, a transceiver, acts as both a transmitter and receiver. Transmitted signals can either be analogue or digital. In an analogue signal, the signal is varied continuously with respect to the information. In a digital signal, the information is encoded as a set of discrete, binary values. During transmission, the in-

formation contained in analogue signals will be degraded by noise, while, unless the noise exceeds a certain threshold, the information contained in digital signals will remain intact. This represents a key advantage of digital signals over analogue signals. A collection of transmitters, receivers or transceivers that communicate with each other is a telecommunication network. Digital networks may consist of one or more routers that route data to the correct user.

Earthquake early warning system (EEWS)

A real-time, modern information system that is able to provide rapid notification of the potential damaging effects of an impending earthquake, through rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern (regional EEWS) or surrounding the target infrastructure (site-specific EEWS). A "regional" EEWS is based on a dense sensor network covering a portion or the entirety of an area that is threatened by earthquakes. The relevant source parameters (event location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence, a ground motion intensity measure at a distant site where a target structure of interest is located. On the other hand, a "site-specific" EEWS consists of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements of amplitude and predominant period on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

Earthquake location An earthquake location specifies the spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized.

Earthquake magnitude The magnitude is a parameter used by seismologists to quantify the earthquake size. The Richter magnitude scale, or more correctly, local magnitude *M_L* scale, assigns a single number to quantify the amount of seismic energy released by an earthquake. It is a base-10 logarithmic scale obtained by calculating the logarithm of the combined horizontal amplitude of the largest displacement from zero on a seismometer output. Measurements have no limits and can be either positive or negative.

Introduced by the Japanese seismologist Aki in 1962, the seismic moment is the present-day physical pa-

rameter used to characterize the earthquake strength. It represents the scalar moment of one of the couples of forces producing the dislocation at an earthquake fault and it is measured from the asymptotic DC level on displacement Fourier spectra of recorded seismic signals.

Probability density function – PDF A function in one or more dimensional space *X* that (i) when integrated over some interval Δx in *X* gives a probability of occurrence of any event within Δx , and (ii) has unit integral over space *X*, where *X* represents a space of possible events.

Seismic data-logger A core element of a digital seismic station, whose aim is to record the analogue signals from seismic sensors and convert them in digital form with an assigned sampling frequency. Ground motion signals acquired by seismic sensors are pre-amplified and anti-aliasing filtered in a data-logger before they are digitalized through an AD (analog-to-digital) converter. The main technical features of a modern data-logger are the number of available channels, the allowed sampling frequencies, the dynamic range, the digitizer clock type, the storage capacity (PCMCIA, internal flash and/or hard disk, USB, ...), network interfaces (ethernet, wireless lan, or ppp) and power consumption.

Seismic hazard The probability that at a given site, a strong motion parameter (generally the peak ground acceleration) exceeds an assigned value in a fixed time period. When the seismic hazard is computed for an extended region it is generally represented as a map. The hazard map is commonly computed for a constant probability level (10%, 5% or 2%) and a given time window (50 years). It represents the spatial variation of the peak ground acceleration (expressed in percentage of gravity *g*) to be exceeded in the given period with the chosen probability level.

Earthquake early warning systems can provide a mean for the evaluation of real-time hazard maps which evolve with time, as new information about source location, magnitude and predicted peak ground motion parameters are available soon after the earthquake occurrence.

Seismic sensors Instruments used to record the ground vibration produced by natural and artificial sources, generally denoted as seismometers. A seismometer measures the relative motion between its frame and a suspended mass. Early seismometers used optics, or motion-amplifying mechanical linkages. The motion was recorded as scratches on smoked glass, or exposures of light beams on photographic paper. In modern

instruments the proof mass is held motionless by an electronic negative feedback loop that drives a coil. The distance moved, speed and acceleration of the mass are directly measured. Most modern seismometers are broadband, working on a wide range of frequencies (0.01–100 Hz). Another type of seismometer is a digital strong-motion seismometer, or accelerometer, which measures soil acceleration. Due to its relatively high dynamic range, the accelerometer can record unsaturated strong amplitude signals at close distances from a large earthquake. This data is essential to understand how an earthquake affects human structures.

Definition of the Subject

The origin of the term “early warning” probably goes back to the first decades of the last century. However, the first practical use of an “early warning” strategy was military and it was developed during the “cold war” years as a countermeasure to the potential threat from inter-continental ballistic missiles. The objective of these systems was to give an alert to target areas as soon as a missile was detected by a radar system or a launch was detected by a satellite system. In this context the term “lead time” was defined as the time elapsing between the detection of the missile and the estimated impact on the target.

In the last decades the use of the term “early warning” greatly expanded. It is used with small, but significant, variations in various types of risks, from epidemiological, to economic, social, and of course all the types of natural and environmental risks.

In fact, in these contexts, including some natural risks such as hydro-geological and volcanic, the warning is not given at the onset of the catastrophic phenomenon, but after the occurrence of some precursory phenomena which can trigger a catastrophic event (for instance intensive rainfall for hydrological risk, earthquakes and/or ground deformation for volcanic risk). The main consequence of this difference is an increase in the probability of issuing false alarms.

The case of earthquake early warning is similar to missile early warning. The alert is given after an earthquake is detected by a network of seismometers. An earthquake early warning is based on the fact that most of the radiated energy is contained in the slower traveling phases (*S*- and surface waves traveling at about 3.5 km/s or less) which arrive at any location with a delay with respect to small amplitude higher velocity phases (*P*-waves, travelling at about 6–7 km/s) or to an electromagnetically transmitted (EM) signal giving the warning.

Introduction

Many regions in the world are affected by natural hazards such as earthquakes, tsunamis, volcanoes, floods, storms, landslides, etc., each of which can have devastating socio-economic impacts. Among these natural events, earthquakes, have been among the most recurrent and damaging hazards during last few decades, resulting in large numbers of casualties, and massive economic losses [30].

The problem of earthquake risk mitigation is faced using different approaches, depending upon the time scale being considered. Whilst over time scales of decades it is of utmost importance that land use regulations and building/infrastructure codes are continuously updated and improved, for time scales of a few years, the main risk mitigation actions are at the level of information and education in order to increase individual and social community awareness about potentially damaging hazards. Over shorter time scales (months to hours), it would naturally be of great benefit to society as a whole if the capability to accurately predict the time, location and size of a potentially catastrophic natural event were available. However, due to the great complexity of the natural processes of concern, such predictions are currently not possible.

On the other hand, on very short time scales (seconds to minutes), new strategies for earthquake risk mitigation are being conceived and are under development worldwide, based on real-time information about natural events that is provided by advanced monitoring infrastructures, denoted as “early warning systems”.

Regional and On-site Early Warning Systems

Earthquake Early Warning Systems (EEWS) are modern, real-time information systems that are able to provide rapid notification of the potential damaging effects of an impending earthquake through the rapid telemetry and processing of data from dense instrument arrays deployed in the source region of the event of concern. Such systems allow mitigating actions to be taken before strong shaking and can significantly shorten the time necessary for emergency response and the recovery of critical facilities such as roads and communication lines.

Advances have been made towards the implementation of operational systems in Japan, Taiwan, and Mexico using two different approaches, i.e., “regional warning” and “onsite warning” [25]. A regional warning system is based on a dense sensor network covering a portion or the entire area that is threatened by earthquakes. The relevant source parameters (earthquake location and magnitude) are estimated from the early portion of recorded signals and are used to predict, with a quantified confidence,

a ground motion intensity measure at a distant site where a target structure of interest is located. Alternatively, “on-site warning” systems consist of a single sensor or an array of sensors deployed in the proximity of the target structure that is to be alerted, and whose measurements on the initial *P*-wave motion are used to predict the ensuing peak ground motion (mainly related to the arrival of *S* and surface waves) at the same site.

Implementation of Early Warning Systems Worldwide

In Japan, since the 1965, the JNR (Japanese National Railway) has developed and operated the Urgent Earthquake Detection and Alarm System (UrEDAS), which is an on-site warning system along the Shinkansen (bullet train) railway. UrEDAS is based on seismic stations deployed along the Japanese Railway with an average distance of 20 km. An alert is issued if the horizontal ground acceleration exceeds 40 cm/s^2 . In the 1996, the UrEDAS was combined with a new seismometer called “compact UrEDAS” [31,32,33].

On the other hand, for about one decade the Japanese Meteorological Agency (JMA) has been developing and experimenting with a mixed single station and network based early warning system to generate immediate alerts after earthquakes with JMA Intensity greater than “lower 5” (approximately $M > 6$) [24]. During a testing period from February 2004 to July 2006, the JMA sent out 855 earthquake early warnings, only 26 of which were recognized as false alarms [40]. On October 1, 2007 the broadcast early warning system developed by the Japanese Meteorological Agency (JMA) became operative. In this system, the first warning is issued 2 s after the first *P* phase detection, if the maximum acceleration amplitude exceeds the threshold of 100 cm/s^2 .

In the United States the first prototype of an early warning system was proposed by Bakun et al. [4] and developed for mitigating earthquake effects in California. It was designed to rapidly detect the Loma Prieta aftershocks and send an alert when the estimated magnitude was greater than 3.7, in order to reduce the risk of the crews working in the damaged area. The system is composed of four components: ground motion sensors deployed in the epicentral area, a central receiver, radio repeaters and radio receivers. The prototypical system worked for 6 months, during which time 19 events with $M > 3.5$ occurred, 12 alerts were issued with only 2 missed triggers and 1 false alarm.

Based on pioneering work by Allen and Kanamori [2] seismologists across California are currently planning real-time testing of earthquake early warning across the

state using the ElarmS (Earthquake Alarms Systems) methodology [1]. The approach uses a network of seismic instruments to detect the first-arriving energy at the surface, the *P*-waves, and translate the information contained in these low amplitude waves into a prediction of the peak ground shaking that follows. Wurman et al. [47] illustrated the first implementation of ElarmS in an automated, non-interactive setting, and the results of 8 months of non-interactive operation in northern California.

Since 1989, in Mexico, the civil association CIRES (Centro de Instrumentacion y Registro Sismico) with the support of Mexico City Government Authorities, developed and implemented the Mexican Seismic Alert System (SAS) [15]. The SAS is composed of (a) a seismic detection network, 12 digital strong motion stations deployed along 300 km of the Guerrero coast, (b) a dual communication system: a VHF central radio relay station and three UHF radio relay stations, (c) a central control system which continuously controls the operational status of the seismic detection and communication system and, when an event is detected, automatically determines the magnitude and issues the alarm, and (d) a radio warning system for broadcast dissemination of the alarm to end users. After 11 years, the SAS system recorded 1373 events in the Guerrero coast, it issued 12 alerts in Mexico city, with only one false alarm.

In Taiwan, the Taiwan Central Weather Bureau (CWB) developed an early warning system based on a seismic network consisting of 79 strong motion stations installed across Taiwan and covering an area of $100 \times 300 \text{ km}^2$ [44]. Since 1995 the network has been able to report event information (location, size, strong motion map) within 1 min after an earthquake occurrence [39]. To reduce the report time, Wu and Teng [44] introduced the concept of a virtual sub-network: as soon as an event is triggered by at least seven stations, the signals coming from the stations less distant than 60 km from the estimated epicenter are used to characterize the event. This system successfully characterized all the 54 events occurred during a test period of 7 months (December 2000 – June 2001), with an average reporting time of 22 s.

In Europe, the development and testing of EEWS is being carried out in several active seismic regions. Europe is covered by numerous high-quality seismic networks, managed by national and European agencies, including some local networks specifically designed for seismic early warning around, for example, Bucharest, Cairo, Istanbul and Naples.

In Turkey, an EEWS is operative, called PreSEIS (pre-seismic shaking), to provide rapid alert for Istanbul and surrounding areas. It consists of 10 strong motion sta-

tions located along the border of the Marmara sea along an arc of about 100 km, close to the seismogenetic zone of the Great Marmara Fault Zone with real time data transmission to Kandilli-Observatory [7,14]. An alarm is issued when a threshold amplitude level is exceeded.

In Romania, the EEWS is based on three tri-axial strong motion sensors deployed in the Vrancea area with a satellite communication link to the Romanian Data Center at NIEP in Bucharest [7,42]. The system is based on first *P* wave detection and prediction of the peak horizontal acceleration recorded in Bucharest, allowing for a warning time of about 25 s.

On 2006 the European Union launched the 3-year project SAFER (Seismic Early Warning for Europe), which is a cooperative scientific program aimed at developing technological and methodological tools that exploit the possibilities offered by real-time analysis of signals coming from these networks for a wide range of actions, performed over time intervals of a few seconds to some tens of minutes. The project includes the participation of 23 research groups from several countries of Europe. The primary aim of SAFER is to develop tools that can be used by disaster management authorities for effective earthquake early warning in Europe and, in particular, its densely populated cities.

The Development of an Early Warning System in Campania Region, Southern Italy

The present article is focused on the description of technologies and methodologies developed for the EEWS under construction in southern Italy.

With about 6 million inhabitants, and a large number of industrial plants, the Campania region (southern Italy), is a zone of high seismic risk, due to a moderate to large magnitude earthquake on active fault systems in the Apenninic belt. The 1980, $M = 6.9$ Irpinia earthquake, the most recent destructive earthquake to occur in the region, caused more than 3000 casualties and major, widespread damage to buildings and infrastructure throughout the region.

In the framework of an ongoing project financed by the Regional Department of Civil Protection, a prototype system for seismic early and post-event warning is being developed and tested, based on a dense, wide dynamic seismic network under installation in the Apenninic belt region (ISNet, Irpinia Seismic Network).

Considering an earthquake warning window ranging from tens of seconds before to hundred of seconds after an earthquake, many public infrastructures and buildings of strategic relevance (hospitals, gas pipelines, railways,

railroads, ...) in the Campania region can be considered as potential EEWS target-sites for experimenting with innovative technologies for data acquisition, processing and transmission based on ISNet. The expected time delay to these targets for the first energetic *S* wave train is around 30 s at about 100 km from a crustal earthquake occurring in the source region. The latter is the typical time window available for mitigating earthquake effects through early warning in the city of Naples (about 2 million inhabitants, including suburbs).

This article illustrates the system architecture and operating principles of the EEWS in the Campania region, focusing on its innovative technological and methodological aspects. These are relevant for a reliable real-time estimation of earthquake location and magnitude which are used to predict, with quantified confidence, ground motion intensity at a distant target site.

The system that we describe in this article uses an integrated approach from real time determination of source parameters to estimation of expected losses.

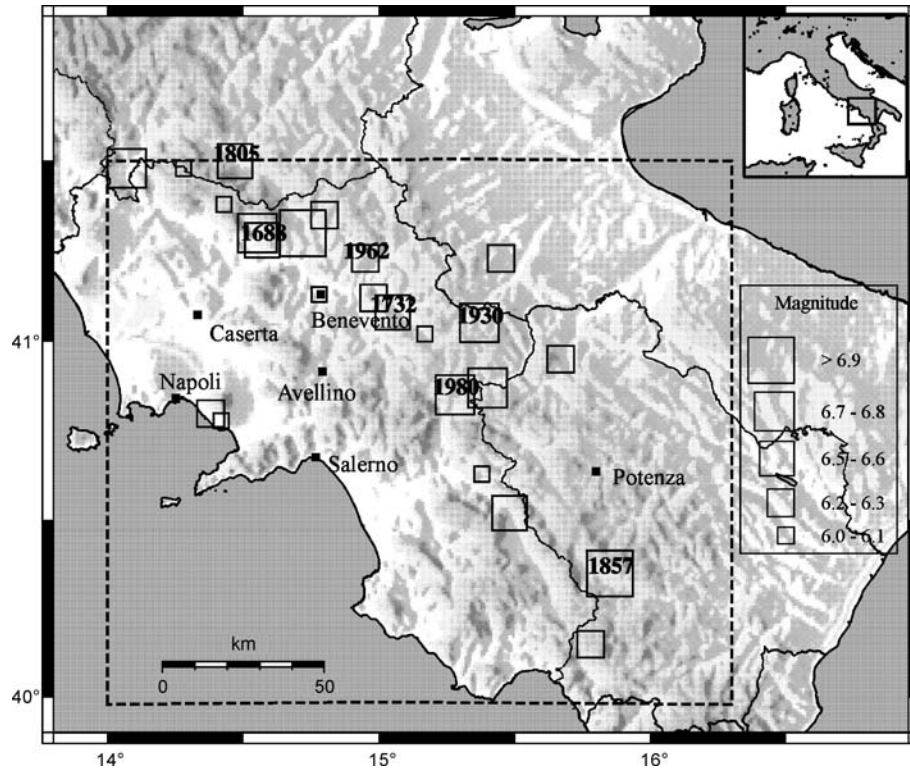
This problem must be dealt in an evolutionary (i. e., time-dependent) and probabilistic framework where probability density functions (PDFs) for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis.

Earthquake Potential and Seismic Risk in the Campania Region

The southern Apennines are an active tectonic region of Italy that accommodates the differential motions between the Adria and Tyrrhenian microplates [23]. The majority of the seismicity in this region can be ascribed to this motion. These earthquakes mainly occur in a narrow belt along the Apennine chain and are associated with young faults, with lengths ranging from 30 to 50 km, and mainly confined to the upper 20 km of the crust [28,41].

Recent stress and seismic data analyzed by [29] using earthquake locations and fault mechanisms show that the southern Apennines are characterized by an extensional stress regime and normal-fault earthquakes. However, the occurrence of recent (e. g., 5 May, 1990, Potenza, M 5.4; 31 October – 1 November, 2002, Molise, M 5.4) and historic (e. g., 5 December, 1456, M 6.5) earthquakes do not exclude other mechanisms such as strike-slip faulting.

There have been numerous large and disastrous events in the southern Apennines, including those which occurred in 1694, 1851, 1857 and 1930. The location of historical earthquakes retrieved from the CFTI (Catalogo dei Forti Terremoti in Italia, Catalogue of Strong Earthquakes



Earthquake Early Warning System in Southern Italy, Figure 1

Location of the main historic earthquakes retrieved from the CFTI database using as region of interest that defined by the external rectangle. The box dimensions are proportional to magnitude. The best constrained historic earthquakes are reported along with their date of occurrence

in Italy) database [6] is shown in Fig. 1. The most recent and well documented event is the complex normal-faulting M 6.9 Irpinia earthquake of 23 November, 1980 [5,43].

As recently indicated in the study by Cinti et al. [9], the southern Apennines has a high earthquake potential with an increasing probability of occurrence for $M \geq 5.5$ earthquakes in the next decade. The new national hazard map (Gruppo di lavoro MPS, 2004), indicates that the main towns of the region fall in a high seismic hazard area, where it is expected that a peak ground acceleration value ranging between 0.15 and 0.25 g will be exceeded in 475 years.

These aspects make the Campania region a suitable experimental site for the implementation and testing of an early warning system. A potential application of an early warning system in the Campania region should consider an expected time delay to the first energetic S wave train varying between 14–20 s at 40–60 km distance to 26–30 s at about 80–100 km, from a crustal earthquake occurring along the Apenninic fault system. Based on those delay times, a large number of civil and strategic infrastructures

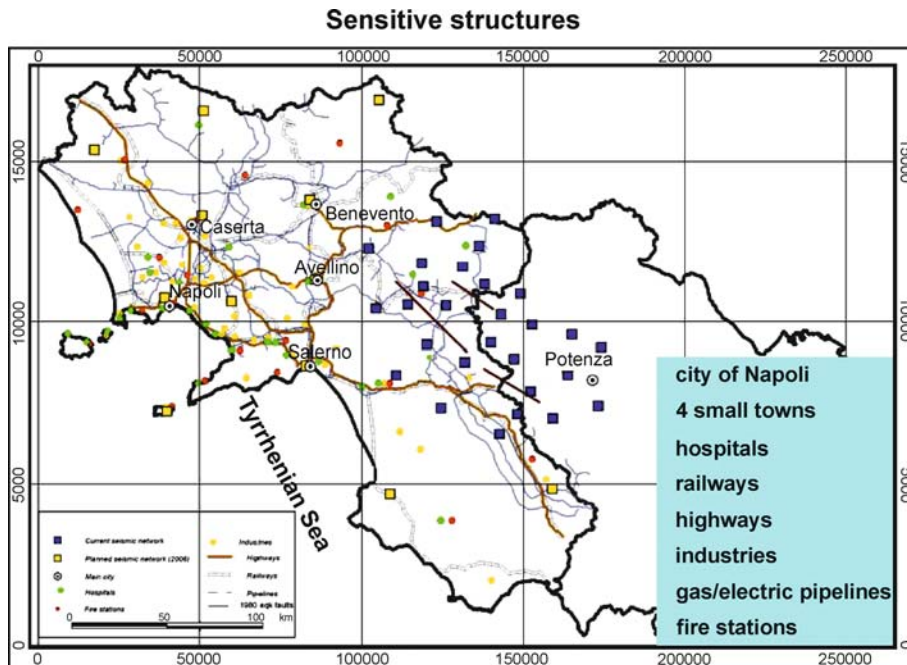
located in the Campania region are eligible for early warning applications, as shown in Fig. 2.

Seismic Network Architecture and Components

The Irpinia Seismic Network (ISNet) is a local network of strong motion, short period and broadband seismic stations deployed along the southern Apenninic chain covering the seismogenic areas of the main earthquakes that occurred in the region in the last centuries, including the $M_s = 6.9$, 23 November 1980 event.

The seismic network is composed of 29 stations organized in six sub-nets, each of them composed of a maximum of 6–7 stations (Fig. 3). The stations of a given sub-net are connected with real-time communications to a central data-collector site called the Local Control Center (LCC).

The different LCCs are linked to each other and to a Network Control Center (NCC) with different types of transmission systems. The whole data transmission system is fully digital over TCP/IP, from the data-loggers, through



Earthquake Early Warning System in Southern Italy, Figure 2

Distribution of the sensitive structures, potential candidates for an early warning system in the Campania-Lucania region



Earthquake Early Warning System in Southern Italy, Figure 3

Topology of the communication system of ISNet showing the extended-star configuration of the seismic network. Symbols explanation: green squares – seismic stations; blue squares – Local Control Centres (LCC); yellow lines – WLAN radio link connecting seismic stations and LCC; white segments – SDH carrier-class radio; red triangles – radio link repeaters; red circle – Network Control Centre RISSC in Naples; yellow squares – main cities

the LCC, to the NCC, located in the city of Naples, 100 km away from the network center.

To ensure a high dynamic recording range, each seismic station is equipped with a strong-motion accelerom-

eter and a three-component velocity meter (natural period = 1 s). In five station locations the seismometers are replaced by broadband (0.025–50 Hz) sensors to guarantee good-quality recording of teleseismic events. Data ac-

quisition at the seismic stations is performed by the innovative data-logger Osiris-6, produced by Agecodagis sarl. The hardware/software characteristics of the system allow it to install self-developed routines to perform real time specific analysis.

The data-loggers are remotely controlled through a configuration tool accessible via TCP/IP, managing sampling rate, gain, application of calibration signal to the resets of disks, GPS, etc. Furthermore, a complete station health status is available, which helps in the diagnosis of component failure or data-logger malfunction. The data-loggers store the data locally or send it to each LCC where the real-time data management system Earthworm (developed at USGS-United State Geological Survey) is operating.

A calibration unit is installed at each seismic station to automatically provide a periodic calibration signal to seismic sensors in order to verify the correct response curve of the overall acquisition chain.

The power supply of the seismic station is provided by two solar panels (120 W peak, with 480 Wh/day), two 130 Ah gel cell batteries, and a custom switching circuit board between the batteries. With this configuration, 72-h autonomy is ensured for the seismic and radio communication equipment. Each site is also equipped with a GSM/GPRS programmable control/alarm system connected to several environmental sensors and through which the site status is known in real time. With SMS (Short Message Service) and through the programmable

GSM controller, the seismic equipment can be completely reset remotely with a power shutdown/restart. The GSM also controls the device start/stop release procedure when the battery goes over/under a predefined voltage level.

Unlike the seismic stations, LCCs, which host the data server and transmission system instruments, are AC power supplied with back-up gel batteries guaranteeing 72-h stand-by power.

Real-Time Data Transmission System

ISNet has a distributed star topology that uses different types of data transmission systems.

The seismic stations are connected via spread-spectrum radio bridges to the LCCs. Data transmission between LCCs from the local control center to the network control center in Naples is performed through different technologies and media types as shown in Table 1.

To transmit waveforms in real time from the seismic stations to the LCCs, a pair of outdoor Wireless LAN bridges operating in the 2.4 GHz ISM band are used. Our tests have shown that these instruments operate continuously without any radio link failure due to adverse weather conditions (snow, heavy rain).

The two primary backbone data communication systems of the central site use Symmetrical High-speed Digital Subscriber Line (SHDSL) technology over a frame-relay protocol. Frame relay offers a number of significant

Earthquake Early Warning System in Southern Italy, Table 1
Specification of the ISNet data communication links

Type	Frequency (GHz)	Bandwidth (Mbps)	# Number of		Comments
			Stations	LCCs	
Spread spectrum Radio	2.45	54	27 ³	–	Throughput around 20–24 Mbps for links between 10–15 km (based on ethernet packets with an average size of 512 bytes).
Ethernet	–	100	2 ³		Stations connected with ethernet cable to LCC infrastructure.
Wireline SHDSL over Frame Relay	–	2.048	–	2	At the central site (RISSC) the CIR ¹ is maximum 1.6 Mbps depending upon number of PVCs ² . At the remote (LCC) site the bandwidth is 640/256 kbps with CIR of 64 kbps in up and download, over ADSL with ATM ABR service class.
Microwave Radio SDH	7	155	–	6	Carrier-class microwave link. Connect six LCC with 155 Mbps (STM-1) truly full bandwidth available. First link constructed for early warning applications.
Microwave Radio HyperLAN/2	5.7	54	–	2	The true usable maximum throughput of HyperLAN/2 is 42 Mbps.

¹ CIR Committed Information Rate.

² PVC permanent virtual circuit.

³ Not included stations hosted by LCCs.

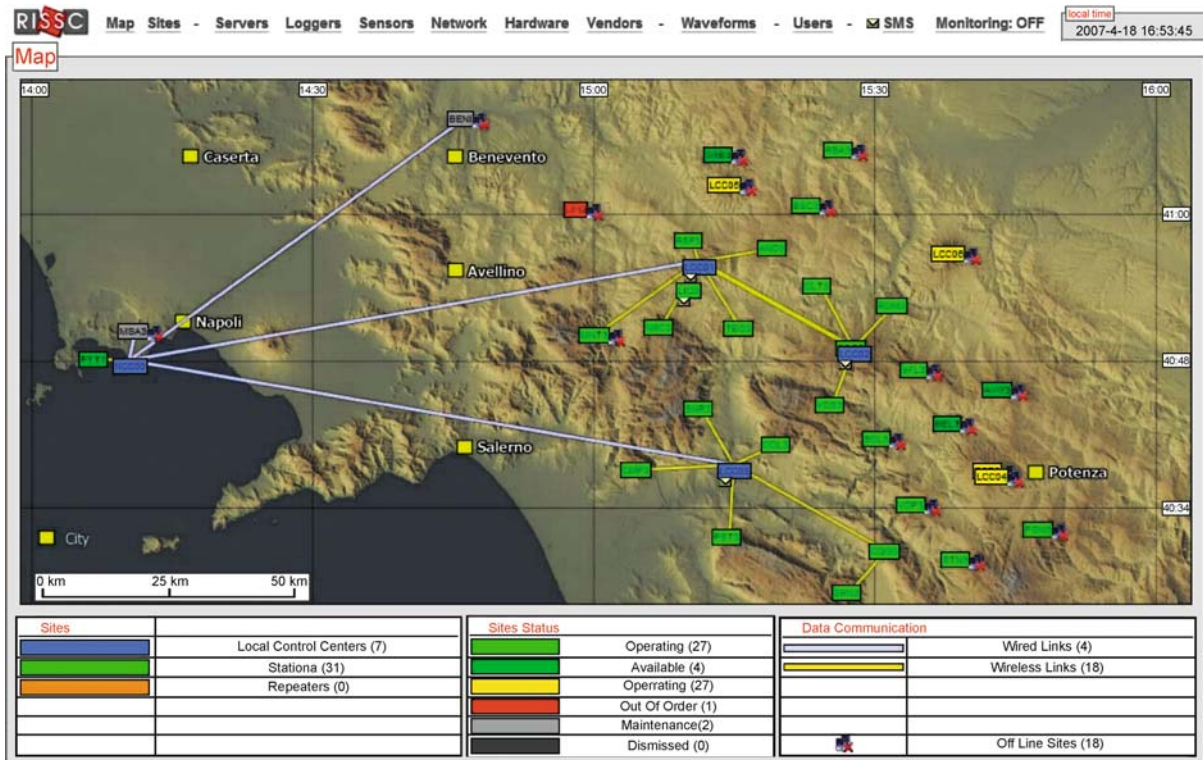
benefits over analogue and digital point-to-point leased lines. With the latter, each LCC requires a dedicated circuit between the LCCs and NCC. Instead, the SHDSL frame relay is a packet-switched network, which allows a site to use a single frame-relay phone circuit to communicate with multiple remote sites through the use of permanent virtual circuits. With virtual circuits, each remote site is seen as part of a single private LAN, simplifying IP address scheme maintenance and station monitoring.

Each seismic site has a real-time data flow of 18.0 kbps (at 125 Hz sampling rate for each physical channel), and the overall data communication bandwidth that is needed is around 540 kbps for 30 stations. ISNet supports this throughput under the worst conditions seen and it has been designed to guarantee further developments, such as the addition of further seismic or environmental sensors, without the need for larger economic and technological investment.

Network Management and Data Archiving

The Network Manager Application and Implementation Overview

As seen in the previous paragraphs, ISNet is a complex infrastructure, and thus needs a suitable software application in order to be effectively managed: a front-end to users and administrators with an interface that is simple to use. To this aim we developed a server-client database-driven application, dubbed *SeismNet Manager*, to keep track of the several components that comprise or are produced by the network, such as stations, devices and recorded data. This application, whose front page is shown in Fig. 4, lets the administrators manage (insert, edit, view and search) the details of (a) seismic stations and Local Control Centers (sites), (b) data communication links between sites (wired or wireless), instruments and devices (sensors, loggers, network hardware), and (c) recorded and computed data (waveforms, events). SeismNet Manager also keeps an



Earthquake Early Warning System in Southern Italy, Figure 4

The front page of SeismNet Devices Manager. This page is meant to convey the state of the whole network at a glance. Each node (station or LCCs) is shown along with its operating state, data links of different types to nearby nodes, whether it's currently on-line or not, along with eventual alarms still pending

historical record of the installations and configurations of the above elements.

All of the mentioned components are handled by leveraging an instrumental database, a flexible repository of information that was implemented by using PostgreSQL, a robust and feature-rich Database Management System available as open source.

The Instrumental Database

The instrumental database is a web-oriented application tool where, at the top level, the network is modeled as a set of sites, with installed loggers, sensors, data acquisition servers, network hardware and generic hardware,

in a given configuration. Each of the mentioned entities is mirrored by a different class of objects in the database, where the relevant details are stored and then presented to the users as interactive web pages. As an example, see the page for a typical seismic station in Fig. 5.

The instrumental database was implemented with a layer of abstraction that lets one easily to perform complex queries and hides the actual implementation details of the underlying structure to a possible client. There are both *stored procedures*, i.e., functions that perform complex tasks given simple inputs, and *views*, i.e., virtual database tables that collect the most important pieces of information about an object, physically scattered in many tables, in a single place and make it possible to easily query,

Status	
State	Operating
Visibility	Private
Online	<input checked="" type="checkbox"/>

Description	
Code (4+ char.)	SNR3
Network	ISNET
Type	Station
Comment	

Location	
Extended Location Name	Senerchia
Longitude E (deg.)	15.1925
Latitude N (deg.)	40.7361
Elevation (m)	998

GSM terminal	
SIM telephone number	+39 3358028209 View SMS

History	
Begin Date (yyyy-mm-dd)	2005-11-11
End Date (yyyy-mm-dd)	

Data Links				
Destination	Location Name	Distance (km)	Technology	Down/Up (Mb/s)
LCC03	Contursi Terme	9.9	Wi-Fi 802.11g	54

Loggers				
Model	IP address	Serial number	Storage medium	Recording type
OSIRIS6	10.37.37.20	370020	CompactFlash	Continuous

Sensors				
Model	Type	Serial number	Connected to	Components
CMG-5T	Accelerometer	T5744	OSIRIS6 370020 channels 0.1.2	TripleComponent
S13J	Velocimeter	V397	OSIRIS6 370020 channel 3	Vertical
S13J	Velocimeter	H490	OSIRIS6 370020 channel 4	NorthSouth
S13J	Velocimeter	H505	OSIRIS6 370020 channel 5	EastWest

[Components History](#)

Network Hardware			
Model	IP address	Type	Serial number
AIR-BR1310G-A-K9-R	192.168.3.37	BRIDGE WIRELESS	FTX0905U0DE

SAR

upload/site39filename_sar.doc

[Satellite Map](#) [Roads Map](#)

Waveforms

Date	Mag	Files
2007-03-15	1.1 MI	6
2007-03-14	2.5 MI	12
2007-03-11	1.8 MI	6
2007-03-08	0.8 MI	6
2007-03-07	3.1 MI	18
2007-03-05	2.1 MI	12
2007-03-03	0.8 MI	6
2007-03-02	2 MI	6
2007-02-28	2.6 MI	6
2007-02-25	3.1 MI	12

740 more file(s)...

Notes & Files

2007-01-12

In data 11-01-06 e' stato sostituito il cavo S13J

2006-11-10

Sostituito cavo S-13J quello attuale permette la calibrazione da remoto e ha resistenza di amo ...

2006-10-09

Installato S13J proveniente da USA. Bloccati i cavi di S13J e GURALP. La stazione osiris è st ...

2006-09-26

Logger OSIRIS Accelerometer: Guralp
CMG-5T Velocimeter: Geotech S13J

Earthquake Early Warning System in Southern Italy, Figure 5

The page relative to a seismic station. This page is a collection of all the pieces of information linked to a particular site: location details and map; some pictures and notes; recently received warning messages; currently installed devices and their configurations and mutual connections; data links to other stations; most recent waveforms recorded. Every device at a site also has an associated installation object, that records the configuration parameters and the physical connections to other nearby devices, valid over a period of time. Some elements, such as the data storage servers and the loggers, also need some further configuration parameters, that are independent of their actual physical installation, for things like firmware release and versions of the software packages run

for example, for all the details of the correctly operating sensors installed one year ago at stations with a working wireless link to a given LCC server.

This abstract interface makes the devices database a central repository for effectively cross-correlating the seismic data recorded at any given place and time with the details of the instrument(s) that recorded them, and the configuration details of the systems that ultimately made them available. The interface approach also makes it easier to change the implementation details without the need to update the web application, or any external client procedures that need to interact with the instrumental database.

Automatic Monitoring of the Devices and Automatic Data Retrieval

All of the details about the network described so far are provided by the administrators of the system and are manually updated every time the configuration of something in the network changes, e.g., after installing a new sensor or replacing some faulty hardware at a station. This man-

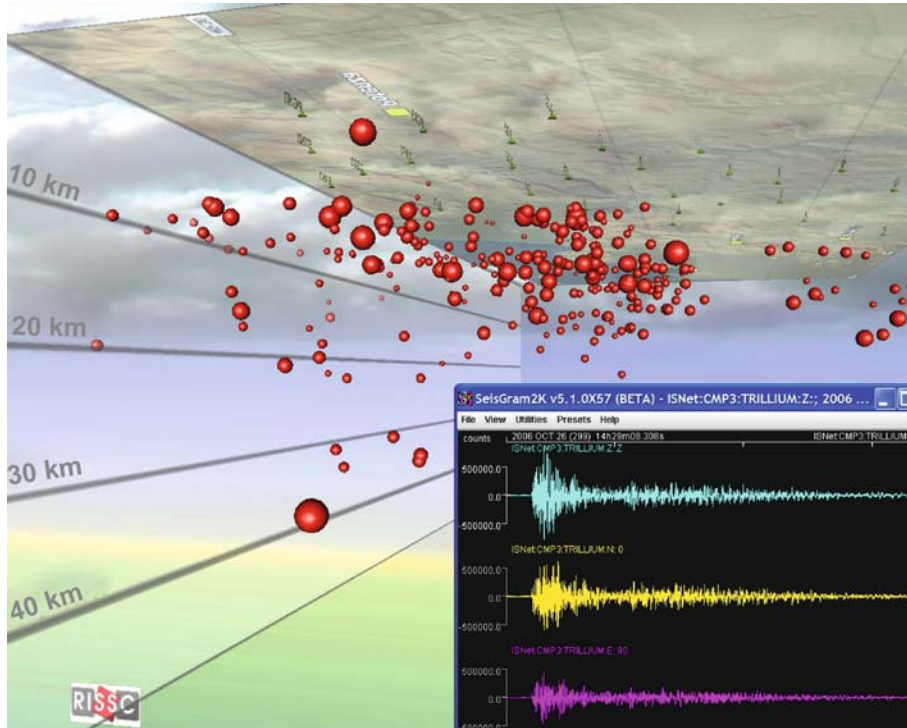
ual input is needed for “dumb” devices, such as sensors. “Smart” devices, i.e., computers with an IP address (loggers, bridges and Earthworm servers), on the other hand, can be queried about their actual configuration from time to time. The web application can plot the temporal evolution of some selected parameters as graphs, spanning a period ranging from hours to years (Fig. 6). This is useful to correlate issues spotted on the recorded seismic data (typically, “holes” in the stream of data) to hardware problems. It is possible to inspect the whole chain of data transfers to pinpoint the source of the problem (e.g., low batteries on a logger due to a faulty inverter, low signal of a wireless connection due to harsh weather conditions).

There are both automatic and manual procedures to insert new events and data files in the system. The automatic procedures make use of several sources of events to process, such as: INGV (Istituto Nazionale di Geofisica e Vulcanologia, Italy) bi-weekly bulletins; INGV real time alerts; our early warning system. Likewise, they exploit several sources of recorded seismic data in order to provide a SAC file (SAC – seismic analysis code from



Earthquake Early Warning System in Southern Italy, Figure 6

Health graphs of the devices. A device can be marked for monitoring and its internal state, or “health”, gets polled at regular intervals. Several of its internal variables are then retrieved and stored into the database, and their temporal evolution can be plotted as a graph. In this case both the internal temperature and CPU load of an OSISRIS data logger are shown, over periods ranging from one hour to one year



Earthquake Early Warning System in Southern Italy, Figure 7

Visualization of the seismic data. This is the graphical presentation of data recorded by ISNet. The waveforms matching the user's search criteria can be viewed on-line via Seisgram2K (where they can also be processed), while the events are rendered via VRML as a fully interactive 3D scene in the browser itself

Lawrence Livermore National Laboratory) spanning the period from just before the arrival time, up to the end of the event.

Sensor data are retrieved from: (1) a repository of files from the internal mass storage of the loggers; (2) a local Earthworm Wave Server that caches older data collected from all the LCCs; (3) the most recent real time recording from the remote Wave Servers. The instrumental database is used to determine which sites/sensors/configurations recorded each event and to fill the headers of the files using the standard SAC format. The waveforms and events database, on the other end, is used by the automatic procedures to know which pieces of data are still missing for already recorded events (due to e.g., the temporary unavailability of one or more seismic data sources) and need to be collected.

The Waveforms and Events Database: Searching and Visualizing the Seismic Data

We also built a waveform and event database, the natural complement to the instrumental database. It keeps track of the events detected by the network and the relative waveforms recorded by the sensors. This database stores ob-

jects for events, origin estimations (time and location), magnitude estimations and waveforms. Several origins can be attached to a single event, as different algorithms and different institutions provide different estimations. Likewise, several magnitude types and estimations are attached to each origin. A waveform object for each sensor that recorded the earthquake is also linked to the event object, and stores a pointer to a SAC file, and its source (site and channel). The latter records are then used to gather, from the instrumental database, the actual details of the instruments that recorded the data.

An interface for searching both events and waveforms is provided, as pictured in Fig. 7. Events can be filtered on origin time and location, magnitude, and distance to the stations. Waveforms can be filtered on station, component, instrument and quality.

Real-Time Earthquake Location and Magnitude Estimation

Real-Time Earthquake Location

Previous Related Studies There are many methodologies for standard earthquake location, performed when

most or all the phase arrival times for an event are available. Standard analysis techniques are generally not suited for early warning applications, since they typically need the seismic event to be fully recorded at several stations, leaving little or no lead time for the warning [25]. For this reason, a different strategy is required, where the computation starts when a few seconds of data and a small number of recording stations are available, and the results are updated with time.

Previous work on earthquake location for early warning includes several approaches to gain constraints on the location at an earlier time and with fewer observations than for standard earthquake location.

In the ElarmS methodology [47], when the first station triggers, the event is temporarily located beneath that station; after a second station trigger the location moves to a point between the two stations, based on the timing of the arrivals; with three or more triggered arrivals, the event location and origin time is estimated using trilateration and a grid search algorithm.

Horiuchi et al. [18] combine standard L2-norm event location, equal differential-time (EDT) location on quasi-hyperbolic surfaces, and the information from not-yet arrived data to constrain the event location beginning when there are triggered arrivals from two stations. The two arrivals times define a hyperbolic surface, which contains the event location. This solution is further constrained by EDT surfaces constructed using the current time (t_{now}) as a substitute for future, unknown arrival times at the stations, which have not yet recorded arrivals. The constraint increases as t_{now} progresses, even if no further stations record an arrival.

Rydelek and Pujol [36], applying the approach of Horiuchi et al. [18], show that useful constraints on an event location can be obtained with only two triggered stations. Cua and Heaton [12], generalized the approach by Rydelek and Pujol in order to start the location with one single triggering station.

The real-time location technique described in this paper is based on the equal differential-time (EDT) formulation [16,27] for standard earthquake location. The EDT location is given by the point traversed by the maximum number of quasi-hyperbolic surfaces, on each of which the difference in calculated travel-time to a pair of stations is equal to the difference in observed arrival times for the two stations. The EDT location determination is independent of origin time and reduces to a 3D search over latitude, longitude and depth. Furthermore, EDT is highly robust in the presence of outliers in the data [27]. This robustness is critical for the problem of earthquake location for seismic early warning, since we

will often work with small numbers of data and may have outlier data such as false triggers, picks from other events, and misidentified picks from energetic, secondary phases.

Assuming that a dense seismic network is deployed around the fault zone, we define as the “*evolutionary approach*” a type of analysis where the estimates of earthquake location and size, and their associated uncertainty, evolve with time as a function of the number of recording stations and of the length of the portion of signal recorded at each station.

A direct implication of the evolutionary strategy is that each algorithm must be capable of *real-time* operation, i. e., its computational time must be smaller than the rate at which data enters the system.

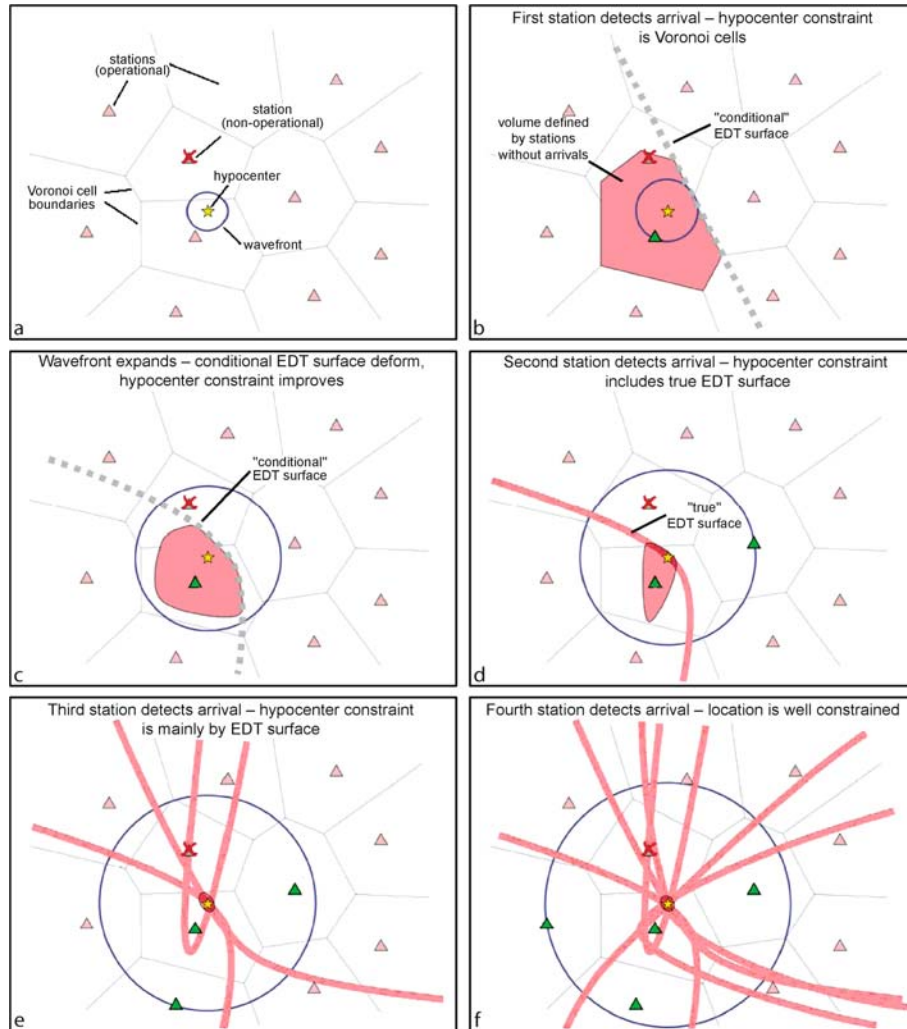
Furthermore, since each algorithm starts processing a limited amount of information, the estimated earthquake parameter must be provided, at each time step, as a *probability density function* (PDF) which incorporates in its definition the uncertainties related both to the model employed and to the available data.

The Real-Time Earthquake Location Method The methodology is related to that of Horiuchi et al. [18], which has been extended and generalized by (a) starting the location procedure after only one station has triggered, (b) using the equal differential-time approach proposed by Font [16] to incorporate the triggered arrivals and the not-yet-triggered stations, (c) estimating the hypocenter probabilistically as a PDF instead of as a point, and (d) applying a full, non-linearized, global-search for each update of the location estimate.

We assume that a seismic network has known sets of operational and non-operational stations (Fig. 8a), that when an earthquake occurs, triggers (first *P*-wave arrival picks) will become available from some of the operational stations, and that there may be outlier triggers which are not due to *P* arrivals from the earthquake of interest.

Let's denote the operational stations as (S_0, \dots, S_N), and consider a gridded search volume V containing the network and target earthquake source regions, and the travel times from each station to each grid point (i, j, k) in V computed for a given velocity model.

The standard EDT approach states that, if the hypocenter (i, j, k) is exactly determined, then the difference between the observed arrival times t_n and t_m at two stations S_n and S_m is equal to the difference between calculated travel times tt_n and tt_m at the hypocentral position, since the observed arrival times share the common earthquake origin time. In other words, the hypocenter must



Earthquake Early Warning System in Southern Italy, Figure 8

Schematic illustration of the evolutionary earthquake location algorithm. For clarity, only a map view with the epicentral location is represented. **a** Given a seismic network with known sets of operational and non-operational stations, we can define a priori the Voronoi cell associated to each station. **b** When the first station triggers, we can define a volume that is likely to contain the location, this volume is limited by conditional EDT surfaces on which the P travel time to the first triggering station is equal to the travel-time to each of the operational but not-yet-triggered stations. **c** As time progresses, we gain additional information from the stations that have not yet triggered, the EDT surfaces move towards and bend around the first triggering station, and the likely-location volume decreases in size. **d** When the second station triggers, we can define a true EDT surface; the hypocenter is on the intersection between this surface and the volume defined by the conditional EDT surfaces, which continues decreasing in size. **e** When a third station triggers, we can define two more true EDT surfaces, further increasing the constraint on hypocenter position. **f** As more stations trigger, the location converges to the standard EDT location composed entirely of true EDT surfaces

satisfy the equality:

$$(tt_m - tt_n)_{i,j,k} = t_m - t_n; \quad m \neq n \quad (1)$$

for each pair of triggering stations S_n and S_m . For a constant velocity model, this equation defines a 3D hyperbolic surface whose symmetry axis goes through the two

stations. Given N triggering stations, $N(N-1)/2$ surfaces can be drawn; the hypocenter is defined as the point crossed by the maximum number of EDT surfaces.

Following an evolutionary approach, the method evaluates, at each time step, the EDT equations considering not only each pair of triggered stations, but also those pairs where only one station has triggered.

Therefore, when the first station, S_n , triggers with an arrival at $t_n = t_{\text{now}}$ (t_{now} is the current clock time), we can already place some limit on the hypocenter position (Fig. 8b). These limits are given by EDT surfaces defined by the condition that each operational but not-yet-triggered station S_l will trigger in the next time instant, $t_l \geq t_n$. That is:

$$(tt_l - tt_n)_{i,j,k} = t_l - t_n \geq 0; \quad l \neq n. \quad (2)$$

On these conditional EDT surfaces, the P travel time to the first triggering station tt_n is equal to the travel-time to each of the not-yet-triggered stations, tt_l , $l \neq n$. These surfaces bound a volume (defined by the system of inequalities) which must contain the hypocenter. In the case of a homogeneous medium with constant P -wave speed, this hypocentral volume is the Voronoi cell around the first recording station, defined by the perpendicular bisector surfaces with each of the immediate neighboring stations.

As the current time t_{now} progresses, we gain the additional information that the not-yet-triggered stations can only trigger with $t_l > t_{\text{now}}$. Thus the hypocentral volume is bounded by conditional EDT surfaces that satisfy the inequality:

$$(tt_l - tt_n)_{i,j,k} \geq \delta t_{n,l}; \quad l \neq n. \quad (3)$$

δt is the time interval between the arrival time at station S_n and the latest time for which we have information from station S_l ,

$$\delta t_{n,l} = t_{\text{now}} - t_n, \quad (4)$$

where t_n is the observed arrival time at station S_n .

The system (3) defines the volume, bounded by the conditional EDT surfaces, in which the hypocenter may be located given that, at current time t_{now} , only the station S_n has triggered. When $\delta t = 0$ the system (3) reduces to the system (2); for $\delta t > 0$, the hypocentral volume will be smaller than the previous one, since the updated, conditional EDT surfaces tend to fold towards and around the first triggered station (Fig. 8c).

We interpret the hypocentral volume in a probabilistic way by defining, for each inequality in (3), a value $p_{n,l}(i, j, k)$ which is 1 if the inequality is satisfied and 0 if not. Then we sum the $p_{n,l}(i, j, k)$ over stations l at each grid point, obtaining a non-normalized probability density $P(i, j, k)$, where $P(i, j, k) = N - 1$ for grid points where all the inequalities are satisfied and a value less than $N - 1$ elsewhere.

When the second and later stations trigger, we first re-evaluate the system (3) for all pairs of triggered stations S_n and all not-yet-triggered stations S_l . Secondly, we

construct standard, true EDT surfaces (see Eq. 2) between each pair S_n, S_m of the triggered stations, by evaluating for each grid point the quantity:

$$q_{n,m}(i, j, k) = \exp \left\{ -\frac{[(tt_n - tt_m)_{i,j,k} - (t_n - t_m)]^2}{2\sigma^2} \right\}; \quad n \neq m. \quad (5)$$

The expression between square brackets at the exponent is the standard EDT Eq. 2 whose solutions are quasi-hyperbolic surfaces; in practice all true EDT surfaces are given a finite width by including the uncertainty σ in the arrival time picking and the travel-time calculation.

The quantity $q_{n,m}(i, j, k)$ has values between 0 and 1. We sum the $q_{n,m}(i, j, k)$ with the $p_{n,l}(i, j, k)$ obtained from the re-evaluation of (4) to obtain a new $P(i, j, k)$.

Starting from P , we define a value:

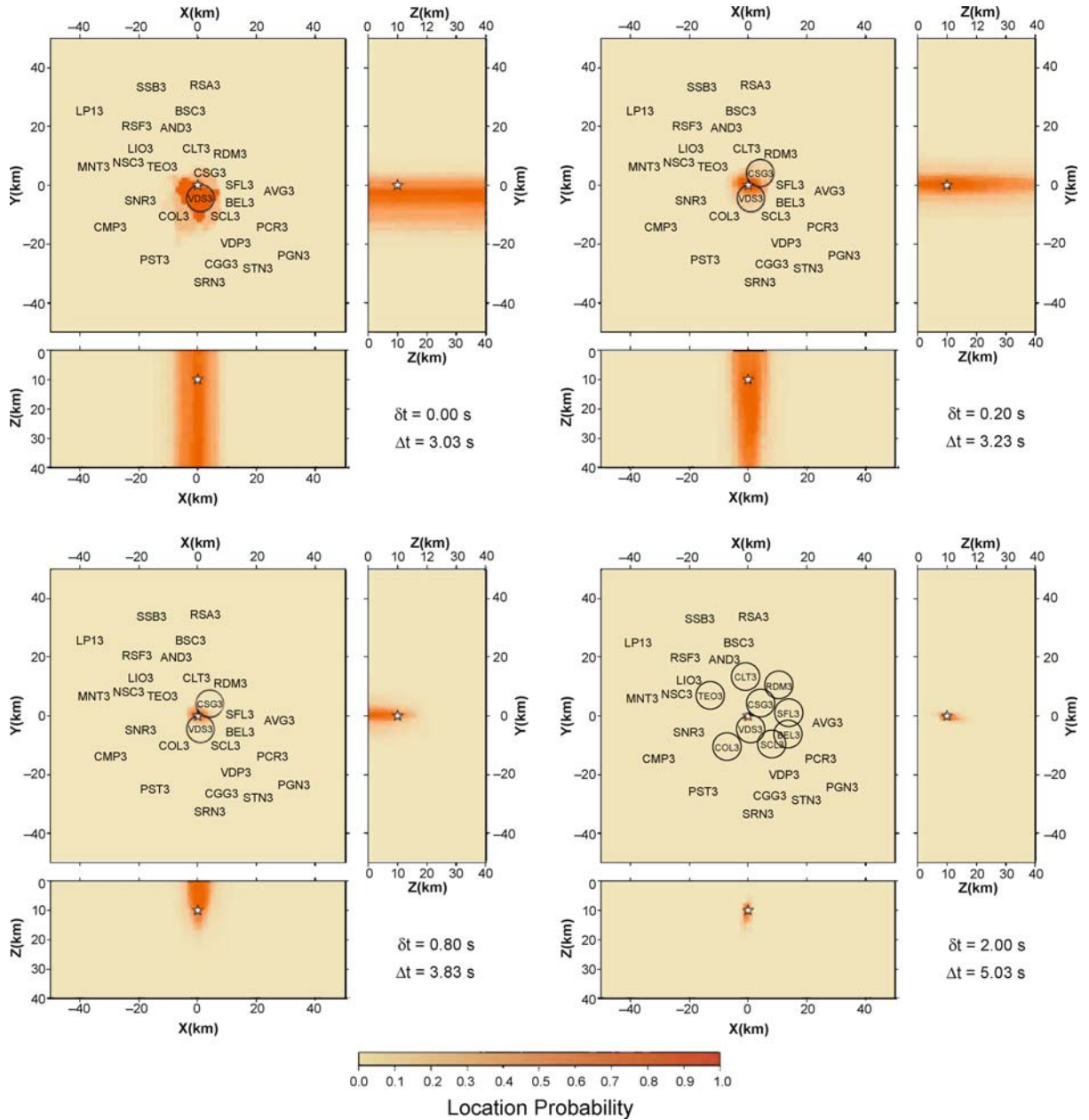
$$Q(i, j, k) = \left(\frac{P(i, j, k)}{P_{\text{max}}} \right)^N, \quad (6)$$

which forms a relative probability density function (PDF, with values between 0 and 1) for the hypocenter location within the grid cell (i, j, k) . The function $Q(i, j, k)$ may be arbitrarily irregular and may have multiple maxima.

At predetermined time intervals, we evaluate (3) and (5) to obtain $Q(i, j, k)$ in the search volume, using the Oct-tree importance sampling algorithm ([13,27], <http://www.alomax.net/nloc/octtree>). This algorithm uses recursive subdivision and sampling of rectangular cells in 3D space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target function values. The Oct-tree search is much faster than a simple or nested grid search (factor 10–100 faster) and more global and complete than stochastic search methods algorithms such as simulated annealing and genetic algorithms [13]. For each grid point, an origin time estimate can be obtained from the observed arrival times and the calculated travel times.

As more stations trigger, the number of not-yet-triggered stations becomes small, and the location converges towards the hypocentral volume that is obtained with standard EDT location using the full set of data from all operational stations (Fig. 8d–f).

If there are uncorrelated outlier data (i.e., triggers that are not compatible with P arrivals from a hypocenter within or near the network), then the final hypocentral volume will usually give an unbiased estimate of the hypocentral location, as with standard EDT location.



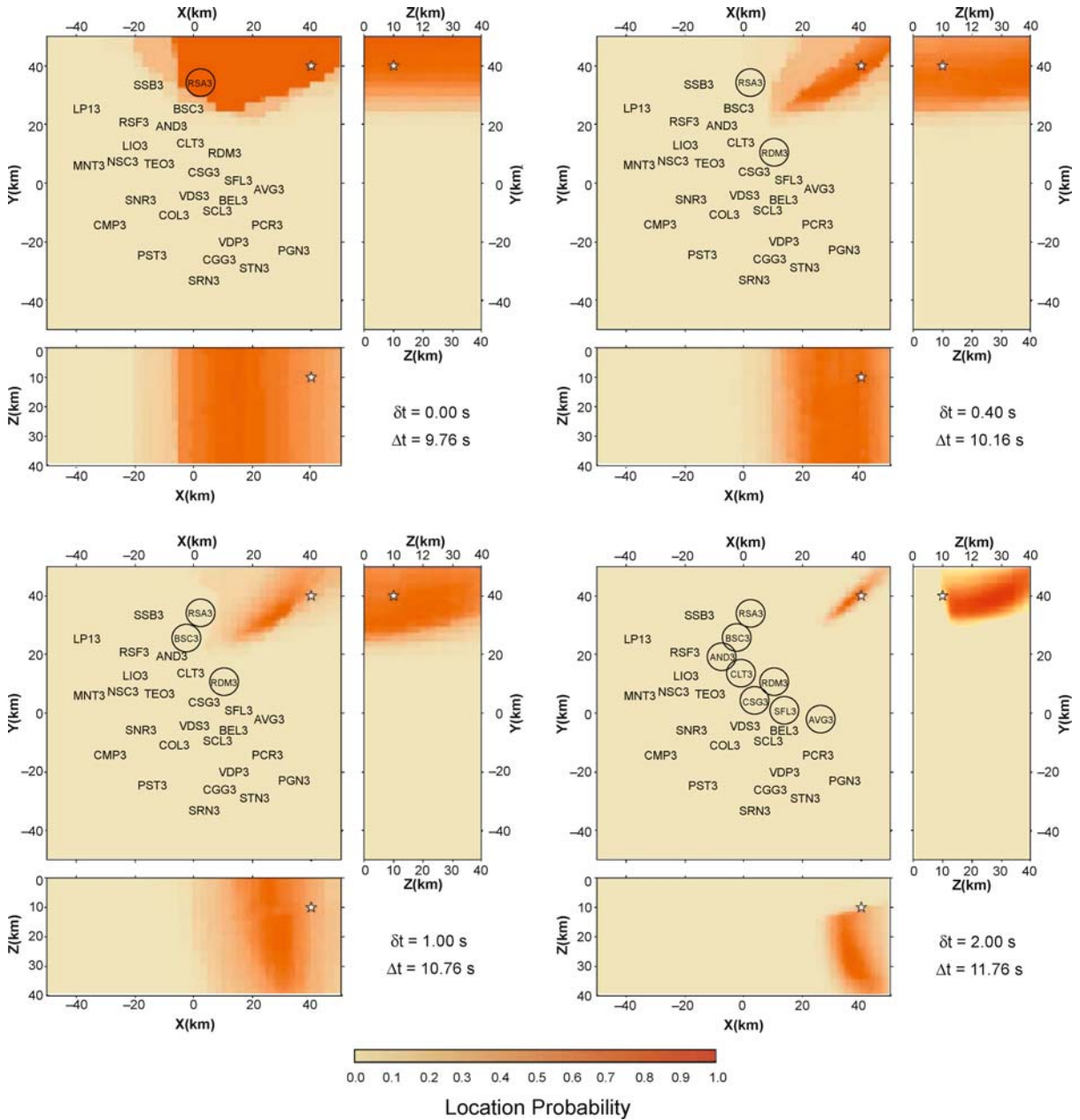
Earthquake Early Warning System in Southern Italy, Figure 9

Location test for a synthetic event occurring at the center of the Irpinia Seismic Network (ISNet). The three orthogonal views show marginal values of the probability function $Q(i, j, k)$. The true hypocenter is identified by a star. δt is the time from the first trigger, Δt is the time from event origin. For each snapshot, stations that have triggered are marked with a circle

However, if one or more of the first arrival times is an outlier, then the earliest estimates of the hypocentral volume may be biased. Synthetic tests have shown that, if N_{out} is the number of outlier data, the bias reduces significantly after about $4 + N_{\text{out}}$ arrivals have been obtained, and then

decreases further with further arrivals, as the solution converges towards a standard EDT location [37].

We performed several synthetic tests using the geometry of the ISNet network. For each simulated event, we computed theoretical arrival picks using travel times ob-



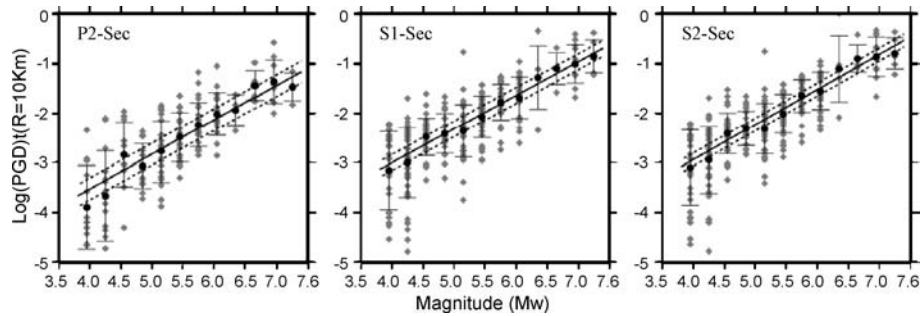
Earthquake Early Warning System in Southern Italy, Figure 10

Location test for a synthetic event occurring outside the ISNet network (see Fig. 9 for explanation)

tained by the finite difference solution of the eikonal equation [35] for a 1D, P -wave velocity model. To reproduce uncertainties introduced by the picking algorithm, we add to each arrival time a random error following a Gaussian distribution with a variance of 0.02 s.

Here we use only P picks since currently most networks have poor capability to perform real-time S pick-

ing. Our tests consider an earthquake occurring at the center of the network at a depth of 10 km (Fig. 9) and an earthquake occurring outside the network at a depth of 10 km (Fig. 10). Each panel in Figs. 9 and 10 is a snapshot at a given time showing the marginal map (i. e., summed over i, j or k) for $Q(i, j, k)$ along the horizontal (x, y) and the two vertical (x, z and y, z) planes. The star shows the



Earthquake Early Warning System in Southern Italy, Figure 11

Correlation between low-pass filtered peak ground motion value and moment-magnitude for earthquakes occurred in the Euro-Mediterranean region (after [49]). The panels show the logarithm of peak ground displacement normalized at a reference distance of 10 km as a function of Mw in time windows of (left) 2 s length from the first *P*-arrival and (middle) 1- and (right) 2-s from the first *S*-arrivals. *P*- and *S*-data are measured on vertical and root-squared sum of horizontal components, respectively. Each panel shows the best fit regression line (solid line) along with 1-WSE limits (dashed lines)

known, synthetic hypocentral location. In the first case, two seconds after the first trigger (5.03 s from the event origin), 9 stations have triggered and the location is already well constrained for early warning purposes.

In the second case, at $\Delta t = 11.76$ s, 2 s after the first event detection, the constraint on the location PDF improves further, but the PDF retains an elongated shape because of the poor azimuthal coverage of the network for this event. The event depth is only constrained by an upper bound, but the depth range includes the true value.

Real-Time Magnitude Estimation Using a Bayesian, Evolutionary Approach

Previous Related Studies The problem of magnitude estimation from early seismic signal has been previously approached and analyzed by different authors.

Nakamura [31] first proposed the correlation between the event magnitude and the characteristic period of *P*-phase defined as the ratio between the energy of the signal and its first derivative.

Allen and Kanamori [2] modified the original Nakamura method and described the correlation between the predominant period and the event magnitude for Southern California events. Lockman and Allen (2007) studied the predominant period – magnitude relations for the Pacific Northwest and Japan. They also investigated the sensitivity of such relations using different frequency bands.

Using a complementary approach, Wu and Kanamori [45] investigated the feasibility of an on-site EEWs for Taiwan region based on prediction of earthquake damage, based on measurements of the predominant period and peak displacement on early *P*-wave signals detected at the network.

Odaka et al. (2003) proposed a single station approach for the real-time magnitude estimation. The authors fit the initial part of waveform envelope and showed a relation between the final event magnitude, the envelope shape coefficient and the maximum *P* amplitude measured in a 3 s time window.

Wu and Zhao [46] and Zollo et al. [49] (Fig. 11) demonstrated the existence of a correlation between the event magnitude and the peak displacement measured a few seconds after the *P* arrival based on massive analysis of Southern Californian and Euro–Mediterranean earthquake records. In particular, Zollo et al. showed that both *P* and *S* wave early phases have the potential for real time estimation of magnitude up to about *M* 7. Zollo et al. [50] and Lancieri and Zollo [26] extended this observation to Japanese earthquake records, showing that a possible saturation effect may exist at about *M* 6.5 for *P* measurements in 2 s windows while it vanishes when a larger, 4 s window is considered. The scaling of displacement peak with magnitude, instead, appears at even shorter (1 s) time lapses after the first *S*-arrival.

Using an alternative method, Simmons [38] proposed a new algorithm based on discrete wavelet transforms able to detect first *P* arrival and to estimate final magnitude analyzing first seconds of *P*-wave.

The Real-Time Magnitude Estimation Method The real time and evolutionary algorithm for magnitude estimation presented in this paper is based on a magnitude predictive model and a Bayesian formulation. It is aimed at evaluating the conditional probability density function of magnitude as a function of ground motion quantities measured on the early part of the acquired signals [19].

The predictive models are empirical relationships which correlate the final event magnitude with the logarithm of quantities measured on first 2–4 s of record.

The first prediction model, based on the predominant period of P -phase (τ_P), has been introduced by Allen and Kanamori [2]. Recently, Wu and Zhao [46] showed the existence of a correlation between magnitude, distance and peak displacement measured in a 2–4 s window after P -phase.

Zollo et al. [49,50] refined this correlation and extended the observation on the peaks measured in 2 s after the S -phase arrival through the analysis of the European and Japanese strong motion data-bases (Ambraseys et al. [3], K-NET www service of NIED – National Research Institute for Earth Science and Disaster Prevention, Japan).

The method therefore assumes that the linear relationship between the logarithm of the observed quantity and magnitude is known, along with standard errors of the predictive models.

At each time step t from the first station trigger, the conditional PDF of magnitude M given the observed data vector $\underline{d} = \{d_1, d_2, \dots, d_n\}$ is expressed via the Bayes theorem as:

$$f(m|\underline{d}) = \frac{f(\underline{d}|m)f(m)}{\int_{M_{\min}}^{M_{\max}} f(\underline{d}|m)f(m)dM}, \quad (7)$$

where $f(m)$ is the a priori distribution which incorporates the information available before the experimental data are collected through a truncated exponential functional form, derived by the Gutenberg–Richter recurrence relationship,

$$f(m) : \begin{cases} \frac{\beta e^{-\beta m}}{e^{-\beta M_{\min}} - e^{-\beta M_{\max}}} & M_{\min} \leq m \leq M_{\max} \\ 0 & m \notin [M_{\min}, M_{\max}] \end{cases}, \quad (8)$$

where $\{\beta, M_{\min}, M_{\max}\}$ depend on the seismic features and on the detection threshold of the seismic network of the considered region.

The conditional probability $f(\underline{d}|m)$ contains all the information concerning the magnitude as retrievable from the data acquired at time t .

Assuming that components of the observed data vector \underline{d} have a lognormal distribution, and that they are stochastically independent and identically distributed random variables of parameters $\mu_{\log(d)}$ and $\sigma_{\log(d)}$, then the likelihood is written as:

$$f(\underline{d}|m) = \prod_{i=1}^v \frac{1}{\sqrt{2\pi}\sigma_{\log(d)}d_i} e^{-\frac{1}{2}\left(\frac{\log(d_i) - \mu_{\log(d)}}{\sigma_{\log(d)}}\right)^2}, \quad (9)$$

where v is the number of stations acquiring at the instant t ; $\mu_{\log(d)}$ and $\sigma_{\log(d)}$ are the mean and the standard deviation of the logs of d_i , respectively..

Substituting Eq. 8 and Eq. 9 into Eq. 7, $f(m|\underline{d})$ results as in Eq. 10 where it depends on data only through $\sum_{i=1}^v \log(d_i)$ and v , which therefore are jointly sufficient statistics for the estimation of magnitude [21]:

$$\begin{aligned} f(m|\underline{d}) &= f\left(m \mid \sum_{i=1}^v \log(d_i)\right) \\ &= \frac{e^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^v \log(d_i)\right) - v\mu_{\log(d)}^2\right) / 2\sigma_{\log(d)}^2} e^{-\beta m}}{\int_{M_{\min}}^{M_{\max}} e^{\left(2\mu_{\log(d)}\left(\sum_{i=1}^v \log(d_i)\right) - v\mu_{\log(d)}^2\right) / 2\sigma_{\log(d)}^2} e^{-\beta m} dM}. \end{aligned} \quad (10)$$

As just outlined, $f(m|\underline{d})$ depends on $\sum_{i=1}^v \log(d_i)$ and on the number of stations triggered, v , at the time of the estimation and, consequently, on the amount of information available. As more stations are triggered, and provide more measures of d , the estimation improves.

The described technique is evolutionary in the sense that $f(m|\underline{d})$ depends on time, i.e., as time passes, additional stations provide new observations (predominant period and/or P -, S -peaks), which are used to refine the probabilistic estimation of magnitude.

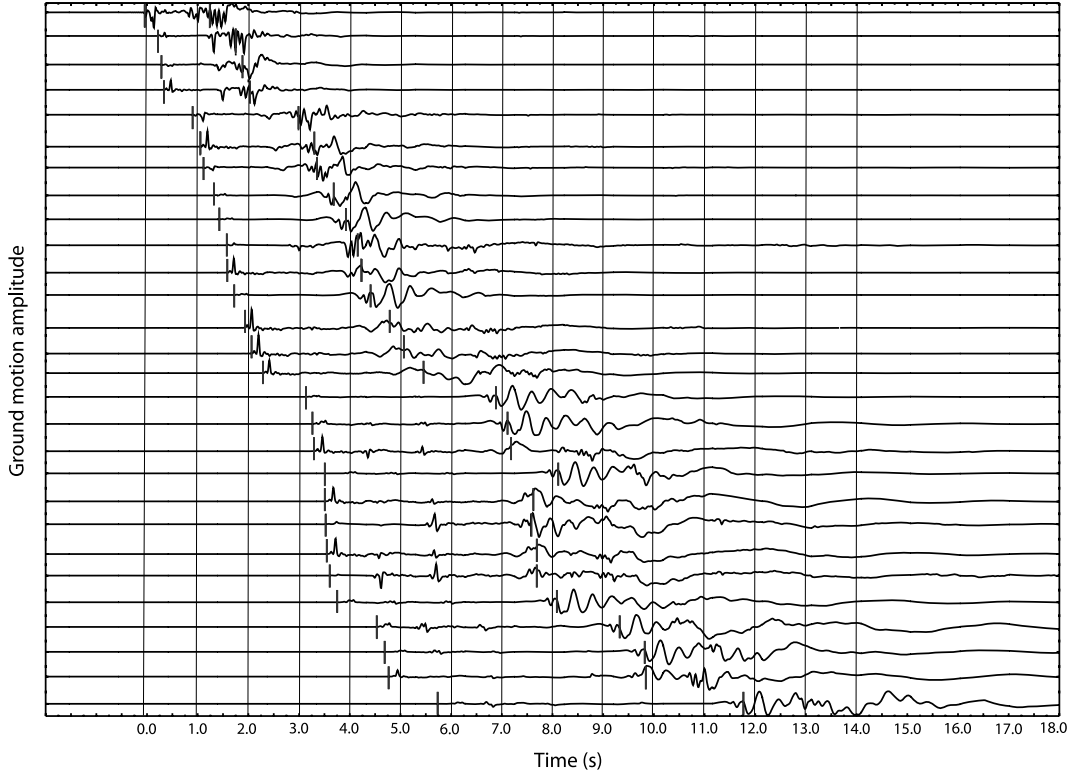
Magnitude Estimation from Peak Displacement Measurements

The empirical relationships between low-pass filtered, initial P - and S -peak displacement amplitudes and moment magnitude (e.g. [49]) can be used as predictive models for the real-time estimation of magnitude using the Bayesian approach described above.

While the P -wave onset is identified by an automatic picking procedure, the S -onset can be estimated from an automatic S -picking or from a theoretical prediction based on the hypocentral distance given by the actual earthquake location. At a given time step after the first P -wave detection at the network, progressively refined estimates of magnitude are obtained from P - and S -peak displacement data. These are preliminarily corrected for distance amplitude effects through an empirical attenuation relationship obtained from available strong motion records [46,49]:

$$f(M, R) = A_{\text{phase}} + B_{\text{phase}}M + C_{\text{phase}} \log(R), \quad (11)$$

where the constants A_{phase} , B_{phase} and C_{phase} are determined through a best-fit regression with a retrieved standard error of $SE_{\text{phase}}^{\text{PMR}}$ and R is the hypocentral distance.



Earthquake Early Warning System in Southern Italy, Figure 12

Synthetic seismograms for a $M 7.0$ earthquake at the center of the network (see Fig. 9). The seismograms are computed using a line source, rupture model (constant rupture velocity) while complete wavefield green's functions in a flat-layered model are computed by using the discrete wavenumber summation method of Bouchon [8]. Each vertical line indicates the 1 s signal packets examined at each time step. This plot allows us to understand seconds after seconds which stations are acquiring and what sort of input (P or S peak) they are giving to the real time system. For example after three seconds to the first P phase picking thirteen stations are acquiring, the 2 s S -phase peak is available at the nearest stations. This observation motivates the use of the S phase information in a real time information. If a dense network is deployed in the epicentral area the nearest station will record the S -phase before the P phase arrives to the far ones, as seen in previous example, and this is perfectly compatible with the real time analysis

Following the procedure described in [49], the relationship (11) is used to correct observed peaks for the distance effect, by normalizing them to a reference distance (e.g., $R = 10$ km) and to determine a new best fit regression between the distance corrected peak value $(PD_{\text{phase}})^{10 \text{ km}}$ and the final magnitude:

$$\log(PD_{\text{phase}}^{10 \text{ km}}) = \log(PD_{\text{phase}}^R) - C_{\text{phase}} \log\left(\frac{R}{10}\right) \quad (12)$$

$$\log(PD_{\text{phase}}^{10 \text{ km}}) = A'_{\text{phase}} + B'_{\text{phase}} M. \quad (13)$$

Assuming a standard error of SE_{phase}^{PM} on peak displacements retrieved from (13) and combining the Eqs. (11) and (13), the mean values and standard deviation of quan-

tity $\log(PD_{\text{phase}})$, can be written as:

$$\begin{aligned} \mu_{\log(PD_{\text{phase}})} &= B'_{\text{phase}} M + A'_{\text{phase}} + C_{\text{phase}} \log\left(\frac{R}{10}\right) \\ \sigma_{\log(PD_{\text{phase}})} &= SE_{\text{phase}}^{PM} + \log\left(\frac{R}{10}\right) \Delta C_{\text{phase}} \\ &\quad + C_{\text{phase}} \frac{1}{R} \Delta R, \end{aligned} \quad (14)$$

where R is estimated with an error of ΔR and ΔC_{phase} is the error on the C_{phase} coefficient in Eq. (12).

The values of coefficients in (14) used for real time magnitude estimates at ISNet are obtained from the regression analysis based on records from the European Strong Motion Database [49] and given in Table 2.

Figure 12 illustrates an example of real time magnitude estimation on a simulated event with $M = 7.0$, whose epicenter is located along the 1980 Irpinia earthquake faulting

Earthquake Early Warning System in Southern Italy, Table 2
Coefficients of the empirical regression relationships between low-pass filtered P and S displacement peaks and magnitude

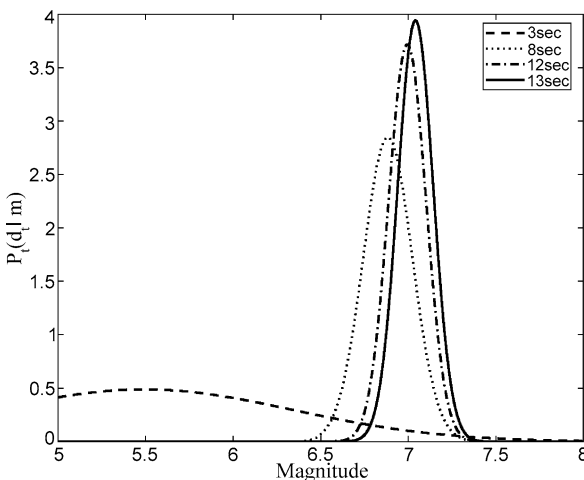
Phase	A'_{phase}	B'_{phase}	C_{phase}	SE_{phase}^{PM}	ΔC_{phase}
2P	-6.31	0.70	-1.05	0.22	0.30
2S	-5.77	0.71	-0.71	0.13	0.16

system. Synthetic seismograms have been computed by using the discrete wave-number method of Bouchon [8] and Coutant (1989) for an extended source model in a flat-layered velocity model.

Figure 13a shows the probability density function defined in Eq. (8) evaluated at each time step. Time zero is assigned to the first P detection at the network. As time evolves the PDF tightens around the predicted magnitude value, indicating a more refined, probabilistic estimate of magnitude.

By defining $F_t(m)$ as the cumulative PDF at time t , it is possible to estimate a magnitude range of variation $[M_{\min}, M_{\max}]$ whose limits are defined based on the shape of the $F_t(m)$ function:

$$\begin{aligned} M_{\min} : \int_{-\infty}^{M_{\min}} f_t(m|d)dm &= \alpha, \\ M_{\max} : \int_{-\infty}^{M_{\max}} f_t(m|d)dm &= 1 - \alpha. \end{aligned} \quad (15)$$



Earthquake Early Warning System in Southern Italy, Figure 13

Application of the method for real time magnitude estimate to a $M 7$ simulated event occurring within the area covered by the ISNet network. *Left panel.* PDF distribution at several time steps measured from the first P -phase picking. *Right top,* magnitude estimation with uncertainties as a function of time. The *dashed line* refers to the actual magnitude value, the errors represent the 95% of confidence bound evaluated as cumulative PDF integral in the 5–95% range. *Right bottom,* probability to exceed magnitude 6.5 and magnitude 7.5 thresholds in function of time. The *dashed line* is the 75% probability level

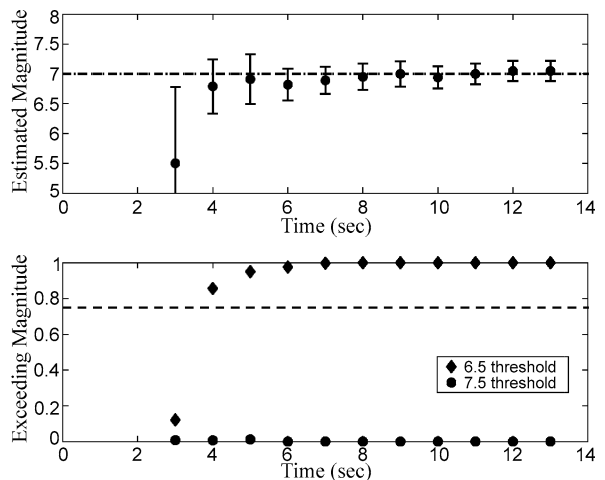
For example, if we assume $\alpha = 1\%$, then M_{\min} and M_{\max} will be, respectively, the $F_t(m)$ evaluated at 0.01 and 0.99.

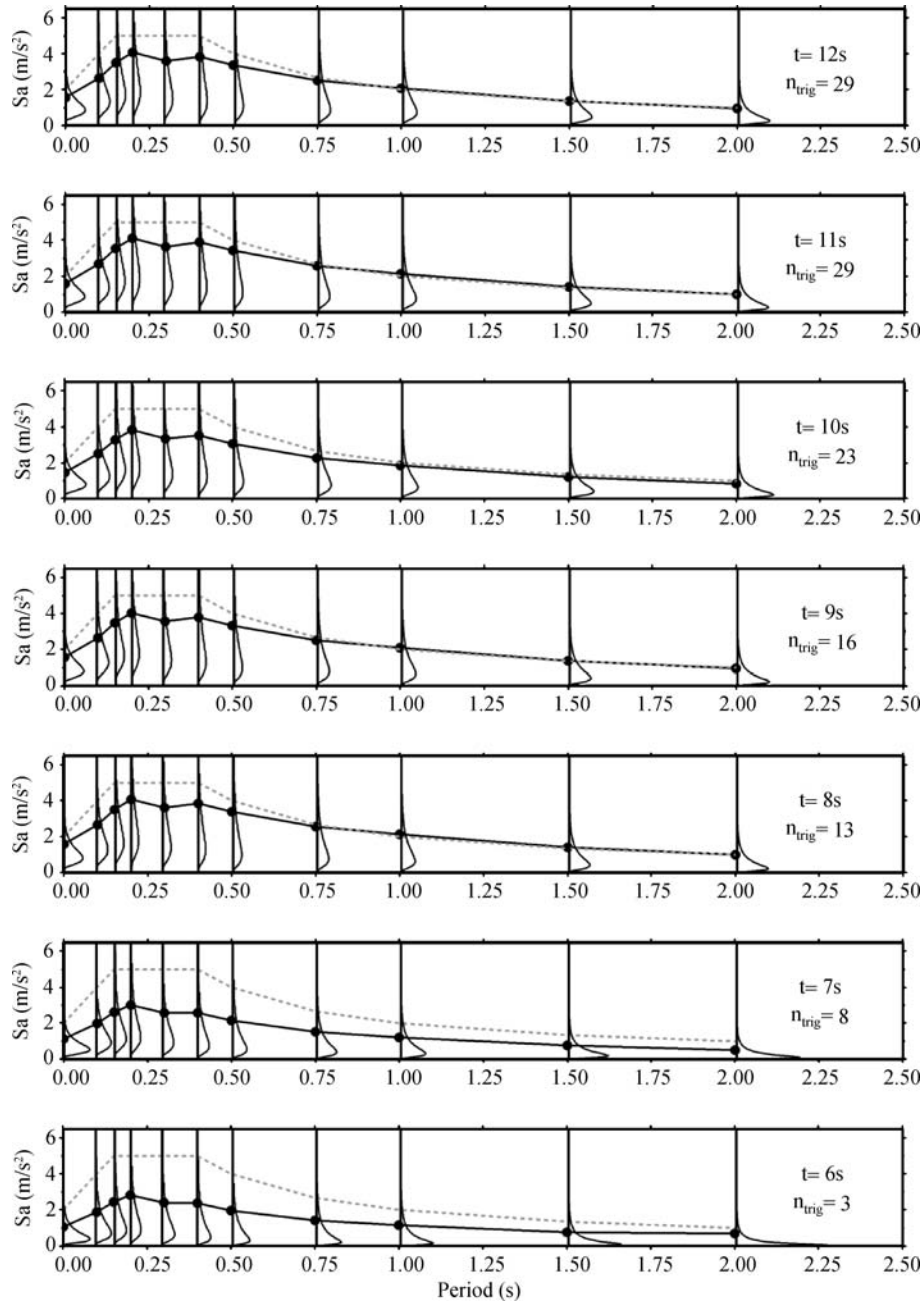
In Fig. 13b the estimates of magnitude uncertainty bounds are reported as a function of time. After three iterations (corresponding to a time of 9 s from the event origin time and 4 s after the first P -phase arrival at the network) the magnitude estimation converges to the true magnitude value. In fact, due to the high density of seismic station in the epicentral area, at that time most of seismic station contributes to the magnitude estimation with peaks read on P -phase windows (Fig. 14), while a further refinement of magnitude estimate is due to the near source S -wave arrivals.

Real-Time Hazard Analysis for Earthquake Early Warning

The Real-Time Hazard Determination

Using the methods previously described for estimating in real-time the event magnitude and location, it is possible to perform a real-time hazard analysis [19]. This analysis is based on the extension of classical Probabilistic Seismic Hazard Analysis (PSHA) proposed by Cornell [11] that is generally used for long-term probabilistic hazard assessment. Classical PSHA integrates data from existing seismic catalogs both in terms of magnitude, location and recorded strong ground motion values in addition to the information concerning seismogenic areas of interest (ex-





Earthquake Early Warning System in Southern Italy, Figure 14

Real-time estimation of spectral ordinates' distributions as function of the number of stations triggered for a M 7.0 event with an epicentral distance of 50 km from the early warning target site. The parameter n_{trig} in the figure is equivalent to the number of stations ν in the text. The acceleration spectrum (*black curve*) was obtained by choosing at each period the spectral value with 20% exceedance probability according to the corresponding distribution, so it is analogous to a uniform hazard spectrum with the exception that it is computed in real-time. The *grey dashed line* is the Italian code spectrum assigned for building design in the target location at the town of Avellino, 40 km distant from the earthquake epicenter, and is reported for comparison purposes (after [10])

pected maximum magnitude, b-value of the Gutenberg Richter relationship, etc.) to provide the hazard curve as the final outcome. Each point on that curve corresponds to the value of a ground motion intensity measure (IM) (e. g., peak ground acceleration, PGA, peak ground velocity, PGV or the spectral acceleration, Sa), having a given probability or frequency of exceedance in a fixed period of time for a site of interest.

The probabilistic framework of the PSHA, specifically the hazard integral, can be used for real-time hazard if the PDFs of magnitude and source-to-site distance are replaced with those depending on the data gathered by the EEWS during the occurrence of a specific earthquake.

This is the case, for example, of the PDF on the source-to-site distance whose statistical moments evolve with real-time earthquake location. As a consequence, this PDF does not depend on the seismic potential of the area of interest (as in the case of the classical PSHA, which accounts for the occurrence of all the earthquake in a fixed range of magnitude), but rather depends on the time evolving event location provided by the EEWS. The same considerations apply to the PDF on the magnitude as described in the following sections whose statistical moment, at a given time, depends on the number of triggered stations at that time.

In this theoretical framework the real-time hazard integral can be written as:

$$f(\text{IM}|\underline{d}, \underline{s}) = \int_M \int_R f(\text{IM}|m, r) f(m|\underline{d}) f(r|\underline{s}) dM dR, \quad (16)$$

where $f(r|\underline{s})$ is the PDF of distance r , which eventually depends only on the triggering sequence of the stations in the network, where $\underline{s} = \{s_1, \dots, s_v\}$ is such a sequence. This renders also the PDF of r time dependent.

Given that for each point in a volume containing the earthquake hypocenter, the probability of that point being coincident with the true hypocenter is calculated via a rapid location technique, a simple geometrical transformation allows one to obtain the probabilistic distribution of the source-to-site distance.

The PDF $f(\text{IM}|m, r)$, is given, for example, by an ordinary attenuation relationship. It is worth to recall that the computed hazard refers to a particular set of triggered stations and, consequently, it depends on the information available at time t from the first detection of the event.

Figure 14 illustrates, as an example, the estimation of spectral acceleration ordinates for different periods, for a M 7.0 event located at an epicentral distance of 50 km from the early warning target site [10].

We note the evolution of Sa predictions via the corresponding PDFs. The different panels correspond to increasing times from the earthquake origin and, therefore, to different numbers of stations triggered.

The False Alarm Issue

Once the EEWS provides a probability distribution of the ground motion intensity measure (IM) at the target site (e. g., peak ground acceleration or velocity), a decisional condition has to be checked in order to decide whether to alert or not.

Several options are available to formulate a decisional rule, for example the alarm may be issued if the probability of the predicted IM exceeding a critical threshold (IM_C) is greater than a reference value (P_c):

$$\text{Alarm if: } \int_0^{\text{IM}_C} f(\text{IM}|\underline{d}, \underline{s}) d(\text{IM}) = P[\text{IM} > \text{IM}_C] > P_c. \quad (17)$$

The efficiency of the decisional rule may be evaluated in terms of false and missed alarms probabilities (known as the “cry wolf” issue, e. g., [20]). The false alarm occurs when, on the basis of the information processed by the EEWS, the alarm is issued while the intensity measure at the site IM_T (T subscript means “true”, indicating the realization of the IM to be distinguished from the prediction IM_C) is smaller than the threshold IM_C . A missed alarm corresponds to not launching the alarm if needed,

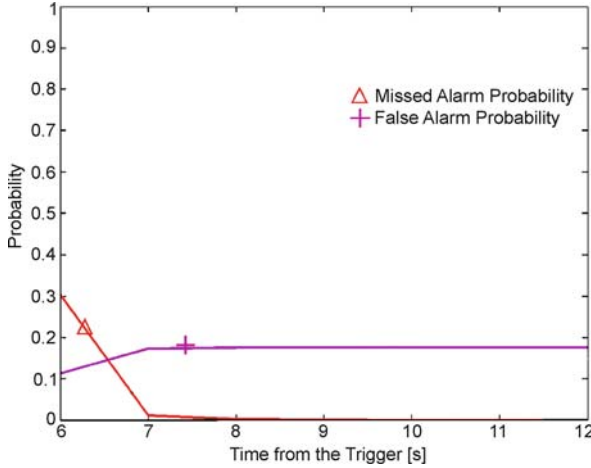
$$\begin{aligned} \text{Missed Alarm} &: \{\text{No Alarm} \cap \text{IM}_T > \text{IM}_C\} \\ \text{False Alarm} &: \{\text{Alarm} \cap \text{IM}_T \leq \text{IM}_C\}. \end{aligned} \quad (18)$$

It has been discussed above how the information and the uncertainties on earthquake location and magnitude are dependent on the number of stations triggered at a certain time.

Therefore, in principle, the decisional rule may be checked at any time after the first station has triggered and, consequently, the false and missed alarm probabilities are also time dependent.

Using the decisional rule of (18) and considering PGA as IM, the time evolution of false/missed alarm probabilities has been simulated for the Campania EEWS, given the occurrence of a M 7 earthquake, and a target site at an epicentral distance of 110 km.

Figure 15 reports the missed and false alarm probabilities as a function of time from the first trigger at the ISNet network.



Earthquake Early Warning System in Southern Italy, Figure 15
Example of estimation of false and missed alarm probabilities as function of the time from the first trigger for a $M 7.0$ event with an epicentral distance of 110 km from the early warning target site. For the decisional rule adopted in this case the threshold is $PGAc = 0.3 \text{ m/s}^2$ and the limit probability is $P_c = 0.2$ (after [19])

A Loss Estimation Approach to Early Warning

Magnitude and distance distributions conditioned to the measurements of the seismic network can also be used for a real-time estimation of risk, which includes losses produced by the earthquake [21]. Based on the real-time risk assessment, a security action aimed at risk mitigation is undertaken if the alarm is issued.

For example, some critical system could shut down or people in buildings may shelter themselves if the warning time is not sufficient to evacuate the dangerous buildings. More complex security measures may be related to the semi-active control of buildings [22].

Therefore, if an EEWS exists, it may trigger a security procedure in case of warning. The estimation of the expected losses for a specific building may be computed, for the case of warning issued and not issued respectively:

$$E^W[L|\underline{d}, \underline{s}] = \int_L \int_{DM} \int_{EDP} \int_{IM} l f^W(l|dm) f(dm|edp) \times f(edp|im) f(im|\underline{d}, \underline{s}) dL dDM dEDP dIM, \quad (19)$$

where $f^W(l|dm)$ is the PDF of the loss (L) given the structural and non-structural damage vector (DM) reflecting the risk reduction in the case of warning; and $f^W(l|dm)$ is the loss function if no alarm is issued (no security action is undertaken); $f(dm|edp)$ is the joint PDF of damages given the Engineering Demand Parameters (EDP), proxy for the structural response; $f(edp|im)$ is the joint PDF of

the EDPs conditioned to a vector of ground motion intensity measures (IM); $f(im|\underline{d}, \underline{s})$ is the real-time hazard expressed by (16) in the case of a scalar IM.

Being able to compute, before the ground motion hits the site, the expected losses in case of warning (W) or not (\bar{W}), is relevant for taking the optimal decision, i. e., to alarm if this reduces the expected losses and to not issue any warning otherwise:

$$\text{to alarm if } E^W[L|\underline{d}, \underline{s}] \leq E^{\bar{W}}[L|\underline{d}, \underline{s}]$$

Optimal decision: (20)

$$\text{to not alarm if } E^W[L|\underline{d}, \underline{s}] > E^{\bar{W}}[L|\underline{d}, \underline{s}]$$

which is a better decisional rule in respect to that of (18).

Computing and comparing expected losses, conditioned to the real-time information coming from the EEWS, in the case of alarming or not, allows the determination of the alarm threshold above which it is convenient to issue the warning according to the optimally maximum criterion.

Assessment of average loss reduction determined by issuing an Early Warning provides a quantitative tool to evaluate the efficiency and feasibility of an EEWS.

Other potential advantages given by this approach are that: (a) the threshold may be set on a statistic (i. e., the summation of the logs) inferred from seismic network measurements, dramatically reducing the required computational effort for real-time decision making; (b) it minimizes the cry wolf issue reducing the probability of false and missed alarms thanks to threshold optimization. In fact, although the number of MA and FA depend on the decisional rule adopted to issued the alarm, the approach developed in Iervolino et al. [20,21,22] avoids explicitly considering the missed and false alarm rates associated with the decision, as the choice to alarm or not is taken based on the expected economic loss (not on the estimation of peak ground motion). In other words, if in computing the expected loss one accounts for the costs of false and missed alarms, there is no need to optimize the *cry wolf* issue, and MA and FA rates are at their values determined by the respective costs, and in this sense are optimal.

Future Directions

We have analyzed and illustrated the main scientific and technological issues related to the implementation and management of an earthquake early warning system under development in the Campania region of southern Italy.

The system is designed for early warning alert notification at distant coastal targets based on a dense, wide-dynamic seismic network (accelerometers, seismometers

and broadband sensors) deployed in the Apenninic belt region (ISNet – Irpinia Seismic Network). It can therefore be classified as a regional Early Warning System consisting of a wide seismic sensor network covering a portion or the entire area which is threatened by a quake's strike.

According to [25], real-time estimates of earthquake location and magnitude are needed for regional warning systems (EEWS), i. e., dense seismic networks covering all or a portion of an area of interest. However the alarm decision in an early warning system is based, rather, on the prediction, with quantified confidence, of a ground motion intensity at a distant target site (where a sensitive structure is located). This problem needs an evolutionary (i. e., time-dependent) and probabilistic frame where pdfs for earthquake location, magnitude and attenuation parameters are combined to perform a real-time probabilistic seismic hazard analysis (e. g., [19]).

Considering the peak displacement amplitude and/or predominant frequency measured in the early portion of *P*-waves, we have shown that suitable probability density functions for the earthquake location and magnitude parameters can be constructed and used for real-time probabilistic assessment of false alarms and loss estimation, which are the key elements based on which automatic actions can be undertaken to mitigate earthquake effects.

Based on the analysis of acceleration records of Euro-Mediterranean and Japanese earthquakes, Zollo et al. [49,50] have shown the advantages of using near source strong motion records for real time estimation of earthquake magnitude. In fact they provide unsaturated recordings of moderate to large earthquakes and, in case of dense station coverage of the source area, the combination of both *P*- and *S*-wave amplitude information can be used to get fast and robust earthquake location and magnitude estimates.

We support the use of *S*-waves recorded in the near-source of an impending earthquake for earthquake early warning, especially in view of the excellent correlation that *S*-peaks show with magnitude up to about $M = 7$ for Euro-Mediterranean and Japanese earthquakes [49,50]. Dense accelerometric networks now operating in Europe, USA, Taiwan, Japan and other seismic regions in the world can provide a sufficient number of records at distances smaller than 20–30 km from potentially damaging crustal earthquakes so that *S*–*P* times are expected to be smaller than 2–3 s. A magnitude estimation using *S*-waves could be therefore available 4–5 s after the first *P*-wave is recorded, which is still useful for sending an alert to distant target sites.

Although relatively few magnitude 7 and larger earthquakes have hit the Apenninic belt, and generally the

Mediterranean region, during the last century, there have been many instances of damaging quakes in the magnitude 6 range.

Earthquake early warning systems have the potential to mitigate the effects of moderate size earthquakes ($M = 6$ –7), which can produce severe damage in densely urbanized areas and places where old structures were not built to current standards. This has been the case for a significant number of earthquakes occurred in the Mediterranean basin during last decades: the 1976 Friuli ($M = 6$ –6.5) and 1997 Colfiorito ($M = 6$) in Italy, 1999 Athens ($M = 5.9$) in Greece, 2002 Nahrin, in Afghanistan ($M = 6.1$), 2003 in Algeria ($M = 6.7$), 2003 Bam ($M = 6.3$) in Iran, 2004 in Morocco ($M = 6.4$).

An earthquake early warning system can be effective for mitigating the effects of moderate earthquakes. For moderate size events, early warning systems could also mitigate earthquake effects in terms of infrastructure operability (e. g., hospitals, firehouses, telecommunication hubs, ...) during the post-event emergency phase and rescue operations. For instance, in tall buildings, the higher floors generally sway much more than those near ground level, so that even a moderate earthquake could cause severe damage to a high rise. Therefore, even at 70–80 km distance from its epicenter, a magnitude 6 quake could affect hospital operating rooms and other critical installations.

Installations as close as 50 km from the epicenter could receive an earthquake warning 10 s prior to the arrival of the more energetic waves (*S* and surface waves) of an earthquake. To take advantage of this brief warning period, automated systems would have to be created that respond instantly to notification alert signals, and they would have to be carefully calibrated to avoid false or missed alarms. Closer to the epicenter, a magnitude 6 or higher earthquake can damage critical infrastructures, such as telephone lines, gas pipelines, highways, and railroads, as well as airport runways and navigation systems. These disruptions would have a domino effect in more distant areas, which could be mitigated by an early warning alert system, based on the earliest primary wave data to arrive at recording stations close to the epicenter.

Finally, we note that earthquake early warning systems can also help mitigate the effects of such earthquake-induced disasters as fires, explosions, landslides, and tsunamis, which can in many cases be more devastating than the earthquake itself. Systems could be installed at relatively low cost in developing countries, where moderate sized earthquakes can cause damage comparable to that caused by much larger earthquakes in developed countries.

Bibliography

Primary Literature

- Allen RM (2007) The ElarmS earthquake early warning methodology and its application across California. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin, pp 21–44. ISBN-13 978-3-540-72240-3
- Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300:786–789. doi:10.1126/science.1080912
- Ambraseys N, Smit P, Douglas J, Margaris B, Sigbjornsson R, Olafsson S, Suhadolc P, Costa G (2004) Internet site for European strong-motion data. *Boll Geofis Teor Appl* 45(3):113–129
- Bakun W, Fischer HF, Jensen E, VanSchaack J (1994) Early warning system for aftershocks. *Bull Seismol Soc Am* 84(2):359–365
- Bernard P, Zollo A (1989) The Irpinia (Italy) 1980 earthquake: detailed analysis of a complex normal fault. *J Geophys Res* 94:1631–1648
- Guidoboni E, Ferrari G, Mariotti D, Comastri A, Tarabusi G, Valensise G (2007) CFTI4Med, Catalogue of Strong Earthquakes in Italy (461 B.C.–1997) and Mediterranean Area (760 B.C.–1500). INGV-SGA. Available from <http://storing.ingv.it/cfti4med/>
- Bose M, Ionescu C, Wenzel F (2007) Earthquake early warning for Bucharest, Romania: Novel and revised scaling relations. *Geophys Res Lett* 34:L07302. doi:10.1029/2007GL029396
- Bouchon M (1979) Discrete wave number representation of elastic wave fields in three-space dimensions, *J Geophys Res* 84:3609–3614
- Cinti FR, Faenza L, Marzocchi W, Montone P (2004) Probability map of the next $M \geq 5.5$ earthquakes in Italy. *Geochim Geophys Geosyst* 5:Q1103. doi:10.1029/2004GC000724.
- Convertito V, Iervolino I, Giorgio M, Manfredi G, Zollo A (2008) Prediction of response spectra via real-time earthquake measurements. *Soil Dyn Earthq Eng* 28(6):492–505. doi:10.1016/j.soildyn.2007.07.006
- Cornell CA (1968) Engineering seismic hazard analysis. *Bull Seismol Soc Am* 59(5):1583–1606
- Cua G, Heaton T (2007) The virtual seismologist (VS) method: A Bayesian approach to earthquake early warning. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin. doi:10.1007/978-3-540-72241-0_7
- Curtis A, Lomax A (2001) Prior information, sampling distributions and the curse of dimensionality. *Geophysics* 66:372–378. doi:10.1190/1.1444928
- Erdik M, Fahjan Y, Ozel O, Alcik H, Mert A, Gul M (2003) Istanbul earthquake rapid response and the early warning system. *Bull Earthquake Eng* 1(1):157–163. doi:10.1023/A:1024813612271
- Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City seismic alert system. *Seismol Res Lett* 66:42–53
- Font Y, Kao H, Lallemand S, Liu C-S, Chiao L-Y (2004) Hypocentral determination offshore Eastern Taiwan using the maximum intersection method. *Geophys J Int* 158(2):655–675. doi:10.1111/j.1365-246X.2004.02317.x
- Grasso V, Allen RM (2005) Earthquake warning systems: Characterizing prediction uncertainty. *Eos Trans AGU* 86(52), Fall Meet. Suppl., Abstract S44B-03
- Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. *Bull Seism Soc Am* 95(2):708–718. doi:10.1785/0120030133
- Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2006) Real time risk analysis for hybrid earthquake early warning systems. *J Earthq Eng* 10(6):867–885
- Iervolino I, Convertito V, Giorgio M, Manfredi G, Zollo A (2007a) The cry wolf issue in seismic early warning applications for the campania region. In: Gasparini P et al (eds) *Earthquake early warning systems*. Springer, Berlin. doi:10.1007/978-3-540-72241-0_11
- Iervolino I, Giorgio M, Manfredi G (2007b) Expected loss-based alarm threshold set for earthquake early warning systems. *Earthq Eng Struc Dyn* 36(9):1151–1168. doi:10.1002/eqe.675
- Iervolino I, Manfredi G, Cosenza E (2007c) Earthquake early warning and engineering applications prospects. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin, doi:10.1007/978-3-540-72241-0_12
- Jenny S, Goes S, Giardini D, Kahle H-G (2006) Seismic potential of Southern Italy. *Tectonophysics* 415:81–101. doi:10.1016/j.tecto.2005.12.003.
- Kamigaichi O (2004) JMA earthquake early warning. *J Japan Assoc Earthq Eng* 3:134–137
- Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Ann Rev Earth Planet Sci* 33:195–214. doi:10.1146/annurev.earth.33.092203.122626
- Lancieri M, Zollo A (2008) A bayesian approach to the real-time estimation of magnitude from the early P- and S-wave displacement peaks. *J Geophys Res*. doi:10.1029/2007JB005386, in press
- Lomax A (2005) A Reanalysis of the hypocentral location and related observations for the great 1906 California earthquake. *Bull Seism Soc Am* 95(3):861–877. doi:10.1785/0120040141
- Meletti C, Patacca E, Scandone P (2000) Construction of a seismotectonic model: The case of Italy. *Pure Appl Geophys* 157:11–35
- Montone P, Mariucci MT, Pondrelli S, Amato A (2004) An improved stress map for Italy and surrounding regions (central Mediterranean). *J Geophys Res* 109:B10410. doi:10.1029/2003JB002703
- Munich Re (eds) (2005) *Environmental report – perspectives – Today's ideas for tomorrow's world*, WKD-Offsetdruck GmbH, München
- Nakamura Y (1988) On the urgent earthquake detection and alarm system (UrEDAS). *Proc 9th World Conf Earthquake Eng VII*, Toyko, 673–678
- Nakamura Y (1989) Earthquake alarm system for Japan railways. *Japan Railway Eng* 109:1–7
- Nakamura Y (2004) Uredas, urgent earthquake detection and alarm system, now and future. *13th World Conference on Earthquake Engineering* 908
- Okada T et al (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bull Seismol Soc Am* 93(1):526–532. doi:10.1785/0120020008
- Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys. J Int.* 105:271–284

36. Rydelek P, Pujol J (2004) Real-time seismic warning with a 2-station subarray, *Bull Seism Soc Am* 94(4):1546–1550. doi:10.1785/012003197
37. Satriano C, Lomax A, Zollo A (2008) Real-time evolutionary earthquake location for seismic early warning. *Bull Seism Soc Am* 98(3):1482–1494. doi:10.1785/0120060159
38. Simons F, Dando JB, Allen R (2006) Automatic detection and rapid determination of earthquake magnitude by wavelet multiscale analysis of the primary arrival. *Earth Planet Sci Lett* 250:214–223. doi:10.1016/j.epsl.2006.07.039
39. Teng TL, Wu Y-M, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong-motion map, effective epicenter, and effective magnitude. *Bull Seism Soc Am* 87(5):1209–1219
40. Tsukada S (2006) Earthquake early warning system in Japan. Proc 6th Joint Meeting UJNR Panel on Earthquake Research, Tokushima, Japan
41. Valensise G, Amato A, Montone P, Pantosti D (2003) Earthquakes in Italy: Past, present and future. *Episodes* 26(3):245–249
42. Wenzel FM et al (1999) An early warning system for Bucharest. *Seismol Res Lett* 70(2):161–169
43. Westaway R, Jackson J (1987) The earthquake of 1980 November 23 in Campania-Basilicata (southern Italy), *Geophys J R Astron Soc* 90:375–443. doi:10.1111/j.1467-6435.1999.tb00581.x
44. Wu Y-M, Teng T (2002) A virtual subnetwork approach to earthquake early warning. *Bull Seismol Soc Am* 92(5):2008–2018. doi:10.1785/0120040097
45. Wu Y-M, Kanamori H (2005) Experiment on onsite early warning method for the Taiwan early warning system. *Bull Seismol Soc Am* 95(1):347–353. doi:10.1785/0120040097
46. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds of *p*-wave amplitude in earthquake early warning. *Geophys Res Lett* 33:L16312. doi:10.1029/2006GL026871
47. Wurman G, Allen RM, Lombard P (2007) Toward earthquake early warning in Northern California. *J Geophys Res* 112:B08311. doi:10.1029/2006JB004830
48. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method, *J Geophys Res* 99(B8):15439–15455
49. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophys Res Lett* 33:L23312. doi:10.1029/2006GL027795
50. Zollo A, Lancieri M, Nielsen S (2007) Reply to comment by P. Rydelek et al on “Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records”. *Geophys Res Lett* 34:L20303. doi:10.1029/2007GL030560.

Books and Reviews

- Berger JO (1985) *Statistical decision theory and Bayesian analysis*. Springer, New York
- Coutant O (1989) *Program de simulation numerique AXITRA*. Rapport LGIT, Grenoble, France
- Gruppo di Lavoro MPS (2004) *Redazione della mappa di pericolosità sismica prevista dall'Ordinanza PCM 3274 del 20 marzo (2003) Rapporto Conclusivo per il Dipartimento della Protezione Civile, INGV, Milano-Roma, aprile (2004) 65 pp + 5 appendici*
- Milne J (1886) *Earthquakes and other earth movements*. Appleton, New York, p 361

Earthquake Engineering, Non-linear Problems in

MIHAILO D. TRIFUNAC

Department of Civil Engineering,

University of Southern California, Los Angeles, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Vibrational Representation of Response

Response in Terms of Wave Propagation –

An Example

Observations of Nonlinear Response

Future Directions

Bibliography

Glossary

Meta-stability of man-made structures is the consequence of their upright construction above ground. For excessive dynamic (earthquake) loads, when the lateral deflection exceeds some critical value (this is normally accompanied by softening nonlinear behavior of the structural members), the overturning moment of the gravity forces becomes larger than the restoring moment, and the structure becomes unstable and moves exponentially toward collapse.

Complex and evolving structural systems are structures with a large number of degrees of freedom and many structural members, which for given loads experience softening nonlinear deformations. During strong excitation, continuous changes (typically decreases) in effective stiffness and time-dependent changes in boundary conditions result in a system whose properties are changing with time.

Soil–structure interaction is a process in which the soil and the structure contribute to mutual deformations while undergoing dynamic response. In time, with continuously changing contact area between the foundation and the soil (opening and closing of gaps), when the deformations are large, soil–structure interaction is characterized by nonlinear geometry and nonlinear material properties in both the soil and in the structure.

Definition of the Subject

Nonlinear problems in structural earthquake engineering deal with the dynamic response of meta-stable, man-

made buildings subjected to strong earthquake shaking. During earthquakes, structures constructed on soft sediments and soils deform together with the underlying soil in the dynamic process called soil–structure interaction. Strong shaking forces the soil–structure systems to evolve through different levels of nonlinear response, with continuously changing properties that depend upon the time history of excitation and on the progression and degree of damage. Thus far, the analyses of this response have used the vibrational approach and lumped mass discrete models to represent real structures. Loss of life and property, however, continue to be high during strong shaking in the vicinity of the faults responsible for earthquakes. This calls for new, more physically refined methods of analysis, which can be based on nonlinear wave propagation, and for balancing of the structural capacities with the power carried by the earthquake waves.

After a brief discussion of the literature on the complex and chaotic dynamics of simple mechanical oscillators, the dynamic characteristics and governing equations in the meta-stable structural dynamics of earthquake engineering are introduced. The nature of the solutions of the governing equations in terms of both the vibrational and the wave representations is discussed, and the dynamic instability, material and geometric nonlinearities, and complexities of the governing equations associated with nonlinear soil–structure interaction are described. Collectively, the examples presented reflect the complex physical nature of meta-stable structural systems that experience nonlinear dynamic response, the characteristics of which change and evolve during earthquake excitation.

Introduction

Earthquake engineering, through a cooperation of structural and geotechnical engineers with seismologists and geologists, aims to develop methods for safer design of man-made structures to withstand shaking near intermediate and large earthquakes. This requires addressing the problems of predictability of the response of complicated nonlinear systems, which is one of the important subjects of modern nonlinear science. Through the studies of the dynamic response, earthquake engineers address complex physical problems and issues with important social implications.

The completeness and beauty of the linear differential equations appear to have led to their dominance in the mathematical training of engineers and scientists during most of the 20th century. The recognition that chaotic dynamics is inherent in all nonlinear physical phenomena, which has created a sense of revolution in applied me-

chanics and physics today, so far has had little if any effect on the research and design of earthquake-resistant structures. In the past, the designs in structural engineering and control systems were kept within the realm of linear system dynamics. However, the needs of modern technology have pushed the design into the nonlinear regimes of large deformations, which has increased the possibility of encountering chaotic dynamic phenomena in structural response. Even a cursory review of papers on chaotic vibrations in mechanical systems leads to the conclusion that chaotic dynamics is not a small, insignificant class of motions and that chaotic oscillations occur in many nonlinear systems and for a wide range of values of the parameters.

If an engineer chooses parameters that produce chaotic output, then he or she loses predictability. However, the chaotic behavior of nonlinear systems does not exclude predictability of the response but rather introduces upper bounds (prediction horizons) [31] and renders the predictions probabilistic. The important question is then over what time-scale are the forecasts reliable, given the current state and knowledge of the system. Another key ingredient for prediction is an adequate physical model. At present, because of the multitude of interacting phenomena and the absence of physically complete equations of motion, there exists no adequate general model of the complete earthquake response process. While the practical outcome of most work in earthquake engineering remains empirically based, the nonlinear methods are gaining popularity, aiming to decipher the governing phenomena and to assess the reliability of the models. It appears now that the broad-based revolution in the worldview of science that begun in the twentieth century will be associated with chaotic dynamics [43]. This revolution should eventually also contribute to better understanding and more complete representation of the response analyses in earthquake engineering.

It has been argued that major changes in science occur not so much when new theories are advanced but when the simple models with which scientists conceptualize a theory are changed [24]. In vibrations, such a conceptual model that embodies the major features of a whole class of problems is the spring-mass system. Lessons emerging from studies of the spring-mass model and several other relevant models can serve as conceptual starting points for generalizations and also as a guide to further studies of more complex models in earthquake engineering and structural dynamics.

Studies of forced vibrations of a pendulum have revealed complex dynamics and chaotic vibrations [14,15]. A simply supported beam with sub-buckling axial com-

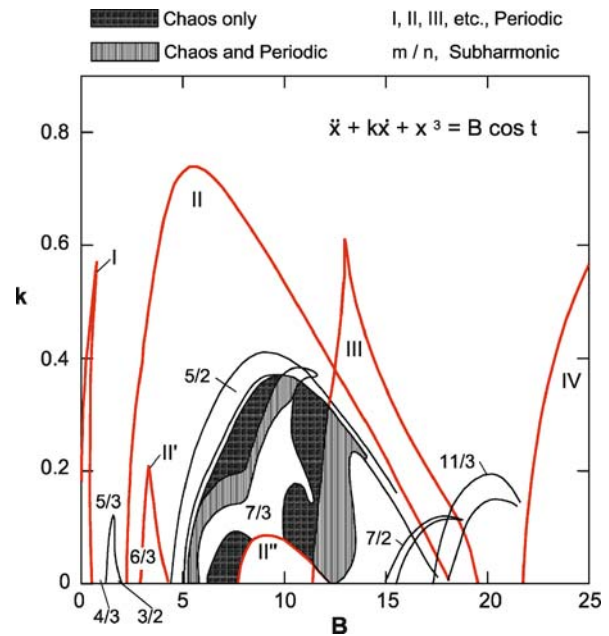
pression modeled by a single mode approximation yields a Mathieu type equation and for certain values of the parameters leads to unstable solutions. When nonlinearities are added, these vibrations result in a limit cycle. A related problem is a classical pendulum with a vibrating pivot support, which also leads to chaotic vibrations [29,34]. Chaotic motions in a double pendulum have been studied by Richter and Scholz [45], and the complex dynamics and chaotic solutions for a spherical pendulum with two degrees of freedom have been described by Miles [35].

Impact-type problems result in explicit difference equations or maps, which can yield chaotic vibrations for certain values of the governing parameters [30]. A mass vibrating in a gap between two stiff springs on either side [17,46,47] is a simple related model, which suggests a starting points for research in nonlinear vibration of piles, and for impact-type interaction of adjacent buildings, excited by strong earthquake ground motion. The reader can find examples of such problems in the description of damage in Mexico City, for example, during several earthquakes [32].

Chaotic motions of an elasto-plastic arch have been studied by Poddar et al. [42]. Forced vibrations of a buckled beam, modeled by the Duffing equation, showed that chaotic vibrations are possible [16]. Forced vibrations described by a Duffing equation with viscous damping and nonlinear (cubic) elastic (stiffening) spring were studied by Ueda [67]. Figure 1 summarizes his results and describes the regions of chaotic, periodic (I, II, etc.), and subharmonic (m/n) motions as functions of the damping and forcing amplitudes. This simple equation, representing a hardening spring system, has direct analogues in the dynamics of piles and in the rocking of buildings, both following the strong-motion phase of earthquake shaking after horizontal gaps have been created between the pile (foundation walls) and the soil [63].

A mechanical system with a nonlinear restoring force and with a control force added to move the system according to some prescribed signal has been studied by Holmes and Moon [19] and Holmes [18]. It was shown that such a system exhibits both periodic limit-cycle oscillation and chaotic motions. Chaotic vibrations in continuous beams have been studied for nonlinear body forces and nonlinear boundary conditions (that depend on the motion), and for motions large enough for the nonlinear terms in the equations of motion to be significant [37,38,39,40,41]. Forced planar vibrations of nonlinear elastica [35,36], were shown to become unstable and exhibit chaotic motions under certain conditions.

The above-mentioned studies imply that there is a conflict in the classical engineering description of the world.



Earthquake Engineering, Non-linear Problems in, Figure 1 Chaos diagram showing regions of chaotic, chaotic and periodic, periodic (I, II, III, etc.), and sub-harmonic ($4/3$, $3/2$, $5/3$, etc.) motions for a nonlinear equation as functions of non-dimensionalized damping and forcing amplitude (from [67])

One aspect of this conflict is the assumption that nature is a deductive system, moving forward in time according to deterministic laws. Another aspect is that a scientist attempting to model portions of the world from finite data projects unverifiable structure onto the local environment. The conflict is that these two views do not match, leaving us with a question: what are models good for? There are many systems in nature that are observed to be chaotic, and for which no adequate physical model exists. Whether a model is adequate or not depends, of course, on the questions asked [7]. Unfortunately, the art of dynamical modeling is often neglected in discussions of nonlinear and chaotic systems, in spite of its crucial importance [1]. In the following, the modeling problem in earthquake engineering will be illustrated using two common approaches to the solution, one based on an equivalent oscillator and the other one using wave representation.

Stochastic processes have been developed to describe irregular phenomena in deterministic systems that are too complicated or have too many variables to be fully described in detail. For example, stochastic processes have been used to model the response of structures to earthquake and wind forces, which are deterministic, and in principle could be completely described. In practice, the

stochastic modeling has been used also as an approximate description of a deterministic system that has unknown initial conditions and may be highly sensitive to the initial conditions. In trying to model real systems, as a result of the modeling process, we sometimes obtain a model that shows very regular behavior, while the real system has very irregular behavior. In that case, random noise is added to the model, but this represents no more than our lack of knowledge of the system structure or the inadequacy of the identification procedure [22].

In earthquake engineering, the complexity of the multi-dimensional real world is reduced to a sub-space, which is defined by (1) the dimensions and properties of the adopted mathematical models, (2) the nature of the adopted boundary conditions, and (3) the method of solution. A linear mechanical system cannot exhibit chaotic vibrations, and for periodic inputs it produces periodic outputs. The chaotic system must have nonlinear elements or properties, which can include, for example, (1) nonlinear elastic or spring elements; (2) nonlinear damping (such as stick-slip friction); (3) backlash, play, or bilinear springs; and (4) nonlinear boundary conditions. The nonlinear effects can be associated with the material properties, with the geometric effects, or both. In the following, the consequences of unorthodox boundary conditions and nonlinear waves in a building will be used to illustrate the extensions and complexities associated with evolving systems. The utility of this complexity can be viewed as the arbiter of the order and randomness.

Vibrational Representation of Response

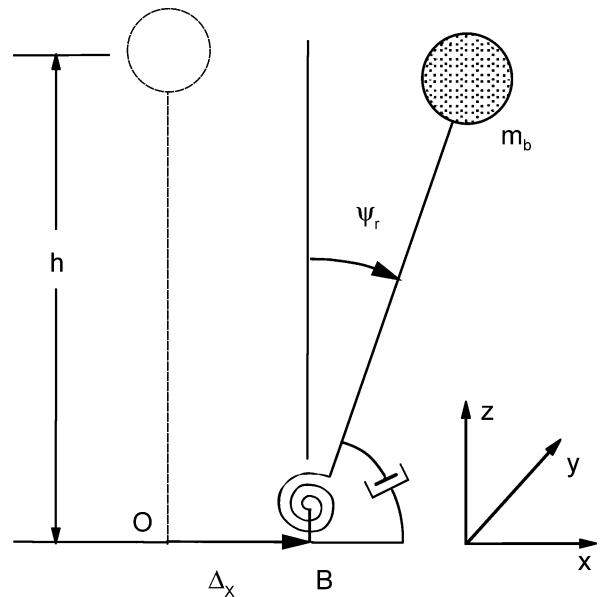
The first modern uses of mechanics in problems of earthquake engineering appeared during the early 1900s, following the earthquake disasters in San Francisco (1906), Messina-Reggio (1908), and Tokyo (1923) and the realization that something needed to be done to prevent such losses of life and property during future events. The first practical steps consisted of introducing the *seismic coefficient* (*shindo* in Japan, and *rapporto sismico* in Italy). This was followed by earthquake-resistant design codes, first adopted in Japan in 1923, and then in California in 1934 [44]. During the same period, there also appeared the first studies of the effects of earthquake shaking on structures in terms of simple mechanical oscillators [48], and in the early 1930s the modern theory based on the response spectrum method was introduced [2,3,4]. These early developments follow the deterministic formulations of Newtonian mechanics and employ linear models and equations of motion.

Elementary Vibrational Representation of Response

The basic model employed to describe the response of a simple structure to only horizontal earthquake ground acceleration, $\ddot{\Delta}_x$, is a single-degree-of-freedom system (SDOF) that experiences rocking ψ_r relative to the normal to the ground surface. The model also assumes that the ground does not deform in the vicinity of the foundation—that is, it neglects the soil–structure interaction (Fig. 2). The rotation ψ_r is restrained by a spring with stiffness K_r and by a dashpot with rocking damping constant C_r , providing the fraction of critical damping ζ_r . The natural frequency of this system is $\omega_r = (K_r/h^2 m_b)^{1/2}$, and for small rocking angles it is governed by the linear ordinary differential equation

$$\ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \omega_r^2 \psi_r = -\ddot{\Delta}_x/h. \quad (1)$$

For any initial conditions, and for arbitrary excitation, this system always leads to a deterministic and predictable response. Equation (1) was used originally to develop the concept of relative response spectrum and continues to this day as the main vehicle in formulation of most earthquake engineering analyses of response [56]. If the gravity force is considered, ω_r in Eq. (1) has to be reduced [5]. The system described by Eq. (1) is meta-stable for ψ_r smaller than its critical value. At the critical value of ψ_r , the over-



Earthquake Engineering, Non-linear Problems in, Figure 2

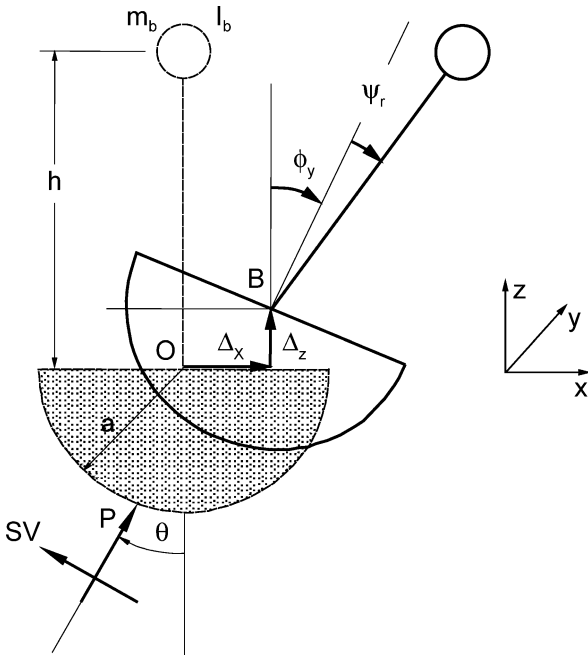
Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum) with equivalent mass m_b and mass-less column of height h , experiencing rocking ψ_r due to horizontal motion of its base Δ_x

turning moment of the gravity force is just balanced by the elastic moment in the restraining spring, and for values greater than the critical value the system becomes unstable.

Advanced Vibrational Representation of Response

In more advanced vibrational representations of the response, additional components of the earthquake excitation, structural dynamic instability, soil–structure interaction, spatial and temporal variations of the excitation, differential motions at different support points, and non-linear behavior of the stiffness K_r can be considered, but the structure usually continues to be modeled by mass-less columns, springs, and dashpots, and with a rigid mass m_b . In the following, we illustrate some of the above-mentioned cases.

Dynamic Instability An example of a simple model that includes instability is shown in Fig. 3. It experiences horizontal, vertical, and rocking excitations, which can result, for example, from incident P and SV waves. The structure



Earthquake Engineering, Non-linear Problems in, Figure 3

Single-degree-of-freedom system (SDOF) representation of a building (inverted pendulum), with equivalent mass m_b , moment of inertia (about O) I_b , and a mass-less column of height h , experiencing relative rocking ψ_r due to horizontal, vertical, and rocking motions of its foundation (Δ_x , Δ_z , and ϕ_y), which result from soil–structure interaction when excited by incident wave motion

is represented by an equivalent single-degree-of-freedom system, with a concentrated mass m_b at height h above the foundation. It has a radius of gyration r_b and a moment of inertia $I_b = m_b r_b^2$ about point O. The degree-of-freedom in the model is chosen to correspond to the relative rocking angle ψ_r . This rotation is restrained by a spring with rocking stiffness K_r and by a dashpot with rocking damping C_r (both not shown in Fig. 3), and the gravitational force $m_b g$ is considered. Taking moments about B results in the equation of motion

$$\ddot{\phi}_y + \ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \omega_r^2 \psi_r = \{ -(\ddot{\Delta}_x/a) \cos(\phi_y + \psi_r) + (\omega_r^2 \varepsilon_g + \ddot{\Delta}_z/a) \sin(\phi_y + \psi_r) \} / \varepsilon, \quad (2)$$

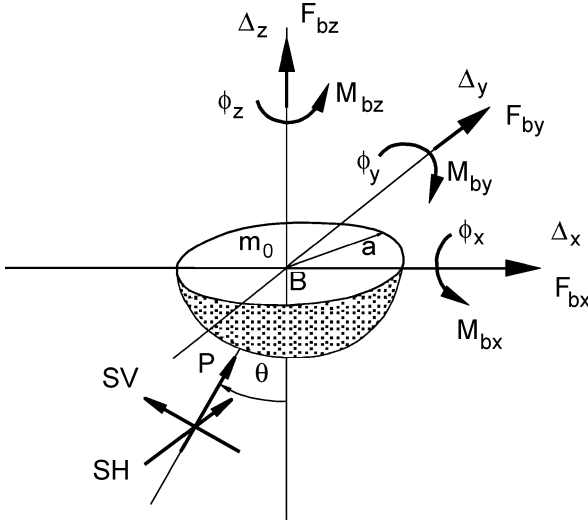
where $\varepsilon = h(1 + (r_b/h)^2)/a$, $\omega_r^2 = K_r/[m(h^2 + r_b^2)]$, ω_r is the natural frequency of rocking, ζ_r is a fraction of critical damping in $2\omega_r \zeta_r = C_r/[m(h^2 + r_b^2)]$, and $\varepsilon_g = 2/\omega_r^2 a$. Equation (2) is a differential equation coupling the rocking of the foundation, ϕ_y , and of the structure, ψ_r , with the horizontal and vertical motions of the foundation. It is a nonlinear equation the solution to which requires numerical analysis. In this example, we will discuss only the case in which $\phi_y + \psi_r$ is small. Then,

$$\ddot{\psi}_r + 2\omega_r \zeta_r \dot{\psi}_r + \{ \omega_r^2 (1 - \varepsilon_g \varepsilon) - \ddot{\Delta}_z / \varepsilon a \} \psi_r = -\ddot{\phi}_y + \{ -\ddot{\Delta}_x / a + (\omega_r^2 \varepsilon_g + \ddot{\Delta}_z / a) \phi_y \} / \varepsilon. \quad (3)$$

For steady-state excitation by incident P and SV waves with frequency ω , Δ_x , ϕ_y , and Δ_z , and therefore the forcing function of Eq. (3), will be periodic. Equation (3) is then a special form of the Hill's equation. Analysis of the stability of this equation can be found in the work of Lee [25]. For general earthquake excitation, Δ_x , ϕ_y , and Δ_z will be determined by the recorded components of motion, and in predictive analyses by simulated ground motions [27,28,70].

In Eq. (3), ϕ_y describes rocking of the foundation to which the structure is attached. In analyses that do not consider soil–structure interaction, ϕ_y will be determined directly by the rocking component of strong ground motion [21,28], and in studies that consider soil–structure interaction ϕ_y will be one of the variables to be determined by the analysis [25].

Soil–Structure Interaction The problem of linear soil–structure interaction embodies the phenomena that result from (1) the presence of an inclusion (foundation, Fig. 4) in the soil [26], and (2) the vibration of the structure supported by the foundation, which exerts dynamic forces on the foundation [25]. Examples and a discussion of the non-linear aspects of soil–structure interaction can be found



Earthquake Engineering, Non-linear Problems in, Figure 4

Six components of motion (three translations and three rotations) $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}$ of point B, and six components of force (three forces and three moments) $\{F_{\text{ext}}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$, that the structure exerts on the foundation at B

in Gicev [9] and in a review of observations of response to earthquake shaking in full-scale structures in Trifunac et al. [63,64,65].

The dynamic response of a rigid, embedded foundation to seismic waves can be separated into two parts. The first part corresponds to the determination of the restraining forces due to the motion of the inclusion, usually assumed to be a rigid body. The second part deals with the evaluation of the driving forces due to scattering of the incident waves by the inclusion, which is presumed to be immobile. This can be illustrated by considering a foundation embedded in an elastic medium and supporting an elastic superstructure. The steady-state harmonic motion of the foundation having frequency ω can be described by a vector $\{\Delta_x, \Delta_y, \Delta_z, \phi_x, \phi_y, \phi_z\}^T$ (Fig. 4), where Δ_x and Δ_y are horizontal translations, Δ_z is vertical translation, ϕ_x and ϕ_y are rotations about horizontal axes, and ϕ_z is torsion about the vertical axis. Using superposition, displacement of the foundation is the sum of two displacements:

$$\{U\} = \{U^*\} + \{U_0\}, \quad (4)$$

where $\{U^*\}$ is the foundation input motion corresponding to the displacement of the foundation under the action of the incident waves in the absence of external forces, and $\{U_0\}$ is the relative displacement corresponding to the dis-

placement of the foundation under the action of the external forces in the absence of incident wave excitation.

The interaction force $\{F_s\}$ generates the relative displacement $\{U_0\}$, which corresponds to the force that the foundation exerts on the soil and that is related to $\{U_0\}$ by $\{F_s\} = [K_s(\omega)]\{U_0\}$, where $[K_s(\omega)]$ is the 6×6 complex stiffness matrix of the embedded foundation. It depends upon the material properties of the soil medium, the characteristics and shape of the foundation, and the frequency of the harmonic motion, and it describes the force-displacement relationship between the rigid foundation and the soil medium.

The driving force of the incident waves is equal to $\{F_s^*\} = [K_s]\{U^*\}$, where the input motion $\{U^*\}$ is measured relative to an inertial frame. The “driving force” is the force that the ground exerts on the foundation when the rigid foundation is kept fixed under the action of the incident waves. It depends upon the properties of the foundation and the soil and on the nature of excitation.

The displacement $\{U\}$ is related to the interaction and driving forces via $[K_s]\{U\} = \{F_s\} + \{F_s^*\}$. For a rigid foundation having a mass matrix $[M_0]$ and subjected to a periodic external force, $\{F_{\text{ext}}\}$, the dynamic equilibrium equation is

$$[M_0]\{\ddot{U}\} = -\{F_s\} + \{F_{\text{ext}}\}, \quad (5)$$

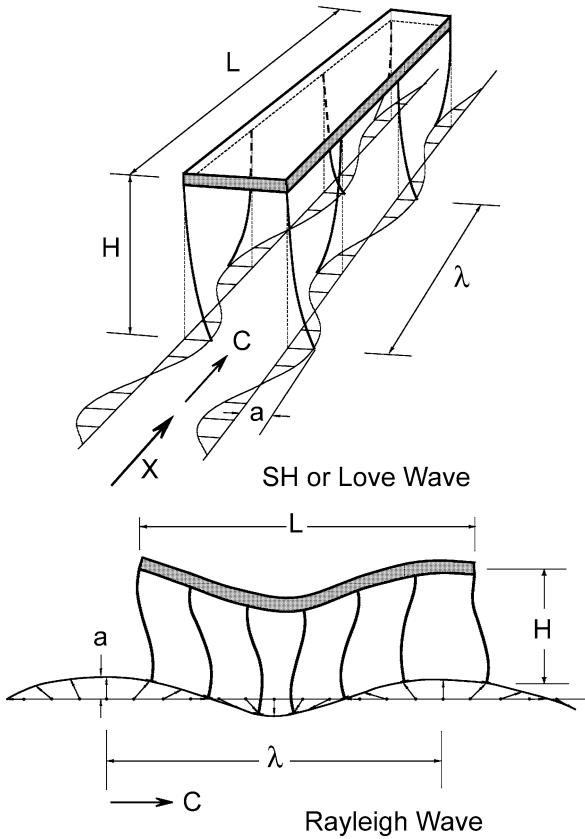
where $\{F_{\text{ext}}\} = \{F_{bx}, F_{by}, F_{bz}, M_{bx}, M_{by}, M_{bz}\}$ is the force the structure exerts on the foundation (Fig. 4). Then, Eq. (5) becomes

$$[M_0]\{\ddot{U}\} + [K_s]\{U\} = \{F_s^*\} + \{F_{\text{ext}}\}. \quad (6)$$

The solution of $\{U\}$ requires the determination of the mass matrix, the impedance matrix, the driving forces, and the external forces [25].

After the mass matrix $[M_0]$, the stiffness matrix $[K_s]$, and the force $\{F_s^*\}$ have all been evaluated, they can be used to determine the foundation displacement $\{U\}$. For in-plane response excited by P and SV waves, for example, the relative response ψ_r is then given by Eq. (3).

Differential Motions Common use of the response spectrum method [56] and many dynamic analyses in earthquake engineering implicitly assume that all points of building foundations move synchronously and with the same amplitudes. This, in effect, implies that the wave propagation in the soil is neglected. Unless the structure is long (e.g., a bridge with long spans, a dam, a tunnel) or “stiff” relative to the underlying soil, these simplifications are justified and can lead to a selection of approximate design forces if the effects of soil-foundation interaction in the presence of differential ground motions can be



Earthquake Engineering, Non-linear Problems in, Figure 5
Schematic representation of the deformation of columns accompanying differential wave excitation of long structures for out-of-plane response (top) and in-plane response (bottom) when SH or Love waves (top) or Rayleigh waves (bottom) propagate along the longitudinal axis of a structure

neglected [6]. Simple analyses of two-dimensional models of long buildings suggest that when $a/\lambda < 10^{-4}$, where a is wave amplitude and λ is the corresponding wavelength, the wave propagation effects on the response of simple structures can be neglected [50].

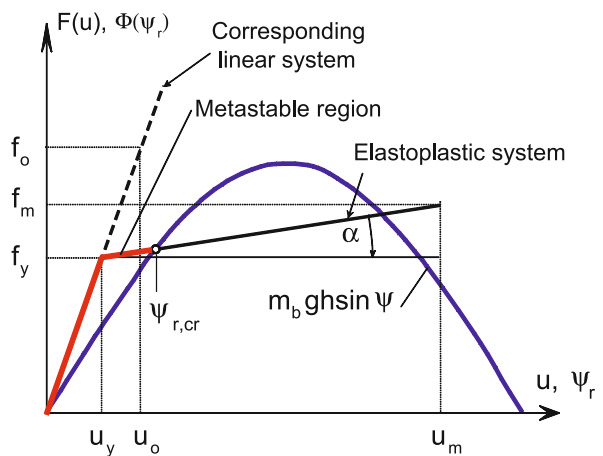
Figure 5 illustrates the “short” waves propagating along the longitudinal axis of a long building or a multiple-span bridge. For simplicity, the incident wave motion has been separated into out-of-plane motion (Fig. 5, top), consisting of SH and Love waves, and in-plane motion (Fig. 5, bottom) consisting of P, SV, and Rayleigh waves. The in-plane motion can further be separated into horizontal (longitudinal), vertical, and rocking components, while out-of-plane motion consists of horizontal motion in the transverse direction and torsion along the vertical axis. Trifunac and Todorovska [61] analyzed the effects of the horizontal in-plane components of differential motion

for buildings with models that are analogous to the sketch in Fig. 5 (bottom), and they showed how the response spectrum method can be modified to include the first-order effects of differential motions. Trifunac and Gicev [59] showed how to modify the spectra of translational motions, into a spectrum that approximates the total (translational and torsional) responses, and how this approximation is valid for strong motion waves an order of magnitude longer than the structure ($\lambda \gg L$).

As can be seen from the above examples the differential motions lead to complex excitation and deformation of the structural members (columns, shear walls, beams, braces), increase the dimensions of the governing differential equations, lead to three-dimensional dynamic instability problems, and can lead to nonlinear boundary conditions. These are all conditions that create an environment in which, even with the most detailed numerical simulations, it is difficult to predict all of the complexities of the possible responses.

Nonlinear Vibrational Analyses of Response

For engineering estimation of the maximum nonlinear response of a SDOF system, u_m , in terms of the maximum linear response, u_0 , it is customary to specify a relation between u_m and u_0 (Fig. 6). By defining the yield-strength reduction factor as $R_y = u_0/u_y$, where u_y is the yielding displacement of the SDOF system equivalent spring, and ductility as $\mu = u_m/u_y$, for the same ground motion the ratio u_m/u_0 is then equal to μ/R_y . Veletsos and New-



Earthquake Engineering, Non-linear Problems in, Figure 6
Bi-linear representation of stiffness (yielding at (u_y, f_y)), overturning moment of gravity force ($m_b g \sin \psi$), critical rocking angle $\psi_{r,cr}$, and meta-stable region ($0 < \psi_r < \psi_{r,cr}$) for an SDOF system

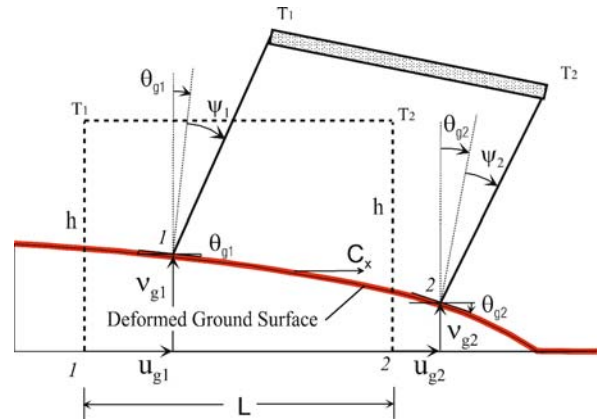
mark [68,69] showed that (1) for a long-period SDOF system when its natural period $T_n = 2\pi/\omega_n$ becomes very long, u_m/u_0 tends toward 1 and R_y approaches μ (equal deformation rule); (2) for the response amplitudes governed mainly by the peak excitation velocities, u_m/u_0 can be approximated by $\mu/\sqrt{2\mu-1}$ and R_y by $\sqrt{2\mu-1}$ (equal strain energy rule); and (3) for a high-frequency (stiff) system when $T_n \sim 0$, $R_y \sim 1$.

Complexities of Simultaneous Action of Dynamic Instability, Nonlinearity, and Kinematic Boundary Conditions – Example The model we illustrate next is an SDOF when it is excited by synchronous horizontal ground motion at its two supports (1 and 2 in Fig. 7), but it behaves like a three-degree-of-freedom (3DOF) system when excited by propagating horizontal, vertical, and rocking ground motions. For such a system, the above classical equal energy and equal displacement rules for SDOF system will not apply.

The goals here are to describe the effects of differential motion on strength-reduction factors R_y of the simple structure shown in Fig. 7 when it is subjected to all of the components of near-source ground motions, and to illustrate the resulting complexities of nonlinear response. Analyses of the consequences of the differences in ground motion at structural supports, caused by non-uniform soil properties, soil–structure interaction, and lateral spreading, for example, will further contribute to the complexities of the response, but these factors will not be discussed here.

The original response spectrum method was formulated using a vibrational solution of the differential equation of an SDOF system excited by synchronous, and only horizontal (one component), ground motion. The consequences of simultaneous action of all six components of ground motion (three translations and three rotations) on the relative response of an SDOF system are still rarely considered in modern engineering design [58], even though it has been 75 years since the original response spectrum method was formulated and about 40 years since it became the principal tool in engineering design [56]. Because the response spectrum method has become an essential part of the design process and of the description of how strong motion should be specified for a broad range of design applications [52], we hope that the present examples will help to further understanding of the complexities of response in more realistic models of structures.

The nature of the relative motion of individual column foundations or of the entire foundation system will depend upon the type of foundation, the characteristics of the soil surrounding the foundation, the type of incident



Earthquake Engineering, Non-linear Problems in, Figure 7

The structure deformed by the wave, propagating from left to right, with phase velocity C_x , for the case of $+v_{gi}$ ("up" motion). Different column rotations ψ_1 and ψ_2 result from different translations and rotations at supports 1 and 2 (from [21])

waves, and the direction of wave arrival, with the motion at the base of each column having six degrees of freedom. In the following example, we assume that the effects of soil–structure interaction are negligible; consider only the in-plane horizontal, vertical, and rocking components of the motion of column foundations; and show selected results of the analysis for a structure on only two separate foundations. We assume that the structure is near the fault and that the longitudinal axis of the structure (X axis) coincides with the radial direction (r axis) of the propagation of waves from the earthquake source, so that the displacements at the base of columns are different as a result of the wave passage alone. We suppose that the excitations at the piers have the same amplitude but different phases and that the phase difference (or time delay) will depend upon the distance between the piers and the horizontal phase velocity of the incident waves.

The simple model we consider, which is described in Fig. 7, represents a one-story structure consisting of a rigid mass, m , with length L , supported by two rigid, mass-less columns with height h , which are connected at the top to the mass and at the bottom to the ground by rotational springs (not shown in Fig. 7). The stiffness of the springs, k_ϕ , is assumed to be elastic-plastic, as in Fig. 6, without hardening ($\alpha = 0$). The mass-less columns are connected to the ground and to the rigid mass by rotational dashpots, c_ϕ , providing a fraction of critical damping equal to 5 percent. Rotation of the columns, $\phi_i = \theta_{gi} + \psi_i$ for $i = 1, 2$, which is assumed to be not small, leads us to consider the geometric nonlinearity. The mass is acted upon by the acceleration of gravity, g , and is excited by differential hori-

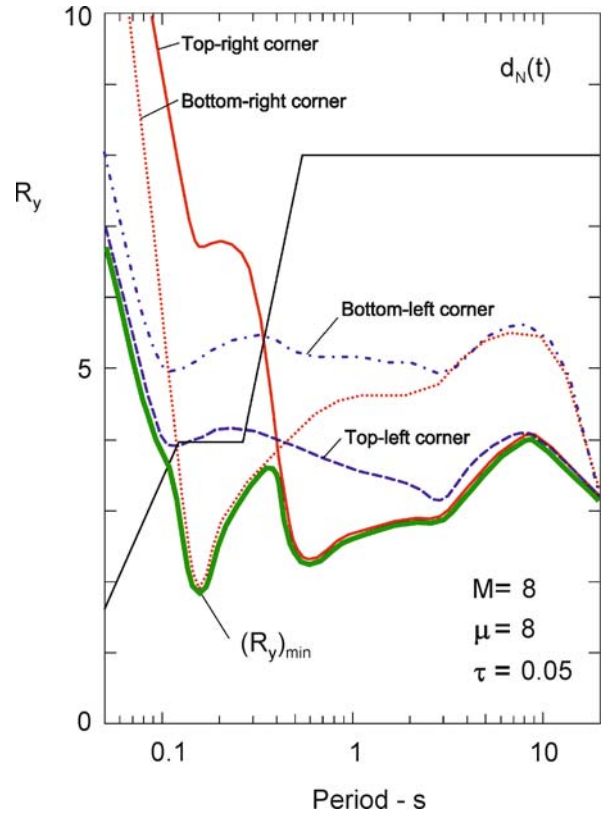
zontal, vertical, and rocking ground motions, u_{gi} , v_{gi} , and θ_{gi} , $i = 1, 2$ (Fig. 7) at the two bases, so that

$$\begin{aligned} u_{g2}(t) &= u_{g1}(t - \tau); & v_{g2}(t) &= v_{g1}(t - \tau); \\ \theta_{g2}(t) &= \theta_{g1}(t - \tau); & \tau &= L/C_x, \end{aligned}$$

with τ being the time delay between the motions at the two piers and C_x the horizontal phase velocity of the incident waves. The functional forms of u_{gi} , v_{gi} , and θ_{gi} are defined by the near-source ground motions [21], and the rocking component of the ground motion is approximated by [28] $\theta_{gi}(t) = -\dot{v}_{gi}(t)/C_x$, where $\dot{v}_{gi}(t)$ is the vertical velocity of the ground motion at the i th column. Of course, in a more accurate modeling, the ratio of the v_{gi} to u_{gi} amplitudes will depend upon the incident angle and the character of incident waves, while the associated rocking θ_{gi} will be described by a superposition of the rocking angles associated with incident body and dispersed surface waves [28].

The yield-strength reduction factor for the system subjected to synchronous ground motion is $R_y = f_0/f_y = u_0/u_y$, where all of the quantities are defined in Fig. 6. In this example, for the assumed model and because of the differential ground motions and rotation of the beams, the relative rotation for the two columns at their top and bottom will be different. Therefore, it is necessary to define the R -factor and ductility for each corner of the system, instead of one factor for the entire system. In all calculations here, we consider the actions of the horizontal, vertical, and rocking components of the ground motion, the effects of gravity force, dynamic instability, and geometric non-linearity. For the structure in Fig. 7, we calculate maximum linear and nonlinear relative rotations at four corners of the system under downward ($-v_{gi}$), radial, and rocking, and upward ($+v_{gi}$), radial and rocking near-source differential ground motions corresponding to a given earthquake magnitude, ductility μ , and for different time delays, τ . Then we plot R_y versus T_n for the four corners of the system.

Figure 8 illustrates typical results for R_y versus the oscillator period for near-source, fault-parallel displacement $d_N(t) = A_N(1 - e^{-t/\tau_N})/2$ [21], with downward vertical ground displacement, magnitude $M = 8$, for a ductility ratio of 8 and a time delay of $\tau = 0.05$ s. It shows the results for the top-left, top-right, bottom-left, and bottom-right corners of the system, assuming wave propagation from left to right (see Fig. 7). For reference and easier comparison with the previously published results, we also plot one of the oldest estimates of R_y versus period, using piecewise straight lines [21]. The curve $(R_y)_{\min}$ shows the minimum values of R_y for $d_N(t)$ motion with $-v_{gi}$, and for $M = 8$, $\mu = 8$, and $\tau = 0.05$ s.



Earthquake Engineering, Non-linear Problems in, Figure 8

Example of the effects of the differential ground motion on the strength-reduction factors R_y at the four corners of the structure in Fig. 7, subjected to horizontal, vertical, and rocking components of the fault-parallel displacement, for downward vertical motion ($-v_{gi}$) for earthquake magnitude $M = 8$, ductility $\mu = 8$, and delay at the right support $\tau = 0.05$ s. The amplitudes of the piecewise straight representation of the classical R_y are shown for comparison [21]. $(R_y)_{\min}$ shows the smallest values of the R -factors, which for the set of conditions in this example are determined by the response at the top left corner (for periods shorter than 0.1 s), at the bottom right corner (for periods between 0.1 and 0.35 s), and at the top right corner (for periods longer than 0.35 s)

For periods longer than 5 to 10 s, R_y curves approach “collapse boundaries” [21]. This is implied in Fig. 8 by the rapid decrease of R_y versus period for periods longer than about 7 s. At or beyond these boundaries, the nonlinear system collapses due to the action of gravity loads and dynamic instability.

The complex results illustrated in Fig. 8 can be simplified by keeping only $(R_y)_{\min}$, since it is only the minimum value of R_y that is needed for engineering design. By mapping $(R_y)_{\min}$ versus period of the oscillator for different earthquake magnitudes, M , different ductilities, μ , and different delay times, τ , design criteria can be formulated

for design of simple structures to withstand near-fault differential ground motions [21]. Nevertheless, the above shows how complicated the response becomes even for as simple a structure as the one shown by the model in Fig. 7, when differential ground motion with all of the components of motion is considered. In this example, this complexity results from simultaneous consideration of material and geometric nonlinearities, dynamic instability, and kinematic boundary conditions.

Response in Terms of Wave Propagation – An Example

The vibrational representation of the solution of response of a multi-degree-of-freedom system subjected to earthquake shaking is frequently simplified by considering only the fundamental and, occasionally, a few of the lowest frequencies of the system. Doing so is analogous to low-pass filtering of the complete solution [56,57], but it can work well when the excitation amplitudes are small and the motions are associated with long waves. However, during strong earthquakes, the ground motion contains large displacement pulses, the duration of which can be shorter than the fundamental period of the structure. For this type

of excitation, the vibrational representation of response and the response spectrum superposition method cease to be suitable and should be replaced by a solution in terms of propagating waves. For short impulsive ground motions, the damage can occur before the wave entering the structure completes its travel up and down the structure, and well before the wave interference can occur—that is, well before the physical conditions can lead to the interference of waves and creation of the mode shapes.

To illustrate the phenomena that can occur during nonlinear wave propagation in a building, we describe horizontal motions, u , in a one-dimensional shear beam, supported by one-dimensional half space and excited by a vertically propagating shear wave described by a half-sine-pulse (Fig. 9). A finite-difference scheme for solution of this problem with accuracy, $O(\Delta t^2, \Delta x^2)$, where Δx and Δt are the space and time increments, leads to the exact solution for $\beta \Delta t / \Delta x = 1$, where β is the velocity of shear waves. For simplicity, the incident displacement in the soil is chosen to be a sinusoidal pulse with the characteristics shown on Fig. 9.

A mesh with different spatial intervals in the soil and in the building will be used. The equation of motion is

$$v_t = (\sigma)_x / \rho, \quad (7a)$$

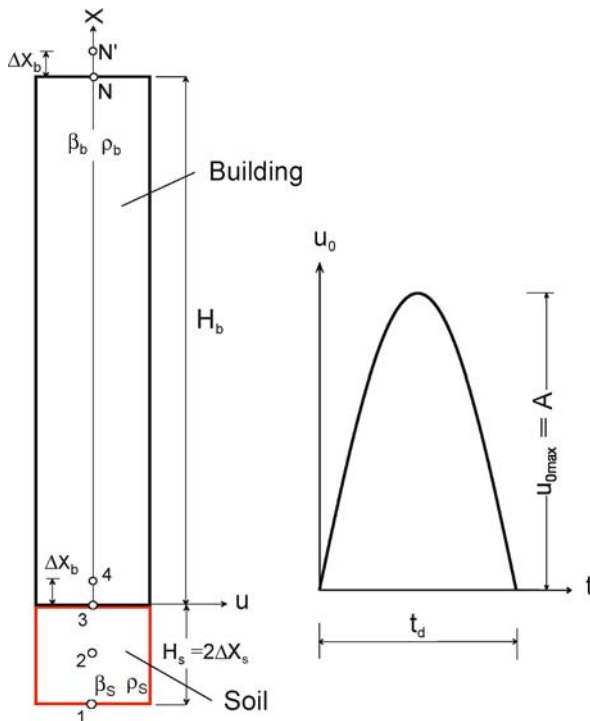
and the relation between the derivative of the strain and the velocity is

$$\varepsilon_t = v_x, \quad (7b)$$

where v, ρ, σ , and ε are particle velocity, density, shear stress, and shear strain, respectively, and the subscripts t and x represent derivatives with regard to time and space.

The domain consists of two materials (Fig. 9): (1) $-\Delta x_s \leq x < 0$ with physical properties ρ_s and μ_s , representing foundation soil, and (2) $0 < x \leq H_b$ with physical properties ρ_b and μ_b for linear response, where ρ_i is the density and μ_i is the shear modulus in the soil ($i = s$) or in the building ($i = b$). $v = \partial u / \partial t$ and $\varepsilon = \partial u / \partial x$ are the velocity and the strain of a particle, and u is out-of-plane displacement of a particle perpendicular to the propagation ray.

It is assumed that the incoming wave is known and that its displacement as a function of time is prescribed at the point 1 in the soil ($x = -2\Delta x_s$). Also, it is assumed that the soil is always in the linear elastic state. The finite difference method for a set of simultaneous equations is used to solve the problem, and spatial intervals are defined by $\Delta x_i = \beta_i \cdot \Delta t$, where β_i is the velocity of shear waves in the soil ($i = s$) or in the building ($i = b$) and Δt is the time step. The transparent boundary adopted for this study, which is described in Fujino and Hakuno [8],



Earthquake Engineering, Non-linear Problems in, Figure 9
Shear beam (building) (*left*) and incoming strong-motion displacement pulse (*right*) in the soil

is a perfect, transparent boundary for one-dimensional waves when $\beta \Delta x / \Delta t = 1$. Point 1 is where the prescribed displacement is applied, and we assume that this displacement travels upward in each time step. Point 2 is the boundary point of the model, where the quantities of motion are updated in each time step, and point 3 is the first spatial point, where the motion is computed using finite differences.

For the linear case at the contact (see point 3 in Fig. 9), one part of the incoming wave is transmitted into the other medium and one is reflected back into the same medium. The corresponding coefficients are obtained from the boundary conditions of continuity of the displacements and stresses at the contact. For a transmitted wave from medium B to medium A, the transmission coefficient is equal to $k_{\text{tr}B \rightarrow A} = 2 / [1 + \rho_a \beta_a / (\rho_b \beta_b)]$. For a reflected wave from medium A back into medium B, this coefficient is $k_{\text{ref}B \rightarrow B} = [1 - \rho_a \beta_a / (\rho_b \beta_b)] / [1 + \rho_a \beta_a / (\rho_b \beta_b)]$. For the opposite direction of propagation, the numerators and the denominators in these fractions exchange places.

Numerical Examples

We consider a shear beam supported by elastic soil, as shown in Fig. 9. The densities of the soil and of the beam are assumed to be the same: $\rho_b = \rho_s = \rho = 2000 \text{ kg/m}^3$. The velocity of the shear waves in the soil is taken as $\beta_s = 250 \text{ m/s}$, and in the building as $\beta_b = 100 \text{ m/s}$.

To describe nonlinear response and the development of permanent deformations in the beam, we introduce two dimensionless parameters: (1) dimensionless amplitude $\alpha = A / (H_b \varepsilon_{yb})$, where A is the amplitude of the pulse (Fig. 9), H_b is the height of the building, and ε_{yb} is the yielding strain in the building, and (2) dimensionless frequency $\eta = H_b / (\beta_b t_d)$, where $\beta_b t_d$ is one half of the wavelength of the wave in the building, β_b is the shear-wave velocity in the building, and t_d is the duration of the half-sine pulse.

To understand the development of the permanent strain in the nonlinear beam, we describe first the solution for the linear beam. The displacement and the strain for the linear beam are:

$$u(x, t) = A \sum_{j=1}^{\infty} k_j \left\{ \sin \frac{\pi}{t_d} \left(t - t_{j-1} - \frac{x}{\beta_b} \right) \left[H \left(t - t_{j-1} - \frac{x}{\beta_b} \right) - H \left(t - t_{j-1} - \frac{x}{\beta_b} - t_d \right) \right] + \sin \frac{\pi}{t_d} \left(t - t_j + \frac{x}{\beta_b} \right) \cdot \left[H \left(t - t_j + \frac{x}{\beta_b} \right) - H \left(t - t_j + \frac{x}{\beta_b} - t_d \right) \right] \right\} \quad (8)$$

and

$$\varepsilon(x, t) = A \frac{\pi}{\beta_b t_d} \sum_{j=1}^{\infty} k_j \left\{ -\cos \frac{\pi}{t_d} \left(t - t_{j-1} - \frac{x}{\beta_b} \right) \cdot \left[H \left(t - t_{j-1} - \frac{x}{\beta_b} \right) - H \left(t - t_{j-1} - \frac{x}{\beta_b} - t_d \right) \right] + \cos \frac{\pi}{t_d} \left(t - t_j + \frac{x}{\beta_b} \right) \cdot \left[H \left(t - t_j + \frac{x}{\beta_b} \right) - H \left(t - t_j + \frac{x}{\beta_b} - t_d \right) \right] \right\} \quad (9)$$

where j is the order number of the passage of the wave on the path bottom-top-bottom in the building, $t_j = 2jH_b/\beta_b$ ($j = 0, 1, 2, 3, \dots$), is the time required for the wave to pass j times over the path bottom-top-bottom (two heights), $k_j = k_t k_r^{j-1}$ is the amplitude factor of the pulse in the soil in its j th passage along the path bottom-top-bottom through the building, and k_t and k_r are coefficients defined by $k_{\text{tr}B \rightarrow A}$ and $k_{\text{ref}B \rightarrow B}$ above.

The odd terms in Eq. (8) and Eq. (9) describe the response to the pulse coming from below, while the even terms describe the response to the pulse arriving from above. For the shear-wave velocities in our example, $k_t = 10/7$ and $k_r = -3/7$. In Eq. (8) the displacement is positive for odd passages and negative for even passages. The displacement and velocity change sign after reflection from the soil-building interface and do not change sign after reflection from the top of the building. The strain changes sign after reflection from the top of the building and does not change sign after reflection from the building-soil interface. The constant that multiplies the series in Eq. (8) in terms of dimensionless amplitude and dimensionless frequency is $A\pi/(\beta_b t_d) = A_\varepsilon = \pi\alpha\eta\varepsilon_{yb}$.

To describe the occurrence of permanent strain, we consider two characteristic points in the building: (1) Point B ($x = 0$) at the soil-building interface (point 3 in the grid, see Fig. 9), and (2) point T ($x = H_b - \beta_b t_b/2$), where the amplitudes of the strain with the same sign meet after reflection from the top of the building. The location of this point is dependent upon the duration (wavelength) of the pulse. The first term in Eq. (8) is one if the argument of the cosine function is equal to $t_d(t - t_0 - x/\beta_b = t_d)$, and the second term is one if the argument of the second cosine function is equal to 0 ($t - t_1 + x/\beta_b = 0$). The position of point T, where the strain amplitude is two times larger than the strain entering the beam, is at $x = H_b - \beta_b t_d/2$, and the time when this occurs is $t = H_b/\beta_b + t_d/2$. From Eq. (9) in the first passage of the pulse, $t < 2H_b/\beta_b$, and only the first term in the series ex-

ists. The strain at point B reaches its absolute maximum at the very beginning, during the entrance of the pulse into the building, and its value is $|\varepsilon_{B\max}^1| = \pi\alpha\eta\varepsilon_{yb}k_t$. If this strain is greater than the yielding strain in the building, ε_{yb} , a permanent strain at the interface will develop, and the condition for occurrence of permanent strain at this point is $|\varepsilon_{B\max}^1| > \varepsilon_{yb}$, or, in terms of the dimensionless parameters,

$$\alpha\eta > (\pi k_t)^{-1} = (\beta_b + \beta_s)/(2\pi\beta_s) = C_B. \quad (10B)$$

At point T (this point does not exist if $t_d > 2H_b/\beta_b$, and it coincides with point B if $t_d = 2H_b/\beta_b$), from Eq. (9), the maximum strain during the first passage occurs at $t = H_b/\beta_b + t_d/2$, and its amplitude is $2A_e \cdot k_t$. The condition for occurrence of the permanent strain is

$$\alpha\eta > (2\pi k_t)^{-1} = (\beta_b + \beta_s)/(4\pi\beta_s) = C_B/2 = C_T. \quad (10T)$$

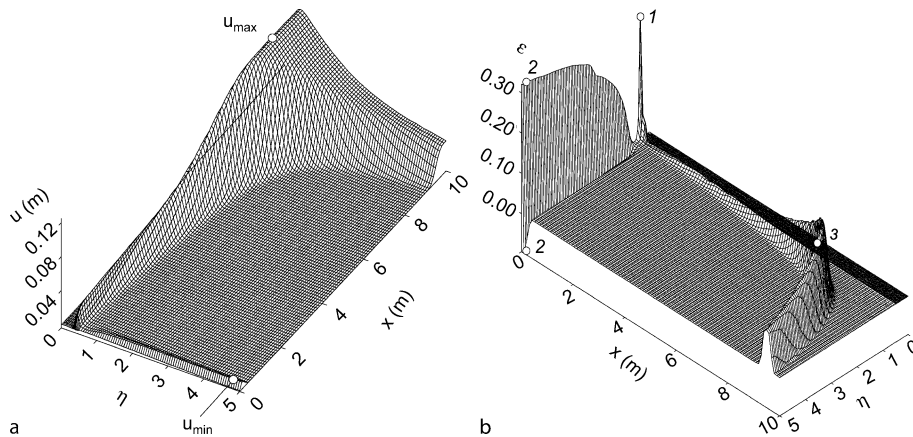
For the shear-wave velocities in our example $C_B = 0.2228$ and $C_T = 0.1114$.

For the above simple model, the occurrence, development, and amplitudes of permanent strains and displacements have been studied by Gicev and Trifunac [10,11]. They found that for large ground-displacement pulses (large α) the maximum permanent strains occur mainly at the interface of the building with the soil, while for smaller amplitudes of pulses permanent strains occur closer to the top of the building. They distinguished three zones of the permanently deformed beam: (1) a permanently deformed zone at the bottom; (2) an intermediate zone, which is not deformed at its bottom part and is deformed in the top part; and (3) a non-deformed zone at the top of the beam. The occurrence and development of these zones depends upon the dimensionless excitation amplitudes and

the dimensionless frequencies, and in particular on the conditions that lead to the occurrence of the first permanent strain (see Eqs. (10B) and (10T)). For large and long strong-motion pulses ($\eta \leq 0.5$; first, the condition in Eq. (10B) is relevant), only zones 1 and 3 are present in the beam. For large amplitudes and short strong-motion pulses, all three zones develop and are present. For smaller excitation amplitudes (when the condition in Eq. (10B) cannot be satisfied for long pulses, and when the condition in Eq. (10T) is satisfied), only zones 2 and 3 exist in the beam. For larger values of η (when the condition in Eq. (10B) is satisfied) all three zones exist.

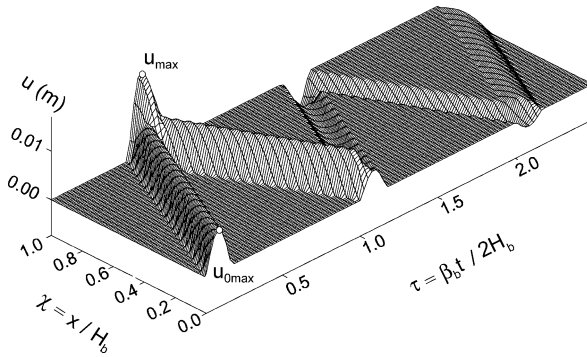
Gicev and Trifunac [10,11] found a similar situation for the occurrence of the maximum strains. For large and long pulses, maximum strain is located at the bottom of the building, and, as the pulses become shorter, peak strains occur at higher positions in the building. For some high frequencies of excitation, the maximum strain again appears at the bottom of the building because the loss of energy due to the development of the permanent strain at the bottom overcomes the effects of the wave reflections from the top of the building (Fig. 10).

Creation of large permanent deformation zones in the building by the incident waves absorbs some or most of the incident wave energy and can reduce or eliminate further wave propagation and the associated energy transport (Figs. 11 and 12). To the extent that the locations of the plastic deformation zones can be controlled by the design process, absorption of the incident-wave energy by structural members may become a new and powerful tool for performance-based design. To take advantage of such possibilities, the governing differential equations must be solved by the wave-propagation method.



Earthquake Engineering, Non-linear Problems in, Figure 10

Permanent displacements ($u_{\max} = 0.126$ m) (left), and permanent strains ($\varepsilon_1 = 0.31$, $\varepsilon_2 = 0.32$, $\varepsilon_3 = 0.20$) (right), along the building versus dimensionless frequency η and for dimensionless amplitude $\alpha = 0.3$



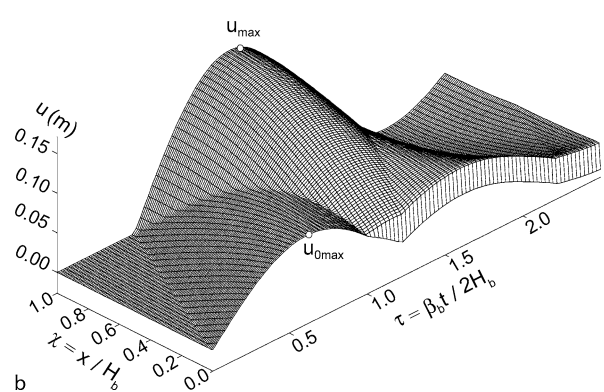
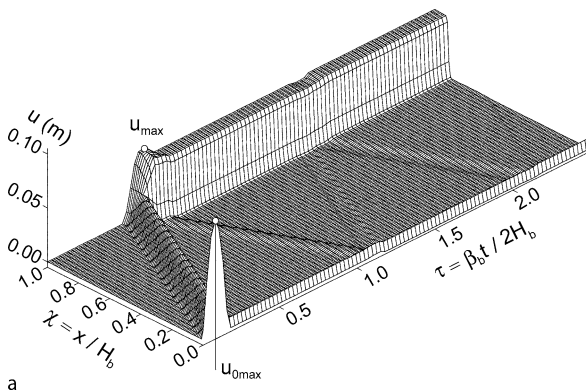
Earthquake Engineering, Non-linear Problems in, Figure 11
Linear displacements along the normalized length of the beam, $\chi = x/H_b$, versus normalized time $\tau = \beta_b t / 2H_b$, for dimensionless pulse amplitude $\alpha = 0.03$ and dimensionless frequency $\eta = 3$

Examples illustrated here show that for excitation of structures by large, near-field displacement pulses failure can occur anywhere in the building before the incident wave has completed its first travel from the foundation to the top of the building and back to the foundation ($2H_b/\beta_b$). Because this travel time is shorter (by 1/2) than the natural period of the structure on the fixed base, it is seen that the common response spectrum method of analysis (based on the vibrational formulation of the solution) cannot provide the required details for the design of structures for such excitation. The complexity of the outcome increases with amplitudes of excitation and depends upon the pulse duration. Because actual strong ground motion in the near field has at least several strong pulses, it can be seen that the complexity in real structures responding to strong earthquake motions will be even greater. In engineering approximation based on the vibrational solution

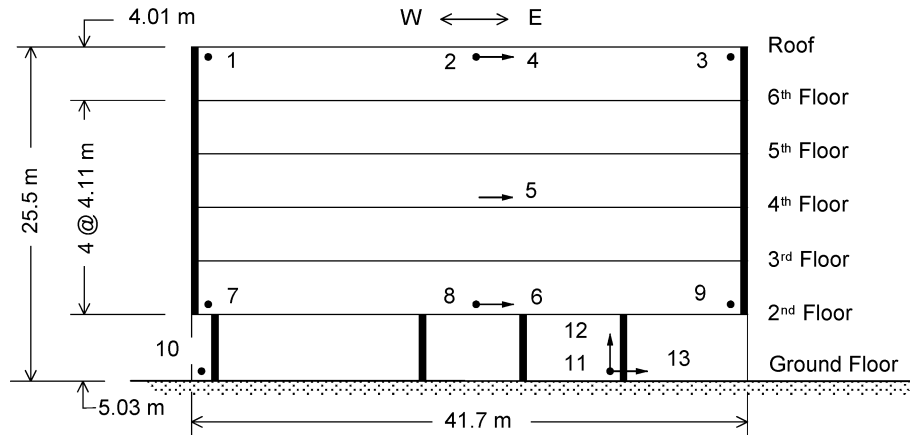
of the problem and on the SDOF models, where the location of ductile response is predetermined by the simple modeling assumptions, this complexity cannot be included because of the modeling constraints. The outcome is that it is virtually impossible for simplified models to identify or to predict the location of damage. In contrast, for properly chosen wave propagation models, prediction and identification of damage is a natural and logical outcome of interaction between excitation and model properties. A good example of this can be found in Gicev and Trifunac [12], who showed how a simple wave-propagation model can predict the actually observed location of damage.

Observations of Nonlinear Response

Invaluable for understanding and proper treatment of the actual nonlinear response, and for validation of vibration monitoring and analysis methods for real-life problems, are earthquake response data from well-instrumented, full-scale structures that have been damaged by an earthquake. Such data are rare and are not always freely available. An example of an instrumented building that has been damaged by an earthquake, and for which information about the damage and strong-motion data on the causative earthquake are available, is the former Imperial County Services Building in El Centro, California, which was severely damaged by the magnitude 6.6 Imperial Valley earthquake of October 15, 1979, and later demolished [23,51]. Its transverse (NS) response was recorded by three vertical arrays (recording channels 1, 3, 7, 9, 10, and 11; see Fig. 13), and its longitudinal (EW) response was recorded by one vertical array (recording channels 4, 5, 6, and 13, also shown in Fig. 13).



Earthquake Engineering, Non-linear Problems in, Figure 12
Nonlinear displacements along the normalized length of the beam, $\chi = x/H_b$, versus normalized time $\tau = \beta_b t / 2H_b$ for dimensionless pulse amplitude $\alpha = 0.3$ and dimensionless frequencies $\eta = 3$ (left) and $\eta = 0.41$ (right)



Earthquake Engineering, Non-linear Problems in, Figure 13

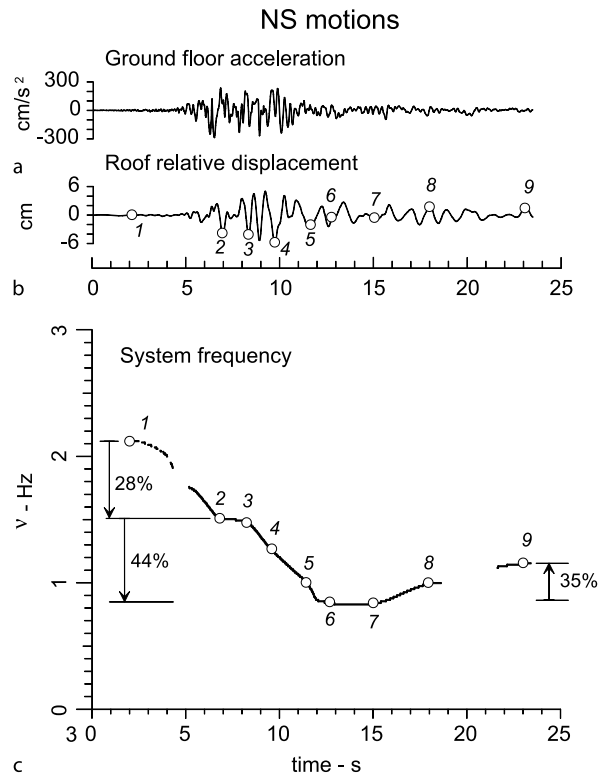
Layout of the seismic monitoring array in the ICS building (dots, without arrows, show the NS recording channels)

For a simplified soil-structure interaction model of a building supported by a rigid foundation, the difference between the roof and base horizontal displacements during earthquake shaking is the sum of the horizontal displacements due to (1) horizontal deformation of the soil, (2) rigid-body rocking of the foundation, and (3) deformation of the structure. The estimated frequency from such data is referred to as system or “apparent” frequency, which differs from the fixed-base frequency of the building. While the fixed-base frequency depends only upon the properties of the structure, the apparent frequency depends also upon the stiffness of the foundation soil. The following relationship holds:

$$\frac{1}{\omega_{\text{sys}}^2} = \frac{1}{\omega_1^2} + \frac{1}{\omega_H^2} + \frac{1}{\omega_R^2}, \quad (13)$$

where $\omega_{\text{sys}} = 2\pi\nu_{\text{sys}}$ is the soil-structure system frequency, ω_1 is the fundamental fixed-base frequency of the structure, and ω_H and ω_R are the horizontal and rocking frequencies, respectively, of a rigid structure on flexible soil [33].

Figure 14c shows that during earthquake shaking (Fig. 14a) the NS frequency of relative system response (Fig. 14b) dropped from $\nu \approx 2.12$ Hz in the early stage of response (at $t \approx 2$ s) to $\nu \approx 1.52$ Hz at $t \approx 6.8$ s ($\Delta\nu \approx 0.6$ Hz, $\Delta\nu/\nu \approx 28\%$), that it was constant during the interval $t \approx 6.8 - 8.5$ s, and that it dropped further to $\nu \approx 0.85$ Hz at $t \approx 12$ s ($\Delta\nu \approx 0.67$ Hz, $\Delta\nu/\nu \approx 44\%$). Then, toward the end of the recorded shaking, the frequency increased to $\nu \approx 1.15$ Hz ($\Delta\nu \approx 0.3$ Hz; $\Delta\nu/\nu \approx 35\%$). Early in the response ($t < 7$ s), the amplitudes of the first story drifts in the building were relatively small ($< 0.5\%$), and the observed decrease of system fre-



Earthquake Engineering, Non-linear Problems in, Figure 14

Time-frequency analysis for the NS response of the ICS building: a ground acceleration, b relative roof response, and c system frequency versus time

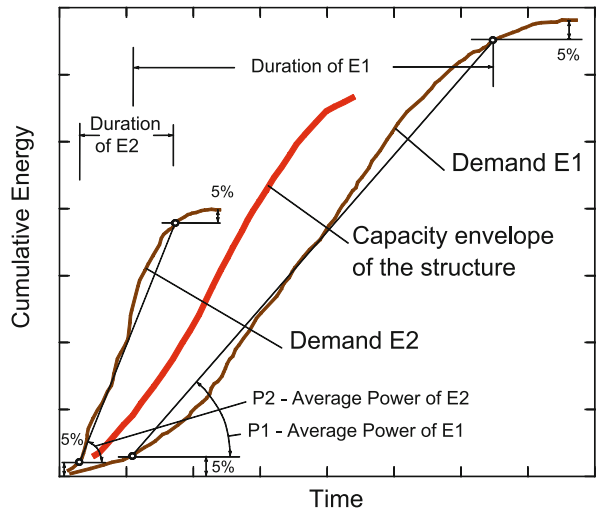
quency is believed to be due to changes in the soil and bonding between the soil and foundation. This was followed by a further decrease in the system frequency of

about 44% (between 8 and 12 s). The first-story drifts in the building were large when this occurred ($>0.5\%$ for NS), and the principal cause for this change is believed to be the damage, with the most severe damage occurring between 8 and 12 s after trigger. Near the end of the shaking, a 35% increase in system frequency was observed, suggesting system hardening, which is believed to be due to changes in the soil [51].

Changes similar to what is shown in Fig. 14c were first observed following the San Fernando earthquake in California in 1971 [66] and then during many subsequent earthquakes. It is known at present that many different factors can contribute to fluctuations of the system frequency, including rainfall, temperature fluctuations, changes in occupancy, remodeling and strengthening of buildings, wind, and earthquakes [49]. The simultaneous action of some of these factors and the associated time-dependent changes in the physical model contribute to complex and evolving system changes that make predictions of the dynamic response difficult.

Future Directions

Well-designed structures are expected to have ductile behavior during the largest credible shaking, and a large energy reserve to at least delay failure if it cannot be avoided. As the structure finally enters large nonlinear levels of response, it absorbs the excess of the input energy through ductile deformation of its components. Thus, it is logical to formulate future earthquake-resistant design procedures in terms of the energy driving this process. From the mechanics point of view, this introduces nothing new, because the energy equations can be derived directly from the dynamic equilibrium equations. The advantage of using energy is that the duration of strong motion, the number of cycles to failure, and dynamic instability all can be addressed directly and explicitly. This, of course, requires scaling of the earthquake source and of the attenuation of strong motion described in terms of its wave energy. Trifunac et al. [65] reviewed the seismological aspects of empirical scaling of seismic wave energy, E_s , and showed how the radiated energy can be represented by the functionals of strong ground motion [53,54,55]. They described the energy propagation and attenuation with distance and illustrated it for the three-dimensional geological structure of the Los Angeles basin during the 1994 Northridge, CA earthquake, then they described the seismic energy flow through the response of soil-foundation-structure systems, analyzed the energy available to excite the structure, and finally examined the relative response of the structure.



Earthquake Engineering, Non-linear Problems in, Figure 15
Schematic comparison of strong-motion power demands E1 and E2 with an envelope of structural power capacity

Power Design

Figure 15 illustrates the cumulative wave energies recorded at a building site during two hypothetical earthquakes, E1 and E2, and presents a conceptual framework that can be used for development of the power design method. E1 results in a larger total shaking energy at the site and has a long duration of shaking, leading to relatively small average power, P_1 . E2 leads to smaller total shaking energy at the site but has short duration and thus greater power, P_2 . The power capacity of a structure cannot be described by one unique cumulative curve, as this depends upon the time history of shaking. For the purposes of this illustration, the line labeled “capacity envelope of the structure” can be thought of as an envelope of all possible cumulative energy paths for the response of this structure. Figure 15 implies that E1 will not damage this structure, but E2 will. Hence, *for a given structure, it is not the total energy of an earthquake event (and the equivalent energy-compatible relative velocity spectrum) but the rate with which this energy arrives and shakes the structure that is essential for the design of the required power capacity of the structure to withstand this shaking and to control the level of damage.*

Trifunac [57] outlined the elementary aspects of such design based on the power of the incident wave pulses. He showed how this power can be compared with the capacity of the structure to absorb the incident wave energy and described the advantages of using the computed power of incident strong motion for design. Power (amplitude

and duration) of the strong near-field pulses will determine whether the wave entering the structure will continue to propagate through the structure as a linear wave or will begin to create nonlinear zones (at first near the top and/or near the base of the structure; Gicev and Trifunac [10,11,12]). For high-frequency pulses, the nonlinear zone, with permanent strains, can be created before the wave motion reaches the top of the structure—that is, before the interference of waves has even started to occur and lead to formation of mode shapes. Overall duration of strong motion [60] will determine the number of times the structure may be able to complete full cycles of response and the associated number of “minor” excursions into the nonlinear response range when the response is weakly non-linear [13], while the presence of powerful pulses of strong motion will determine the extent to which the one-directional quarter period responses [57] may lead to excessive ductility demand, leading to dynamic instability and failure, precipitated by the gravity loads [20]. All of these possibilities can be examined and quantified deterministically by computation of the associated power capacities and power demands for different scenarios, for given recorded or synthesized strong-motion accelerograms, or probabilistically by using the methods developed for Uniform Hazard Analysis [52].

Bibliography

- Beltrami E (1987) *Mathematics for Dynamic Modeling*. Wiley, New York
- Biot MA (1932) Vibrations of buildings during earthquakes. In: *Transient Oscillations in Elastic System*. Ph D Thesis No. 259, Chapter II. Aeronautics Department, California Institute of Technology, Pasadena
- Biot MA (1933) Theory of elastic systems vibrating under transient impulse with an application to earthquake-proof buildings. *Proc Natl Acad Sci* 19(2):262–268
- Biot MA (1934) Theory of vibration of buildings during earthquakes. *Z Angew Math Mech* 14(4):213–223
- Biot MA (2006) Influence of foundation on motion of blocks. *Soil Dyn Earthq Eng* 26(6–7):486–490
- Bycroft GN (1980) Soil-foundation interaction and differential ground motions. *Earthq Eng Struct Dyn* 8(5):397–404
- Crutchfield JP (1992) Knowledge and meaning. In: Lam L, Naroditsky V (eds) *Modeling Complex Phenomena*. Springer, New York, pp 66–101
- Fujino Y, Hakuno M (1978) Characteristics of elasto-plastic ground motion during an earthquake. *Bull Earthq Res Inst Tokyo Univ* 53:359–378
- Gicev V (2005) Investigation of soil-flexible foundation-structure interaction for incident plane SH waves. Ph D Dissertation, Department of Civil Engineering. University Southern California, Los Angeles
- Gicev V, Trifunac MD (2006) Rotations in the transient response of nonlinear shear beam. Department of Civil Engineering Report, CE 06–02. University Southern California, Los Angeles
- Gicev V, Trifunac MD (2006) Non-linear earthquake waves in seven-story reinforced concrete hotel. Dept. of Civil Engineering, Report, CE 06–03. University Southern California, Los Angeles
- Gicev V, Trifunac MD (2007) Permanent deformations and strains in a shear building excited by a strong motion pulse. *Soil Dyn Earthq Eng* 27(8):774–792
- Gupta ID, Trifunac MD (1996) Investigation of nonstationarity in stochastic seismic response of structures. Dept. of Civil Eng. Report, CE 96–01. University of Southern California, Los Angeles
- Gwinn EG, Westervelt RM (1985) Intermittent chaos and low-frequency noise in the driven damped pendulum. *Phys Rev Lett* 54(15):1613–1616
- Hackett K, Holmes PJ (1985) Josephson Junction, annulus maps, Birkhoff Attractors, horseshoes and rotation sets. Center for Applied Math Report, Cornell University, Ithaca
- Holmes PJ (1979) A nonlinear oscillator with a strange attractor. *Philos Trans R Soc London A* 292:419–448
- Holmes PJ (1982) The dynamics of repeated impacts with a sinusoidally vibrating table. *J Sound Vib* 84:173–189
- Holmes PJ (1985) Dynamics of a nonlinear oscillator with feedback control. *J Dyn Syst Meas Control* 107:159–165
- Holmes PJ, Moon FC (1983) Strange Attractors and Chaos in Nonlinear Mechanics. *J Appl Mech* 50:1021–1032
- Husid R (1967) Gravity effects on the earthquake response of yielding structures. Ph D Thesis, California Institute of Technology, Pasadena
- Jalali R, Trifunac MD (2007) Strength-reduction factors for structures subjected to differential near-source ground motion. *Indian Soc Earthq Technol J* 44(1):285–304
- Kapitaniak T (1991) *Chaotic Oscillations in Mechanical Systems*. Manchester University Press, Manchester
- Kojic S, Trifunac MD, Anderson JC (1984) A post-earthquake response analysis of the Imperial County Services building in El Centro. Report, CE 84–02. University of Southern California, Department of Civil Engineering, Los Angeles
- Kuhn T (1962) *The Structure of Scientific Revolutions*. The University of Chicago Press, Chicago
- Lee VW (1979) Investigation of three-dimensional soil-structure interaction. Department of Civil Engineering Report, CE 79–11. University of Southern California, Los Angeles
- Lee VW, Trifunac MD (1982) Body wave excitation of embedded hemisphere. *ASCE, EMD* 108(3):546–563
- Lee VW, Trifunac MD (1985) Torsional accelerograms. *Int J Soil Dyn Earthq Eng* 4(3):132–139
- Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. *Int J Soil Dyn Earthq Eng* 6(2):75–89
- Levin PW, Koch BP (1981) Chaotic behavior of a parametrically excited damped pendulum. *Phys Lett A* 86(2):71–74
- Lichtenberg AJ, Lieberman MA (1983) *Regular and Stochastic Motion*. Springer, New York
- Lighthill J (1994) Chaos: A historical perspective. In: Newman WI, Gabrielov A, Turcotte D (eds) *Nonlinear Dynamics and Predictability of Geophysical Phenomena*. Geophysical Monograph 83, vol 18. IUGG, American Geophysical Union, Washington DC, pp 1–5

32. Lomnitz C, Castanos H (2006) Earthquake hazard in the valley of Mexico: entropy, structure, complexity, Chapter 27. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg
33. Luco JE, Wong HL, Trifunac MD (1986) Soil-structure interaction effects on forced vibration tests. Department of Civil Engineering Report, No. 86-05. University of Southern California, Los Angeles
34. McLaughlin JB (1981) Period-doubling bifurcations and chaotic motion for a parametrically forced pendulum. *J Stat Phys* 24(2):375-388
35. Miles J (1984) Resonant motion of spherical pendulum. *Physica* 11D:309-323
36. Miles J (1984) Resonantly forced motion of two quadratically coupled oscillators. *Physica* 13D:247-260
37. Moon FC (1980) Experiments on chaotic motions of a forced nonlinear oscillator: Strange attractors. *ASME J Appl Mech* 47:638-644
38. Moon FC (1980) Experimental models for strange attractor vibration in elastic systems. In: Holmes PJ (ed) *New Approaches to Nonlinear Problems in Dynamics*. SIAM, Philadelphia, pp 487-495
39. Moon FC, Holmes PJ (1979) A magnetoelastic strange attractor. *J Sound Vib* 65(2):275-296; A magneto-elastic strange attractor. *J Sound Vib* 69(2):339
40. Moon FC, Holmes WT (1985) Double Poincaré sections of a quasi-periodically forced, chaotic attractor. *Phys Lett A* 111(4):157-160
41. Moon FC, Shaw SW (1983) Chaotic vibration of beams with nonlinear boundary conditions. *J Nonlinear Mech* 18:465-477
42. Poddar B, Moon FC, Mukherjee S (1986) Chaotic motion of an elastic-plastic beam. *J Appl Mech ASME* 55(1):185-189
43. Rasband SN (1990) *Chaotic Dynamics of Nonlinear Systems*. Wiley, New York
44. Reitherman R (2006) The effects of the 1906 earthquake in California on research and education. *Earthq Spectra* 52(22):S207-S236
45. Richter PH, Scholtz HJ (1984) Chaos and classical mechanics: The double pendulum. In: Schuster P (ed) *Stochastic Phenomena and Chaotic Behavior in Complex Systems*. Springer, Berlin, pp 86-97
46. Shaw SW (1985) The dynamics of a harmonically excited system having rigid amplitude constraints, parts 1, 2. *J Appl Mech* 52(2):453-464
47. Shaw S, Holmes PJ (1983) A periodically forced piecewise linear oscillator. *J Sound Vib* 90(1):129-155
48. Sorrentino L (2007) The early entrance of dynamics in earthquake engineering: Arturo Danusso's contribution. *ISCT J* 44(1):1-24
49. Todorovska MI, Al Rjoub Y (2006) Effects of rainfall on soil-structure system frequency: Examples based on poroelasticity and comparison with full-scale measurements. *Soil Dyn Earthq Eng* 26(6-7):708-717
50. Todorovska MI, Trifunac MD (1990) Note on excitation of long structures by ground waves. *ASCE, EMD* 116(4):952-964 (Errata in 116:1671)
51. Todorovska MI, Trifunac MD (2007) Earthquake Damage Detection in the Imperial County Services Building I: the Data and Time-Frequency Analysis. *Soil Dyn Earthq Eng* 27(6):564-576
52. Todorovska MI, Gupta ID, Gupta VK, Lee VW, Trifunac MD (1995) Selected topics in probabilistic seismic hazard analysis. Department of Civil Engineering Report, No. CE 95-08. University of Southern California, Los Angeles
53. Trifunac MD (1989) Dependence of Fourier spectrum amplitudes of recorded strong earthquake accelerations on magnitude, local soil conditions and on depth of sediments. *Earthq Eng Struct Dyn* 18(7):999-1016
54. Trifunac MD (1993) Long-period Fourier amplitude spectra of strong motion acceleration. *Soil Dyn Earthq Eng* 12(6):363-382
55. Trifunac MD (1994) Q and High-Frequency Strong-Motion Spectra. *Soil Dyn Earthq Eng* 13(3):149-161
56. Trifunac MD (2003) 70th Anniversary of Biot Spectrum, 23rd Annual ISET Lecture. *Indian Soc Earthq Technol* 1(40):19-50
57. Trifunac MD (2005) Power design method. *Proc. of Earthquake Engineering in the 21st Century to Mark 40th Anniversary of IZIS-Skopje*, 28 Aug-1 Sept. Skopje and Ohrid, Macedonia
58. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures, Ch 39. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake Source Asymmetry, Structural Media, and Rotation Effects*. Springer, Heidelberg
59. Trifunac MD, Gicev V (2006) Response Spectra for Differential Motion of Columns, Paper II: Out-of-Plane Response. *Soil Dyn Earthq Eng* 26(12):1149-1160
60. Trifunac MD, Novikova EI (1994) State-of-the-art review on strong motion duration. 10th European Conf on Earthquake Eng, Vienna, 28 Aug - 2 Sep 1994 vol I. AA Balkema, Rotterdam, pp 131-140
61. Trifunac MD, Todorovska MI (1997) Response spectra and differential motion of columns. *Earthq Eng Struct Dyn* 26(2): 251-268
62. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building I: Fourier analysis. *J Struct Eng ASCE* 127(5):517-526
63. Trifunac MD, Ivanovic SS, Todorovska MI (2001) Apparent periods of a building II: Time-frequency analysis. *J Struct Eng ASCE* 127(5):527-537
64. Trifunac MD, Hao TY, Todorovska MI (2001) Response of a 14-story reinforced concrete structure to nine earthquakes: 61 years of observation in the Hollywood storage building. Department of Civil Engineering Report, CE 01-02. University of Southern California, Los Angeles
65. Trifunac MD, Hao TY, Todorovska MI (2001) On energy flow in earthquake response, Department of Civil Engineering Report, No. CE 01-03. University of Southern California, Los Angeles
66. Udwadia FE, MD Trifunac (1974) Time and amplitude dependent response of structures. *Earthq Eng Struct Dyn* 2:359-378
67. Ueda Y (1980) Steady motions exhibited by Duffing's equation. In: Holmes PJ (ed) *A picture book of regular and chaotic motions. New Approaches to Nonlinear Problems in Dynamics*. SIAM, Philadelphia
68. Veletsos AS, Newmark NM (1960) Effect of inelastic behavior on the response of simple systems to earthquake motions. *Proc 2nd World Conf on Earthquake Engineering*, Jul 1960, vol II. Science Council of Japan, Tokyo, pp 859-912
69. Veletsos AS, Newmark NM (1964) Response spectra for single-degree-of-freedom elastic and inelastic systems. Report No. RTD-TDR-63-3096, vol III. Air Force Weapons Lab, Albuquerque
70. Wong HL, Trifunac MD (1979) Generation of artificial strong motion accelerograms. *Int J Earthq Eng Struct Dyn* 7(6): 509-527

Earthquake Forecasting and Verification

JAMES R. HOLLIDAY, JOHN B. RUNDLE,
DONALD L. TURCOTTE
University of California, Davis, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Earthquake Forecasting
Forecast Verification
Future Directions
Bibliography

Glossary

Binary forecast A type of forecast where earthquakes are forecast to occur in certain regions and forecast not to occur in other regions.

Continuum forecast A type of forecast where the likelihood of an earthquake throughout an entire region is specified.

Failure to predict Earthquake event that occurs where no earthquake are forecasted to occur.

False alarm Earthquake event that is forecasted to occur at a specific location at a specific time but does not occur.

PDF Probability Density Function – A probability density function is any function $f(x)$ that describes the probability density in terms of the input variable x such that $f(x)$ is greater than or equal to zero for all values of x and the total area of the function is 1.

Definition of the Subject

Forecasts of likely future events are used in almost every field: from forecasting tomorrow's weather to modeling the rise and fall of financial indices to predicting the growth of cancerous cells in human tissue. Generally, these forecasts are created under the belief that having a forecast – regardless of the level of complexity of the underlying models – is more desirable than not having a forecast. That is, human nature prefers the foreseeable over the unexpected. It is therefore important to verify the forecast and measure its skill, or “goodness”, and its value, or “usefulness”. The process of testing a given forecast with past trend data is the study of forecast verification. Forecast verification allows for a precise and repeatable (as op-

posed to relative or subjective) judgment of a forecasting model.

Introduction

Earthquakes are the most feared of natural hazards because they generally occur without warning. Hurricanes can be tracked, floods develop gradually, tornados are caused by measurable atmospheric conditions, and volcanic eruptions are preceded by a variety of precursory phenomena. Earthquakes, however, occur suddenly and often without precursory indicators. There have been a wide variety of approaches applied to the forecasting of earthquakes [30,31,37,40,56,65]. These approaches can be divided into two general classes. The first approach is based on empirical observations of precursory changes. Examples include precursory seismic activity, precursory ground motions, and many others. The second approach is based on statistical patterns of seismicity. Neither approach has been able to provide reliable short-term forecasts (days to months) on a consistent basis.

Although short-term predictions are not available (see Table 1), long-term seismic-hazard assessments can be made. It is also possible to assess the long-term probability of having an earthquake of a given magnitude in a given region. These assessments are primarily based on the hypothesis that future earthquakes will occur in regions where past earthquakes have occurred [14,35]. Specifically, the rate of occurrence of small earthquakes in a region can be analyzed to assess the probability of occurrence of much larger earthquakes. While some earthquakes occur in plate interiors – a specific example is the three large (magnitude ~ 7.7) earthquakes that occurred near New Madrid, Missouri in 1810 and 1811 – the large majority of all earthquakes occur in the vicinity of plate boundaries. A number of large cities are located very close to plate boundaries. Examples include Tokyo, Los Angeles, San Francisco, Seattle, Lima, Jakarta, and Santiago. Much of China is a diffuse plate boundary, and major earthquakes have caused devastating losses of life throughout this region. A recent example was in the 1976 Tangshan earthquake with some 500,000 deaths.

A major goal for earthquake forecasting is to quantify the risk of occurrence of an earthquake of a specified magnitude, in a specified area, and in a specified time window. This is routinely done and results in the creation of hazard maps. Another goal is to specifically forecast or predict earthquakes. The fundamental question is whether forecasts of the time and location of future earthquakes can be accurately made. It is accepted that long term hazard maps of the expected rate of occurrence of earthquakes are rea-

Earthquake Forecasting and Verification, Table 1

Warning times, scientific bases, and scientific feasibility for various types of earthquake predictions and estimates of long-term potential [62]

Term	Warning Time	Scientific Basis	Feasibility
Immediate alert	0 to 20 seconds	Speed of electro -magnetic waves \gg speed of seismic waves	Good
Short-term prediction	Hours to weeks	Accelerating aseismic slip, foreshocks for some events	Unknown
Mid-term prediction	1 month to 10 years	Changes in seismicity, strain, chemistry, and fluid pressure	Fair
Long-term prediction	10 to 30 years	Time remaining in cycle of large shocks, increase in regional shocks	Good
Long-term potential	> 30 years	Long-term rate of activity, plate tectonic setting	Very good

sonably accurate. But is it possible to do better? Are there precursory phenomena that will allow earthquakes to be forecast?

Earthquake Forecasting

Chaos and Forecasting

One of the reasons earthquakes are difficult to accurately forecast is the underlying complexity of the fault system. Earthquakes are caused by displacements on preexisting faults. Most earthquakes occur at or near the boundaries between the near-rigid plates of plate tectonics. Earthquakes in California are associated with the relative motion between the Pacific plate and the North American plate. Much of this motion is taken up by displacements on the San Andreas fault, but deformation and earthquakes extend from the Rocky Mountains on the east into the Pacific Ocean on the west. Clearly this deformation and the associated earthquakes are extremely complex.

It is now generally accepted that earthquakes are examples of deterministic chaos [66]. Some authors [16,17] have argued that this chaotic behavior precludes the prediction of earthquakes. Weather systems, however, are also chaotic, yet short-term forecasts are routinely made. Weather forecasts are probabilistic in the sense that weather cannot be predicted exactly. One such example is the track of a hurricane. Probabilistic forecasts of hurricane tracks are made every year; sometimes they are extremely accurate while at other times they are not. Another example of weather forecasting is the forecast of El Niño events. Forecasting techniques based on pattern recognition and principle components of the sea surface temperature fluctuation time series have been developed that are quite successful in forecasting future El Niños, but again they are probabilistic in nature [11]. It has also been argued [62] that chaotic behavior does not preclude the probabilistic forecasting of future earthquakes. The belief is that the chaos and nonlinearity in earthquakes arise mainly during unstable sliding in large events. Thus, predictions are possible before large earthquakes, but take a fi-

nite amount of time for the system to recover after large earthquakes.

Unobservable Dynamics

Another reason earthquakes are difficult to accurately forecast is that the true dynamics driving the system are simply unobservable and unmeasurable. As discussed above, earthquake faults occur in topologically complex, multi-scale networks that are driven to failure by external forces arising from plate tectonic motions [66]. The basic problem is that the details of the true space-time, force-displacement dynamics are in general unobservable, except in a few selected locations such as deep drill holes [52] or in a very crude, time-averaged sense such as the World Stress Map [81]. In order to completely describe the system, the true dynamics would have to be observable for all space and at all times. In fault systems these unobservable dynamics are usually encoded [59] in the time evolution of the Coulomb failure function, $CFF(x, t)$:

$$CFF(x, t) = \tau(x, t) - \mu_s \sigma_N(x, t), \quad (1)$$

where $\tau(x, t)$ is the shear stress at point x and at time t , μ_s is the coefficient of static friction, and $\sigma_N(x, t)$ is normal stress at point x and at time t . The space-time patterns associated with the time, location, and magnitude of the earthquakes, however, are observable. This leads to a focus on understanding the observable, multi-scale, apparent dynamics [52] of earthquakes in an attempt to infer the underlying dynamics.

Empirical Approaches

Empirical approaches to earthquake prediction rely on local observations of precursory phenomena in the vicinity of the earthquake to be predicted. It has been suggested that one or more of the following phenomena may indicate a future earthquake [30,31,37,40,56,65]:

1. precursory increase or decrease in seismicity in the vicinity of the origin of a future earthquake rupture,

2. precursory fault slip that leads to surface tilt and/or displacements,
3. electromagnetic signals,
4. chemical emissions, and
5. changes in animal behavior.

Examples of successful near-term predictions of future earthquakes based solely on empirical observations have been rare. A notable exception was the prediction of the $M = 7.3$ Haicheng earthquake in northeast China that occurred on 4 February 1975. This prediction led to the evacuation of the city which undoubtedly saved many lives. The Chinese reported that the successful prediction was based on foreshocks, groundwater anomalies, and animal behavior. Unfortunately, a similar prediction was not made prior to the magnitude $M = 7.8$ Tangshan earthquake that occurred on 28 July 1976 [68]. Official reports placed the death toll in this earthquake at 242,000, although unofficial reports placed it as high as 655,000.

In order to thoroughly test for the occurrence of direct precursors the United States Geological Survey (USGS) initiated the Parkfield (California) Earthquake Prediction Experiment in 1985 [1,30]. Earthquakes on this section of the San Andreas had occurred in 1857, 1881, 1901, 1922, 1934, and 1966. It was expected that the next earthquake in this sequence would occur by the early 1990s, and an extensive range of instrumentation was installed. The next earthquake in the sequence finally occurred on 28 September 2004. No precursory phenomena were observed that were significantly above the background noise level. Although the use of empirical precursors cannot be ruled out, the future of those approaches does not appear to be promising at this time.

Statistical Approaches

A variety of studies have utilized variations in seismicity over relatively large distances to forecast future earthquakes. The distances are large relative to the rupture dimension of the subsequent earthquake. These approaches are based on the concept that the earth's crust is an activated, or driven, thermodynamic system [52]. Among the evidence for this behavior is the continuous level of background seismicity in all seismographic areas. About a million magnitude two earthquakes occur each year on our planet. In southern California about a thousand magnitude two earthquakes occur each year. Except for the aftershocks of large earthquakes, such as the 1992 $M = 7.3$ Landers earthquake, this seismic activity is essentially constant over time. If the level of background seismicity varied systematically with the occurrence of large earth-

quakes, earthquake forecasting would be relatively easy. This, however, is not the case.

While there is yet no indication of a universal earthquake indicator, there is increasing evidence that there are systematic precursory variations in some aspects of regional seismicity at least some of the time. For example, it has been observed that there is a systematic variation in the number of magnitude $M = 3$ and larger earthquakes prior to at least some magnitude $M = 5$ and larger earthquakes, and a systematic variation in the number of magnitude $M = 5$ and larger earthquakes prior to some magnitude $M = 7$ and larger earthquakes. The spatial regions associated with this phenomena tend to be relatively large, suggesting that an earthquake may resemble a phase change with an increase in the "correlation length" prior to an earthquake [5,26]. A specific example is the sequence of earthquakes that preceded the 1906 San Francisco earthquake [61]. This seismic activation has been quantified as a power law increase in seismicity prior to earthquakes [4,5,6,7,8,9,10,26,34,38,46,55,79]. There have also been reports of anomalous quiescence in the source region prior to a large earthquake, a pattern that is often called a "Mogi Donut" [30,40,76,77]. Unfortunately, these studies have all been performed retrospectively and their successes have depended on knowing the location of the subsequent earthquake.

There are two fundamentally different approaches to assessing the probabilistic risk of earthquake occurrence using statistical methods. The first of these is fault based, where the statistical occurrence of earthquakes is determined for mapped faults. The applicable models are known as renewal models and a tectonic loading of faults is included. The second approach is seismicity based, where the risk of future earthquakes is based on the past seismicity in the region. These are also known as cluster models and include the epidemic type aftershock sequence (ETAS) model and the branching aftershock sequence (BASS) model.

Fault Based Models Fault based models consider the earthquakes that occur on recognized (i.e., previously known) active faults. These models are also known as renewal models. Renewal models assume that the stress on an individual fault is "renewed" by the tectonic drive of plate tectonics. The simplest renewal model would be that of a single planar strike-slip fault subjected to a uniform rate of strain accumulation (plate motion). In this case, "characteristic" earthquakes would occur periodically. Clearly the earth's crust is much more complex with faults present at all scales and orientations. This complexity leads to chaotic behavior and statistical variability.

An important question is whether the concept of quasi-periodic “characteristic” earthquakes is applicable to tectonically active areas. There is extensive evidence that characteristic earthquakes do occur quasi-periodically on major faults. Many studies have been carried out to quantify the recurrence time statistics of these characteristic earthquakes [43,45,67]. Recurrence time statistics can be characterized by a mean value, μ , and a coefficient of variation, C_v . The coefficient of variation is the ratio of the standard deviation to the mean. Mathematically, $C_v = 0$ for periodic characteristic earthquakes and $C_v = 1$ for a random distribution of recurrence times. Ellsworth et al. [13] reviewed many examples of recurrence time statistics and concluded that $C_v \approx 0.5$ for characteristic earthquakes. Many probability distribution functions have been proposed for recurrence times, including the Weibull, lognormal, Brownian passage time, and gamma distributions.

Two major renewal simulation models have been developed. The first is “Virtual California” [49,50,53]. This is a geometrically realistic numerical simulation of earthquakes occurring on the San Andreas fault system and includes all major strike-slip faults in California. The second model is the “Standard Physical Earth Model” (SPEM) developed by Ward [69] and applied to characteristic earthquakes associated with subduction at the Middle American trench. This model was further developed and applied to the entire San Andreas fault system by Goes and Ward [18], to the San Andreas system in southern California by Ward [70], and to the San Andreas system in northern California by Ward [71].

Both simulation models utilize backslip, with the accumulation of a slip deficit on each fault segment prescribed using available data. The backslip represents the tectonic drive. Both models “tune” the prescribed static friction to give recurrence times that are consistent with available data. In both models fault segments are treated as dislocations when characteristic earthquakes occur, and all fault segments interact with each other elastically utilizing dislocation theory. These chaotic interactions result in statistical distributions of recurrence times on each fault. The resulting coefficients of variation are measures of this interaction.

Yakovlev et al. [78] utilized the Virtual California model to test alternative distributions of recurrence times. They concluded that the Weibull distribution is preferable and based its use on its scale invariance. The hazard rate is the probability that a characteristic earthquake will occur at a given time after the last characteristic earthquake. The Weibull distribution is the only distribution that has a power-law (scale-invariant) hazard function. In

the same study, Yakovlev et al. [78] found that the coefficient of variation of the recurrence times of 4606 simulated great earthquakes on the northern San Andreas fault is $C_v = 0.528$. Goes and Ward [18] using the SPEM simulator found that $C_v = 0.50 - 0.55$ on this fault. The two simulations are quite different, so the statistical variability appears to be a robust feature of characteristic earthquakes. A similar simulation model for New Zealand has been given by Robinson and Benites [47,48].

Renewal models have also formed the basis for three formal assessments of future earthquake probabilities in California. These assessments were carried out by the United States Geological Survey [72,73,74,75]. A major problem with renewal models is that large earthquakes in nature often occur on faults that were not previously recognized. Recent examples in California include the 1952 Kern County earthquake, the 1971 San Fernando Valley earthquake, the 1992 Landers earthquake, the 1994 Northridge earthquake, and the 1999 Hector Mine earthquake. At the times when these earthquakes occurred, the associated faults were either not mapped or were considered too small to have such large earthquakes. To compensate for this problem, renewal models often include a random level of background seismicity unrelated to recognized faults.

Seismicity Based Models An alternative approach to probabilistic seismic hazard assessment and earthquake forecasting is to use observed seismicity. The universal applicability of Gutenberg–Richter frequency-magnitude scaling allows the rate of occurrence of small earthquakes to be extrapolated to estimate the rate of occurrence and location of large earthquakes. This type of extrapolation played an important role in creating the national seismic hazard map for the United States [15].

A more formalistic application of this extrapolation methodology is known as a relative intensity (RI) forecast. This type of forecast was made on a world wide basis by Kossobokov et al. [35] and to California by Holliday et al. [23]. A related forecasting methodology is the pattern informatics (PI) method [22,24,25,51,63,64]. This method was used by Rundle et al. [51] to forecast $m = 5$ and larger earthquakes in California for the time period 2000–2010. This forecast successfully predicted the locations of 16 of the 18 large earthquakes that have subsequently occurred.

Keilis-Borok [31,33] and colleagues utilized patterns of seismicity to make formal intermediate term earthquake predictions. The most widely used algorithm, M8, has been moderately successful in predicting the times and locations of large earthquakes. More recently, this group has used chains of premonitory earthquakes to make interme-

mediate term predictions [32,58]. Again, moderate success was achieved.

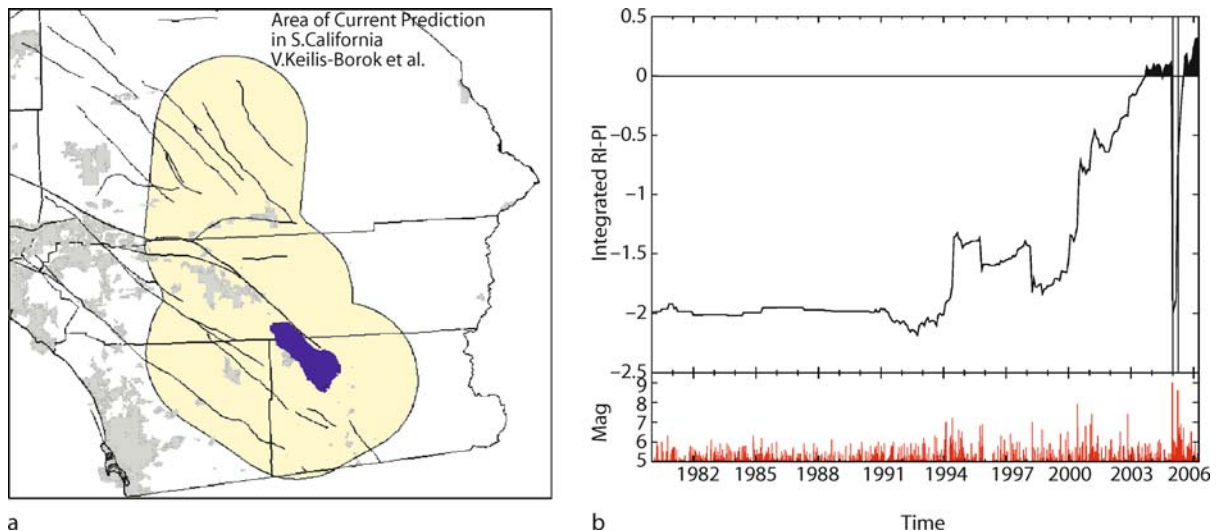
It has also been proposed that there is an increase in the number of intermediate sized earthquakes prior to a large earthquake [26]. This phenomenon is known as accelerating moment release (AMR) and is due primarily to an increase in the number of intermediate-size events that occur within a characteristic distance of the main shock and that scale with magnitude. AMR is characterized by a decrease in the rate of regional seismicity followed by a rapid rebound back to historic levels. Sammis and Bowman [54] have proposed a number of physical models to explain AMR. These include:

1. an analogy with critical phase transitions where the correlation length of the stress field rapidly increases as the system nears the critical point,
2. an erosion of a stress shadow from some previous, large event, and
3. a slow, silent earthquake propagating upward on a ductile extension loading the seismogenic crust above.

The existence of such a seismicity pattern does, however, appear to require a certain regional fault system structure

and density. Simulation models using a hierarchical distribution of fault sizes match this pattern well, but other types of fault distributions may also support AMR [26]. Conversely, some real-world fault distributions may not support AMR as a predictive tool. The AMR approach has shown considerable success retrospectively [5,10,55] but has not evolved into a successful prediction algorithm as of yet.

Seismicity based models are often referred to as clustering models. That is, clusters of small earthquakes indicate the future occurrence of larger earthquakes. The RI, PI, and AMR models clearly belong to this class. Other approaches in this class are the epidemic type aftershock sequence (ETAS) model [21,29,42,44] and the branching aftershock sequence (BASS) model. These are statistical models based on applicable scaling laws: Gutenberg–Richter scaling relation and the modified Båth's law for the scaling relation of magnitudes, Omori's law for the distribution of earthquake times, and a modified form of Omori's law for the distribution of earthquake locations. Clustering by definition is not a random process. A rationale for the application of clustering models is that the clustering is related to families of foreshocks, main shocks, and aftershocks.



Earthquake Forecasting and Verification, Figure 1

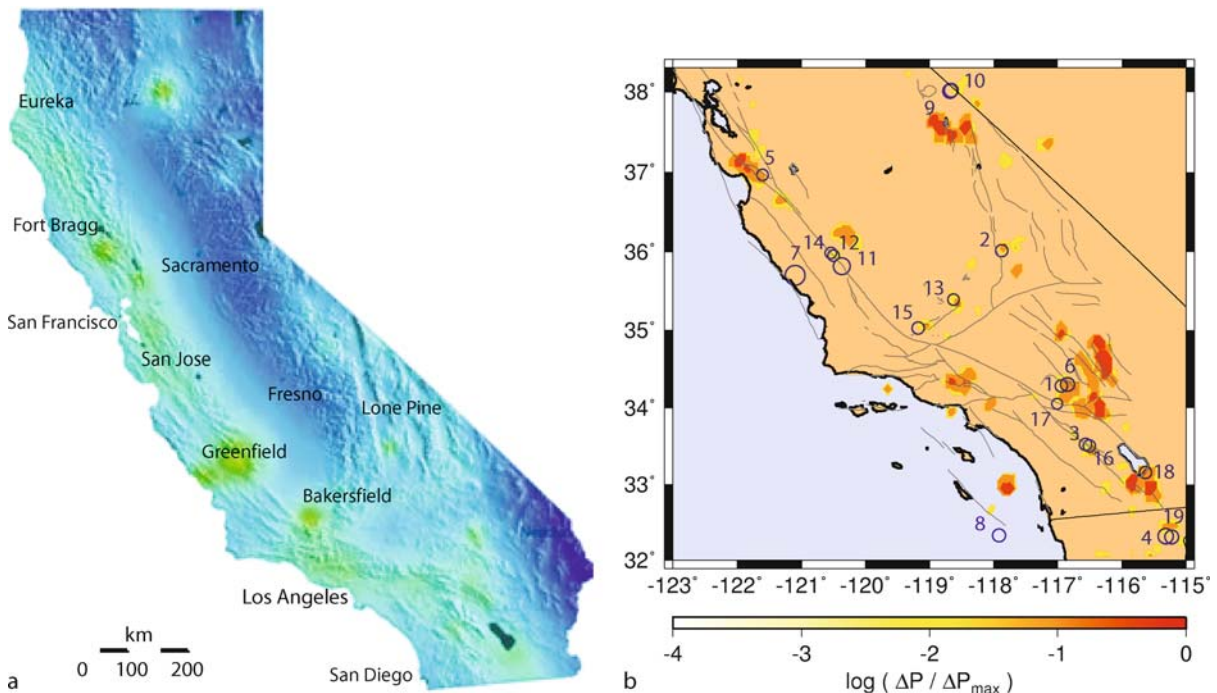
Plots of sample binary earthquake forecasts. **a** Prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm (image courtesy Dr. Vladimir Keilis-Borok). This is a binary forecast since it forecasts an earthquake to occur within the *shaded* area during the time period and not to occur in the *non-shaded* region. Ultimately no magnitude 6.4 or greater earthquake occurred in the test region during the forecast interval. **b** Retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005 (image courtesy Dr. James Holliday). This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive. Note that the magnitude 9.0 Sumatra-Andaman earthquake occurred 18 months after the time series became positive

Types of Forecasts

Binary Forecasts The simplest type of earthquake forecast is a binary forecast. An earthquake is forecast to occur in a certain regions and forecast not to occur in other regions. This is analogous to the issuance of tornado warnings. Examples of two binary forecast maps are presented in Fig. 1. The plot on the left is for a prediction for a magnitude 6.4 or greater earthquake to occur between 5 January 2004 and 4 September 2004, within a 12,440 sq. miles area of southern California using the M8 algorithm. This map is a binary forecast since it forecasts an earthquake to occur within the shaded area during the time period and not to occur in the non-shaded region. The plot on the right is a retrospective prediction for a magnitude 8.0 or greater earthquake to occur in the Sumatra region between 1 June 2003 and 1 June 2005. This is a time-dependent binary forecast since large earthquakes are forecasted to occur in the test region within a two year window once the time series becomes positive and not to occur when the time series is negative.

Continuum Forecasts The alternative to binary forecasts is a continuum forecast. The likelihood of an earthquake throughout the entire region is specified. This would be analogous to temperature forecasts in the atmospheric sciences. Examples of two continuum forecast maps are presented in Fig. 2. The plot on the left is a time-dependent map produced in real time by the USGS Earthquake Hazards Program giving the probability of strong shaking at any location in California within a given 24-hour period. The plot on the right is a forecast map giving the probability for large (magnitude greater than 5.0) earthquakes in southern California using the Pattern Informatics method.

Any continuum forecast can be converted into a binary forecast through the use of a hard threshold. Spatial regions where the likelihood value is greater than the threshold are taken to be regions where earthquakes are forecasted to occur. Spatial regions where the likelihood value is less than the threshold value are taken to be regions where earthquakes are forecasted not to occur.



Earthquake Forecasting and Verification, Figure 2

Plots of sample continuum earthquake forecasts. **a** Time-dependent map giving the probability of strong shaking at any location in California within a given 24-hour period (image courtesy USGS Earthquake Hazards Program). **b** Forecast map giving the probability for large ($m > 5$) earthquakes in southern California using the Pattern Informatics method (image courtesy Dr. Kristy Tiampo). Circles mark the locations of large earthquakes which occurred after the forecast creation. Both of these are continuum forecasts since they present a continuous likelihood for earthquakes to occur throughout the entire test region

Forecast Verification

Continuum Forecasts

Likelihood Tests The standard approach for testing the hypothesis that a probability measure can forecast future earthquakes is the maximum likelihood test [2,19,24,28,51,57,63]. The likelihood \mathbf{L} is a probability measure that can be used to assess the quality of one forecast measure over another. Typically, one computes the log-likelihood $\mathcal{L} \equiv \log(\mathbf{L})$ for the proposed forecast measure \mathbf{L} . Models with higher (less negative) log-likelihood values are said to perform better than models with lower (more negative) log-likelihood values. In these types of likelihood tests, a probability density function (PDF) is required. Two different PDFs are commonly used: a global, Gaussian model and a local, Poissonian model.

Tiampo et al. [63] calculated likelihood values by defining $P[\mathbf{x}]$ to be the union of a set of N Gaussian density functions $p_G(|\mathbf{x} - \mathbf{x}_i|)$ [2] centered at each location \mathbf{x}_i . Each individual Gaussian density has a standard deviation σ equal to the width of their coarse-grained lattice cell and a peak value equal to the calculated probability divided by σ^2 . $P[\mathbf{x}(e_j)]$ was then interpreted as a probability measure that a future large event e_j would occur at location $\mathbf{x}(e_j)$:

$$P[\mathbf{x}(e_j)] = \sum_i \frac{P_i}{\sigma^2} e^{-\frac{|\mathbf{x}(e_j) - \mathbf{x}_i|^2}{\sigma^2}}. \quad (2)$$

If there are J future events, the normalized likelihood \mathbf{L} that all J events are forecast is

$$\mathbf{L} = \prod_j \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]}. \quad (3)$$

Furthermore, the log-likelihood value \mathcal{L} for a given calculation can be calculated and used in ratio comparison tests:

$$\mathcal{L} = \sum_j \log \frac{P[\mathbf{x}(e_j)]}{\sum_i P[\mathbf{x}_i]}. \quad (4)$$

The second model commonly used is based on work performed by the Regional Earthquake Likelihood Models (RELM) group [57]. For each coarse-grained lattice cell i an expectation value λ_i is calculated by scaling the local probability value P_i by the number of earthquakes that occurred over all space during the forecasted time period:

$$\lambda_i = n \cdot P_i, \quad (5)$$

where n is the number of future events. Note that for any future time interval (t_2, t_3) , n could in principle be estimated by using the Gutenberg–Richter relation. For each

bin an observation value ω_i is also calculated such that ω_i contains the number of future earthquakes that actually occurred in cell i . Note that $\sum_i \omega_i = n$. For the RELM model, it is assumed that earthquakes are independent of each other. Thus, the probability of observing ω_i events in cell i with expectation λ_i is the Poissonian probability

$$p_i(\omega_i|\lambda_i) = \frac{\lambda_i^{\omega_i}}{\omega_i!} e^{-\lambda_i}. \quad (6)$$

The log-likelihood \mathcal{L} for observing ω earthquakes at a given expectation λ is defined as the logarithm of the probability $p(\omega|\lambda)$, thus

$$\mathcal{L}(\omega|\lambda) = \log p(\omega|\lambda) = -\lambda + \omega \log \lambda - \log(\omega!). \quad (7)$$

Since the joint probability is the product of the individual cell probabilities, the log-likelihood value for a given calculation is the sum of $\mathcal{L}(\omega_i|\lambda_i)$ over all cells i :

$$\mathcal{L} = \sum_i \mathcal{L}(\omega_i|\lambda_i) = \sum_i (-\lambda_i + \omega_i \log \lambda_i - \log(\omega_i!)). \quad (8)$$

Most tests of earthquake forecasts have emphasized the likelihood test [24,28,51,64]. These tests have the significant disadvantage that they are overly sensitive to the least probable events. For example, consider two forecasts. The first perfectly forecasts 99 out of 100 events but assigns zero probability to the last event. The second assigns zero probability to all 100 events. Under a log-likelihood test, both forecasts will have the same skill score of $-\infty$. Furthermore, a naive forecast that assigns uniform probability to all possible sites will always score higher than a forecast that misses only a single event but is otherwise superior. For this reason, likelihood tests are more subject to unconscious bias. Other methods of evaluating earthquake forecasts are suggested by Vere-Jones [20] and Holliday et al. [25].

Information Metrics One such alternative is the use of information metrics. Using methods from information theory [12], it is possible to calculate the entropy, H , of a forecast map. Entropy can be considered a measure of disorder (e.g. randomness) or “surprise”, hence maps with lower entropy contain more useful information than maps with higher entropy. We define entropy as

$$H(z) = - \sum_{i=1}^N p(\mathbf{x}_i; z) \log p(\mathbf{x}_i; z), \quad (9)$$

where

$$p(\mathbf{x}_i; z) = \begin{cases} P(\mathbf{x}_i) & P(\mathbf{x}_i) \geq z \\ 0 & P(\mathbf{x}_i) < z \end{cases}, \quad (10)$$

and the probabilities are scaled such that $\sum_{i=1}^N p(\mathbf{x}_i) = 1$. This definition allows a measurement of entropy as a function of some lower threshold z . A small, non-zero value for z allows for the measurement of the entropy above and relative to background noise.

Binary Forecasts

ROC Analysis The standard approach to the evaluation of a binary forecast is the use of a relative operating characteristic (ROC) diagram [39,60]. This method evaluates the performance of the forecast method relative to random guessing by constructing a plot of the fraction of failures to predict (events that occur where no event is forecast) against the fraction of false alarms (events that are forecast to occur at a location but do not occur) for an ensemble of forecasts. Molchan [41] has used a modification of this method to evaluate the success of intermediate term earthquake forecasts.

The binary approach has a long history, over 100 years, in the verification of tornado forecasts [39]. These forecasts take the form of a tornado forecast for a specific location and time interval, each forecast having a binary set of possible outcomes. For example, during a given time window of several hours duration, a forecast is issued in which a list of counties is given with a statement that one or more tornadoes will or will not occur. A 2×2 contingency table is then constructed, the top row contains the counties in which tornadoes are forecast to occur and the bottom row contains counties in which tornadoes are forecast to not occur. Similarly, the left column represents counties in which tornadoes were actually observed, and the right column represents counties in which no tornadoes were observed.

With respect to earthquakes, binary forecasts take exactly this form. A time window is proposed during which the forecast of large earthquakes having a magnitude above some minimum threshold is considered valid. An example might be a forecast of earthquakes larger than $M = 5$ during a period of five or ten years duration. A map of the seismically active region is then completely covered ("tiled") with boxes of two types: boxes in which the epicenters of at least one large earthquake are forecast to occur and boxes in which large earthquakes are forecast to not occur. In other types of forecasts, large earthquakes are given some continuous probability of occurrence from 0% to 100% in each box [28]. These forecasts can be converted to the binary type by the application of a threshold value. Boxes having a probability below the threshold are assigned a forecast rating of non-occurrence during the time window, while boxes having a probability above

Earthquake Forecasting and Verification, Table 2

Schematic contingency table for categorical forecasts of a binary event

Forecast	Observed		
	Yes	No	Total
Yes	a	b	$a + b$
No	c	d	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d = N$

the threshold are assigned a forecast rating of occurrence. A high threshold value may lead to many failures to predict, but few false alarms. The level at which the threshold is set is then a matter of public policy specified by emergency planners, representing a balance between the prevalence of failures to predict and false alarms.

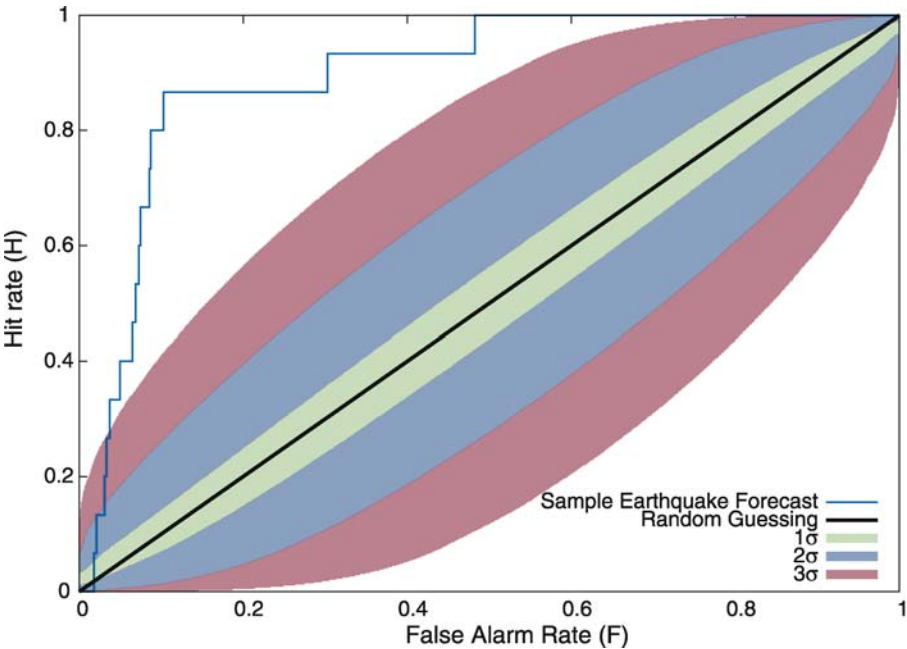
Contingency Tables An extensive review on forecast verification in the atmospheric sciences has been given by Jolliffe and Stephenson [27]. The wide variety of approaches that they consider are directly applicable to earthquake forecasts as well. Verification of earthquake forecasts proceeds in exactly the same way as for, say, tornado forecasts when using these approaches. For a given forecast, the contingency table (see Table 2) is constructed. Values for the table elements a (Forecast=yes, Observed=yes), b (Forecast=yes, Observed=no), c (Forecast=no, Observed=yes), and d (Forecast=no, Observed=no) are obtained from the forecast map. The fraction of alarm space, also called the probability of forecast of occurrence, is $r = (a + b)/N$, where the total number of boxes is $N = a + b + c + d$. The hit rate is $H = a/(a + c)$ and the false alarm rate is $F = b/(b + d)$. From these quantities a number of descriptive, performance, and skill measures can be constructed [27]. Table 3 lists a few possible measures.

ROC Curves The standard ROC diagram [23,27] is a plot of the points $\{H, F\}$ calculated for a binary forecast (see Fig. 3). If the forecast was converted from continuum map, H and F are plotted as the lower (conversion) threshold is varied. A perfect forecast of occurrence (perfect order, no fluctuations) would consist of two line segments, the first connecting the points $(H, F) = (0, 0)$ to $(H, F) = (1, 0)$, and the second connecting $(H, F) = (1, 0)$ to $(H, F) = (1, 1)$. A curve of this type can be described as maximum possible hits ($H = 1$) with minimum possible false alarms ($F = 0$). Another type of perfect forecast consists of two lines connecting the points $(0, 0)$ to $(0, 1)$ and $(0, 1)$ to $(1, 1)$, a perfect forecast of non-occurrence.

Earthquake Forecasting and Verification, Table 3

Table of various descriptive and performance measures that can be calculated from the binary contingency table for an earthquake forecast

Name	Definition	Definition (H , F , and τ)	Range
Fraction of alarm space τ	$\tau = (a + c)/N$	τ	$[0, 1]$
Hit rate H	$H = a/(a + c)$	H	$[0, 1]$
False alarm rate F	$F = b/(b + d)$	F	$[0, 1]$
False alarm ratio FAR	$FAR = b/(a + b)$	$FAR = (1 + \frac{\tau}{1-\tau} \frac{H}{F})^{-1}$	$[0, 1]$
Miss rate ν	$\nu = c/(a + c)$	$\nu = 1 - H$	$[0, 1]$
Peirce's skill score PSS	$PSS = \frac{ad-bc}{(b+d)(a+c)}$	$PSS = H - F$	$[-1, 1]$
Yule's Q	$Q = \frac{ad-bc}{ad+bc}$	$Q = \frac{H-F}{H(1-F)+F(1-H)}$	$[-1, 1]$
Peirce Area A	$A = \int PSS$	$A = \int HdF - \frac{1}{2}$	$[-\frac{1}{2}, \frac{1}{2}]$



Earthquake Forecasting and Verification, Figure 3

Sample relative operating characteristic (ROC) diagram. Shown is a plot of hit rates, H , as a function of false alarm rates, F , for a sample earthquake forecast (blue) and random guessing (black). Confidence intervals for the one-, two- and three- σ levels are shown as well [23,80]

The line $H = F$ occupies a special status, and corresponds to a completely random forecast [23,27] (maximum disorder, maximum fluctuations) where the false alarm rate is the same as the hit rate and no information is produced by the forecast. Points above this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose $H-F$ curves lies the highest is said to outperform the others. Often, however, competing forecasts will have intersecting curves. In this case, forecasts

are said outperform each other only in specific ranges and only for specific choices of the lower (conversion) threshold value.

ν - τ Curves An alternative diagram [41] is a plot of the points $\{\nu, \tau\}$ for a binary forecast map. In this case a perfect forecast of occurrence would consist of two line segments, the first connecting the points $(\nu, \tau) = (0, 0)$ to $(\nu, \tau) = (0, 1)$, and the second connecting $(\nu, \tau) = (0, 1)$ to $(\nu, \tau) = (1, 1)$. A curve of this type can be described as

minimum possible missed events ($\nu = 0$) with minimum possible alarm space ($\tau = 0$).

As with the H - F curve, the line $\nu = \tau$ corresponds to a completely random forecast. Points below this line are said to have performed better than simple random guessing. If competing forecasts are plotted on the same graph, the forecast whose ν - τ curves lies the lowest is said to outperform the others. As can be verified from Table 3, ν - τ curves offer the same information as H - F curves and are identical in the range $a \ll d$.

Future Directions

It is actually quite surprising that immediate local precursory phenomena are not seen. Prior to a volcanic eruption, increases in regional seismicity and surface movements are generally observed. For a fault system, the stress gradually increases until it reaches the frictional strength of the fault and a rupture is initiated. It is certainly reasonable to hypothesize that the stress increase would cause increases in background seismicity and aseismic slip. In order to test this hypothesis the Parkfield Earthquake Prediction Experiment was initiated in 1985. The expected Parkfield earthquake occurred beneath the heavily instrumented region on 28 September 2004. No local precursory changes were observed [3,36]. In the absence of local precursory signals, the next question is whether broader anomalies develop, and in particular whether there is anomalous seismic activity.

Assuming precursors do exist and can be exploited to create earthquake forecasts, testing and verification of the usefulness of the forecasts is the necessary next step. A forecast method that predicts all earthquakes but carries a high false alarm rate is likely to be useless as a public warning tool. Similarly, a forecast method that issues warnings for only a small fraction of actual earthquakes but never issues false alarms is likely to be a poor tool for catastrophe preparation. Forecast verification techniques can be used to find the middle ground that is most useful.

As a final warning, researchers must be careful not to create artificial skill in their forecasts. Since nature provides us only with one earthquake record (the actual history of a given test region), the quality of a new forecast system is often assessed on the same data set used to create it. This can potentially lead to an optimistic bias in any skill scores. This is a particular problem if the score itself is used to calibrate the method, either directly or indirectly. Development of realistic earthquake simulators and cross-testing against test regions with similar fault structures can help protect against this.

Bibliography

Primary Literature

1. Bakun WH, Lindh AG (1985) The Parkfield, California, earthquake prediction experiment. *Science* 229:619–624
2. Bevington PR, Robinson DK (1992) *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill, New York
3. Borchardt RD, Johnston MJS, Glassmoyer G, Dietel C (2006) Recordings of the 2004 parkfield earthquake on the general earthquake observation system array: Implications for earthquake precursors, fault rupture, and coseismic strain changes. *Bull Seismol Soc Am* 96(4b):73–89
4. Bowman DD, King GCP (2001) Accelerating seismicity and stress accumulation before large earthquakes. *Geophys Res Lett* 28:4039–4042
5. Bowman DD, Ouillon G, Sammis CG, Sornette A, Sornette D (1998) An observational test of the critical earthquake concept. *J Geophys Res* 103:24359–24372
6. Bowman DD, Sammis CG (2004) Intermittent criticality and the Gutenberg–Richter distribution. *Pure Appl Geophys* 161:1945–1956
7. Brehm DJ, Braille LW (1998) Intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone. *Bull Seismol Soc Am* 88:564–580
8. Brehm DJ, Braille LW (1999) Intermediate-term earthquake prediction using the modified time-to-failure method in southern California. *Bull Seismol Soc Am* 89:275–293
9. Buffe CG, Nishenko SP, Varnes DJ (1994) Seismicity trends and potential for large earthquakes in the Alaska–Aleutian region. *Pure Appl Geophys* 142:83–99
10. Buffe CG, Varnes DJ (1993) Predictive modeling of the seismic cycle of the greater San Francisco Bay region. *J Geophys Res* 98:9871–9883
11. Chen D, Cane MA, Kaplan A, Zebian SE, Huang D (2004) Predictability of El Niño in the past 148 years. *Nature* 428:733–736
12. Cover TM, Thomas JA (1991) *Elements of Information Theory*. Wiley-Interscience, New York
13. Ellsworth WL, Mathews MV, Nadeau RM, Nishenko SP, Reasenberg PA, Simpson RW (1999) A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities. Open-File Report 99-522, US Geological Survey
14. Frankel AF (1995) Mapping seismic hazard in the central and eastern United States. *Seismol Res Lett* 60:8–21
15. Frankel AF, Mueller C, Barnhard T, Perkins D, Leyendecker EV, Dickman N, Hanson S, Hopper M (1996) National seismic hazard maps. Open-File Report 96-532, US Geological Survey
16. Geller RJ (1997) Earthquake prediction: A critical review. *Geophys J Int* 131:425–450
17. Geller RJ, Jackson DD, Kagen YY, Mulargia F (1997) Earthquakes cannot be predicted. *Science* 275:1616–1617
18. Goes SDB, Ward SN (1994) Synthetic seismicity for the San Andreas fault. *Annali Geofis* 37:1495–1513
19. Gross S, Rundle JB (1998) A systematic test of time-to-failure analysis. *Geophys J Int* 133:57–64
20. Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. *Pure Appl Geophys* 162:1229–1253. doi:10.1007/s00024-004-2667-2
21. Helmstetter A Is earthquake triggering driven by small earthquakes? *Phys Rev Lett* 91:0585014

22. Holliday JR, Chen CC, Tiampo KF, Rundle JB, Turcotte DL, Donnellan A (2007) A RELM earthquake forecast based on pattern informatics. *Seis Res Lett* 78(1):87–93
23. Holliday JR, Nanjo KZ, Tiampo KF, Rundle JB, Turcotte DL (2005) Earthquake forecasting and its verification. *Nonlinear Process Geophys* 12:965–977
24. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Modification of the pattern informatics method for forecasting large earthquake events using complex eigenvectors. *Tectonophysics* 413:87–91. doi:10.1016/j.tecto.2005.10.008
25. Holliday JR, Rundle JB, Tiampo KF, Klein W, Donnellan A (2006) Systematic procedural and sensitivity analysis of the pattern informatics method for forecasting large ($M \geq 5$) earthquake events in southern California. *Pure Appl Geophys* 163:2433–2454. doi:10.1007/s00024-006-0131-1
26. Jaumé SC, Sykes LR (1999) Evolving towards a critical point: A review of accelerating seismic moment/energy release prior to large and great earthquakes. *Pure Appl Geophys* 155:279–306
27. Jolliffe IT, Stephenson DB (2003) *Forecast Verification*. Wiley, Chichester
28. Kagan YY, Jackson DD (2000) Probabilistic forecasting of earthquakes. *Geophys J Int* 143:438–453
29. Kagan YY, Knopoff L (1981) Stochastic synthesis of earthquake catalogs. *J Geophys Res* 86(4):2853–2862
30. Kanamori H (2003) Earthquake prediction: An overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake & Engineering Seismology*. Academic Press, Amsterdam, pp 1205–1216
31. Keilis-Borok V (2002) Earthquake predictions: State-of-the-art and emerging possibilities. *Ann Rev Earth Planet Sci* 30:1–33
32. Keilis-Borok V, Shebalin P, Gabrielov A, Turcotte D (2004) Reverse tracing of short-term earthquake precursors. *Phys Earth Planet Int* 145:75–85
33. Keilis-Borok VI (1990) The lithosphere of the earth as a nonlinear system with implications for earthquake prediction. *Rev Geophys* 28:19–34
34. King GCP, Bowman DD (2003) The evolution of regional seismicity between large earthquakes. *J Geophys Res* 108:2096
35. Kossobokov VG, Keilis-Borok VI, Turcotte DL, Malamud BD (2000) Implications of a statistical physics approach for earthquake hazard assessment and forecasting. *Pure Appl Geophys* 157:2323–2349
36. Lindh AG (2005) Success and failure at Parkfield. *Seis Res Lett* 76:3–6
37. Lomnitz C (1994) *Fundamentals of Earthquake Prediction*. Wiley, New York
38. Main IG (1999) Applicability of time-to-failure analysis to accelerated strain before earthquakes and volcanic eruptions. *Geophys J Int* 139:F1–F6
39. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) *Forecast Verification*. Wiley, Chichester, pp 37–76
40. Mogi K (1985) *Earthquake Prediction*. Academic Press, Tokyo
41. Molchan GM (1997) Earthquake predictions as a decision-making problem. *Pure Appl Geophys* 149:233–247
42. Ogata Y (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *J Am Stat Assoc* 83:9–27
43. Ogata Y (1999) Seismicity analysis through point-process modeling: a review. *Pure Appl Geophys* 155:471–507
44. Ogata Y, Zhuang J (2006) Space-time ETAS models and an improved extension. *Tectonophysics* 413:13–23. doi:10.1016/j.tecto.2005.10.016
45. Rikitake T (1982) *Earthquake forecasting and warning*. D. Reidel Publishing Co, Dordrecht
46. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. *Geophys J Int* 140:568–576
47. Robinson R, Benites R (1995) Synthetic seismicity models of multiple interacting faults. *J Geophys Res* 100:18229–18238
48. Robinson R, Benites R (1996) Synthetic seismicity models for the Wellington Region, New Zealand: implications for the temporal distribution of large events. *J Geophys Res* 101:27833–27844
49. Rundle JB, Rundle PB, Donnellan A (2005) A simulation-based approach to forecasting the next great San Francisco earthquake. *Proc Natl Acad Sci* 102(43):15363–15367
50. Rundle JB, Rundle PB, Donnellan A, Fox G (2004) Gutenberg–Richter statistics in topologically realistic system-level earthquake stress-evolution simulations. *Earth Planets Space* 55(8):761–771
51. Rundle JB, Tiampo KF, Klein W, Martins JSS (2002) Self-organization in leaky threshold systems: The influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting. *Proc Natl Acad Sci USA* 99(Suppl 1):2514–2521
52. Rundle JB, Turcotte DL, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41(4):1019. doi:10.1029/2003RG000135
53. Rundle PB, Rundle JB, Tiampo KF, Donnellan A, Turcotte DL (2006) Virtual California: fault model, frictional parameters, applications. *Pure Appl Geophys* 163:1819–1846
54. Sammis CG, Bowman DD (2006) Competing models for accelerating moment release before large earthquakes. 5th Annual ACES International Workshop, Maui, Hawaii, USA
55. Sammis CG, Bowman DD, King G (2004) Anomalous seismicity and accelerating moment release preceding the 2001–2002 earthquakes in northern Baha California, Mexico. *Pure Appl Geophys* 161:2369–2378
56. Scholz CH (2002) *The Mechanics of Earthquakes & Faulting*, 2nd edn. Cambridge University Press, Cambridge
57. Schorlemmer D, Jackson DD, Gerstenberger M (2003) Earthquake likelihood model testing. <http://moho.ess.ucla.edu/~kagan/sjg.pdf>. Accessed 8 Oct 2004
58. Shebalin P, Keilis-Borok V, Zaliapin I, Uyeda S, Nagao T, Tsybin N (2004) Advance short-term prediction of the large Tokachi-oki earthquake, September 25, $M = 8.1$: A case history. *Earth Planets Space* 56:715–724
59. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
60. Swets JA (1973) The relative operating characteristic in psychology. *Science* 182:990–1000
61. Sykes LR, Jaumé SC (1990) Seismic activity on neighboring faults as a long-term precursor to large earthquakes in the San Francisco Bay area. *Nature* 348:595–599
62. Sykes LR, Shaw BE, Scholz CH (1999) Rethinking earthquake prediction. *Pure Appl Geophys* 155:207–232
63. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Eigenpatterns in southern California seismicity. *J Geophys Res* 107(B12):2354. doi:10.1029/2001JB000562

64. Tiampo KF, Rundle JB, McGinnis S, Gross SJ, Klein W (2002) Pattern dynamics and forecast methods in seismically active regions. *Pure Appl Geophys* 159:2429–2467
65. Turcotte DL (1991) Earthquake prediction. *Ann Rev Earth Planet Sci* 19:263–281
66. Turcotte DL (1997) *Fractals & Chaos in Geology & Geophysics*, 2nd edn. Cambridge University Press, Cambridge
67. Utsu T (1984) Estimation of parameters for recurrence models of earthquakes. *Earthquake Res Insti-Univ Tokyo* 59:53–66
68. Utsu T (2003) A list of deadly earthquakes in the world: 1500–2000. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake & Engineering Seismology*. Academic Press, Amsterdam, pp 691–717
69. Ward SN (1992) An application of synthetic seismicity in earthquake statistics: the Middle America trench. *J Geophys Res* 97(B5):6675–6682
70. Ward SN (1996) A synthetic seismicity model for southern California: cycles, probabilities, and hazard. *J Geophys Res* 101(B10):22393–22418
71. Ward SN (2000) San Francisco Bay Area earthquake simulations: a step toward a standard physical earthquake model. *Bull Seismo Soc Am* 90(2):370–386
72. Working Group on California Earthquake Probabilities (1988) Probabilities of large earthquakes occurring in California on the San Andreas fault. Open-File Report 88-398, US Geological Survey
73. Working Group on California Earthquake Probabilities (1990) Probabilities of large earthquakes in the San Francisco Bay region, California. Circular 1053, US Geological Survey
74. Working Group on California Earthquake Probabilities (1995) Seismic hazards in southern California: probable earthquakes, 1994–2024. *Seis Soc Am Bull* 85:379–439
75. Working Group on California Earthquake Probabilities (2003) Earthquake probabilities in the San Francisco Bay Region, 2002–2031. Open-File Report 2003-214, US Geological Survey
76. Wyss M (1997) Nomination of precursory seismic quiescence as a significant precursor. *Pure Appl Geophys* 149:79–114
77. Wyss M, Habermann RE (1988) Precursory seismic quiescence. *Pure Appl Geophys* 126:319–332
78. Yakovlev G, Turcotte DL, Rundle JB, Rundle PB (2006) Simulation-based distributions of earthquake recurrence times on the San Andreas fault system. *Bull Seis Soc Am* 96:1995–2007
79. Yang W, Vere-Jones D, Li M (2001) A proposed method for locating the critical region of a future earthquake using the critical earthquake concept. *J Geophys Res* 106:4151–4128
80. Zechar JD, Jordan TH (2005) Evaluation techniques for alarm-based forecasts. *EOS Trans. AGU. Fall meeting*
81. Zoback ML (1992) First- and second-order patterns of stress in the lithosphere: The world stress map project. *J Geophys Res* 97:11703–11728

Books and Reviews

- Jolliffe IT, Stephenson DB (2003) *Forecast Verification*. Wiley, Chichester
- Turcotte DL, Schubert G (2002) *Geodynamics*. Cambridge University Press, Cambridge
- Wilks DS (1995) *Statistical Methods in the Atmospheric Sciences*. Academic Press, San Diego

Earthquake Location, Direct, Global-Search Methods

ANTHONY LOMAX¹, ALBERTO MICHELINI²,
ANDREW CURTIS³

¹ ALomax Scientific, Mouans-Sartoux, France

² Istituto Nazionale di Geofisica e Vulcanologia, Roma, Italy

³ ECOSSE (Edinburgh Collaborative of Subsurface Science and Engineering), Grant Institute of GeoSciences, The University of Edinburgh, Edinburgh, United Kingdom

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Earthquake Location Problem](#)

[Location Methods](#)

[Illustrative Examples](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Arrival time The time of the first measurable energy of a seismic phase on a seismogram.

Centroid The coordinates of the spatial or temporal average of some characteristic of an earthquake, such as surface shaking intensity or moment release.

Data space If the data are described by a vector **d**, then the data space **D** is the set of all possible values of **d**.

Direct search A search or inversion technique that does not explicitly use derivatives.

Earthquake early-warning The goal of earthquake early-warning is to estimate the shaking hazard of a large earthquake at a nearby population center or other critical site before destructive S and surface waves have reached the site. This requires that useful, probabilistic constraint on the location and size of an earthquake is obtained very rapidly.

Earthquake location An earthquake location specifies a spatial position and time of occurrence for an earthquake. The location may refer to the earthquake hypocenter and corresponding origin time, a mean or centroid of some spatial or temporal characteristic of the earthquake, or another property of the earthquake that can be spatially and temporally localized. This term also refers to the process of locating an earthquake.

Epicenter The point on the Earth's surface directly above a hypocenter.

Error A specified variation in the value assumed by a variable. See also *uncertainty*.

Global search A search or inversion that samples throughout the prior *pdf* of the unknown parameters.

Hypocenter The point in three-dimensional space of initial energy release of an earthquake rupture or other seismic event.

Importance sampling A sampling procedure that draws samples following the posterior *pdf* of an inverse, optimization or other search problem. Since these problems involve initially unknown, posterior *pdf* functions, importance sampling can only be performed approximately, usually through some adaptive or learning procedure as sampling progresses.

Inverse problem, inversion The problem of determining the parameters of a physical system given some data. The solution of an inverse problem requires measurements of observable quantities of the physical system, and the mathematical expression (the forward problem) that relates the parameters defining the physical system (model space) to the data (data space). In inverse problems, estimates of the unknown parameters in the model space and of their uncertainties are sought from the combination of the available information on the model parameters (prior *pdf*), the data and the forward problem.

Likelihood function A non-normalized *pdf*.

Misfit function A function that quantifies the disagreement between observed and calculated values of one or more quantities. See *objective function*.

Model space If the model parameters are described by a vector \mathbf{m} , then model space \mathcal{M} is the set of all possible values of \mathbf{m} .

Objective function A function expressing the quality of any point in the model space. Inversion and optimization procedures use an objective function to rank and select models. Usually objective functions are defined in terms of misfit functions, and for probabilistic inversion the objective function must be a *pdf* or likelihood function.

Origin time The time of occurrence of initial energy release of an earthquake rupture or other seismic event.

Prior pdf A *pdf* that expresses the information on the unknown parameters available before an inverse problem is solved. For an earthquake location, the prior *pdf* is often a simple function (e.g., boxcar) of three spatial dimensions and time. See also *Inverse problem*.

Probability density function – pdf A function in one or more dimensional space \mathbf{X} that (i) when integrated

over some interval $\Delta\mathbf{x}$ in \mathbf{X} gives a probability of occurrence of any event within $\Delta\mathbf{x}$, and (ii) has unit integral over space \mathbf{X} , where \mathbf{X} represents a space of possible events. An earthquake location *pdf* is often a 3-dimensional probability density function over all possible spatial locations or a 4-dimensional probability density function over all possible spatial locations and times of occurrence.

Posterior pdf A *pdf* that expresses the information about the unknown parameters available after inversion. The posterior *pdf* for an earthquake location is often a function of the three spatial dimensions and the origin time of the hypocenter parameters; this function may be complicated. See also *Inverse problem*.

Ray path A local minimum-time path between a source and receiver of idealized, infinite frequency wave energy of a specified wave type (e.g., P or S).

Receiver or station Synonyms for an observation point where ground motion is detected and a seismogram recorded.

Seismic phase A distinct packet of energy from a seismic source. Usually refers to a specified wave type (e.g. P or S) satisfying a particular physics of wave propagation.

Seismicity The distribution in space and time of seismic event locations.

Seismogram An analogue or digital recording of the ground motion at a point (receiver or station) in the Earth. Also called a waveform.

Source A general term referring to an earthquake, explosion or other release of seismic energy as a physical phenomenon localized in space and time.

Station See *receiver*.

Travel time The time that a signal, e.g. elastic wave energy of a seismic phase, takes to propagate along a ray path between two points in a medium.

Uncertainty Random variation in the values assumed by a variable. See also *error*.

Definition of the Subject

An earthquake location specifies the place and time of occurrence of energy release from a seismic event. A location together with a measure of size forms a concise description of the most important characteristics of an earthquake. The location may refer to the earthquake's epicenter, hypocenter, or centroid, or to another observed or calculated property of the earthquake that can be spatially and temporally localized. A location is called *absolute* if it is determined or specified within a fixed, geographic coordinate system and a fixed time base (e.g., Coordinated

Universal Time, UTC); a location is called *relative* if it is determined or specified with respect to another spatio-temporal object (e. g., an earthquake or explosion) which may have unknown or uncertain absolute location.

For rapid hazard assessment and emergency response, an earthquake location provides information such as the locality of potential damage or the source region of a possible tsunami, and a location is required to calculate most measures of the size of an earthquake, such as magnitude or moment. Locations are required for further analysis and characterization of the event, for studies of general patterns of seismicity, to calculate distributions of stress and strain changes around the earthquake, for assessing future earthquake hazard, and for basic and applied seismological research.

Since earthquakes occur deep in the Earth, their source locations must be inferred indirectly from distant observations, and earthquake location is thus a remote-sensing problem. Most commonly an earthquake location is determined by the match or misfit between observed arrival times of seismic wave-energy at seismic stations, and predictions of these arrival times for different source locations using a given elastic-wave speed model; this is an inverse problem. Essentially, many potential locations (place and time) are examined and those for which some measure of misfit between predicted and measured arrival times is smallest are retained as best estimates of the true location.

Many numerical location methods involve linearization of the equations relating the predicted arrival times to the location through Taylor expansion involving partial derivatives; these are called *linearized* methods. Methods that do not involve linearization are called *nonlinearized* or *direct-search* methods. The term *nonlinear* is used ambiguously in geophysics to refer to linearized-iterated and to nonlinearized methods. In this chapter we focus on nonlinearized, direct-search methods, and to avoid ambiguity we identify them with the term *direct-search*.

Direct-search location can be performed through graphical analysis, regular or stochastic searches over a space of possible locations, and other algorithms. Direct-search earthquake location is important because, relative to linearized methods, it is easy to apply with realistic earth models which may have abrupt and complicated velocity variations in three-dimensions, it places little restriction on the form of the measure of misfit, it is stable (i. e., does not suffer numerical convergence problems) when the observations are insufficient to fully constrain the spatial location or origin time, and it can produce comprehensive, probabilistic solutions which indicate the full location uncertainty, often a complex function of space and time.

Conversely, the primary advantage of linearized location methods is that they are much less demanding computationally than direct-search methods.

Introduction

Most commonly, an earthquake location is determined using observed seismic-phase arrival-times and associated uncertainties, and predicted travel times in a given wave-speed model. Ideally, the location procedure will determine a 4-dimensional, posterior probability density function, or location *pdf*, over all possible solutions (spatial locations and origin times). This location *pdf* quantifies the agreement between predicted and observed arrival times in relation to all uncertainties, and forms a complete, probabilistic solution. In practice, however, an earthquake location is often specified as some optimal solution (a point in space and time) with associated uncertainties.

The earliest, formal earthquake locations using seismic-phase arrival-time observations employed direct-search procedures such as graphical methods (e. g., [37]) or simple grid searches (e. g., [52]). The advent of digital computers in the 1960's lead to the use of iterated linearized approaches based mainly on Geiger's method [17]. Since the 1980's, the increasing power of digital computers has made large-scale, grid and stochastic direct searches practical for routine earthquake location. Direct-search methods are now used routinely in research and earthquake monitoring (e. g., [22,23,31,32,46,48,61]).

In principle, direct-search methods can be applied to locate the relative positions of ensembles of events, and for joint epicentral determination (e. g., [47]) to simultaneously determine multiple earthquake locations and station corrections related to errors in the velocity model. However, the high-dimensionality of such problems makes direct-search solution difficult and computationally demanding; at the present time these problems are usually performed through large scale, linearized procedures. For these reasons, we mainly consider here absolute location of individual events.

In this article we describe the earthquake location problem and direct-search methods used to perform this location, and we present a number of examples of direct-search location. We do not compare different direct-search location methods or compare direct-search to linearized algorithms, but instead focus on illustrating important features and complexity in earthquake location results. For this reason we emphasize direct, global-search, probabilistic location, which produces general and complete solutions that best illuminate these features and complexity.

The Earthquake Location Problem

An Inherently Nonlinear Problem

In a homogeneous medium with wave speed v and *slowness* defined to be $u = 1/v$, the arrival time, t_{obs} , at an observation point $x_{\text{obs}}, y_{\text{obs}}, z_{\text{obs}}$ of a signal emitted at origin time t_0 from a source location at x_0, y_0, z_0 is,

$$t_{\text{obs}} = t_0 + u \left[(x_{\text{obs}} - x_0)^2 + (y_{\text{obs}} - y_0)^2 + (z_{\text{obs}} - z_0)^2 \right]^{1/2}, \quad (1)$$

This expression shows that a change in the spatial position of the source introduces a nonlinear change in t_{obs} , even in the simplest possible medium. When the speed v and hence slowness u are inhomogeneous in space, the arrival time at the observation point becomes,

$$t_{\text{obs}} = t_0 + \int_{\mathbf{r}_0(s)} u(\mathbf{r}_0) ds, \quad (2)$$

where $\mathbf{r}_0(s)$ denotes a point at distance s along ray path \mathbf{r}_0 between source and receiver locations. Equation (1) is a special case of (2) that has straight source-receiver ray paths. Equation (2) is nonlinear since a change in the source location changes the ray path over which the integral is calculated. Thus, earthquake location, which maps arrival times into spatial location and origin time, is inherently a nonlinear problem.

The Observed Data

Data used to constrain earthquake locations are usually derived from seismograms recorded at seismic stations distributed around the earthquake source area, usually at or near the surface of the Earth. The derived data for earthquake location include arrival times, polarization angles, or array slownesses and azimuths. For earthquake location there are three important aspects of this data determination:

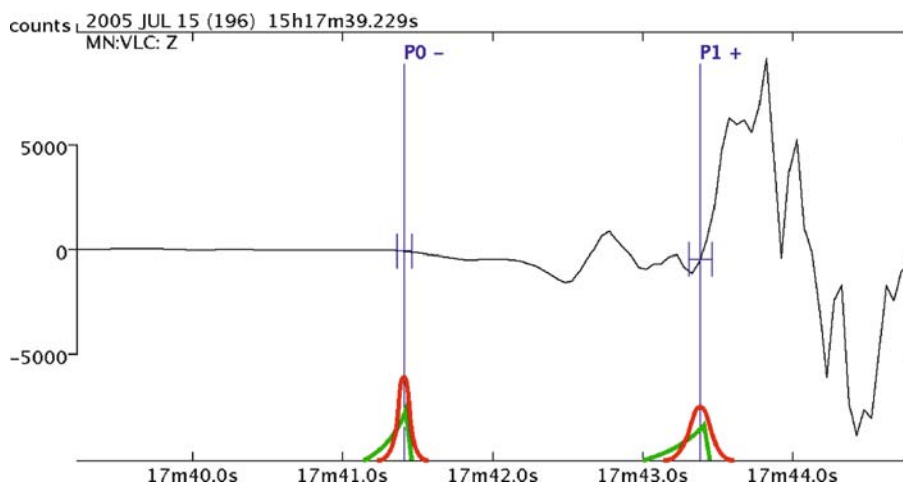
- 1) choosing locations for the stations (before data have been collected),
- 2) deriving data and associated uncertainties from the seismograms, and
- 3) association of the derived data into subsets of data corresponding to unique events.

The first important aspect of data determination is choosing station locations with the goal of constraining as tightly as possible event locations for a given source area; this is classified as a problem of “experimental design” in the field of statistics. The design problem must be resolved prior to

data collection and so is posed in terms of *expected* data, and *expected* location results. We describe experimental design techniques in more detail later, after introducing and discussing the location solution on which such designs depend.

Once stations are installed and have recorded seismograms from earthquakes of interest, a data set must be extracted that is sensitive to the event source location, and which we can associate with some physics (e.g., of P or S waves) and paths of wave propagation. Most commonly for earthquake location the data set will be phase arrival times and associated uncertainties picked manually or automatically from seismograms (Fig. 1). It is often easy to detect and pick arrivals manually since the human eye can identify a change in amplitude or frequency in the signal even in the presence of significant noise. The picking of the S phase is sometimes more difficult because it arrives in the P coda and can be preceded by S to P or other converted phases; this is a common problem with recordings at local (e.g., up to about 100 km) and near-regional (e.g., up to about 300 km) distances, especially if horizontal component seismograms are not available. The automatic detection, identification and picking of P and S arrivals is much more difficult, especially in the presence of high noise levels. However, automatic detection and picking is faster and, for the case of initial P phases or other phases with characteristic forms, can produce a more consistent data set than manual processing. Automatic arrival detection relies on identifying temporal variations in energy, frequency content, polarization or other characteristics of the signal which are anomalous relative to their background or noise level. Often the detection and picking algorithms are applied to filtered and processed time-series in order either to reduce noise, or to augment the signal in pre-set or dynamically-determined frequency bands or polarization directions. See [18] for an approach that exploits neural networks for phase identification and picking, and [82] for a review and systematic comparison of several approaches to automatic detection and picking.

The data used for earthquake location (e.g., arrival times) must have associated uncertainty estimates otherwise the location uncertainty and a probabilistic solution (i.e., location *pdf*) can not be calculated. Most generally, a vector \mathbf{d} that describes the data takes values from a data space \mathbf{D} , and $p(\mathbf{d})$ denotes the *pdf* representing uncertainty in \mathbf{d} . The uncertainty in arrival time data should include not only an estimate of the uncertainty in the picked phase arrival time, but also uncertainty in which phase (e.g., P or S) is associated with the pick. If there are multiple expected phase arrivals close to the picked arrival time of a phase, then ideally these should all be taken as can-



Earthquake Location, Direct, Global-Search Methods, Figure 1

A short waveform segment (~ 5 sec) showing the first P wave arrivals from a small earthquake in Northern Italy recorded on a vertical component seismogram at a nearby station. Automatic arrival pick times (vertical blue lines) and uncertainty estimates (blue error bars) are shown for two phases, a first arriving P phase (P0) and secondary P arrival (P1). The red curves show the data pdf functions representing these arrival pick times and uncertainties for an event location procedure where the data $P(d)$ is approximated by a normal distribution. The green curves show irregular, asymmetric pdf functions that may more accurately represent the uncertainty in the phase arrival times; if such pdf functions were routinely estimated during arrival picking, they could be used for direct-search location without major difficulty

didate phase types for the arrival. Also, the pick uncertainty of each phase may be best described by a pdf that is asymmetric in time, since usually a latest-possible time for a pick is much easier to define than the earliest time (Fig. 1). True data uncertainty pdf 's are therefore generally multi-modal, and can be quite complex to calculate and parametrize. In practice, an enumerated quality indication or, at best, a simple normal distribution (Gaussian uncertainty) is used to describe the picking error, and the phase association is usually fixed (e.g. to P or S) so corresponding uncertainties are ignored. In many cases these simplified data uncertainty estimates will lead to bias or increased error in the resulting event locations.

The third important aspect of data determination is the association of the derived data into sets of data for unique events. For example, this association may entail the assignment of each observed arrival time within a specified time window to a unique event, forming the minimum possible number of events and corresponding arrival time sets required to explain observed data. This association procedure can be very difficult, especially with automatic systems and when there are signals from multiple seismic events that are close or overlapping in time (e.g., [24]), and we do not address this issue further here. In the following, except for an examination of outlier data, we implicitly assume that location is performed with a data set that is already associated to a unique event.

The Velocity or Slowness Model

The velocity or slowness model specifies seismic wave-speeds in the region of the Earth containing the sources, the receivers and the ray paths between the sources and receivers. Equation (2) is nonlinear with respect to source location, but also with respect to slowness since a change in the slowness distribution of the medium changes the ray path. The velocity structure is sometimes estimated through coupled, simultaneous inversions for velocity structure and event locations (commonly called seismic tomography), but these are very large inverse problems solved mainly with linearized methods. Usually for earthquake location the velocity model is taken as known and fixed.

Often, for computational convenience or due to lack of information, the velocity model is parametrized with velocity varying only with depth. This is commonly called a laterally homogeneous or 1-dimensional (1D) model. Such a model may consist of one or more layers of constant or vertical-gradient wave-speeds. For work at a local or near-regional scale the layers may be horizontal and flat; for larger, regional or global scale problems the layers should be spherically symmetric shells to represent the curvature of the Earth. When more information on the velocity structure is available, a 3D model may be used in order to increase the accuracy of the ray paths and

travel times, and hence of the locations, relative to a 1D model. All models, whether 1D or 3D, are described by a limited number of parameters and include some form of spatial averaging or interpolation with respect to the true Earth. Although 3D models can potentially represent velocity variations in the Earth more accurately than 1D models, in practice the velocities in 3D models can locally be poorly constrained and have large errors. It is therefore often important to consider several different possible 1D and 3D velocity structures in a location study, either to test the sensitivity of the locations to errors in velocity, or to better estimate the travel-time uncertainties and produce a more meaningful location *pdf*. In principle the use of diverse velocity models poses no difficulties with direct-search location methods.

The Travel-Time Calculation

The theoretical seismic wave travel-times through a given velocity model between any particular source and receiver locations are required by most location algorithms. The calculation of travel times is commonly referred to as forward modeling, because inverse theory need not be invoked. There are three basic classes of methods to calculate the travel times: full-waveform methods, ray methods, and Huygens wavefront or eikonal methods.

Full-waveform methods (e.g., [1]) produce complete synthetic seismograms from which predicted travel times can be extracted. These methods include frequency-wavenumber or modal-summation techniques which are valid for a broad range of frequencies and can produce exact waveforms, but which are only applicable for relatively simple velocity structures. Numerical techniques such as finite elements and finite differences can accurately model full wave phenomena in complicated structures, but these methods typically require large computing resources and computing time. Currently, full-waveform methods are rarely used to determine predicted travel times for earthquake location because these times can be obtained directly and more efficiently with ray and eikonal methods.

Ray methods (e.g., [1,7,75]) provide travel times and the path, or ray, traveled by high-frequency waves, and can be applied to complicated and 3D velocity structures. With simple model parametrizations such as flat layers with constant or gradient velocity, ray paths and travel times can be determined very rapidly with analytical or semi-numerical algorithms. For these and more complicated models, shooting, or ray tracing techniques generate rays by iteratively solving of a set of ray-tracing equations starting in a specified direction at the source or receiver location. The ray that passes through a specified end point is

found by a search over the direction at the starting point; this search can be time consuming or unstable. In addition, shooting methods do not produce diffracted arrivals (e.g., “head waves” from the Mohorovičić discontinuity) which are often the first arriving signal at near-regional distances and are thus critical for earthquake location. Two-point, ray bending and perturbation techniques rely on Fermat’s principle of least time: an initial guess at the ray between two points is perturbed repeatedly to attain a minimum travel time and corresponding ray between the points. These techniques perform best with smooth models, but do produce diffracted arrivals. In general, except for analytical or semi-numerical algorithms in simple models, ray methods are too computationally expensive for direct-search location, which usually requires evaluation of travel times between a very large number of source and receiver positions. However, some ray bending methods (e.g., [39,77]) are efficient enough to be used in direct-search location when a relatively small number of source and receiver positions need to be examined.

Wavefront, eikonal and graph-based methods [75] provide travel-times of the first arriving, high frequency waves including diffracted arrivals, and are efficient and applicable with complicated, 3D velocity structures. In effect, these methods propagate wavefronts through a velocity model with repeated application of Huygen’s principle, by considering a large number of virtual sources (*Huygens sources*) along each wavefront. At time t these sources emit circular wavelets which expand for a small time Δt through the (constant) local, medium velocity. The locus of the first arriving circular wavelets defines the new wavefront location at time $t + \Delta t$. The synthetic travel time of the first-arriving energy at the receiver is the time at which a wavefront first touches the receiver. In practice, this problem is solved on a computer either by replicating this “wavefront marching” process (e.g., [50,66]), or by finding a numerical solution to the eikonal equation (e.g., [44,79]), or by graphical analysis (e.g., [38]). Though wavefront, eikonal and graph-based methods produce directly only the travel time of the first-arriving signal, information about the path traveled by the signal can be derived numerically from the travel-time field or from ray-tracing, and travel-times of secondary arrivals can be obtained through multi-stage calculations (e.g., [44,51]).

Wavefront, eikonal and graph-based methods can efficiently generate the travel-times from one point in a gridded velocity model to all other points in the model. This makes these methods particularly useful for direct-search location, which may test a large number of possible source positions widely distributed in space. For this purpose, the travel times from each seismic station to all points in the

model can be pre-calculated and stored in computer disk files or in memory; obtaining the travel time from a station to any other point then reduces to a simple lookup (e. g., [35,38]).

A Complete Solution – Probabilistic Location

Consider a vector \mathbf{d}_{obs} of observed data (e. g., arrival times) that takes values in a data space \mathbf{D} , and let $p(\mathbf{d})$ be the *pdf* over \mathbf{D} describing the data uncertainty in \mathbf{d}_{obs} due to measurement and processing uncertainties. Similarly, let \mathbf{m} denote the vector of source location parameters (spatial coordinates and origin time) which take values from parameter space \mathbf{M} . Let $p(\mathbf{m})$ be the prior *pdf* representing all information available about the location before (prior to) using the data \mathbf{d}_{obs} ; $p(\mathbf{m})$ might include knowledge of the known, active fault zones in the area, or might specify the bounds of a region within which we know the event occurred from damage reports, or of a region containing the network of stations that recorded the event. Also consider the forward problem (e. g., travel time calculation) relating \mathbf{m} to a vector of predicted data (e. g., arrival times), \mathbf{d}_{calc} . In general the forward problem may also be uncertain, for example due to uncertainties in velocity structure, so we use $F(\mathbf{d}, \mathbf{m})$ to denote the *pdf* of the relationship between \mathbf{d}_{calc} and \mathbf{m} as constrained by the forward problem.

As an example of F , it is commonly assumed that for each particular \mathbf{m} , the corresponding predicted data \mathbf{d}_{calc} are given by a function $\mathbf{f}(\mathbf{m})$ with negligible errors. Then the conditional *pdf* $F(\mathbf{d}|\mathbf{m})$ (the probability distribution of \mathbf{d} when \mathbf{m} is fixed at a particular value) is described by $F(\mathbf{d}|\mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})]$ where δ is the Dirac delta-function. Also, the forward problem is often assumed to place minimum possible constraint on parameters \mathbf{m} ; the *pdf* describing this state of information about \mathbf{m} is called the homogeneous distribution, represented by $\mu(\mathbf{m})$. No *pdf* exists that describes zero information, but *some* information about \mathbf{m} always exists in practice (the positivity of parameter values, for example); $\mu(\mathbf{m})$ describes that minimum state of information. In that case the forward problem is given by [73,74],

$$F(\mathbf{d}, \mathbf{m}) = \delta[\mathbf{d} - \mathbf{f}(\mathbf{m})] \mu(\mathbf{m}) . \quad (3)$$

A solution to the earthquake location problem is found by combining the information in the observed data, $p(\mathbf{d})$, the prior *pdf*, $p(\mathbf{m})$, and the ability of the forward problem to predict the observed data, $F(\mathbf{d}, \mathbf{m})$ [73,74]. This is achieved in a probabilistic framework by constructing a *pdf* Q describing the state of posterior (post-experimental) infor-

mation by:

$$Q(\mathbf{d}, \mathbf{m}) = k \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})p(\mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} , \quad (4)$$

where the constant k normalizes Q to unit integral over $\mathbf{D} \times \mathbf{M}$ and $\mu(\mathbf{d}, \mathbf{m})$ is the homogeneous distribution over data \mathbf{d} and parameters \mathbf{m} . Equation (4) contains all information (from the prior knowledge, data and physics) that could have a bearing on location \mathbf{m} , and defines a joint *pdf* between parameters \mathbf{m} and data \mathbf{d} . The final, posterior state of information about location parameters \mathbf{m} is given by integrating over the data \mathbf{d} to obtain the marginal posterior *pdf*,

$$Q(\mathbf{m}) = k p(\mathbf{m}) \int_{\mathbf{D}} \frac{p(\mathbf{d})F(\mathbf{d}, \mathbf{m})}{\mu(\mathbf{d}, \mathbf{m})} d\mathbf{d} . \quad (5)$$

Equation (5) is the general, probabilistic solution to the inverse problem of event location from the available data since it describes the uncertainty in event location \mathbf{m} given all available information. It is usual to call the integral in (5) the *likelihood* function $L(\mathbf{m})$, which gives a (non-normalized) measure of how good any model \mathbf{m} is in explaining the observed data $p(\mathbf{d})$.

As mentioned earlier it is often the case that $p(\mathbf{d})$ for the observed data is approximated by a Gaussian distribution, described by mean \mathbf{d}_0 and covariance matrix \mathbf{C}_d . Assuming that the uncertainties in the forward problem F relating \mathbf{d} and \mathbf{m} are negligible results in the form of F in Eq. (3). It is also usually assumed that \mathbf{d} and \mathbf{m} are independent and hence that $\mu(\mathbf{d}, \mathbf{m})$ can be written $\mu(\mathbf{d})\mu(\mathbf{m})$; $\mu(\mathbf{d})$ is usually taken to be constant. With these simplifications, used by many current direct-search location procedures, the (non-normalized) likelihood function is given by,

$$L(\mathbf{m}) = \exp \left\{ -\frac{1}{2} [\mathbf{d}_0 - \mathbf{f}(\mathbf{m})]^T \mathbf{C}_d^{-1} [\mathbf{d}_0 - \mathbf{f}(\mathbf{m})] \right\} . \quad (6)$$

With the above simplifications a maximum likelihood origin time, t_0 , can be determined analytically from weighted means of the observed arrival times and the predicted travel times (e. g., [74]), and if the observed and predicted times are uncorrelated we arrive at a likelihood function,

$$L(\mathbf{x}) = \exp \left\{ -\frac{1}{2} \sum_i \frac{[T_i^o - T_i^c(\mathbf{x})]^2}{\sigma_i^2} \right\} , \quad (7)$$

where \mathbf{x} is the spatial part of \mathbf{m} , T_i^o are observed travel times, T_i^c are the calculated travel times for observation i (i. e., T_i^c represents the travel time, rather than arrival time,

part of $\mathbf{f}(\mathbf{m})$), and σ_i summarizes the associated standard deviation of uncertainty in T_i^o and T_i^c .

Though not normalized, $L(\mathbf{x})$ is sufficient to provide the *relative* probability of any location \mathbf{m} being the best estimate of the event location given the available data measurements. Since in practice integrating over all of $\mathbf{D} \times \mathbf{M}$ to find normalizing constant k in Eq. (5) is often computationally intractable, the product of the prior, spatial location information $p(\mathbf{x})$ (i. e., the spatial part of $p(\mathbf{m})$) and the non-normalized likelihood $L(\mathbf{x})$ is usually taken as the objective function for inversion and searching in direct-search location algorithms. If $L(\mathbf{x})$ is determined through-out the prior *pdf* $p(\mathbf{x})$ through a global-search, then Eq. (5) can be normalized approximately after location. In the following text and examples, we refer to such an approximately normalized function, $p(\mathbf{x}) L(\mathbf{x})$, as a location *pdf*.

The likelihood function in Eq. (5) is entirely defined by the probabilistic error processes involved. However, often it is desirable to change the approximations employed in deriving Eqs. (6) and (7) from Eq. (5), in order to remove biases or instability in the solution. The approximation in Eq. (6) uses the exponential of an L2-norm misfit function (the term in braces $\{ \}$ in Eq. (6) or (7)) to represent the *pdf* of the data error variation, but because data used for location often contain outliers it is often considered that an L1 norm or other L_p norm ($p < 2.0$) is more appropriate (e. g., [69]), where L_p -norm $|\mathbf{x}| = \sqrt[p]{\sum |x_i|^p}$. Earthquake location problems formulated with an L_p norm (or indeed other kinds of likelihood functions – see Eq. (8) below), can be solved relatively easily with direct-search methods, which, unlike linearized methods, do not require determination of partial derivatives of the likelihood or objective function with respect to event location.

An alternative to L_p -likelihood functions that is very robust in the presence of outliers is given by the equal differential-time (EDT) formulation [16,31,84]. For the EDT case, the location likelihood is given by,

$$L(\mathbf{x}) = \left[\sum_{a,b} \frac{1}{\sqrt{\sigma_a^2 + \sigma_b^2}} \cdot \exp \left(- \frac{\{ [T_a^o - T_b^o] - [TT_a^c(\mathbf{x}) - TT_b^c(\mathbf{x})] \}^2}{\sigma_a^2 + \sigma_b^2} \right) \right]^N, \quad (8)$$

where \mathbf{x} is the spatial part of \mathbf{m} , T_a^o and T_b^o are the observed arrival times and TT_a^c and TT_b^c are the calculated travel times for two observations a and b ; the sum is taken over all pairs of observations, and N is the total number of observations. Standard deviations σ_a and σ_b summarize the

assigned uncertainties on the observed arrival times and calculated travel times, where it is assumed that the observed and the calculated times are uncorrelated.

In Eq. (8), the first and second terms in brackets in the exponent are, respectively, the differences between the observed arrival times and the differences between the calculated travel times. The exponent is the difference between these two terms, and thus the exponential has a maximum value of 1 which occurs at points \mathbf{x} where the two differences are equal (hence, the name “equal differential time”). Such points \mathbf{x} best satisfy the two observations a and b together, and, in general, the set of \mathbf{x} where the exponential is nonzero forms a “fat,” curved surface in 3D space. Because the summation over observations is outside the exponential, the EDT location *pdf* has its largest values for those points \mathbf{x} where the most pairs of observations are satisfied and thus is far less sensitive to outlier data than L_p norms which seek to best satisfy all of the observations simultaneously. Note that the EDT likelihood function $L(\mathbf{x})$ does not require calculation of an origin time t_0 ; this reduces the hypocenter search to a purely 3-parameter problem and contributes to the robustness of the EDT method. Nevertheless, a compatible estimate of t_0 can be calculated for any hypocenter point \mathbf{x} .

Ultimately, the full solution to the probabilistic location problem is a posterior *pdf* which includes as comprehensive as possible uncertainty information over parameters \mathbf{m} . This may include multiple “locally-optimal” solutions, e. g., $Q(\mathbf{m})$ or $p(\mathbf{x}) L(\mathbf{x})$ may have multiple maxima, and may have a highly irregular form. Some studies of seismicity and seismotectonics make explicit use of a probabilistic representation of seismic event locations (e. g., [22,31,46]).

Experimental Design Methods – Choosing Receiver Locations

As noted earlier, it is important to position stations so as to constrain as tightly as possible the event locations for a given source area. The location inverse problem solution in Eqs. (5), (7) or (8) is constrained by prior information on location $p(\mathbf{m})$, by observed data $p(\mathbf{d})$, and by forward-problem physics relating \mathbf{d} and \mathbf{m} . One way to significantly influence the form of this inverse problem, and hence uncertainty in its solution, is to change the data we record. Thus, we alter both $p(\mathbf{d})$ and the forward-problem physics, $F(\mathbf{d}, \mathbf{m})$.

For seismic location problems we may change the data by employing experimental design methods to choose or change the locations of seismic receivers. The goal of the design procedure is to place receivers such that the loca-

tion information described by solution $Q(\mathbf{m})$ is expected to be maximized. This is a “macro-optimization” problem where, prior to the occurrence of an earthquake, we optimize the design of the inverse problem that we expect to solve after an earthquake has occurred.

The design is varied such that it maximizes an objective function. This is usually taken to be the expected value of some approximation to the unique measure of information that was discovered by Shannon [67],

$$J(\mathbf{R}) = E_{\mathbf{m}_t} \{I[Q(\mathbf{m}); \mathbf{R}, \mathbf{m}_t]\} \quad (9)$$

where \mathbf{R} is a vector describing the design (e. g., receiver locations), $I[Q(\mathbf{m}); \mathbf{R}, \mathbf{m}_t]$ is the information contained in the resulting posterior *pdf* $Q(\mathbf{m})$ for design \mathbf{R} when the true parameters (e. g., event location) is \mathbf{m}_t , and the statistical expectation $E_{\mathbf{m}_t}$ is taken over all possible \mathbf{m}_t which (according to our prior knowledge) are expected to be distributed according to the prior distribution $p(\mathbf{m})$. $J(\mathbf{R})$ should be maximized.

Within the expectation in Eq. (9), the design criterion $J(\mathbf{R})$ takes account of all possible potential true event locations \mathbf{m}_t , their prior probability of occurrence $p(\mathbf{m})$, and the corresponding data (including their uncertainties) that are expected to be recorded for each location (the latter are included within $Q(\mathbf{m})$). To calculate the expectation usually requires integration over a far greater proportion of the model and data spaces, \mathbf{M} and \mathbf{D} respectively, than need be considered when solving the inverse problem after a particular event has occurred (since then $p(\mathbf{d})$ and hence $Q(\mathbf{m})$ are fixed, and $p(\mathbf{m})$ is more tightly constrained). Consequently, experimental design is generally far more computationally costly than solving any particular inverse problem post-event.

For this reason, design methods invoking linearized approximations to the model-data relationship $F(\mathbf{m}, \mathbf{d})$ (e. g., Eq. (10) below) have been employed by necessity in the past [11,13,49,70], or indeed non-probabilistic methods have been employed (e. g., [9,10,36,71]). Truly non-linearized design methods have been developed for location problems only relatively recently [12,78,80]. Historically, however, station network geometry has been defined more by heuristics (rules of thumb) and geographical, logistical and financial constraints, with design theory only recently being deployed.

Location Methods

Once data have been recorded and prior *pdf*'s defined, a solution such as that of Eqs. (5), (7) or (8) must be evaluated throughout the prior *pdf*, $p(\mathbf{x})$, to identify one or more “locally-optimal” solutions, or, preferably, to obtain

a full probabilistic location *pdf*. This evaluation generally requires direct-search optimization and search techniques, which we discuss below. We first digress and summarize linearized location procedures, which typically determine a single optimal hypocenter along with a simplified and approximate representation of the location *pdf* (e. g., a confidence ellipsoidal centered on the estimated hypocenter and origin time).

Linearized Location Methods

With linearized methods the arrival time expression (2), which is nonlinear with respect to the spatial location $\mathbf{m} = (x, y, z)$, is approximated by a Taylor series expansion around some prior estimate $\mathbf{m}_0 = (x_0, y_0, z_0)$ of the spatial location:

$$f(\mathbf{m}) = f(\mathbf{m}_0) + (\mathbf{m} - \mathbf{m}_0)f'(\mathbf{m}_0) + \frac{(\mathbf{m} - \mathbf{m}_0)^2}{2}f''(\mathbf{m}_0) + O[(\mathbf{m} - \mathbf{m}_0)^3] \quad (10)$$

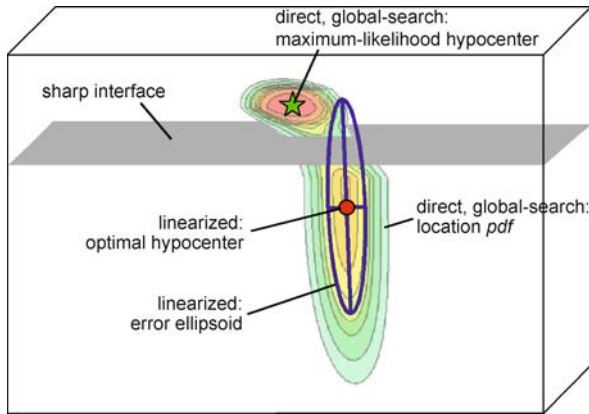
where $f(\mathbf{m})$ is the forward problem that calculates an arrival time d_{calc} given a location \mathbf{m} (e. g., $f(\mathbf{m})$ might represent the right hand side of Eq. (2) directly). A linear vector-matrix inverse problem is obtained if we approximate the forward problem for all d_{calc} by using only the first two terms of the Taylor series. The resulting vector-matrix equation may be solved using linear algebraic methods. This process is called *linearized inversion*.

Usually, this linearized inversion is iterated: the prior estimate \mathbf{m}_0 is set equal to the newly-found, best-fit location, the problem is re-linearized around this new estimate using Eq. (10), and the new linear problem solved again. This method may be repeated (iterated) many times, as needed to attain some convergence criteria.

Linearized methods produce a single, best-fit (e. g., maximum likelihood) hypocenter and origin time location, and associated, linearly-estimated uncertainties, such as a multi-dimensional, normal-distribution centered on the best-fit hypocenter and origin time. However, this linearized solution is often a poor representation of the complete solution *pdf* (Fig. 2 and see examples), and it may be unstable when the *pdf* is irregular or has multiple peaks due to insufficient or outlier data, velocity model complexities, and other causes (e. g., [5,34]).

Direct-Search Location Methods

The earliest, formal earthquake locations from phase arrival time observations used nonlinearized procedures. Milne [37] describes and applies several graphical and algebraic methods to determine earthquake locations. These



Earthquake Location, Direct, Global-Search Methods, Figure 2
 Schematic diagram comparing linearized and direct-search locations for the case where the complete location *pdf* is moderately complicated, with two maxima. This example arises from the case of a location at the limits of the recording network and near a sharp, horizontal interface in the velocity model between lower velocities above and higher velocities below. The colored, contoured form shows the true location *pdf*, as should be determined by a complete, probabilistic, direct-search location procedure. A linearized location that iterates from an initial trial location below the sharp interface will find an optimal hypocenter near the secondary, local maximum of the location *pdf*, below the interface. The linearized error ellipsoid, based on the curvature of the misfit function at this optimal hypocenter, reflects the form of this secondary maximum only. The linearized location procedure never identifies or explores the primary maximum of the *pdf* above the sharp interface, and produces incorrect error information above this interface (i. e. the uppermost part of the error ellipsoid). A probabilistic, direct, global-search procedure can determine the complete location *pdf* and identify correctly the maximum likelihood hypocenter located above the sharp interface

include a perpendicular bisector method for the case of 3 or more simultaneous arrival time observations (related to the modern arrival order or bisector method), a method of hyperbolae based on the differences in arrival times at pairs of stations (related to the modern EDT method) and a method using the differences in arrival times of different wave types at individual stations. The latter is a generalization of the method of circles using *S-P* times, in which the distance from a station to the source is, for given *P* and *S* velocity models, a function of the difference of the *S* and *P* arrival times; an epicenter can be constrained with such *S-P* based distances from 3 stations. Reid [52] determined a hypocenter location for the great 1906 California earthquake through a coarse, systematic grid search over velocity, position along the causative fault and depth, solving for the origin time and wave velocity by least-squares at each grid point.

The arrival order or bisector method [2,42] is a non-linear, geometrical approach that uses the constraint that if a phase arrival is earlier at station A than at station B, then the event is closer to A than to B (assuming the velocity model is such that arrival order implies distance order). Applying this constraint to all pairs of stations defines a convex region containing the event location. This method is useful for obtaining some constraint on the location of events far outside of an observing station network, and for rapidly and robustly obtaining starting locations for linearized methods.

Most other modern, direct-search earthquake location methods (excluding graphical methods that are now mainly used for illustrative and educational purposes) are based on deterministic or stochastic searches which may be exhaustive or directed and evolutionary. These searches are used to explore or map likelihood functions such as those given in Eqs. (5), (7) or (8). When these searches gather and retain information globally, throughout the prior *pdf* $p(\mathbf{x})$, they can produce a complete, probabilistic location *pdf*. Otherwise, searches may determine a global or local maximum of the location *pdf*, or may explore the neighborhood around these optimal points to locally estimate the *pdf* and obtain uncertainty information.

Regular, Deterministic Search Regular and deterministic searches, such as grid-searches, nested grid-searches and stochastic, “crude” Monte-Carlo searches (e. g., [20,62]) use global and well-distributed sampling of the model space and thus can estimate the complete location *pdf*. All of these approaches are computationally demanding for problems with many unknown parameters, large parameter spaces, or time consuming forward calculations, because the number of models that must be tested can be very large. These methods have been successfully applied to the determination of optimal hypocenters (i. e., [14,26,61,69]), and to probabilistic location (i. e., [6,34,38,83]), but their inefficiency may impose unacceptable limitations on the number of events that can be considered, or on the size of the search volume.

Directed Search Directed, stochastic search techniques include evolutionary, adaptive global search methods such as the genetic algorithm [19,59] and simulated annealing [28,53,72]. The simplex method is a directed, deterministic search technique that is nonlinearized and can be used for earthquake location (e. g., [48]). Most of these methods were developed for optimization or the identification of some very good solutions, which is equivalent to identifying a global or local maximum of the location *pdf*. In general, these methods do not explore the prior

pdf $p(\mathbf{x})$ in a manner that can produce complete, probabilistic solutions to inverse problems. For example, the genetic algorithm performs global searching and may be one of the most efficient stochastic methods for optimization, but it does not use well distributed sampling (the sampling tends to converge rapidly to the region of a locally optimum solution). Similarly, in the simulated annealing, random-walk method the interaction of its variable “temperature” parameter and step size with the local structure of the misfit function can lead to convergence and stalling near a locally optimum solution, and a sample distribution that is neither well nor globally distributed. Both the genetic algorithm and simulated annealing can be tuned to sample more broadly and in the limit become crude Monte Carlo searches, but this removes the main advantage of these methods – that of rapid stochastic optimization.

Though not directly applicable to complete, probabilistic location, directed search algorithms are useful for direct-search, earthquake hypocenter estimation because of their efficiency (e. g., [4,48,60,61]).

Importance sampling The efficiency of a Monte Carlo algorithm used to estimate properties of a target (misfit or likelihood) function can be increased by choosing a sampling density which follows the target function as closely as possible [20,30,45]. Techniques that follow this rule are referred to as importance sampling methods, and were originally developed in physics for fast and accurate numerical integration of multi-dimensional functions. The target function is unknown, however, and consequently the optimum importance sampling distribution cannot be determined a priori. Instead, improved efficiency is attained by adjusting (or adapting or evolving) the sampling by using information gained from previous samples so that the sampling density tends towards the target function as the search progresses [30,40,45,65]. For example, importance sampling to determine an earthquake location *pdf* or likelihood function (e. g., Eqs. (5), (7) or (8)), can be obtained by beginning with a sampling that follows the prior *pdf*, $p(\mathbf{m})$, and then adjusting the sampling as the search progresses so that the sampling density approaches the posterior, location *pdf*.

Importance sampling techniques that can be used to find complete, probabilistic solutions to inverse problem include the VEGAS algorithm [30], the Metropolis algorithm [40], the neighborhood algorithm [55] and, for three-dimensional problems, oct-tree [33]. Other importance sampling methods are discussed in Hammersley and Handscomb [20] and in Press et al. [45] in the context of numerical integration.

The VEGAS algorithm [30,45] performs importance sampling by accumulating appropriate sampling distributions independently for each parameter as the sampling proceeds. This method can give very good estimates of an individual or a joint marginal *pdf*, but it loses efficiency if the target function includes strong correlation between parameters or if it is independent of some parameters [45]. In addition, the VEGAS algorithm may be difficult or impossible to implement with prior information, such as smoothness constraints, that introduces correlation between parameters. Consequently, this algorithm may not be appropriate for some geophysical problems, including earthquake location, when the location parameters are often correlated or poorly resolved.

The Metropolis or Metropolis-Hastings algorithm (e. g., [40]) is similar to simulated annealing but with a constant temperature parameter. The Metropolis algorithm performs a random walk in the model space, testing at each step nearby trial samples which are accepted or rejected after evaluation of the forward problem according to a likelihood $L(\mathbf{m})$. In [40] it is shown that this algorithm samples from the posterior *pdf* of the problem and is therefore an importance sampling method. They show that, in the limit of a very large number of trials, it will not become permanently “trapped” near local maxima and consequently will produce global sampling. Also, because it is a random walk technique, the Metropolis algorithm can perform well even if the volume of the significant regions of the posterior *pdf* is small relative to the volume of the prior *pdf*. However in practical application, with a finite number of samples, this algorithm can become trapped in strong local maxima of the posterior *pdf* if this function is complicated. The Metropolis algorithm has been applied to earthquake location in 3D structures [34,35].

Another recently developed importance sampling technique used in geophysics is the neighborhood algorithm [55,56,57], applicable to high dimensional model spaces. Given an existing set of samples of the objective function, the neighborhood algorithm forms a conditional *pdf* using an approximate Voronoi cell partition of the space around each sample. The algorithm generates new samples through a uniform random walk within the Voronoi cells of the best fitting models determined so far. This algorithm is applied to the 4D hypocenter location problem in [27,58].

The oct-tree importance-sampling method [33] uses recursive subdivision and sampling of rectangular cells in three-dimensional space to generate a cascade structure of sampled cells, such that the spatial density of sampled cells follows the target *pdf* values. The relative probability

that an earthquake location lies within any given cell i is approximately,

$$P_i = V_i L(\mathbf{x}_i) , \quad (11)$$

where V_i is the cell volume and \mathbf{x}_i is the vector of coordinates of the cell center. Oct-tree importance-sampling is used to determine a location *pdf* by first taking a set of samples on a coarse, regular grid of cells throughout the search volume. This is followed by a recursive process which takes the cell k that has the highest probability P_k of containing the event location, and subdividing this cell into 8 child cells (hence the name oct-tree), from which 8 new samples of the *pdf* are obtained. These samples are added to a list of all previous samples, from which the highest probability cell is again identified according to Eq. (11). This recursive process is continued until a predetermined number of samples are obtained, or until another termination criterion is reached.

For most location problems, including those with a complicated location *pdf*, the oct-tree recursive subdivision procedure converges rapidly and robustly, producing an oct-tree structure of cells specifying location *pdf* values in 3D space. This oct-tree structure will have a larger number of smaller cells in areas of higher probability (lower misfit) relative to areas of lower *pdf* value and thus the oct-tree method produces approximate importance-sampling without the need for complex geometrical constructs such as Voronoi cells. Oct-tree sampling can be used with the L2-norm likelihood function in Eq. (7) or the EDT likelihood function in Eq. (8), since both require searching over three-dimensional spatial locations only. Oct-tree sampling has been applied to earthquake location in 3D structures [22,23,31,32]; we use this sampling method to determine locations in the examples presented below. Though limited to determination of the 3D, spatial location, this recursive sampling procedure can be extended to 4D to allow determination of the origin time.

Illustrative Examples

We illustrate the concepts described in the previous sections using an M3.3 earthquake that occurred in the Garfagnana area of Northern Tuscany, Italy, on March 5, 2007 at 20:16 GMT. The earthquake was recorded by stations of the Italian National Seismic Network (INSN) at distances from less than 10 km to more than 300 km. We use manually picked *P* and *S* phase arrival times from the INSN bulletin with Gaussian uncertainties (standard deviations from 0.01 to 0.1 s), and a 1-D velocity model similar to the standard model used by INSN for routine

earthquake location in Italy. We perform all event locations with the probabilistic location program NonLin-Loc [31,34,35] (<http://www.alomax.net/nlloc>; NLL hereafter), using the oct-tree sampling algorithm (Sect. “Location Methods”) to perform a global-search within a parameter space \mathbf{M} formed by a rectangular volume 360 km on each side and from the Earth’s surface to 35 km depth (except as noted in figure captions). We use the L2-norm (Eq. (7)) or EDT (Eq. (8)) likelihood functions to obtain location *pdf*’s in 3D space and corresponding maximum likelihood origin times.

In order to describe the location problem and the solution quality for each of the examples presented below we focus on geometrical properties of the location *pdf*, which represents most completely the results of probabilistic, direct, global-search methodologies. We also consider the maximum likelihood hypocenter, defined as the point in space of the maximum value of the location *pdf*, and the corresponding origin time. We examine statistics of the quality of the solutions using the half-lengths of three principal axes of a 68% confidence error ellipsoid approximation to the location *pdf*, l_{ell} , the weighted, root-mean-square of the arrival residual (observed – calculated) times, *rms*, and a relative measure of the volume of the high likelihood region of the location *pdf*, V_{pdf} , given by,

$$V_{\text{pdf}} = \int_{\mathbf{M}} \frac{\text{pdf}(\mathbf{x})}{\text{pdf}^{\text{max}}} dV , \quad (12)$$

where pdf^{max} is the maximum value of the location *pdf* in \mathbf{M} . We also make use of standard measures of the experimental design quality (i.e, stations coverage) including the *gap* – the largest angle between the epicenter and two azimuthally adjacent stations used for location, and the distance Δ_0 from the hypocenter to the closest station. These indicators are summarized in Table 1 for the examples presented here.

These examples are meant to show important features and complexity in earthquake location results, not to compare different direct-search location methods or to compare direct-search to linearized algorithms. However, because linearized earthquake location has been and remains an important and widely used tool, we indicate for each example the location results obtained with a linearized algorithm, Hypoellipse [29]. Hypoellipse uses a least-squares, L2-norm and produces a 68% confidence ellipsoid for the hypocenter location. For well constrained locations this ellipsoid should closely match the *pdf* of our probabilistic, L2-norm locations; we plot the Hypoellipse ellipsoid for cases where it differs notably from the probabilistic location, L2-norm *pdf*.

Earthquake Location, Direct, Global-Search Methods, Table 1

Summary of results and quality indicators for the example locations. R_{pdf} is the radius of a sphere with volume V_{pdf} ; l_{ell}^1 , l_{ell}^2 , l_{ell}^3 are the half-lengths of the error ellipsoid axes; N_p is the number of phases used for the location; Δ_0 is the distance to the closest station

		Example	Lat (°)	Lon (°)	Depth (km)	rms (s)	gap (°)	Δ_0 (km)	N_p	V_{pdf} (km ³)	R_{pdf} (km)	l_{ell}^1 (km)	l_{ell}^2 (km)	l_{ell}^3 (km)
	Ideal	1a	44.208	10.295	10.98	0.399	89	9.1	20	26	1.8	1.6	2.2	3.2
		1b	44.208	10.295	10.98	0.000	63	9.1	50	7	1.2	1.1	1.3	2.1
Station distribution	Few stations	2a	44.163	12.267	47.44	0.000	335	232	2	742001	82.3	30.9	53.9	223
		2b	44.172	9.565	77.60	0.001	192	25.8	4	21600	25.3	5.9	43.2	81.9
		2c	44.208	10.295	14.40	0.000	173	64.5	3	2011	7.8	4.8	7.3	20.2
		2d	44.208	10.295	10.98	0.000	99	9.1	8	66	2.5	2.1	2.7	4.9
	Side	3	44.207	10.296	11.03	0.004	251	29.0	19	444	4.7	3.4	4.7	11.0
	Far	4a	44.207	10.296	11.03	0.006	103	106	46	234	3.8	2.0	2.4	17.8
		4b	44.208	10.295	10.98	0.000	103	106	50	66	2.5	1.7	2.2	8.1
	Experimental design	5a	44.215	10.290	15.13	0.014	229	9.9	6	1806	13.4	8.3	13.7	59.1
		5b	44.208	10.295	7.83	0.007	89	36.3	6	381	6.6	3.1	3.7	11.5
Outlier	L2-norm	6a	44.207	10.296	11.03	0.007	135	9.00	10	172	3.5	2.9	3.7	6.7
		6b	44.120	10.267	9.02	0.813	156	10.5	10	172	3.5	3.0	4.2	9.0
	EDT	6c	44.221	10.305	9.94	0.006	132	9.4	10	167	3.4	4.5	7.7	15.2
		6d	44.215	10.304	10.02	0.540	133	9.0	10	275	4.0	10.7	30.8	42.9
	Early warning	7a	44.139	10.194	24.79	0.012	307	15.6	3	628090	53.1	18.7	81.6	104
		7b	44.210	10.302	10.57	0.015	250	8.8	4	33908	20.1	18.4	30.0	105
		7c	44.208	10.295	11.12	0.008	227	9.1	5	1704	7.4	5.3	13.9	30.2
		7d	44.207	10.296	11.03	0.007	135	9.0	10	172	3.5	2.9	3.7	6.7
Incorrect velocity model	L2-norm	8a	44.208	10.295	10.98	0.000	63	9.1	50	15	1.5	1.3	1.7	2.9
		8b	44.160	10.244	2.69	0.808	66	11.4	50	17	1.6	1.3	1.7	3.0
	EDT	8c	44.220	10.305	9.98	0.000	63	9.4	50	11	1.4	1.2	1.5	2.6
		8d	44.192	10.280	7.20	0.795	64	9.3	50	167	3.4	2.8	3.9	6.7

Example 1: An Ideal Location

To construct an ideal, reference location and synthetic data set for the 2007 Italian earthquake we first locate the event using the earliest 20 observed P or S arrival times (Fig. 3; Table 1; Example 1a). Next, we subtract the arrival residuals for this location from the corresponding times for all observations and relocate the event with the earliest 50 of these “corrected” times (Fig. 3; Table 1; Example 1b). This procedure results in an ideal, synthetic data set and a location problem that are equivalent to the case of no “a posteriori” picking error and no travel-time error (i. e., no velocity model error). For this problem the quality of the solution and the shape of the resulting location pdf reflect primarily the station geometry and corresponding ray take-off angles about the source.

The reference location (Fig. 3; Table 1; Example 1b) has $rms = 0$ s, $gap = 63^\circ$, $\Delta_0 \sim 9$ km, $V_{pdf} \sim 7.0$ km³ and $l_{ell} = 1.05, 1.32$ and 2.05 km. The rms is necessarily zero because we used residuals as time corrections, while the other indicators and the near-ellipsoidal form of the location pdf show a well constrained location. The location

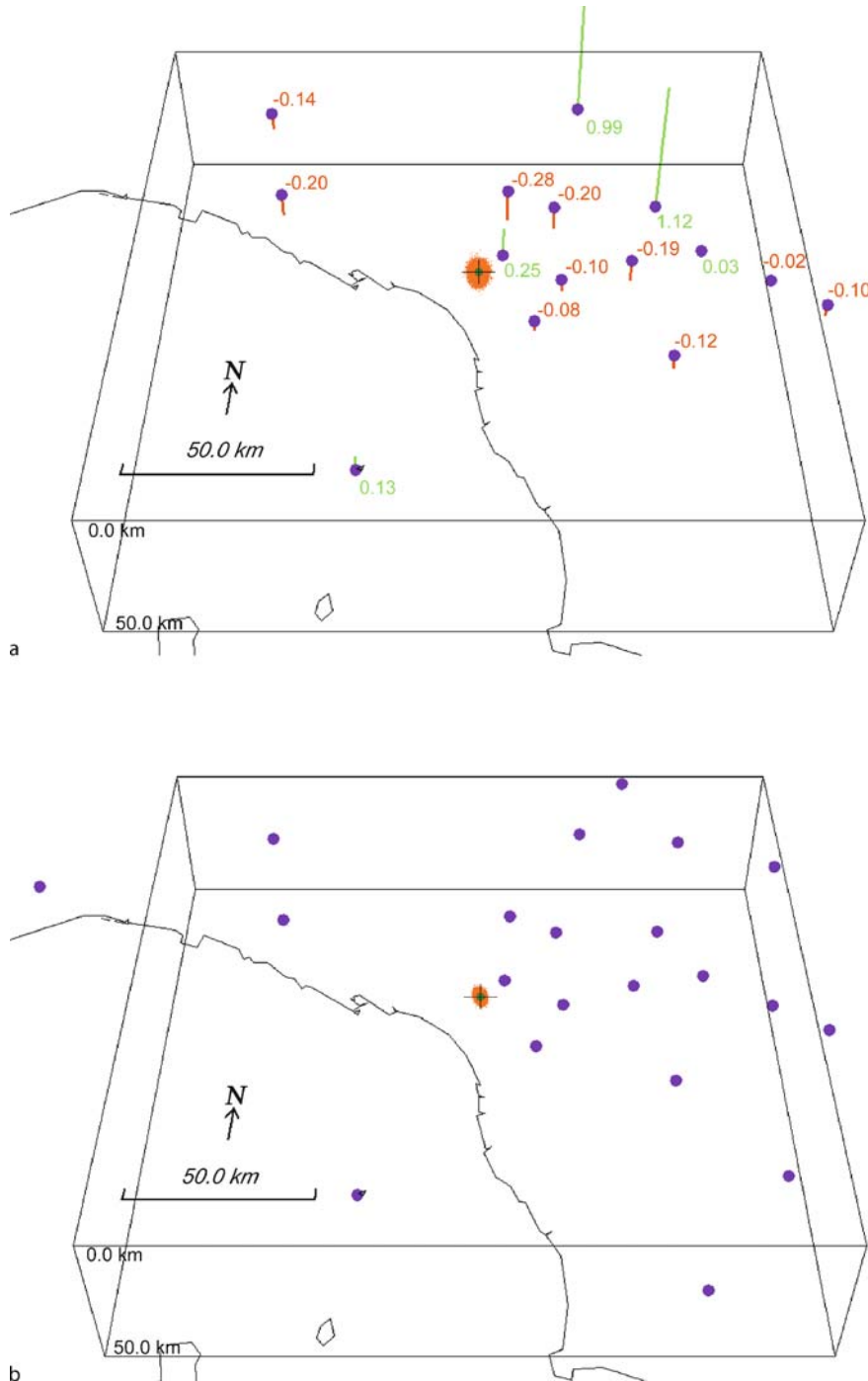
is well constrained by the data because stations are available at a wide range of distances and azimuths. In particular, the presence of a station nearly above the event, and of both P and S -wave arrival times for the closer stations, give good depth constraint.

Examples 2–5: Station Distribution

In the next examples we show locations for three cases with poor station distribution about the source:

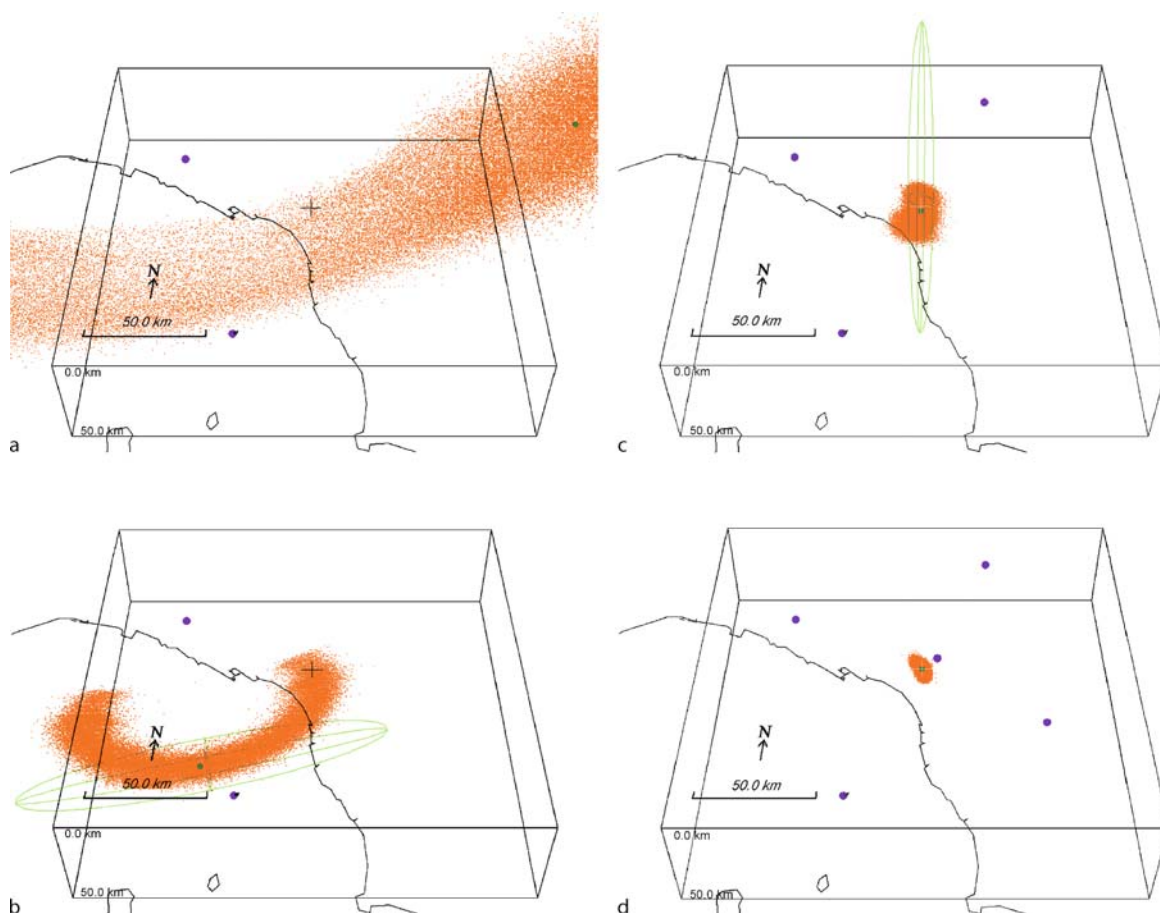
- 1) few available stations;
- 2) stations all to one side of the event; and
- 3) no data for stations near or above the source. In addition we illustrate the application of experimental design techniques to improve the station distribution.

Example 2: Few Available Stations We first examine relocations of the 2007 Italian earthquake obtained with different numbers of P and S arrival times selected from the ideal, synthetic data set (Fig. 4; Table 1; Example 2a–d). With only two stations and 2 arrivals (2 P phases) the location pdf is a fat, near-vertical, planar surface with an elon-



Earthquake Location, Direct, Global-Search Methods, Figure 3

Example 1: An ideal location. a Location obtained using the first 20, observed *P* or *S* arrival times; b location obtained using the first 50, *P* or *S* corrected arrival times from the ideal, synthetic data set. The elements shown in these and the following figures are: stations used for location (blue dots, in some cases stations fall outside the plotted region); location pdf (red cloud of points showing an importance sample drawn from the pdf); maximum likelihood hypocenter (green dot); ideal, synthetic location (black cross); *P* arrival residuals at each station: positive (green, up-going bars) and negative (red, down-going bars), numbers indicate residual value in sec. The Hypoellipse linearized locations and ellipsoids do not differ significantly from the direct-search locations shown in this figure



Earthquake Location, Direct, Global-Search Methods, Figure 4

Example 2: Few available stations. Locations obtained using progressively (a–d) a larger number of arrival observations. **a** 2P phases (2 stations); **b** 2P and 2 S phases (2 stations); **c** 3P phases (3 stations); **d** 5P and 3S phases (5 stations). For the locations in **a** and **b** the oct-tree search is performed to 100 km depth. In this and the following figures the 68% Hypoellipse ellipsoid is shown with green lines. Hypoellipse linearized location: does not converge for the location in panel **a**; ellipsoid differs markedly from the direct-search location *pdf* in panels **b** and **c**; and does not differ markedly from the direct-search location *pdf* in panel **d**

gated, boomerang shape trending perpendicular to the line connecting the two receivers (Fig. 4a). With the addition of *S* arrivals from the same stations (4 arrivals – 2 *P* and 2 *S* phases) the location *pdf* is greatly reduced in volume, and has the form of an annulus oriented roughly perpendicular to the line connecting the two receivers (Fig. 4b). The annular form of this *pdf* results from the intersection of the boomerang shape *pdf* produced by the 2 *P* phases (Fig. 4a) and two hemispherical *pdf*'s centered on each station. Each of these hemispherical *pdf*'s would be produced by location using only the *P* and the *S* reading from either station; this is the probabilistic analogue to the method of circles using *S-P* times.

With three stations (3 arrivals – 3P phases) the location *pdf* forms one mass and its volume is further reduced. This location *pdf* retains an irregular, curved shape result-

ing from poor constraint of one spatial dimension that trades off with origin time (Fig. 4c). For all of these locations the problem is effectively underdetermined – the data cannot constrain all three hypocentral coordinates and origin time. In these cases a linearized location algorithm may not converge and would be unable to represent properly the effective location uncertainties. As more data are added, the location *pdf* progressively reduces in size and complexity, and with the addition of a station close to and above the source (8 arrivals – 5 *P* and 3 *S* phases), the location *pdf* has a compact, near ellipsoidal form indicating some constraint on all hypocentral coordinates and origin time (Fig. 4d). This location is similar to that obtained with the complete, ideal data set (Fig. 3b), though the location *pdf* remains much larger than that of the ideal case which has arrival times from many more stations.

Example 3: Stations to One Side of the Event – Large Gap

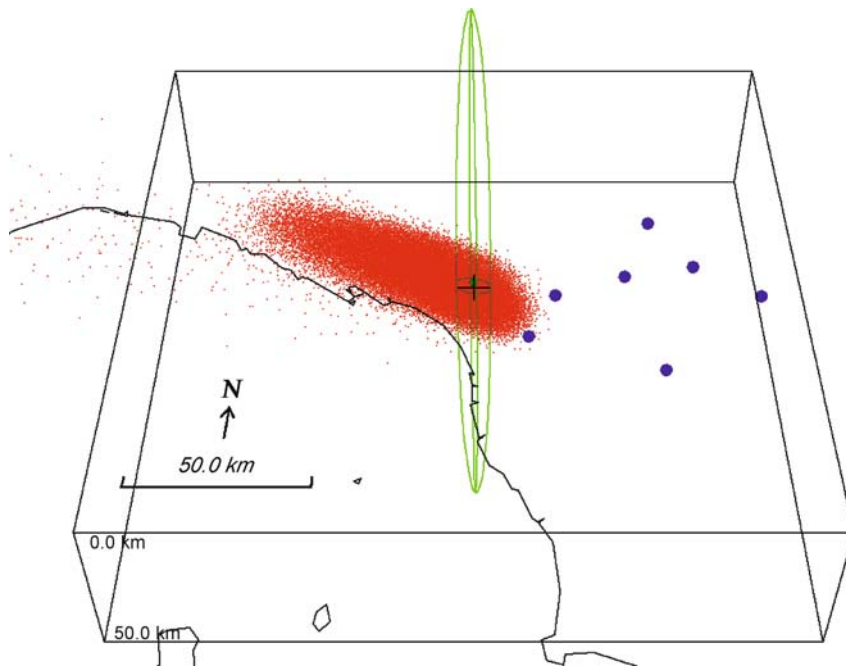
Next, we examine the case of earthquakes occurring outside of the recording network with an example using P arrival times from stations only to the southeast of the earthquake (Fig. 5; Table 1; Example 3). The location pdf is large and elongated in a northwest-southeast direction oriented towards the centroid of the available stations because the lack of stations to the northwest (and use of P times only) allows a strong trade-off between potential hypocenter locations along this direction and origin time. In contrast, there is some constraint of the pdf to the northeast and southwest due to the aperture of the available stations. The poor station distribution and potential lack of constraint is clearly indicated by the large gap value for this location, $\text{gap} = 251^\circ$. One or more good quality S readings can reduce the elongation of the pdf .

Example 4: Stations Far From the Event – Vertically Elongated PDF

We next show an example where the nearest recording stations are far from the earthquake, relative to its depth, and either P arrival times only or both P and S arrival times are available (Fig. 6; Table 1; Examples 4a–b). With this station geometry the seismic rays leave the source region with approximately the same dip-

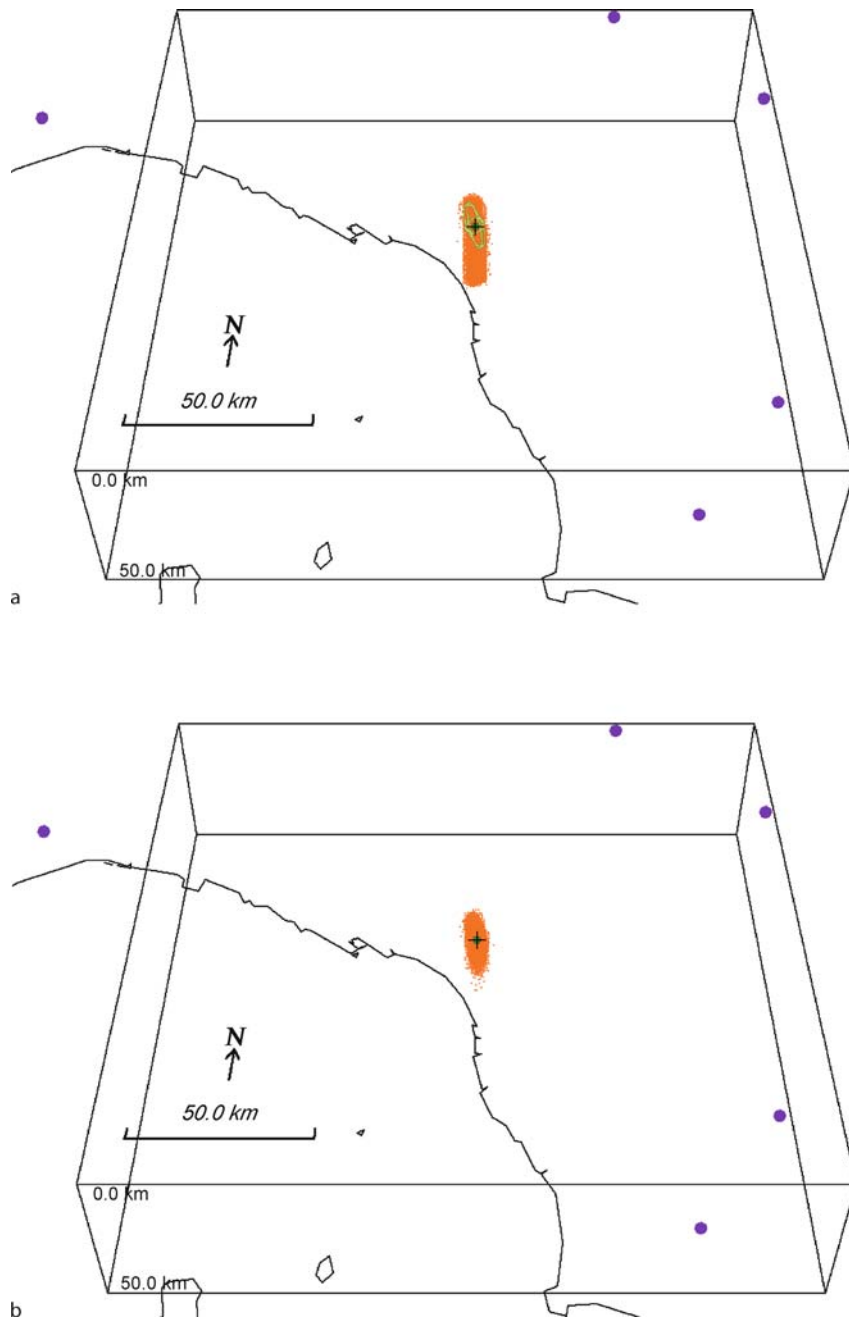
direction to all stations. Consequently a change in source depth gives about the same change in predicted travel times to all stations. This change in travel time is indistinguishable from a change in origin time (c.f., Eqs. (1) or (2), leading to a strong trade-off between origin time and depth. Consequently the location pdf has a vertically elongated shape which, for the case of P arrivals only (Fig. 6a), extends throughout the entire search range in depth indicating no depth constraint. For a linearized location algorithm this location problem can be effectively underdetermined, though most linearized algorithms can fix the hypocenter depth artificially in order to obtain a stable epicentral location. The addition of S arrival times (Fig. 6b) improves the depth constraint to some extent, although the location pdf remains highly elongated in the vertical direction. The lack of close stations and potential lack of constraint is clearly indicated by the large Δ_0 value for this location, $\Delta_0 \approx 106$ km.

This case is common with sparse networks and with shallow sources. Reducing the vertical extent of the pdf requires stations at distances of the order of the source depth or less. The addition of one or more good quality S readings, especially at the closest stations, would further improve the depth constraint.



Earthquake Location, Direct, Global-Search Methods, Figure 5

Example 3: Stations to one side of the event. A location example with P -wave arrival times at 7 stations only to the southeast of the event. The Hypoellipse ellipsoid differs markedly from the direct-search location pdf in this figure



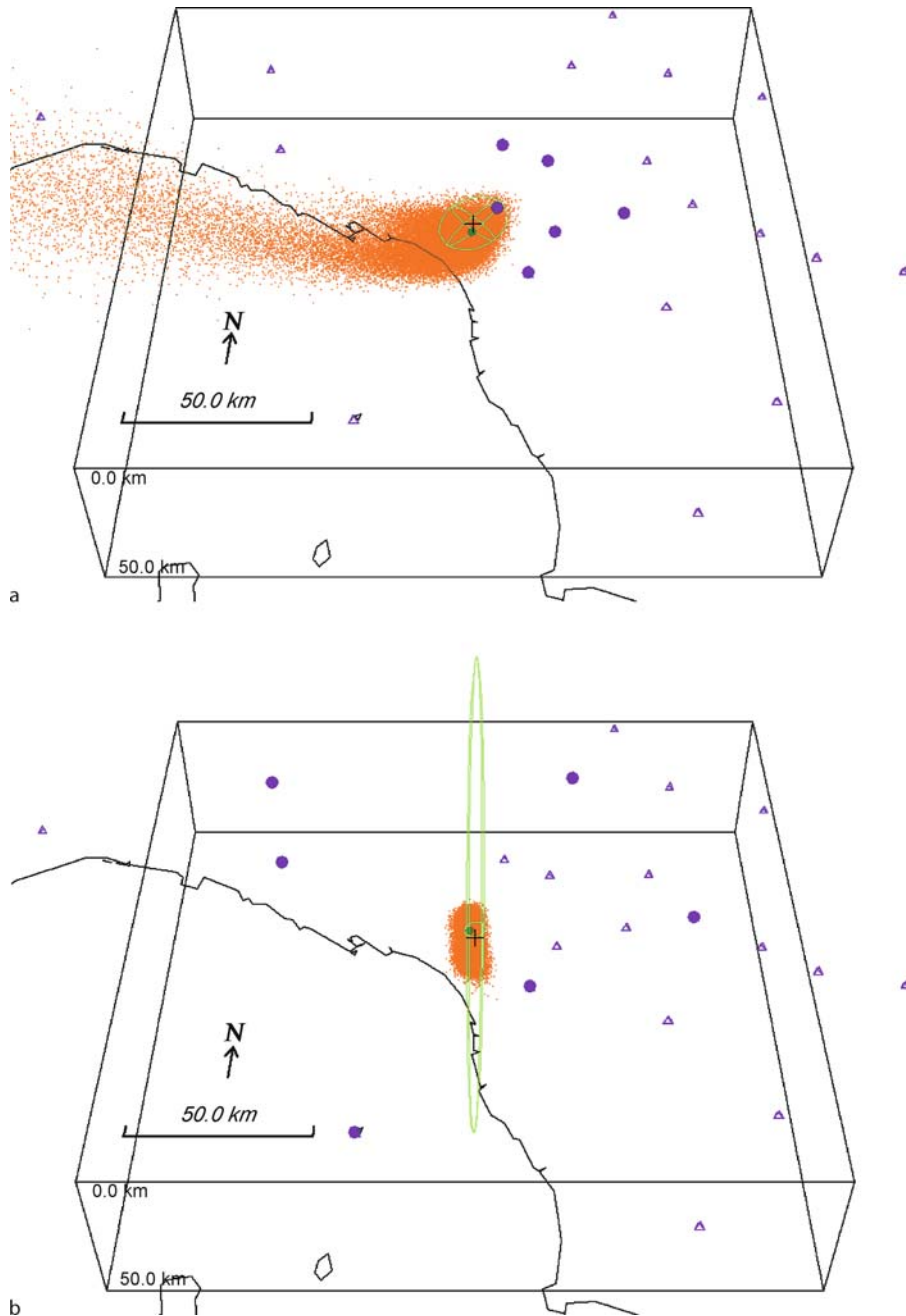
Earthquake Location, Direct, Global-Search Methods, Figure 6

Example 4: Stations far from the event. A location example using stations far from the epicenter, with **a** P arrival times only, **b** both P and S arrival times. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location *pdf* in panel **a**; and does not differ markedly from the direct-search location in panel **b**

Example 5: Stations Selection with Experimental Design

Next, we illustrate the application of experimental design techniques to station selection (Fig. 7; Table 1; Example 5). Considering a case similar to Example 3, which has 6 sta-

tions to one side of the event giving poor constraint on the location, we determine an optimal set of 6 stations to best constrain the location. To do this we apply a linearized design method [13] to select an optimal subset of 6



Earthquake Location, Direct, Global-Search Methods, Figure 7

Example 5: Stations selection with experimental design. A location example showing a Location using the stations with the first 6 available arrival times, **b** location using an optimal set of 6 stations as determined with a linearized experimental-design method. Available stations not used or selected are shown with *open triangle symbols*. Hypoellipse linearized location: ellipsoids differ markedly from the direct-search location *pdf*'s in panels **a** and **b**

of the available INSN stations to best constrain an event at the (known) location produced by the ideal, synthetic data (Example 1b).

The design procedure does not simply select the 6 closest receivers to the source (i. e. first 6 available arrival times, Fig. 7a), but instead selects receivers distributed

around, and to a large distance away from the source (Fig. 7b). This choice can be understood as balancing the distribution of directions (azimuth and inclination) that the rays leave the source to the selected receivers, a direct result of the use of the linearized approximations to the model-data relationship Eq. (10) in the linearized design method [13]. This method is based on selecting stations based on the similarity between the rows of the location kernel matrix of the linearized problem; the approach does not differ significantly from that of Uhrhammer [76] based on the condition number of the same matrix. The improvement in station distribution in azimuth is indicated by the small gap value for this location, $\text{gap} = 89^\circ$. The resulting location *pdf* (Fig. 7b) is compact and symmetric relative to the location *pdf* obtained from the first 6 stations recording the *P* phases (Fig. 7a), and the maximum likelihood hypocenter is close to the ideal location hypocenter.

Example 6: Incorrect Picks and Phase Identification – Outlier Data

For a given hypocenter location, an outlier arrival time has a residual that is much greater than its nominal error. Data outliers are common with automatic phase arrival picking algorithms, with *S* arrival picks, for small events, distant stations, or other cases where the signal to noise ratio is low, and for early instrumental data where large timing errors are common. In many cases, such as automatic earthquake monitoring and early warning systems, it is important to have robust location procedures that are influenced as little as possible by the presence of outliers. One way to achieve this is to use robust likelihood functions such as EDT Eq. (8). In the example below, we compare the performance of EDT and the more commonly used L2-norm likelihood functions.

This example uses only stations near the source, and arrival times from ideal, synthetic data sets for both the L2-norm and the EDT likelihood functions. We add 3 s to the *P* arrival time at two stations to generate outlier data, and examine L2-norm and EDT locations without and with the outlier data (Fig. 8; Table 1; Examples 6a–d). The L2-norm location with the outlier data (Fig. 8b) does not identify and isolate the two outlier *P*-arrivals but instead mixes information from these arrivals with the other data resulting in relatively large, non-zero residuals for all arrivals. This results in a bias of about 10 km in the maximum likelihood hypocenter location relative to the ideal location hypocenter, while the location *pdf* for the L2-norm locations with and without outlier data have about the same size and form, but have little over-

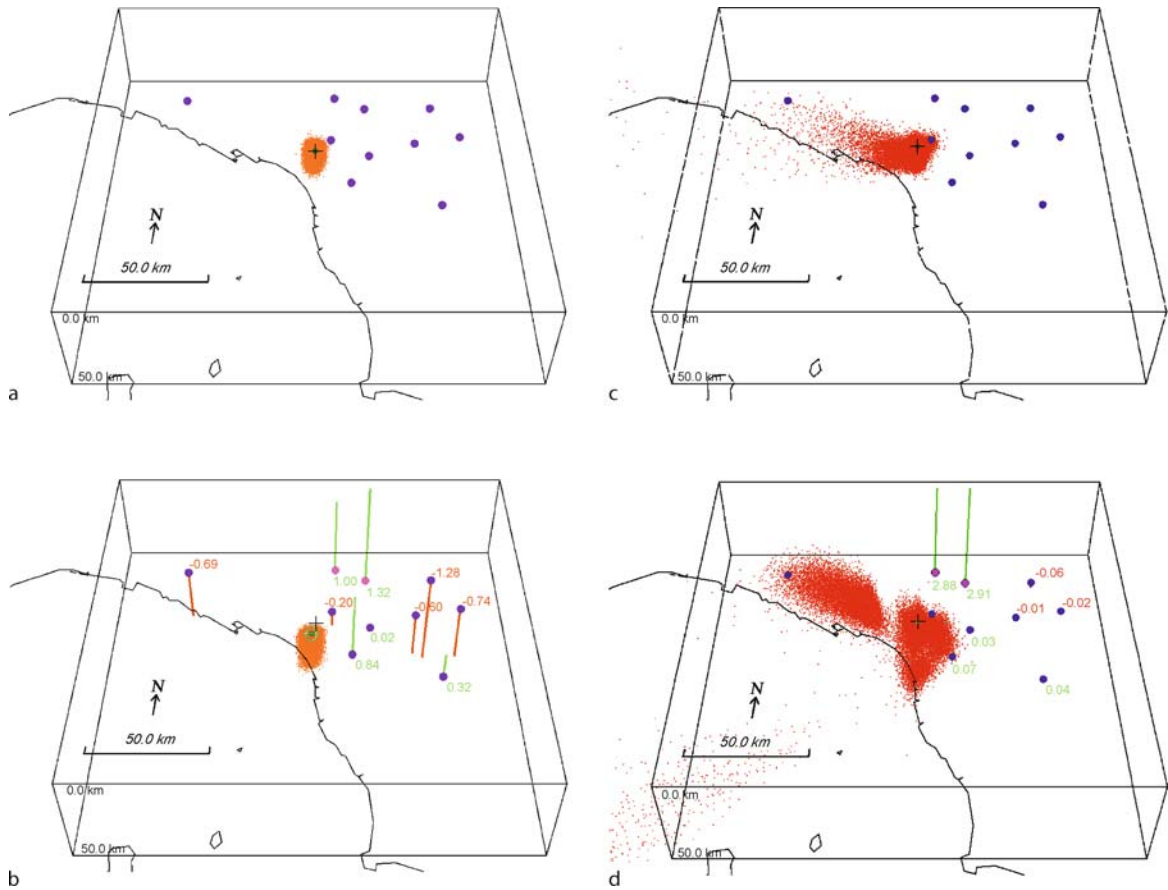
lap (Figs. 8a and 8b). Thus the L2-norm solution gives no clear indication of the presence of outlier data, or that the solution may be biased. In contrast, the EDT location for the data set containing the outliers (Fig. 8d) correctly identifies the two outlier arrivals (the EDT residuals for these two outlier data are both about 2.9 s) and strongly down-weights them (from 1.2 to 0.17 posterior weight), while producing small residuals (< 0.08 s) for the remaining arrival, as would be the case without outlier data. The maximum likelihood hypocenters for the EDT locations with and without outlier data are almost identical, but the location *pdf*'s are very different (Fig. 8c and 8d). With outlier data, the *pdf* has an irregular shape and several distinct parts, reflecting the inconsistency of the data set to constrain a unique event location. For the outlier locations, a potential problem with the data set is indicated by the large *rms* values with both L2-norm and EDT, and with EDT alone, by the asymmetry in residuals, the irregular *pdf* shape, and the large V_{pdf} and I_{ell} values.

This result shows that location in the presence of outlier data can be remarkably stable with the EDT likelihood function, which is easy to implement with direct-search location techniques. In contrast, the same location with the commonly used, L2-norm likelihood function is biased, while presenting few indicators of this bias.

Example 7: Earthquake Early-Warning Scenario

Location for earthquake early-warning must be performed rapidly and in an evolutionary manner starting with the first available phase arrivals. In this example we examine the ability of direct-search location to obtain robust and useful location information using *P* arrivals from the first stations that record the Northern Italian event (Fig. 9; Table 1; Examples 7a–d).

Within about 6 seconds after the origin time, t_0 , three *P* readings are available. Location with these readings produces an extensive location *pdf* that fills the southwest quadrant of the search region (Fig. 9a); this *pdf* does not provide useful constraint on the location, but is robust in that it includes the true location. Progressive addition of more arrival time data (Fig. 9b and 9c) reduces the size of the location *pdf*. With 5 arrivals, at about 7 s after t_0 (Fig. 9c), the maximum likelihood location is close to that of the ideal, synthetic location and the location *pdf* is well delimited, although elongated towards the west because no arrivals are yet available from stations in that direction. By 13 s after t_0 (Fig. 9d), 10 *P* arrivals are available and the location *pdf* is now compact and symmetrical, primarily because a station to the northwest is included. This *pdf* has small enough V_{pdf} and I_{ell} values to provide useful, prob-



Earthquake Location, Direct, Global-Search Methods, Figure 8

Example 6: Incorrect picks and phase identification – outlier data. Locations using ten P -wave arrival times with L2-norm and a no outliers, **b** two arrival-time outliers, and with EDT and **c** no outliers, **d** two arrival-time outliers. The stations with outlier arrivals are shown with violet dots. Note the small pdf of L2-norm regardless of the outliers and, in contrast, the ability of EDT to detect the outliers (see text). The Hypoellipse ellipsoid differs markedly from the direct-search location pdf in panel **b**. Hypoellipse not compared to EDT locations in panels **c** and **d**

abilistic constraint on the location for early-warning purposes at a regional scale, while the maximum likelihood hypocenter is effectively the same as that of the ideal location. In practical application, direct-search location results similar to those illustrated here can be obtained within a delay of less than 1 sec after the readings are available (e. g. [64]).

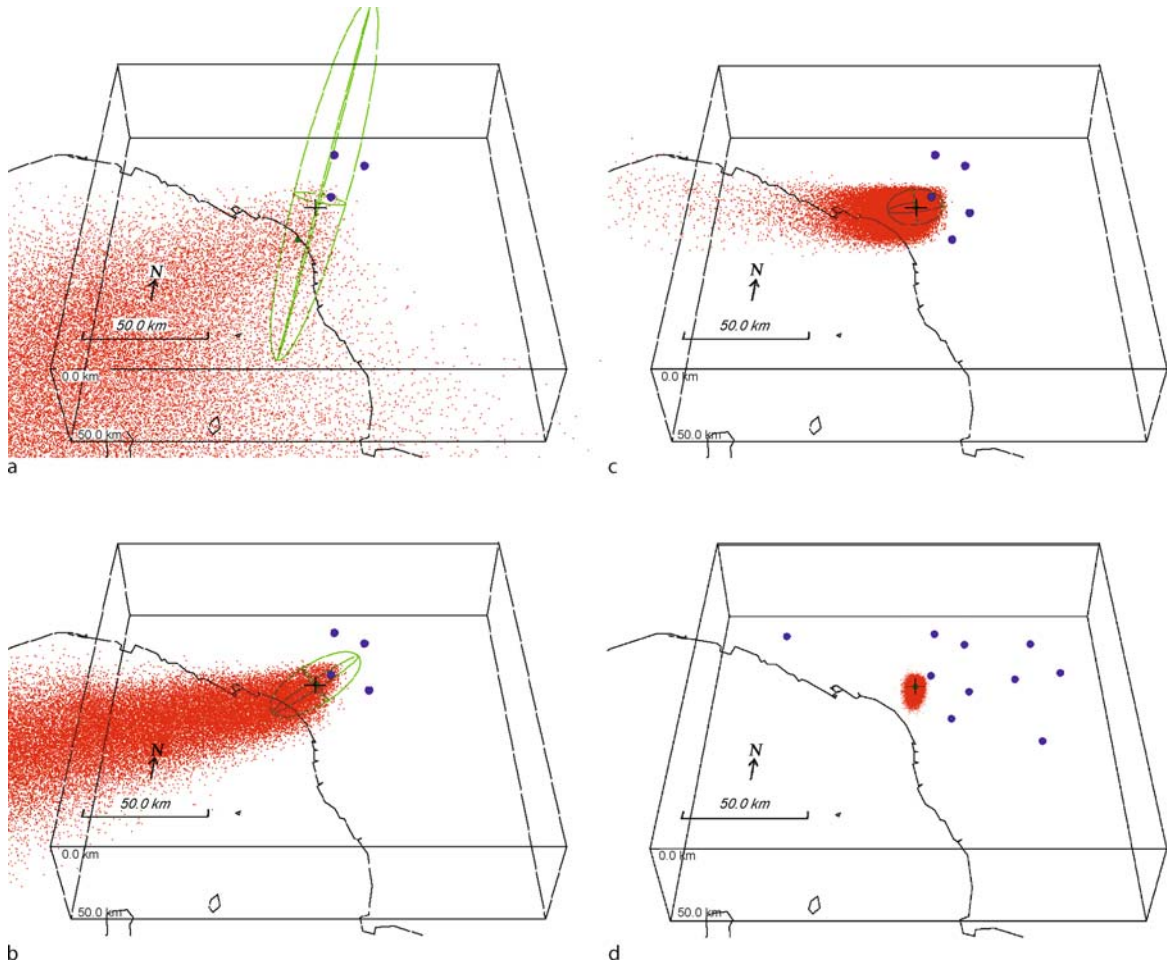
Example 8: Incorrect Velocity Model

Any velocity model used for earthquake location is an approximation to the true Earth and thus will in general produce erroneous predicted travel times. The magnitude of error in the travel times depends on many factors, but will in general be larger for more distant stations and with increased complexity in the true Earth structure. We examine the effect of an incorrect velocity models by repeating

the ideal location (Example 1a and b) with and without the “corrected” times, and using 50 P arrivals (the ideal location was determined using the first 20 P or S arrivals). We examine locations using the L2-norm and EDT likelihood functions (Fig. 10; Table 1; Examples 8a–d).

The locations with time corrections (Fig. 10a and 10c) simulate the unrealizable case of perfect knowledge of the velocity structure. With both the L2-norm and EDT the location results show zero residuals, compact location pdf s and a maximum likelihood hypocenter that necessarily matches exactly the corresponding ideal location. We note, however, that the L2-norm and EDT “ideal” locations differ slightly because they are derived from noisy, real data, and they use different likelihood functions.

The locations without time corrections (Fig. 10b and 10d) use the true observed data (i. e., travel times through the true Earth) and thus show the effect of an in-



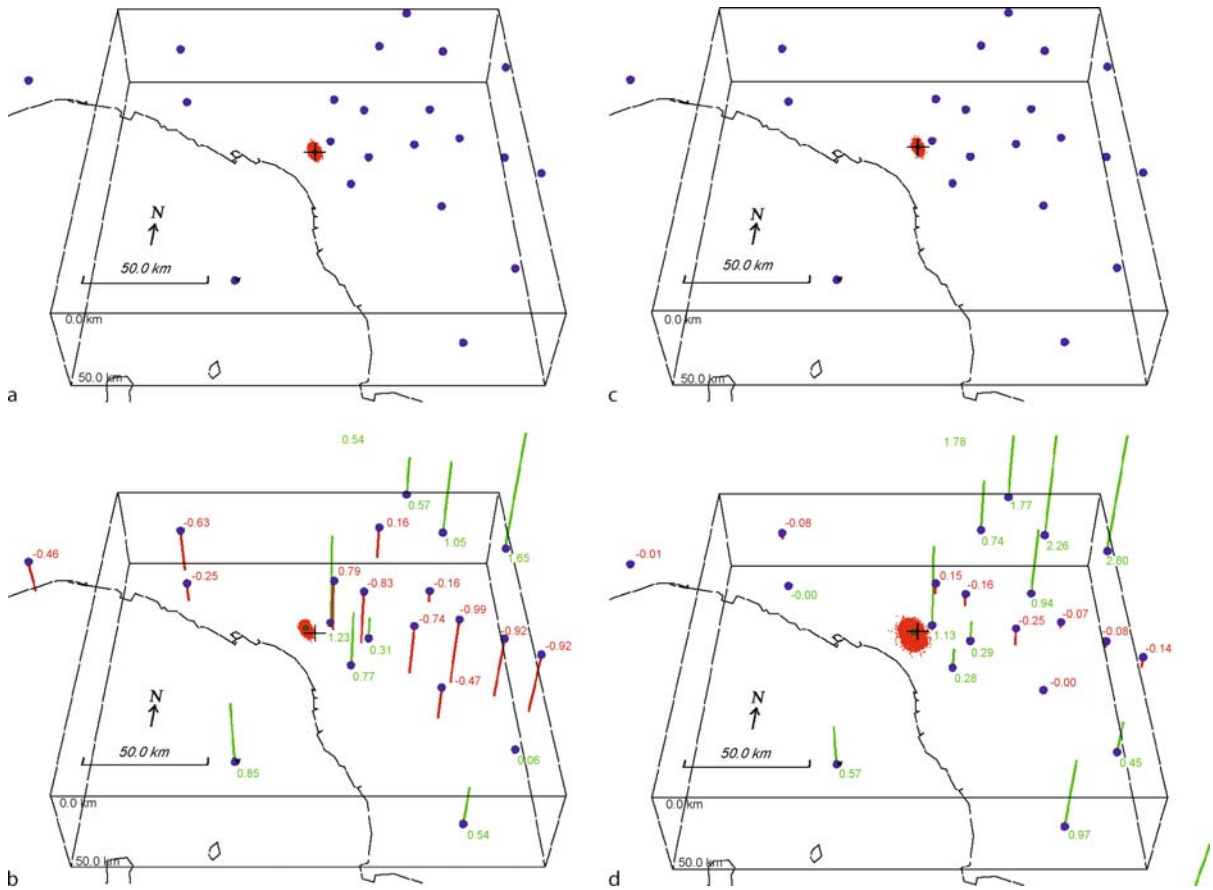
Earthquake Location, Direct, Global-Search Methods, Figure 9

Example 7: Earthquake early-warning scenario. Progressive location using a 3, b 4, c 5 and d 10 stations. Hypoellipse linearized location: ellipsoid differs markedly from the direct-search location *pdf*'s in panel a, b and c; and does not differ markedly from the direct-search location in panel d

correct velocity model (i.e., the 1-D velocity model used for location). This is shown by the pattern of positive and negative residuals obtained with both the L2-norm and EDT. The L2-norm location without time corrections has a balanced distribution of positive and negative residuals and, relative to the L2-norm location with corrections, a similar size location *pdf* and a biased maximum likelihood hypocenter. In contrast, the EDT location without corrections has more positive than negative residuals and, relative to the EDT location with corrections, a larger location *pdf* and nearly identical, unbiased maximum likelihood hypocenter. For these locations, a potential problem with the velocity model is indicated by the large residuals and *rms* values with both L2-norm and EDT, and, with EDT, by the asymmetry in residuals, the irregular *pdf*

shape, and the large V_{pdf} and I_{ell} values, as with the outlier data example (Example 6).

In effect, locations with an incorrect velocity model and with outlier data are mathematically similar, though in the former case all or most residuals may be large while in the latter case only a few will be large. It is difficult to distinguish between the two cases with the L2-norm because this algorithm seeks to best satisfy *all* of the observations simultaneously (cf., Eq. (7)) by balancing the distribution of positive and negative residual (cf., Fig. 8b and Fig. 10b). Thus, relative to the residuals corresponding to the correct location, the L2-norm solution damps and hides larger residuals at the expense of increasing small residuals. In contrast, EDT seeks to best satisfy the *most* pairs of observations (cf., Eq. (8)) and imposes no inher-



Earthquake Location, Direct, Global-Search Methods, Figure 10

Example 8: Incorrect velocity model. Locations using 50 P arrivals with the L2-norm and a time corrections, **b** no time corrections, and with EDT and **c** time corrections, **d** no time corrections. The locations without the ideal time corrections show the effect of an incorrect velocity model. The Hypoellipse linearized locations and ellipsoids do not differ markedly from the direct-search locations shown in this figure

ent constraint on the distribution of residuals. Thus with EDT the difference in number, magnitude and distribution of large residuals – few and large for the outlier case, many of similar magnitude and spatially correlated for the incorrect velocity model case – allows one, in principle, to distinguish between the two cases (cf., Fig. 8d and Fig. 10d). In addition, the size and complexity of the location *pdf*'s generally increases more rapidly with EDT than with the L2-norm as the solution quality decreases. Thus, with both the outlier and incorrect velocity model cases, the location results with the EDT likelihood function are more informative than with the L2-norm. However, location with the EDT likelihood function can become unstable (e.g. define only a local maximum of the *pdf*) for cases where the outlier data or velocity model errors lead to extreme complexity in the topology of the EDT location *pdf*.

Future Directions

There are various ways that direct, global-search location methodologies may evolve in the future. For example, the stability and completeness of the location and location *pdf* could be improved with the use of more complete data uncertainties, expressed as a *pdf*. These *pdf*'s may typically be irregular and asymmetric, and difficult to determine and parametrize. Currently, enumerated quality indications or, at best, simple normal distributions (describing Gaussian uncertainty) are used to describe the picking error.

Similarly, we have shown that earthquake location depends inherently on the velocity model adopted, but that no realistic uncertainties are associated with this model. Differences between the velocity model and the true Earth can result in complicated differences in ray-paths and

travel-times, which will depend strongly on the source and receiver positions. These complications, combined with the lack of knowledge about the true Earth, makes estimating true travel time uncertainties effectively impossible. However, it can be assumed that changes become progressively larger with increasing ray-path length. This effect could be accounted for approximately by travel-time uncertainties that increase with the ray-length or travel time. Instead of using a velocity model to generate travel times, another approach is to derive the required times from tables of empirically determined or corrected travel times (e.g. [41,43]). With this approach the travel-time uncertainties are estimated from timing information, with little or no direct use of velocity structures or ray paths.

We have described and illustrated the importance of the source-receiver geometry for locating earthquakes, notably with regards to constraining a compact and symmetric location *pdf*. Thus, improved constraint on event locations can be achieved through prior use of survey design techniques to select station sites. In a related manner, after an event occurs, these techniques could be employed dynamically to weight the available arrival times used for location with respect to the geometry of the available stations around the likely source region.

The demand for rapid, real-time location and earthquake early warning requires improvements in the integration, speed, quality and robustness of the phase arrival picking, phase association and event location procedures. Currently, development is progressing on integrated procedures which are evolutionary and probabilistic, using, for example, robust likelihood functions such as EDT and information from not-yet-triggered stations (e.g., [8,21,54,63,64]).

A current problem in direct-search location is how to describe in a standardized and compact way the sometimes topologically-complex location *pdf*. For example, such a description is needed if the *pdf* is to be included in standard earthquake catalogs and for rapid dissemination of probabilistic location information for earthquake early warning. More generally, making full use of the extensive information in direct-search location solutions will require new methods and procedures to store, distribute and analyze the location *pdf*, maximum likelihood hypocenter, arrival residuals and weights, and other statistics and quality indicators of the solutions.

The continuing increase in computer speed will allow application of direct-search inversion methods to relative location of ensembles of events and for joint epicentral determination in the near future. The use of these methods will be important to explore more completely the vast so-

lution space and better determine the error and resolution for such high-dimensional inverse problems.

The continuing increase in computer speed will also make practical earthquake location techniques using waveform recordings directly, without the intermediate stage of extracting phase arrival times. In these techniques, continuous waveform data streams are matched to synthetic Green's functions within a global-search over possible source locations and source parameters. This type of approach is used to locate previously unidentified earthquakes using low amplitude surface waves on off-line, continuous, broadband waveforms [15,68], and for automatic, real-time estimation of moment tensors and location from continuous broadband data streams (e.g., [25]). Waveform methods will likely be applied to earthquake location on local and regional scales as faster computers and more accurate 3D velocity models become available [81]; related applications using simple ray or acoustic theories to generate the Green's functions show promising results (e.g., [3]).

Bibliography

Primary Literature

1. Aki K, Richards PG (1980) Quantitative Seismology. Freeman, New York
2. Anderson K (1981) Epicentral location using arrival time order. Bull Seism Soc Am 71:541–545
3. Baker T, Granat R, Clayton RW (2005) Real-time Earthquake Location Using Kirchhoff Reconstruction. Bull Seism Soc Am 95:699–707
4. Billings SD (1994) Simulated annealing for earthquake location. Geophys J Int 118:680–692
5. Buland R (1976) The mechanics of locating earthquakes. Bull Seism Soc Am 66:173–187
6. Calvert A, Gomez F, Seber D, Barazangi M, Jabour N, Ibenbrahim A, Demnati A (1997) An integrated geophysical investigation of recent seismicity in the Al-Hoceima region of North Morocco. Bull Seism Soc Am 87:637–651
7. Červený V (2001) Seismic Ray Theory. Cambridge University Press, Cambridge
8. Cua G, Heaton T (2007) The Virtual Seismologist (VS) Method: a Bayesian Approach to Earthquake Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) Earthquake Early Warning Systems. Springer, Berlin
9. Curtis A (1999) Optimal experiment design: Cross-borehole tomographic examples. Geophys J Int 136:637–650
10. Curtis A (1999) Optimal design of focussed experiments and surveys. Geophys J Int 139:205–215
11. Curtis A (2004) Theory of model-based geophysical survey and experimental design Part A – Linear Problems. Lead Edge 23(10):997–1004
12. Curtis A (2004) Theory of model-based geophysical survey and experimental design Part B – Nonlinear Problems. Lead Edge 23(10):1112–1117

13. Curtis A, Michelini A, Leslie D, Lomax A (2004) A deterministic algorithm for experimental design applied to tomographic and microseismic monitoring surveys. *Geophys J Int* 157:595–606
14. Dreger D, Uhrhammer R, Pasyanos M, Frank J, Romanowicz B (1998) Regional and far-regional earthquake locations and source parameters using sparse broadband networks: A test on the Ridgecrest sequence. *Bull Seism Soc Am* 88:1353–1362
15. Ekström G (2006) Global Detection and Location of Seismic Sources by Using Surface Waves. *Bull Seism Soc Am* 96:1201–1212. doi:10.1785/0120050175
16. Font Y, Kao H, Lallemand S, Liu CS, Chiao LY (2004) Hypocentral determination offshore Eastern Taiwan using the Maximum Intersection method. *Geophys J Int* 158:655–675
17. Geiger L (1912) Probability method for the determination of earthquake epicenters from the arrival time only (translated from Geiger's 1910 German article). *Bull St Louis Univ* 8:56–71
18. Gentili S, Michelini A (2006) Automatic picking of P and S phases using a neural tree. *J Seism* 10:39–63. doi:10.1007/s10950-006-2296-6
19. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
20. Hammersley JM, Handscomb DC (1967) Monte Carlo Methods. Methuen, London
21. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An Automatic Processing System for Broadcasting Earthquake Alarms. *Bull Seism Soc Am* 95:708–718
22. Husen S, Smith RB (2004) Probabilistic Earthquake Relocation in Three-Dimensional Velocity Models for the Yellowstone National Park Region, Wyoming. *Bull Seism Soc Am* 94:880–896
23. Husen S, Kissling E, Deichmann N, Wiemer S, Giardini D, Baer M (2003) Probabilistic earthquake location in complex three-dimensional velocity models: Application to Switzerland. *J Geophys Res* 108:2077–2102
24. Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. *US Geol Surv Open-File Rep* pp 94–621
25. Kawakatsu H (1998) On the real-time monitoring of the long-period seismic wavefield. *Bull Earthq Res Inst* 73:267–274
26. Kennett BLN (1992) Locating oceanic earthquakes – the influence of regional models and location criteria. *Geophys J Int* 108:848–854
27. Kennett BLN (2006) Non-linear methods for event location in a global context. *Phys Earth Planet Inter* 158:46–54
28. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
29. Lahr JC (1999) HYPOELLIPSE: A Computer Program for Determining Local Earthquake Hypocentral Parameters, Magnitude, and First-Motion Pattern (Y2K Compliant Version) 1999 Version 1.0. US Geological Survey Open-File Report, pp 99–23. http://jclahr.com/science/software/hypoellipse/hypoel/hypoman/hypomst_pdf.pdf
30. Lepage GP (1978) A new algorithm for adaptive multi-dimensional integration. *J Comput Phys* 27:192–203
31. Lomax A (2005) A Reanalysis of the Hypocentral Location and Related Observations for the Great 1906 California Earthquake. *Bull Seism Soc Am* 91:861–877
32. Lomax A (2008) Location of the Focus and Tectonics of the Focal Region of the California Earthquake of 18 April 1906. *Bull Seism Soc Am* 98:846–860
33. Lomax A, Curtis A (2001) Fast, probabilistic earthquake location in 3D models using oct-tree importance sampling. *Geophys Res Abstr* 3:955. www.alomax.net/nllloc/octtree
34. Lomax A, Virieux J, Volant P, Berge C (2000) Probabilistic earthquake location in 3D and layered models: Introduction of a Metropolis-Gibbs method and comparison with linear locations. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
35. Lomax A, Zollo A, Capuano P, Virieux J (2001) Precise, absolute earthquake location under Somma-Vesuvius volcano using a new 3D velocity model. *Geophys J Int* 146:313–331
36. Maurer HR, Boerner DE (1998) Optimized and robust experimental design. *Geoph J Int* 132:458–468
37. Milne J (1886) *Earthquakes and Other Earth Movements*. Appleton, New York
38. Moser TJ, van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *J Geophys Res* 97:6563–6572
39. Moser TJ, Nolet G, Snieder R (1992) Ray bending revisited. *Bull Seism Soc Am* 82:259–288
40. Mosegaard K, Tarantola A (1995) Monte Carlo sampling of solutions to inverse problems. *J Geophys Res* 100:12431–12447
41. Myers SC, Schultz CA (2000) Improving Sparse Network Seismic Location with Bayesian Kriging and Teleseismically Constrained Calibration Events. *Bull Seism Soc Am* 90:199–211
42. Nicholson T, Gudmundsson Ó, Sambridge M (2004) Constraints on earthquake epicentres independent of seismic velocity models. *Geophys J Int* 156:648–654
43. Nicholson T, Sambridge M, Gudmundsson Ó (2004) Three-dimensional empirical traveltimes: construction and applications. *Geophys J Int* 156:307–328
44. Podvin P, Lecomte I (1991) Finite difference computations of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *Geophys J Int* 105:271–284
45. Press WH, Flannery BP, Saul AT, Vetterling WT (1992) *Numerical Recipes*, 2nd edn. Cambridge Univ Press, New York
46. Presti D, Troise C, De Natale G (2004) Probabilistic Location of Seismic Sequences in Heterogeneous Media. *Bull Seism Soc Am* 94:2239–2253
47. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
48. Rabinowitz N (2000) Hypocenter location using a constrained nonlinear simplex minimization method. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
49. Rabinowitz N, Steinberg DM (2000) A statistical outlook on the problem of seismic network configuration. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
50. Rawlinson N, Sambridge M (2004) Wave front evolution in strongly heterogeneous layered media using the fast marching method. *Geophys J Int* 156:631–647
51. Rawlinson N, Sambridge M (2004) Multiple reflection and transmission phases in complex layered media using a multi-stage fast marching method. *Geophys* 69:1338–1350
52. Reid HF (1910) *The Mechanics of the Earthquake*. Vol II of: *The California Earthquake of 18 April 1906*. Report of the State Earthquake Investigation Commission, Lawson AC

- (Chairman). Carnegie Institution of Washington Publication, vol 87 (reprinted 1969)
53. Rothman DH (1985) Nonlinear inversion, statistical mechanics, and residual statics estimation. *Geophysics* 50:2784–2796
 54. Rydelek P, Pujol J (2004) Real-Time Seismic Warning with a Two-Station Subarray. *Bull Seism Soc Am* 94:1546–1550
 55. Sambridge M (1998) Exploring multi-dimensional landscapes without a map. *Inverse Probl* 14:427–440
 56. Sambridge M (1999) Geophysical inversion with a Neighbourhood algorithm, vol I. Searching a parameter space. *Geophys J Int* 138:479–494
 57. Sambridge M (1999) Geophysical inversion with a neighbourhood algorithm, vol II. Appraising the ensemble. *Geophys J Int* 138:727–746
 58. Sambridge M (2003) Nonlinear inversion by direct search using the neighbourhood algorithm. In: *International Handbook of Earthquake and Engineering Seismology*, vol 81B. Academic Press, Amsterdam, pp 1635–1637
 59. Sambridge M, Drijkoningen G (1992) Genetic algorithms in seismic waveform inversion. *Geophys J Int* 109:323–342
 60. Sambridge M, Gallagher K (1993) Earthquake hypocenter location using genetic algorithms. *Bull Seism Soc Am* 83:1467–1491
 61. Sambridge M, Kennett BLN (1986) A novel method of hypocentre location. *Geophys J R Astron Soc* 87:679–697
 62. Sambridge M, Mosegaard K (2002) Monte Carlo Methods In Geophysical Inverse Problems. *Rev Geophys* 40:1009–1038
 63. Satriano C, Lomax A, Zollo A (2007) Optimal, Real-time Earthquake Location for Early Warning. In: Gasparini P, Gaetano M, Jochen Z (eds) *Earthquake Early Warning Systems*. Springer, Berlin
 64. Satriano C, Lomax A, Zollo A (2007) Real-time evolutionary earthquake location for seismic early warning. *Bull Seism Soc Am* 98:1482–1494
 65. Sen M, Stoffa PL (1995) Global optimization methods in geophysical inversion. Elsevier, Amsterdam, p 281
 66. Sethian JA (1999) Level set methods and fast marching methods. Cambridge University Press, Cambridge
 67. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423
 68. Shearer PM (1994) Global seismic event detection using a matched filter on long-period seismograms. *J Geophys Res* 99:13,713–13,735
 69. Shearer PM (1997) Improving local earthquake locations using the L1 norm and waveform cross correlation: Application to the Whittier Narrows, California, aftershock sequence. *J Geophys Res* 102:8269–8283
 70. Steinberg DM, Rabinowitz N, Shimshoni Y, Mizrahi D (1995) Configuring a seismographic network for optimal monitoring of fault lines and multiple sources. *Bull Seism Soc Am* 85:1847–1857
 71. Stummer P, Maurer HR, Green AG (2004) Experimental Design: Electrical resistivity data sets that provide optimum subsurface information. *Geophysics* 69:120–139
 72. Tarantola A (1987) Inverse problem theory: Methods for data fitting and model parameter estimation. Elsevier, Amsterdam
 73. Tarantola A (2005) Inverse Problem Theory and Methods for Model Parameter Estimation. SIAM, Philadelphia
 74. Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys Res* 50:159–170
 75. Thurber CH, Kissling E (2000) Advances in travel-time calculations for three-dimensional structures. In: Thurber CH, Rabinowitz N (eds) *Advances in Seismic Event Location*. Kluwer, Amsterdam
 76. Uhrhammer RA (1980) Analysis of small seismographic station networks. *Bull Seism Soc Am* 70:1369–1379
 77. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. *Bull Seism Soc Am* 77:972–986
 78. van den Berg J, Curtis A, Trampert J (2003) Bayesian, nonlinear experimental design applied to simple, geophysical examples. *Geophys J Int* 55(2):411–421. Erratum: 2005. *Geophys J Int* 161(2):265
 79. Vidale JE (1988) Finite-difference calculation of travel times. *Bull Seism Soc Am* 78:2062–2078
 80. Winterfors E, Curtis A (2007) Survey and experimental design for nonlinear problems. *Inverse Problems* (submitted)
 81. Wither M, Aster R, Young C (1999) An automated local and regional seismic event detection and location system using waveform correlation. *Bull Seism Soc Am* 8:657–669
 82. Withers M, Aster R, Young C, Beiriger J, Harris M, Moore S, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bull Seism Soc Am* 88:95–106
 83. Wittlinger G, Herquel G, Nakache T (1993) Earthquake location in strongly heterogeneous media. *Geophys J Int* 115:759–777
 84. Zhou H (1994) Rapid 3-D hypocentral determination using a master station method. *J Geophys Res* 99:15439–15455

Books and Reviews

- Gasparini P, Gaetano M, Jochen Z (eds) (2007) *Earthquake Early Warning Systems*. Springer, Berlin
- Lee WHK, Stewart SW (1981) *Principles and applications of microearthquake networks*. Academic Press, New York
- Thurber CH, Rabinowitz N (eds) (2000) *Advances in Seismic Event Location*. Kluwer, Amsterdam

Earthquake Magnitude

PETER BORMANN, JOACHIM SAUL

GeoForschungsZentrum Potsdam, Potsdam, Germany

Article Outline

Glossary

Definition of the Subject

Introduction to Common Magnitude Scales:

Potential and Limitations

Common Magnitude Estimates

for the Sumatra 2004 M_w 9.3 Earthquake

Magnitude Saturation and Biases

Due to Earthquake Complexity

Proposals for Faster Magnitude Estimates

of Strong Earthquakes

Future Requirements and Developments

Bibliography

Glossary

Technical terms that are written in the text in italics are explained in the Glossary.

Corner frequency The frequency f_c at which the curve that represents the Fourier amplitude spectrum of a recorded seismic signal abruptly changes its slope (see Fig. 5). For earthquakes, this frequency is related to the fault size, rupture velocity, rupture duration and stress drop at the source. Also the frequency at which the magnification curve of a recording system (e.g., Fig. 3) changes its slope.

Dispersion Frequency-dependence of the wave propagation velocity. Whereas seismic body-waves show virtually no dispersion, it is pronounced for seismic surface waves. It causes a significant stretching of the length of the surface-wave record and the rather late arrival of its largest amplitudes (Airy phases) from which the surface-wave magnitude M_S and the mantle magnitude M_m , respectively, are determined.

Earthquake size A frequently used, but not uniquely defined term. It may be related – more or less directly – to either the geometric-kinematic size of an earthquake in terms of area and slip of the fault or to the *seismic energy* radiated from a seismic source and its potential to cause damage and casualty (moment or energy *magnitude*).

Earthquake source In general terms, the whole area or volume of an *earthquake* rupture where seismic body waves are generated and radiated outwards. More specifically, one speaks either of the *source mechanism* or the source location. The latter is commonly given as earthquake hypocenter (i.e. the location at the source depth h from where the seismic rupture, collapse or explosion begins) or as the point on the Earth's surface vertically above the hypocenter, called the epicenter. Earthquakes at $h < 70$ km are shallow, those at larger depth either intermediate (up to $h = 300$ km) or deep earthquakes ($h = 300$ – 700 km). The determination of the geographical coordinates latitude φ , longitude λ , and focal depth h , is the prime task of seismic source location. However, for extended seismic sources, fault ruptures of great earthquakes in particular, the hypocenter is generally not the location of largest fault slip and/or seismic moment/energy release and the epicenter is then also not the location where the strongest ground shaking is felt. The locations of largest effects may be dozens of kilometers in space and many seconds to minutes in time away from the hypocenter or epicenter, respectively.

Fundamental modes The longest period oscillations of the whole Earth with periods of about 20 min (spheroidal mode), 44 min. (toroidal mode) and some 54 min (“rugby” mode), excited by great earthquakes.

Magnitude A number that characterizes the relative *earthquake size*. It is usually based on measurement of the maximum motion recorded by a seismograph (sometimes for waves of a particular type and frequency) and corrected for the decay of amplitudes with epicenter distance and source depth due to geometric spreading and attenuation during wave propagation. Several magnitude scales have been defined. Some of them show *saturation*. In contrast, the moment magnitude (M_w), based on the concept of *seismic moment*, is uniformly applicable to all earthquake sizes but is more difficult to compute than the other types, similarly the energy magnitude, M_e , which is based on direct calculation of the *seismic energy* E_s from broadband seismic records.

Saturation (of magnitudes) Underestimation of *magnitude* when the duration of the earthquake rupture significantly exceeds the seismic wave period at which the magnitude is measured. The shorter this period, the earlier respective magnitudes will saturate (see relation (13) and Figs. 4 and 5).

Seismic energy Elastic energy E_s (in joule) generated by, and radiated from, a seismic source in the form of seismic waves. The amount of E_s is generally much smaller than the energy associated with the non-elastic deformation in the seismic source (see *seismic moment* M_o). The ratio $E_s/M_o = (\Delta\sigma/2\mu) = \tau_a/\mu$, i.e., the seismic energy released per unit of M_o , varies for earthquakes in a very wide range between some 10^{-6} and 10^{-3} , depending on the geologic-tectonic environment, type of *source mechanism* and related stress drop $\Delta\sigma$ or apparent stress τ_a .

Seismic moment M_o A special measure of earthquake size. The moment tensor of a shear rupture (see *earthquake source*) has two non-zero eigenvalues of the amount $M_o = \mu \bar{D} F_a$ with μ -shear modulus of the ruptured medium, \bar{D} -average source dislocation and F_a -area of the ruptured fault plane. M_o is called the scalar seismic moment. It has the dimension of Newton meter (Nm) and describes the total non-elastic (i.e., ruptural and plastic) deformation in the seismic source volume. Knowing M_o , the moment *magnitude* M_w can be determined via Eq. (11).

Source mechanism Depending on the orientation of the earthquake fault plane and slip direction in space, one discerns different source mechanisms. Strike-slip faults are vertical (or nearly vertical) fractures along

which rock masses have mostly shifted horizontally. Dip-slip faults are inclined fractures. If the rock mass above an inclined fault moves down (due to lateral extension) the fault is termed normal, whereas, if the rock above the fault moves up (due to lateral compression), the fault is termed reverse (or thrust). Oblique-slip faults have significant components of both slip styles (i. e., strike-slip and dip-slip). The greatest earthquakes with the largest release of seismic moment and the greatest potential for generating tsunamis are thrust faults in subduction zones where two of Earth's lithosphere plates (e. g., ocean–continent or continent–continent) collide and one of the two plates is subducted underneath the overriding plate down into the Earth's mantle. Different source mechanisms are characterized by different radiation patterns of seismic wave energy.

Transfer function The transfer function of a seismic sensor-recorder system (or of the Earth medium through which seismic waves propagate) describes the frequency-dependent amplification, damping and phase distortion of seismic signals by a specific sensor-recorder (or medium). The modulus (absolute value) of the transfer function is termed the amplitude-frequency response or, in the case of seismographs, also magnification curve (see Fig. 3).

Definition of the Subject

Besides earthquake location (i. e., the determination of the geographical coordinates of the epicenter, the hypocenter depth and the origin time; for definition of these terms see *earthquake source* in the Glossary), the *magnitude* is the most frequently determined and commonly used parameter to characterize an earthquake. Despite its various imperfections, it provides important information concerning the earthquake source spectrum at the period where the magnitude is measured and current source theories (cf. [3]) allow one to understand differences in the source spectra of different earthquakes in terms of source dimension and stress drop, i. e., the difference between the stress level before and after the earthquake. Via various empirical relations, magnitudes enable estimates of the *seismic moment* and the *seismic energy* released by the earthquake. These parameters are important in the discussion of various global problems such as the seismic slip rates between lithosphere plates and the excitation of Chandler Wobble [25]. Besides these more academic issues, magnitude values have an immense practical value in providing:

- a) Rapid simple parameter estimates of the strength of an earthquake that can help to realistically assess the re-

lated ground shaking or tsunami potential and thus assist efficient disaster management response;

- b) Mass data in earthquake catalogs and data banks, covering long time periods over many decades – and hopefully centuries in future, which allows one to assess the seismic activity and related hazards of Earth's regions and their possible variability in space and time. This is not only of high scientific interest, but also the very basis for realistic long-term disaster preparedness and risk mitigation efforts.

The term magnitude and the basic method of its determination were introduced by Charles F. Richter in 1935 [71]. He intended to compare the relative *earthquake size* in southern California in terms of differences in the maximum amplitudes A recorded at a network of seismic stations that were equipped with standard short-period Wood–Anderson (WA) torsion seismometers.

The WA seismometer response is depicted in Fig. 3 and Fig. 1 shows a WA record and magnitude measurement example. In order to make amplitudes recorded by stations at different epicentral distances D from the earthquake comparable, Richter had to compensate for the amplitude decay with D using an appropriate correction term $-A_o(D)$. Since the strength and thus the radiated amplitudes of earthquakes vary in a wide range Richter defined his local magnitude scale M_L , determined from records at source distances up to 600 km, as follows:

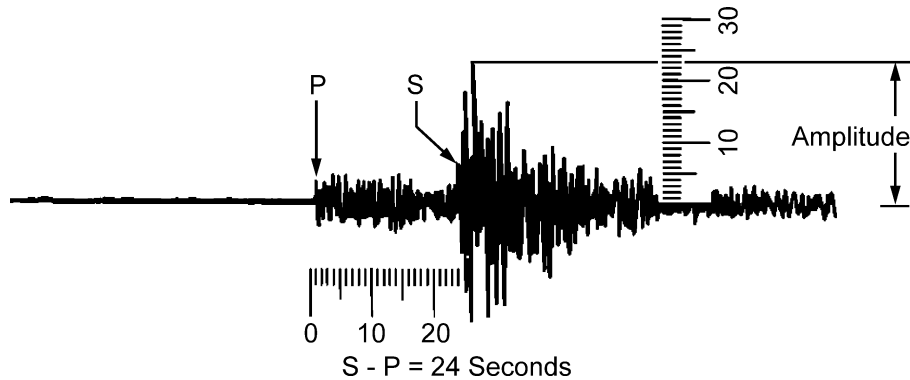
“The magnitude of any shock is taken as the logarithm of the maximum trace amplitude, expressed in microns, with which the standard short-period torsion seismometer ... would register that shock at an epicentral distance of 100 km.”

Thus:

$$M_L = \log A_{\max} - \log A_o(D) . \quad (1)$$

According to the above definition, an amplitude of $1 \mu\text{m}$ in a WA record at a distance $D = 100 \text{ km}$ from the epicenter would correspond to $M_L = 0$. Amplitude means in (1) and the following text either the center-to-peak or half of the peak-to-trough amplitude.

Wood–Anderson (WA) seismographs record horizontal short-period ground motions with an amplification of only about 2080 times [82]. Modern electronic seismographs may achieve magnifications larger than 10^6 and thus are able to record local earthquakes with even negative magnitudes, down to about -2 . The largest values determined with the M_L scale are around seven. Later it was found that all magnitudes derived from short-period waves (typically with periods $T < 3 \text{ s}$) show *saturation*



Earthquake Magnitude, Figure 1

Record of a short-period Wood-Anderson seismograph (frequency-magnification curve see Fig. 3) of a local earthquake. *P* marks the onset of the first arriving longitudinal *P* wave, and *S* the onset of the much stronger secondary, transverse polarized shear wave. Note the long tail of coda-waves following *S*. From the time difference $S - P = 24$ s follows a hypocentral distance $R = 190$ km. The maximum record amplitude is $A_{\max} = 23$ mm. Applying the amplitude-distance correction $-\log A_0(190 \text{ km}) = 3.45$ according to Richter [72] results in a magnitude $M_L = 4.8$

(see Glossary, Fig. 4 and Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”). Therefore, it was necessary to develop complementary magnitude scales that use medium to long-period ($T \approx 5 \text{ s} - 30 \text{ s}$) as well as very long-period waves ($T \approx 50 \text{ s} - 3000 \text{ s}$) in order to enable less or non-saturating magnitude estimates (see Sect. “Introduction to Common Magnitude Scales: Potential and Limitations”). For the so far strongest instrumentally recorded earthquake (Chile 1960) a value of $M = 9.5$ was determined that way. Accordingly, instrumental seismic monitoring currently covers the magnitude range of about $-2 \leq M < 10$. This roughly corresponds to ruptures of some millimeters to more than 1000 km long. They radiate approximately the same amount of seismic wave energy E_s as well-contained underground explosions with yields ranging from a few milligrams (10^{-9} t) to several 10 to 100 Gt ($1 \text{ Gt} = 10^9 \text{ t}$) Trinitrotoluol (TNT) equivalent, thus covering about 20 orders in energy. Earthquakes with magnitudes around four may cause only minor local damage, those with magnitudes > 6 heavy damage, and those with magnitudes > 7 already widespread devastating damage. Shallow submarine earthquakes with magnitudes > 7 may generate significant local tsunamis with damage potential to nearby shores whereas those with magnitudes > 8.5 may stimulate ocean-wide tsunamis causing destruction and casualties even at shores thousands of kilometers away from such earthquakes.

In order to measure and classify *earthquake size* in the wide range of magnitudes from about -2 to < 10 and satisfy specific requirements in research and application which are based on magnitude data, it was indispensable

to develop different magnitude scales that are complementary, but properly scaled to the original Richter M_L . Thus, there exists today a host of magnitude scales applicable in a wide range of source distances from less than 1 km up to more than 10,000 km. These scales, their specifics, potential and limitations are discussed in detail (with many reference given) in Chapter 3 of the IASPEI New Manual of Seismological Observatory Practice [6]. The early pioneers of magnitude scales, Beno Gutenberg and Charles Richter, had hoped that different magnitude scales could be cross-calibrated to yield a unique value for any given earthquake (cf. [25,30]). In their joint book [29] “Seismicity of the Earth” (1954; first edition 1949) and later in Richter’s [72] famous text book “Elementary Seismology” as well as in Duda [22] only one magnitude value M was given per earthquake. However, this approach proved only partially realistic under certain conditions and within limited magnitude ranges because of the often significant differences in measurement procedures as well as period and bandwidth ranges used in later magnitudes scales. Decades later it took significant efforts (cf. [1,2,25]) to reconvert these M values, which turned out to be not even compatible (cf. [25]) into their original body or surface wave magnitudes in order to get values that agree with the original definition of these specific magnitude scales and can be compared with current data of the same type.

In general, such magnitude conversion relations strongly depend on initial data errors and the type of least-square regression procedure applied [11,14]. Moreover, the latter have often not been interpreted and used in a correct way. This may result in the case of noisy mag-

nitude data for events at the upper and lower end of the investigated magnitude range, in conversion errors of more than 0.5 magnitude units (m.u.) with serious consequences on seismic hazard estimates based on such converted magnitudes (cf. [7,11,14,15]). Moreover, magnitude values determined within the *saturation* range of a given scale cannot reliably be converted via empirical regression relations into the equivalent magnitude values of another less or non-saturating magnitude scale (see Fig. 4 and [44]). Furthermore, some magnitudes relate best to the released *seismic energy* while others are scaled to the static *seismic moment*, i. e., they measure equally important but fundamentally different physical aspects of the source and the radiated seismic waves and may differ by sometimes more than 1 m.u. Thus there is no way to characterize *earthquake size* in all its different aspects by just a single magnitude value. Proper interpretation and use of different types of magnitude data, however, requires one to understand the physics behind such values and how these may be affected by the complexity and duration of the earthquake rupture process. Further, this necessitates one to discriminate unambiguously the different types of magnitude values by using a unique nomenclature and to assure that magnitude values published with a given nomenclature have been determined with an internationally agreed standard procedure. With this in mind, the most important magnitude scales and related problems are summarized in Sects. “[Introduction to Common Magnitude Scales: Potential and Limitations](#)” and “[Common Magnitude Estimates for the Sumatra 2004 \$M_w\$ 9.3 Earthquake](#)”.

Introduction to Common Magnitude Scales: Potential and Limitations

Magnitude Scales Used in the Local and Regional Distance Range ($D < 2000$ km)

The original Richter local magnitude scale for Southern California [71] has been further developed since its invention [38]. In its expanded form (with the nomenclature M_L common in the United States), the following relation now holds:

$$M_L = \log_{10}(A_{\max}) + 1.11 \log_{10} R + 0.00189 R - 2.09 \quad (2)$$

with R = distance from the station to the hypocenter in kilometers and A_{\max} = maximum trace amplitude in nanometers (instead of μm in a WA record). This amplitude is measured on the output from a horizontal-component seismograph that is filtered so that the response of the seismograph/filter system replicates that of a WA standard seismograph but with a static magnification of one. The

underlying procedure of M_L determination according to relation (2) was adopted by the International Association of Seismology and Physics of the Earth's Interior (IASPEI) in 2004 as the standard procedure for determining local magnitudes in the distance range up to typically less than 1000 km [42]. For earthquakes in the Earth's crust of regions with attenuation properties that differ from those of coastal California, and for measuring M_L with vertical component seismographs, the standard equation takes the form:

$$M_L = \log_{10}(A_{\max}) + F(R) + G \quad (3)$$

where $F(R)$ is an R -dependent calibration function and G a constant which have to compensate for different regional attenuation and/or for any systematic biases of amplitudes measured on vertical instead on horizontal seismographs. Examples of regional M_L calibration functions developed for different parts of the world have been compiled by Bormann (Chap. 3, p. 26, and DS 3.1 in [6]).

A few decades ago, analog seismic records prevailed. They had a rather limited dynamic range of only some 40 dB. This caused record traces often to go off-scale when stronger seismic events were recorded at local or regional distances. Then A_{\max} could not be measured. Yet, it was found that the duration \mathbf{d} of the coda that follows A_{\max} with exponentially decaying amplitudes (see Fig. 1) increases with magnitude and distance D . On this basis, local duration magnitude formulas of the following general form

$$M_d = a + b \log \mathbf{d} + cD \quad (4)$$

have been developed with a , b and c being coefficients to be determined locally. When using only recordings at distances $D < 100$ km the distance term cD is not even needed. However, crustal structure, scattering and attenuation conditions vary from region to region. Moreover, the resulting specific equations will also depend on the chosen definition for \mathbf{d} , the local signal-to-noise (SNR) conditions and the sensor sensitivity at the considered seismic station(s) of a network. Therefore, M_d scales have to be determined locally for a given source-network configuration and scaled to the best available amplitude-based M_L scale.

Nowadays digital recorders with large usable dynamic range of about 140 dB are common. Thus even sensitive modern broadband seismographs remain on scale when recording local or regional earthquakes up to $M \approx 7$. This reduces the need for M_d scales. Moreover, the increasing availability of modern strong-motion (SM) recorders with comparably large dynamic range, which will not clip even in the case of very strong nearby earthquakes, have led to

the development of (partially) frequency-dependent M_L^{SM} scales. They are usually based on the calculation of synthetic WA seismograph outputs from strong-motion accelerograms [35,54].

Also, amplitudes of short-period L_g waves with periods around 1 s are sometimes used to determine magnitudes, termed $m_b(L_g)$. L_g waves travel with group velocities of 3.6 to 3.2 km/s and arrive after the (secondary, shear) S wave onset (Fig. 1). They propagate well in continental platform areas. Recently, the IASPEI [42] adopted a measurement procedure for $m_b(L_g)$ as international standard, which had been developed for eastern North America [62] with the aim to improve yield estimates of Nevada Test Site explosions. However, as for all other local or regional magnitude scales, the calibration term is strongly influenced by the local/regional geologic-tectonic conditions in the Earth's crust and requires a proper scaling to this standard, when applied to other areas than eastern North America.

Tsuboi developed for the Japan Meteorological Agency (JMA) in 1954 [79] a magnitude formula for shallow earthquakes (depth $h < 60$ km) that have been recorded at epicentral distances D up to 2000 km:

$$M_{JMA} = \log_{10} A_{\max} + 1.73 \log_{10} D - 0.83. \quad (5)$$

A_{\max} is the largest ground motion amplitude (in μm) in the total event record of a seismograph with an eigenperiod of 5 s. If horizontal seismographs are used then $A_{\max} = (A_{NS}^2 + A_{EW}^2)^{1/2}$ with A_{NS} and A_{EW} being half the maximum peak-to-trough amplitudes measured in the two horizontal components. This formula was devised to be equivalent to the medium to long-period Gutenberg–Richter [29] magnitude M . Therefore, M_{JMA} agrees rather well with the seismic moment magnitude M_w . The average difference is less than 0.1 in the magnitude range between 4.5 and 7.5 but becomes > 0.5 for $M_w > 8.5$ (see Fig. 4). Katsumata [49,50] has later modified the M_{JMA} formula for earthquakes deeper than 60 km.

Another, more long-period regional moment magnitude scale, termed M_{wp} , has been developed in Japan as well [80]. It provides quick and less saturating magnitude estimates for tsunami early warning. Velocity-proportional records are twice integrated and approximately corrected for geometrical spreading and an average P -wave radiation pattern (see *source mechanism*) to obtain estimates of the scalar seismic moment M_0 at each station. Usually the first maximum in the integrated displacement trace, called “moment history” $M_0(t)$, is assumed to represent M_0 . From these M_0 values moment magnitudes M_w are then calculated for each station according to

Eq. (11) and averaged. M_{wp} results from adding an empirically derived correction of 0.2 m.u. to the averaged station M_w [80]. Finally, a magnitude-dependent correction is applied to M_{wp} [86] in order to get an even better estimate of the recognized “authoritative” Global Centroid Moment Tensor magnitude M_w (GCMT) which is calculated according to the Harvard procedure [23] and now published under [41].

The M_{wp} concept was originally developed for earthquakes at $5^\circ \leq D^\circ \leq 15^\circ$, but can be applied for $M_w < 7.5$ (down to about $M_w \approx 5$) even to shorter local distances as long as this distance is significantly larger than the rupture length. Later the M_{wp} procedure has been adopted for application to records of deep and teleseismic earthquakes as well [81]. M_{wp} estimates are standard routine in Japan, at the Alaska and the Pacific Tsunami Warning Centers (ATWC and PTWC), and the National Earthquake Information Center (NEIC) of the United States Geological Survey (USGS). However, each of these centers use slightly different procedures. Values for most strong earthquakes are usually available some 10 to 15 min after the origin time (OT). On average M_{wp} data scale well with M_w . Exceptions, however, are extremely slow or very large complex earthquakes. Then M_{wp} is usually too small, up to about 1 m.u.

In recent years great attention is paid to the development of even more rapid earthquake early warning systems (EWS). They aim at event location and magnitude estimates from the very first few seconds of broadband acceleration, velocity or displacement records and within about 10 to 30 s after origin time (OT) of strong damaging earthquakes on land. These data are to be used for instantaneous public alarms and/or automatically triggered risk mitigation actions after strong earthquakes with damage potential. The goal is to minimize the area of “blind zones” which are left without advanced warning before the arrival of the S waves which have usually the largest strong-motion amplitudes (see Fig. 1). This necessitates very dense and robust local seismic sensor networks within a few tens of kilometers from potentially strong earthquake sources. Such networks are at present available only in very few countries, e. g. in Japan, Taiwan, Turkey, and Italy.

Their principles of rapid magnitude estimates differ from those mentioned above and below and the data analysis from such systems is largely based on still much debated concepts such as the hypothesis of the deterministic nature of earthquake rupture [66,73]. Data presented in [66] seem to suggest that in the range $3.0 < M$ (not specified) < 8.4 the magnitude can be estimated with an average absolute deviation of 0.54 m.u. from the maximum period within the initial 4 s of the first arriving

(primary, longitudinal) P wave when many low-pass filtered velocity records within 100 km from the epicenter are available. However, for $M > 6$ the systematic increase of these greatly scattering periods becomes rather questionable. When analyzing waveforms of the Japanese Hi-net seismic network [73], it could not be confirmed that such a dominant frequency scaling with magnitude exists. Also Kanamori [46], together with Nakamura [60,61], one of the fathers of this idea, expressed much more caution about the prospects of this method after he had run, together with Wu [89], an experiment with the Taiwan EWS. For each event they analyzed the first 3 s of at least eight P -wave records at epicentral distances < 30 km. They knew that: "... the slip motion is in general complex and even a large event often begins with a small short-period motion, followed by a long-period motion. Consequently, it is important to define the average period during the first motion." (termed τ_c in [46,89]). However, after applying the τ_c concept to the Taiwan EWS they concluded: "For EWS applications, if $\tau_c < 1$ s, the event has already ended or is not likely to grow beyond $M > 6$. If $\tau_c > 1$ s, it is likely to grow, but how large it will eventually become, cannot be determined. In this sense, the method provides a threshold warning". Thus it seems that these new concepts work reasonably well only for earthquakes with $M < 6.5$ and thus total rupture durations that are according to Eq. (13) on average not more than about 2–3 times the measurement time windows of 3 s or 4 s used in [46,66,89]. Nakamura and Saita [61] reported data from a much smaller set of events ($N = 26$) recorded at local distances in the range $4.6 < M < 6.9$. We calculated the average absolute deviation of their rapid UrEDAS system magnitudes (0.47 m.u.) from the official magnitudes M_{JMA} published later by the Japan Meteorological Agency. This error decreases to 0.32 m.u. when only earthquakes with magnitudes up to $M_{JMA} = 6.0$ are considered. This seems to support our assessment that the reliability of real-time EMS magnitudes decreases rapidly if the analyzed time window is much shorter than the rupture duration.

Magnitude Scales Used in the Teleseismic Distance Range ($D > 2000$ km)

Ten years after the introduction of the local magnitude M_L , Beno Gutenberg [26,27,28] extended the concept of magnitude determination to teleseismic distances larger than about 1000–2000 km. He used both records of seismic waves that propagate along the Earth's surface (or near to it with a period-dependent penetration depth) and waves which travel through the Earth. Accordingly, the former are termed surface waves and the latter body waves. For

the surface-wave magnitude Gutenberg [28] gave the following relation:

$$M_S = \log_{10} A_{H\max} + 1.656 \log D^\circ + 1.818 \quad (6)$$

with $A_{H\max}$ = maximum "total" horizontal displacement amplitude of surface-waves in μm for periods around 20 ± 2 s measured in the distance range $15^\circ < D^\circ < 130^\circ$ ($1^\circ = 111,195$ km).

While the original Richter M_L and Gutenberg M_S magnitudes were calculated from the maximum ground displacement amplitudes, Gutenberg [26,27,30] proposed to determine the body-wave magnitudes m_B from the relation:

$$m_B = \log_{10} (A/T)_{\max} + Q(D^\circ, h), \quad (7)$$

i. e., by measuring the maximum ratio of ground displacement amplitude A (in μm) divided by the related period T (in s). A/T is equivalent to measuring the maximum ground motion velocity $A_{v\max}/2\pi$ which is proportional to the square root of seismic energy, i. e. $\sqrt{E_s}$. Thus the magnitude becomes a measure of the elastic kinetic wave energy radiated by an earthquake. Only in this way comparable magnitude data could be obtained for different types of body waves and measurements at different sites. Another great advantage of m_B is that it permits magnitude estimates also from intermediate and deep earthquake sources, which produce only weak or no surface waves at all. Empirical relationships permit estimating E_s (in units of Joule) from body-wave magnitude m_B [30]

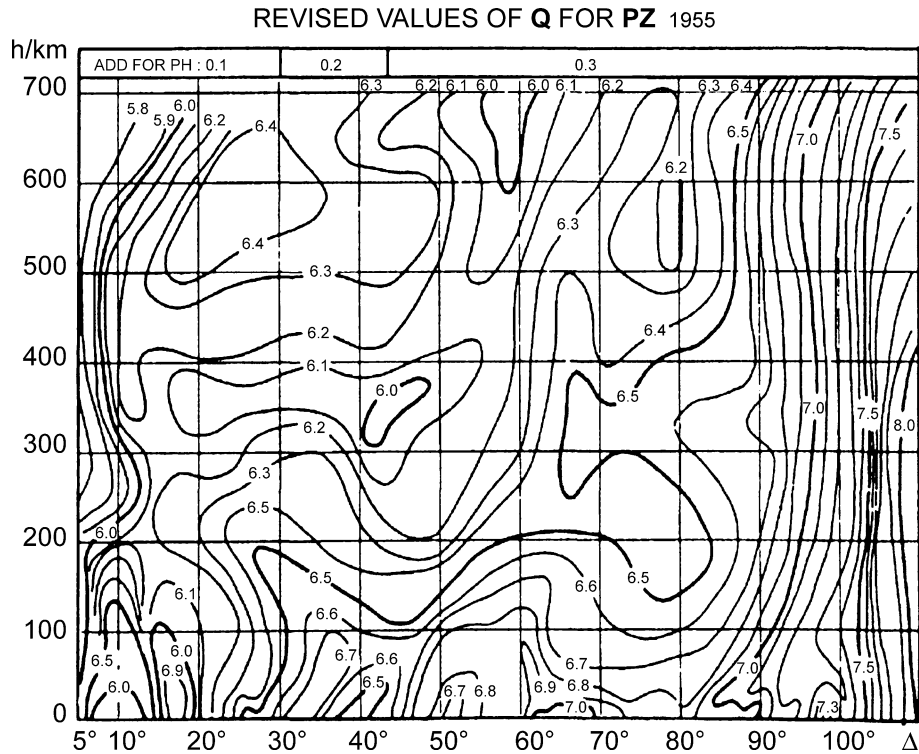
$$\log_{10} E_s = 2.4 m_B - 1.2 \quad (8)$$

or surface-wave magnitude M_S [72]

$$\log_{10} E_s = 1.5 M_S + 4.8. \quad (9)$$

Accordingly, an increase by 1 m.u. in m_B and M_S corresponds to an increase of radiated seismic energy by about 250 and 30 times, respectively.

Revised empirical distance-depth corrections for the calibration of body-wave magnitudes, so-called Q -functions, were published in 1956 by Gutenberg and Richter [30]. They are given as separate tables and charts for the body-wave phases P , PP (a P wave reflected at the surface of the Earth about the half way between source and station) and S . They are still in use, especially Q_{PV} for calibrating amplitude measurements made on vertical component P -wave records (Fig. 2). However, for epicenter distances between 5° and 20° these calibration values are not reliable enough for global application. In this range



Earthquake Magnitude, Figure 2

Calibration values $Q(D^\circ, h)$ for vertical (Z) component P -wave amplitudes depending on epicentral distance $D^\circ = \Delta$ and source depth h as used in the calculation of body-wave magnitudes m_b and m_B according to Gutenberg and Richter, 1956 [30]

the wave propagation is strongly affected by regional variations of the structure and properties of the Earth's crust and upper mantle. And for $D > 100^\circ$ the P -wave amplitudes decay rapidly because of the propagation of P waves is influenced by the Earth's core (so-called core shadow). Therefore, in agreement with current IASPEI recommendations [42], m_B and its short-period complement m_b (see below), should be determined by using Q_{PV} only between $21^\circ \leq D \leq 100^\circ$.

These body-wave magnitude calibration functions had been derived from amplitude measurements made mostly on medium-period broadband displacement records which dominated during the first half of the 20th Century at seismological stations. Their period-dependent magnification curve resembled more or less that of the classical standard seismograph type C shown in Fig. 3, although for some of these instruments the roll-off of the amplification occurred already at periods $T > 10$ s.

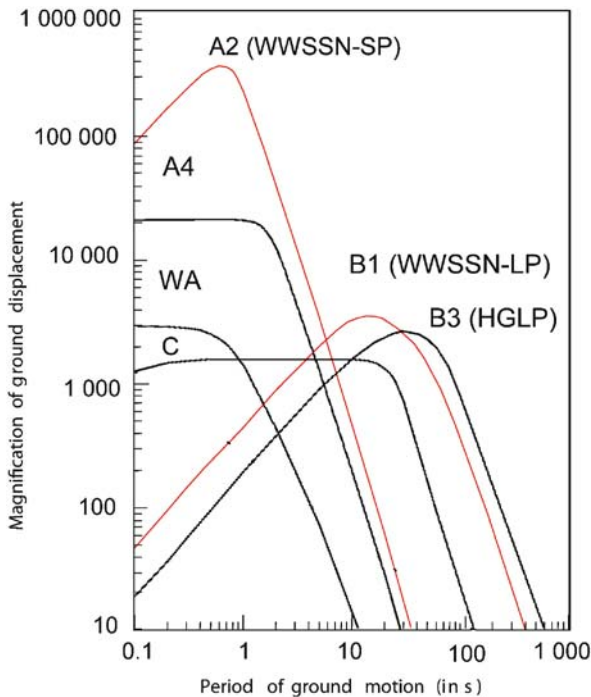
Another, so-called Prague–Moscow formula for surface-wave magnitudes was proposed in 1962 by Vaněk et al. [84]. It is based on the measurement of $(A/T)_{\max}$ in records of shallow earthquakes ($h < 60$ km) in wide pe-

riod and distance ranges ($3 \text{ s} < T < 30 \text{ s}$; $2^\circ \leq D^\circ \leq 160^\circ$):

$$M_S = \log_{10}(A/T)_{\max} + 1.66 \log_{10} D^\circ + 3.3. \quad (10)$$

This relationship, which is – as Eq. (7) – more directly related to E_s , was adopted by the IASPEI in 1967 as international standard.

The NEIC adopted Eq. (10), but continues to limit the range of application to distances between $20^\circ \leq D^\circ \leq 160^\circ$ and displacement amplitudes in the very limited period range as in formula (6) although Soloviev [74] had shown already in 1955 that $(A/T)_{\max}$ is a stable quantitative feature of surface waves whatever the period of their maximum at all epicentral distances. Also theory has confirmed [63] that using the ratio (A/T) is a partial and ad hoc compensation for a large number of frequency-dependent terms ignored in (10). In fact, the periods at the surface-wave maximum used for M_S determination vary in a wide range between some 3 s and 25 s and show – despite large scatter – a clear distance dependence [84,87]. Therefore, several authors [36,70] showed that using Eq. (10) only for amplitude readings around 20 s results in system-

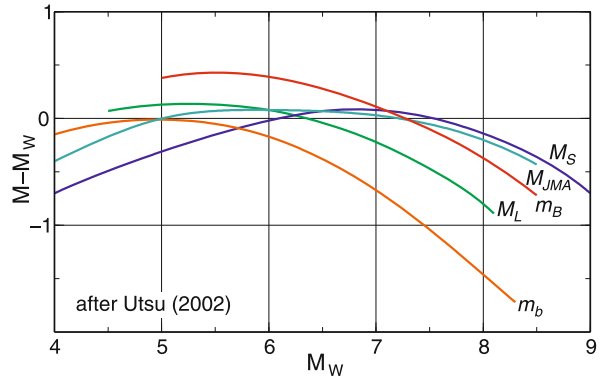


Earthquake Magnitude, Figure 3

Magnification of ground displacement amplitudes by common standard types of seismographs. WA = Wood–Anderson seismograph; WWSSN-SP and WWSSN-LP = short-period and long-period seismographs used in the former United States World-Wide Seismograph Standard Network; HGLP = US type of High Gain Long Period seismographs; A2, A3, B1, B3 and C = standard types of seismographs according to Willmore [87]. Reprint from [6] with © granted by IASPEI

atic distance-dependent biases. However, their proposed revised calibration functions for 20 s waves are not yet used in routine practice at international seismological data centers.

The formulas (6) and (10) had originally been developed for horizontal component amplitude readings. Beginning in the 1960s, however, more and more long-period and broadband vertical component instruments became available and are now commonly used for magnitude determination from surface waves. This procedure is easier and better defined than measuring and combining the amplitude measurements made in two horizontal components, yields on average values that are largely comparable with the Gutenberg M_S [25] and has recently been adopted as IASPEI [42] standard. Herak et al. [37] published theoretical and observed depth corrections for $M_S(20)$ when determined according to (10). These corrections allow determination of more reliable surface-wave magnitudes for earthquakes in all depth ranges and improve significantly



Earthquake Magnitude, Figure 4

Average relationships between different common types of magnitudes and the moment magnitude M_w . Modified from Fig. 1 in [83]

the relationship between M_S and the seismic moment M_0 .

In the 1960s, the United States deployed a World-Wide Standard Seismograph Network (WWSSN) equipped with short-period (SP) and long-period (LP) seismographs of limited bandwidth (cf. Fig. 3). This network had two priority tasks. Firstly, to significantly increase the signal-to-noise ratio of the seismic records by narrow-band short-period filtering, thus improving the global detection threshold for teleseismic events down to magnitudes around 4–4.5 and the location accuracy for seismic events. Secondly, to realize an effective discriminator between underground nuclear explosions (UNE) and natural earthquakes based on the ratio of a short-period body-wave and a long-period surface-wave magnitude. Natural earthquakes have a much longer source duration (seconds to minutes) than explosions of comparable size (typically milliseconds). Also, at comparable seismic moment magnitude, UNEs radiate significantly more high-frequency energy (see dotted curve in Fig. 5) than earthquakes. Therefore, a better discrimination of the two types of events was achieved by measuring the P -wave amplitude only at periods < 3 s (typically around 1 s) and calculating a short-period P -wave magnitude termed m_b . In contrast, the Gutenberg m_B is based on measuring A_{\max} at periods T usually between 2 s and 30 s. Further, during the first two decades of the WWSSN, the P -wave amplitude was not – as required by Gutenberg’s procedure for m_B determination – always measured at the maximum of the whole P -wave train (whose length depends on the source duration and thus on the magnitude itself) but initially within the first five half-cycles only and later by USGS requirement in the first 5 s of the record.

Because of the short source duration of explosions, their P -waves will always reach maximum amplitudes

within such a short time-interval. However, P waves radiated by large earthquakes of much longer source duration will reach their maximum amplitude usually much later in the rupture process. For magnitude 6 the average rupture duration is on average already 6 s, and may increase to about 600 s for the strongest earthquakes (cf. relation (13)). Both effects together, plus the fact, that m_b was still computed using the Q_{PV} function derived for mainly medium-period P waves, which are much less affected by frequency-dependent attenuation than 1 Hz P waves, resulted in a systematic underestimation of the earthquake size for magnitudes larger than 5 and a *saturation* of m_b at around 6.5.

In the late 1970s, the NEIC switched back to a longer time window of about 15 s and more recently, with an automatic procedure, to a window covering the first 10 cycles of short-period teleseismic P waves. In the case of strong earthquakes this window may later be extended interactively up to 60 s. This mitigates to some extent the saturation of m_b . However, no m_b -values larger than 7.2 have ever been measured with this procedure. On the other hand m_b yields rather reliable values for magnitudes < 5 when the corner frequency of the average source spectra falls within the passband of the short-period seismograph or is even more high frequency (cf. Figs. 3, 4, 5). For magnitudes < 5 m_B can usually no longer be determined because of too small signal-to-noise ratio (SNR) in broadband records. Then m_b is often the only available teleseismic estimator of earthquake size for small earthquakes.

Most seismic stations and networks worldwide adopted the US procedure for m_b measurement and – with the exception of Russia, China and their former allies – completely abandoned measuring m_B as originally defined. This change in attitude was stimulated by the fact that the NEIC, which serves in fact as one of the leading international data centers for seismology, did not accept reported P -wave amplitudes other than those obtained from short-period measurements. Some stations, national and global data centers continue (at least up to 2008) to measure for m_b the maximum amplitude of P exclusively within the first 5 s after the P -wave *first arrival*, such as the China Earthquake Network Center and the International Data Center (IDC) of the Comprehensive Test-Ban Treaty Organization (CTBTO) in Vienna.

Because of these inconsistencies in m_b and M_S determination and the proven merits of both broadband m_B and M_S (see also [11]) the IASPEI Working Group on Magnitude Measurements recommended that in future:

- a) m_b is always determined from A_{\max} at periods $T < 3$ s within the whole P -wave train;

- b) The band-limited magnitudes m_b and $M_S(20)$ be complemented by true broadband magnitudes m_B and $M_S(BB)$. The latter two will be obtained by measuring $A_{v\max}$ on unfiltered velocity broadband records and thus always include the maximum velocity amplitudes of the source spectrum in the magnitude range of interest (cf. Fig. 5). This will link these two broadband magnitudes to the seismic energy released by an earthquake, more closely than the common band-limited magnitudes.

These recommendations have been adopted by the IASPEI Commission on Seismic Observation and Interpretation (CoSOI) in 2005 as new magnitude measurement standards. More details about the new measurement procedures for m_b , m_B , $M_S(20)$ and $M_S(BB)$ are given on the CoSOI web site [42]. Beginning in 2007 they are gradually implemented at the main seismological data centers and networks.

Since all magnitudes discussed so far show more or less pronounced *saturation* for large earthquakes (cf. Fig. 4 and [44]) a non-saturating magnitude, termed M_w , has been proposed [31,43,69]. The moment magnitude M_w is derived from the scalar *seismic moment* M_o via the relation

$$M_w = (2/3)(\log_{10} M_o - 9.1) . \quad (11)$$

M_o has the dimension of Newton meter (Nm) and expresses the total inelastic “work” required for rupturing and displacing the considered earthquake fault. It can be determined either by waveform analysis and inversion in the time domain or by measuring the spectral amplitude $u_{0p,s}$ of the low-frequency level (plateau) of the displacement spectrum of P or S waves (cf. Fig. 5) via the relationship

$$M_o = 4\pi r \rho v_{p,s}^3 u_{0p,s} / R_{\theta,\phi}^{p,s} \quad (12)$$

with r = hypocenter distance, ρ = average density of rocks in the source and receiver area, $v_{p,s}$ = average velocity of the P or S waves from the source to the receiver area and $R_{\theta,\phi}^{p,s}$ = a factor correcting the observed seismic amplitudes for the influence of the radiation pattern of the given *source mechanism*, which is different for P and S waves.

M_o is expected to show no *saturation*, provided that the amplitude level is measured only at periods significantly larger than the magnitude-dependent *corner period* of the seismic source spectrum (cf. Fig. 5). In Sects. “Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake” and “Magnitude Saturation and Biases Due to Earthquake Complexity” we will show, however, that in-

correct determination of M_0 may still result in an underestimation of the earthquake size. Since M_w is derived from M_0 it is related to the tectonic effect of earthquakes, i. e., to the product of rupture area and average fault slip and thus also relevant to assess the tsunami potential of strong shallow marine earthquakes. An example is the off-shore Nicaragua earthquake of 2 September 1992. Its $m_b = 5.3$ was too weak to alert the people ashore, some 70–120 km away from the source area. However, its $M_w = 7.6$ was much larger and caused a damaging local tsunami with almost 200 casualties.

Yet, M_0 and thus M_w do not carry any direct information about the dominant frequency content and thus of the seismic energy released by the earthquake (cf. Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”). In fact, relation (11) was derived by assuming constant stress drop and an average ratio of $E_s/M_0 = 5 \times 10^{-5}$ on the basis of elastostatic considerations and empirical data [43] and then replacing in Eq. (9) M_S by M_w .

As source theory has advanced and broadband digital data have become readily available, the radiated seismic energy E_s could be computed explicitly rather than from an empirical formula. Boatwright and Choy (cf. [5,16]) developed such an algorithm for computing E_s as well as a related energy magnitude M_e which agrees with M_w for $E_s/M_0 = 2 \times 10^{-5}$. E_s is computed by integrating squared velocity-proportional broadband records over the duration of the P -wave train, corrected for effects of geometrical spreading, frequency-dependent attenuation during wave propagation and source radiation pattern. According to [16], the radiated seismic energy may vary for a given seismic moment by two to three orders of magnitude. Further, it was found that a list of the largest events is dominated by earthquakes with thrust mechanisms when size is ranked by moment, but dominated by strike-slip earthquakes when ranked by radiated seismic energy. Choy and Kirby [18] gave a striking example for differences between M_e and M_w for two Chile earthquakes in 1997 which occurred in the same area but with different *source mechanisms*. One was interplate-thrust with $M_w = 6.9$ and relatively low $M_e = 6.1$, whereas the other was intraslab-normal with $M_w = 7.1$ and rather large $M_e = 7.6$. The first earthquake had a low potential to cause shaking damage and was felt only weakly in a few towns. In contrast, the second one caused widespread damage, land- and rock-slides, killed 300 people and injured 5000. Thus, M_w , although it theoretically does not saturate, may strongly underestimate or overestimate the size of an earthquake in terms of its potential to cause damage and casualties. Shaking damage is mainly controlled by the relative amount of

released high-frequency energy at $f > 0.1$ Hz which is better measured by M_e .

The quantity $\tau_a = \mu E_s/M_0$ is termed apparent stress [90]. It represents the dynamic component of stress acting on the fault during slip, which is responsible for the generation of radiated kinetic seismic wave energy E_s . On average it holds that $\tau_a \approx 2\Delta\sigma$ (with $\Delta\sigma$ = stress drop = difference between the stress in the source area before and after the earthquake rupture). Both τ_a and $\Delta\sigma$ depend strongly on the seismotectonic environment, i. e., the geologic-tectonic conditions, fault maturity and type of earthquake *source mechanisms* prevailing in seismically active regions [16,17,18,19]. However, $M_e \approx M_w$ holds only for $\tau_a \approx 0.6$ MPa.

Another important teleseismic magnitude is called mantle magnitude M_m . It uses surface waves with periods between about 60 s and 410 s that penetrate into the Earth’s mantle. The concept has been introduced by Brune and Engen [13] and further developed by Okal and Talandier [64,65]. M_m is firmly related to the seismic moment M_0 . Best results are achieved for $M_w > 6$ at distances > 15 – 20° although the M_m procedure has been tested down to distances of 1.5° [77]. However, at $D < 3^\circ$ the seismic sensors may be saturated in the case of big events. Also, at short distances one may not record the very long periods required for unsaturated magnitude estimates of very strong earthquakes, and for $M_w < 6$, the records may become too noisy at very long-periods. A signal-to-noise ratio larger than 3 is recommended for reliable magnitude estimates. M_m determinations have been automated at the PTWC and the CPPT [39,85] so that estimates are available in near real-time within about 10 min after OT from near stations, however typically within about half an hour, plus another few minutes for great earthquakes measured at the longest periods. Since M_m is determined at variable very long periods this magnitude does not – or only marginally – saturate even for very great, slow or complex earthquakes.

Common Magnitude Estimates for the Sumatra 2004 M_w 9.3 Earthquake

On 26 December 2004, the great Sumatra–Andaman Island earthquake with a rupture length of more than 1000 km occurred. It caused extensive damage in Northern Sumatra due to strong earthquake shaking. Moreover, it generated an Indian Ocean-wide tsunami with maximum run-up heights of more than 10 m. In total, this event claimed more than 200,000 victims and caused widespread damage on the shores of Sumatra, Thailand, India and Sri Lanka that were reached by the tsunami wave

within some 15 min to about two hour's time. This earthquake put the current procedures for magnitude determination to a hard test both in terms of the reliability and compatibility of calculated values and the timeliness of their availability to guide early warning and disaster management activities. Here we address only seismological aspects, not the additional major problems of inadequate global monitoring and insufficient regional communication and disaster response infrastructure. The earliest magnitudes reported by or made available to the Pacific Tsunami Warning Center (PTWC) were:

- $m_b > 7$, about 8 min after origin time (OT);
- $M_{wp} = 8.0$, available after some 12 minutes at the PTWC (including a magnitude-dependent correction [86]);
- somewhat later in Japan $M_{wp} = 8.2$ after magnitude-dependent correction [48];
- $M_m \geq 8.5$ at the PTWC about 45 min after OT, hours later upgraded to $M_m = 8.9$ by using mantle surface waves with longer periods ($T \approx 410$ s);
- a first surface-wave magnitude estimate $M_S = 8.5$, some 65 min after OT;
- $M_w = 8.9$ (later revised to 9.0) released by Harvard Seismology more than 6 h after OT.

Other available measurements were: $m_b = 5.7$ and $M_S = 8.3$ by the IDC of the CTBTO, $m_b = 7.0$, $M_S = 8.8$, $M_e = 8.5$ and another long-period P -wave based $M_w = 8.2$ by the NEIC. All these values were too small and mostly available only after several hours or days (e.g., IDC data). Weeks later, after the analysis of Earth's *fundamental modes* with periods up to 54 min and wavelength of several 1000 km, the now generally accepted value $M_w = 9.3$ was published [76]. Why were the other magnitude values all too low and/or too late?:

- m_b NEIC suffers from the combined effect of both spectral and time-window dependent saturation that we will discuss in more detail in Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”;
- m_b IDC is even more affected by these saturation effects, because of the very short measurement time window of only 5 s after the first P -wave onset. In the case of the Sumatra 2004 earthquake, the first P -wave maximum occurred after some 80 s and another, with comparable amplitude, after about 330 s (cf. Fig. 8). Further, prior to m_b measurement, the IDC broadband data are filtered with a more narrow-band response peaked at even higher frequencies (3–4 Hz) than at NEIC (≈ 2.5 Hz) [11];

- The reported surface-wave magnitudes ranged between $M_S = 8.3$ (IDC), 8.8 (NEIC and Japan Meteorological Agency) and 8.9 (Beijing), i.e., some of them are close to the moment magnitudes. However, because of the late arrival of long-period teleseismic surface waves, good estimates are usually not available within 1–2 h after OT. This leaves a sufficient tsunami warning lead time only for shores more than 1000–2000 km away from the source.
- The NEIC P -wave moment magnitude $M_w = 8.2$ was too small because its procedure is, similar as for M_{wp} determinations, based on relatively short-period (typically $T < 25$ s) P -wave recordings and a single-source model (cf. Sect. “Magnitude Saturation and Biases Due to Earthquake Complexity”).
- The preliminary $M_e = 8.5$, computed a few hours after the December 2004 earthquake agreed with the final M_e computed later using a more robust method [20]. Another algorithm simulating a near-real-time computation would have yielded $M_e = 8.3$. Yet M_e , by its very nature as an energy magnitude and because of the relation $E_s = \Delta\sigma/2\mu M_e$, will generally be smaller than M_w for slow, long duration earthquakes with low stress drop. This is often the case for shallow thrust earthquakes in subduction zones. Extreme examples are four well-known slow tsunami earthquakes of 1992 (Nicaragua; $M_w = 7.6$, $\Delta M_e = -0.9$), 1994 (Java; $M_w = 7.8$, $\Delta M_e = -1.3$), 2000 (New Britain Region; $\Delta M_w = 7.8$, $\Delta M_e = -1.0$) and 2006 (Java; $M_w = 7.7$, $\Delta M_e = -0.9$) [40].

Magnitude Saturation and Biases Due to Earthquake Complexity

Currently, the most common magnitude scales, especially those based on band-limited short-period data, still suffer *saturation*, e.g., the magnitudes m_b , M_L , m_B and M_S , which are typically measured at periods around 1 s, 2 s, 5–15 s and 20 s, respectively begin to saturate for moment magnitudes M_w larger than about 5.5, 6.5, 7.5 and 8.0. Earthquakes with $m_b > 6.5$, $M_L > 7.0$, $m_B > 8.0$ and $M_S > 8.5$ are rare been found due to saturation (cf. Fig. 4 and [44]). Magnitude saturation has two causes: spectral saturation and saturation due to insufficient time-window length for the amplitude measurements. Source complexity may cause additional biases between different magnitude scales.

Spectral Saturation of Magnitudes

Spectral saturation occurs when the magnitude-dependent *corner frequency* f_c (for energy-related magnitudes) or

the low-frequency plateau of displacement amplitudes (for moment magnitude) fall outside of the passband range of the seismographs *transfer function* or magnification curve (Fig. 3) of the recording seismograph or of the filter applied to broadband data before magnitude measurements are made. The reason for spectral saturation can best be explained by way of idealized average “source spectra” of ground displacement $u(f)$ and ground velocity $v(f)$ that have been corrected for the instrument response, for the decay of the wave amplitudes due to attenuation that is caused by internal friction and scattering of the seismic waves at heterogeneities of the Earth and for amplification effects at the surface of the receiver site. For better understanding such source spectra have been multiplied in Fig. 5 by the factor $4\pi\rho v_{p,s}^3/R_{\theta,\phi}^{p,s}$ given in Eq. (12) in order to get for the displacement amplitudes $u_0 = \text{constant}$ at $f < f_c$ the related scalar seismic moment M_0 (Fig. 5, left) and its time-derivative, the so-called moment rate (Fig. 5, right).

The shape of the source spectrum can be interpreted as follows: The critical wavelength, which corresponds to f_c , is $\lambda_c = v_{p,s}/f_c = c_{m1}\pi R = c_{m2}(L \times W)^{1/2}$ with $v_{p,s}$ – velocity of the P or S waves in the source region, depending on whether f_c relates to a P -wave or an S -wave spectrum, R – radius of a circular fault rupture model, L – length and W – width of a rectangular fault rupture model; c_{m1} and c_{m2} are model dependent constants. For very long fault ruptures, i.e., $L \gg W$, one can even write $\lambda_c = c_{m3}L$. Thus, λ_c is proportional to the linear dimension of the fault. For $f < f_c$, λ_c becomes larger than the fault. Rupture details along the fault can then no longer be resolved and the fault is “seen” by these long wavelengths just as a point source. Therefore, all frequencies $f < f_c$ have the same displacement amplitudes. Accordingly, M_0 , which is proportional to the fault area and the average slip over the fault, has to be determined either in the spectral domain from the low-frequency asymptote u_0 to the displacement spectrum or in the time domain by fitting synthetic long-period waves with $f < f_c$ to observed ones that have been low-pass filtered in the same frequency range.

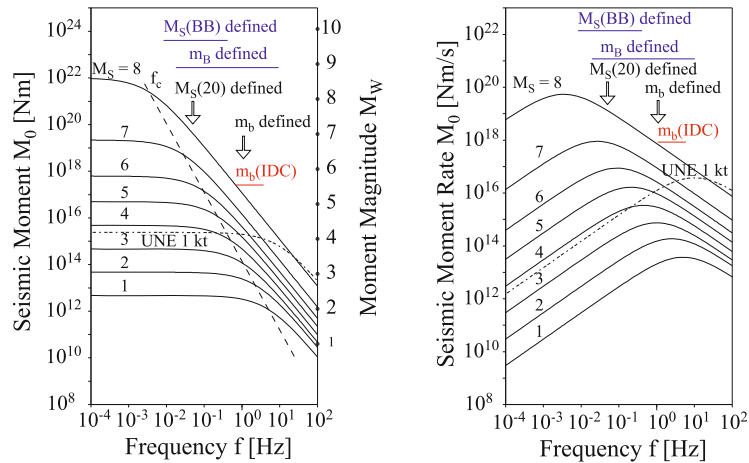
For radiated frequencies $f > f_c$ with $\lambda < \lambda_c$, the shape of the spectrum changes drastically. Related displacement amplitudes are then excited by successively smaller patches of the rupture plane. The area of the rupture elements decreases with the second order of their linear dimension. Accordingly, the generated displacement amplitudes are $A_d \sim f^{-2}$, while the related velocity amplitudes $A_v = A_d 2\pi f$ decay only $\sim f^{-1}$. In the seismological literature this is usually called the ω^{-2} rupture model [3], based on the concept of similarity, which implies a constant stress drop independent of source size. More com-

plicated rupture models yield a high-frequency amplitude decay $\sim \omega^{-3}$ [33,34] and even more rapid decays have sometimes been found in empirical data (up to 5th order). Steeper than ω^{-2} amplitude decay would further amplify the spectral saturation of magnitude data discussed below.

The Harvard standard procedure for M_0 determination assumes a single point source model with a prescribed, triangular moment-rate function in the time domain (as an approximation to moment-rate curves such the ones shown in Fig. 7) as well a minimum period of 200 s for strong earthquakes with magnitudes > 8 . Assuming an average rupture velocity of 2.5 km/s, this period would correspond to a wavelength of 500 km. This is much shorter than the total rupture length of more than 1100 km for the great Sumatra 2004 earthquake and explains why $M_w(\text{HRV}) = 9.0$ was smaller than the moment magnitude $M_w = 9.3$ determined by using fundamental Earth’s modes with periods of 1000 s and more [76].

The relationship between the two currently most common magnitudes, m_b and $M_S(20)$, can be understood with reference to Fig. 5. m_b is measured in the period range $0.5 < T < 3$ s, typically around 1 s. This corresponds approximately to the *corner frequencies* of earthquakes with $M_S \approx 3$ to 4.5. According to Utsu [83] this is equivalent to an m_b between about 3.5 and 5.0. For $M_S < 4.5$ or $m_b < 5$, m_b is thus likely to be determined from amplitude measurements near or below the *corner frequency* of the source spectrum. In that case m_b is a good measure of seismic moment. However, for larger magnitudes m_b samples spectral amplitudes well above f_c , resulting in systematically too small m_b values as compared to M_S and M_w . For great earthquakes this difference may reach 2 m.u. (Fig. 4). In contrast, $M_S(20)$ is measured at periods around 20 s and thus saturates much later at values between about 8.5 to 9.

However, these arguments only hold on average. The stress drop $\Delta\sigma$ of individual events may vary by about 2 to 3 orders, as apparent stress τ_a , especially for earthquakes with $M_w < 7.5$ [16,17]. According to the relation $M_0 = (16/7)\Delta\sigma R^3$ given by Keilis-Borok [52] this may change source radii R and associated f_c by about one order. As an example, the dotted curve in Fig. 5 shows the approximate seismic source spectrum for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT which corresponds to a magnitude $m_b \approx 4$. Its source volume is much smaller than that of an earthquake with same seismic moment. Hence the *corner frequency* of its source spectrum is not around 1 Hz but around 10 Hz. This is the reason why m_b determined from UNE records does not saturate, even for the strongest UNE ever tested with $m_b \approx 7$. Moreover, Fig. 5 also illustrates that an earthquake and an UNE



Earthquake Magnitude, Figure 5

“Source spectra” of ground displacement (*left*) and velocity (*right*) for an average single rupture seismic shear source, scaled on the left ordinates to seismic moment M_0 (*left diagram*) and moment rate (*right diagram*), respectively. The black spectral lines have been scaled according to Aki [3] to integer surface-wave magnitudes M_S between 1 and 8. For reference the respective integer moment magnitude values M_w between 1 and 10, calculated according to Eq. (11), have been marked with equidistant dots on the right-side ordinate of the *left diagram*. The broken line shows the increase of the corner frequency f_c with decreasing seismic moment of the event, the dotted curve gives the approximate “source spectrum” for a well contained underground nuclear explosion (UNE) of an equivalent yield of 1 kt TNT. Note the plateau in the displacement spectrum towards low frequencies (corresponding to $u_0 = \text{constant}$ for $f < f_c$), from which M_0 is determined according to Eq. (11) when using the frequency-domain approach. For $f > f_c$ the amplitudes decay $\sim f^{-2}$. The open arrows point to the center frequencies on the abscissa at which the 1 Hz body-wave magnitude m_b and the 20 s surface-wave magnitude $M_S(20)$, respectively, are determined and the blue horizontal interval bars mark the range of frequencies within which the maximum P-wave and Rayleigh-wave amplitudes for m_b and M_S (BB) should be measured according to the new IASPEI standards [37]. In contrast, the red bar marks the frequency range of maximum velocity-proportional magnification of the bandpass filter between 1 Hz and 4 Hz which is used for m_b determination at the IDC.

with seismic moment around 4×10^{15} Nm and $M_w \approx 4$ have different maximum seismic moment-rate release at about 4×10^{15} and 4×10^{16} Nm/s, respectively. The latter corresponds to 100 times higher seismic energy release or to an energy magnitude M_e that is 1.3 m.u. larger. Large differences have also been observed amongst earthquakes, e.g., the Balleny Island earthquake of 25.03.1998 had $M_w(\text{HRV}) = 8.1$ and $M_e(\text{NEIC}) = 8.8$. The opposite will happen in the case of low stress drop earthquakes propagating with very low rupture velocity [38]. The Java tsunami earthquake of 17 July 2006 was a striking example with $M_e = 6.8$, $m_B = 7.0$ and $M_w = 7.7$.

Similar observations had already been made in the 1970s when comparing m_b and M_S values of identical events. This prompted the Russian scientist Prozorov to propose a “creepex” parameter $c = M_S - a \times m_b$ (with $a = \text{constant}$ to be determined empirically for different source types and stress drop conditions). It aims at discriminating between normal, very slow (creeping) and explosion-like (fast rupture, high stress drop) earthquakes. World-wide determination of this parameter for earthquakes in different regions revealed interesting relations of c to source-geometry and tectonic origin [51]. Simi-

lar systematic regional differences were also reported for $M_S - M_w$ [24,67] and $M_e - M_w$ [16,19], suggesting systematic regional differences in stress drop.

Magnitude Saturation Due to Insufficient Time-Window Length for Amplitude Measurement

The second reason for magnitude saturation is insufficient time-window length for measuring $(A/T)_{\text{max}}$ in seismic records. It is most relevant when determining body-wave magnitudes, but it has been a subject of controversy, misconceptions and disregard of earlier recommendations for decades. The reason is that in teleseismic seismograms the P-wave group does not always appear sufficiently well separated in time from later phase arrivals such as the depth phases pP and sP . These do not directly travel from the seismic source at depth h to the recording station but travel first to the Earth’s surface above the source and from there, after reflection or conversion from S to P, propagate back into the Earth. Depending on h , which may vary from a few kilometers up to 700 km, and the type of depth phase recorded, they may arrive from a few seconds up to about 4.5 min after the onset of direct P. Depending on the radi-

ation pattern of the *source mechanism*, some stations may even record the depth phases with larger amplitudes than the direct *P* wave. This is one of the concerns that led many researchers to propose measuring the *P*-wave amplitudes for magnitude measurements within a short time window after the *P* onset. On average, however, the depth phases have smaller amplitudes than *P* and will not bias m_b estimates at all. If, however, a seismic station is situated near to the nodal line of the so-called focal sphere, corresponding to strongly reduced *P*-wave radiation in these directions, the amplitude of the depth phase is a better estimator for the body-wave energy radiated by this seismic source and thus of its corresponding magnitude.

Two or three more phases of longitudinal waves may arrive close to the direct *P* at teleseismic distances between 20° and 100° . These include *PcP*, which results from *P*-wave energy reflected back from the surface of the Earth's core at 2900 km depth, and the phases *PP* and *PPP*, which are *P* waves that have been reflected back from the Earth's surface once at half-way or twice at 1/3- and 2/3-way between the seismic source and the recording station, respectively. However, in short-period records the amplitudes of *PP* and *PPP* are generally smaller and those of *PcP* even much smaller than the amplitudes of direct *P* waves. These later arrivals will therefore, never bias m_b estimates. Yet on broadband records *PP* may sometimes have equal or even slightly larger amplitudes than primary *P*. However, *P* and *PP* phases are usually well separated by more than 1 min (up to 4 min) and not likely misinterpreted. Only for rare large earthquakes with $M > 7.5$ the rupture duration and related *P*-wave radiation may extend into the time window where *PP* should arrive. But even then, wrongly taking PP_{\max} for P_{\max} , the bias in m_b estimate will not exceed 0.2 m.u. and usually be much smaller.

This experience from extensive seismogram analysis practice led Bormann and Khalturin [7] to state in 1974:

... "that the extension of the time interval for the measurement of $(A/T)_{\max}$ up to 15 or 25 sec., as proposed ... in the Report of the first meeting of the IASPEI Commission on Practice (1972) ... is not sufficient in all practical cases, especially not for the strongest earthquakes with $M > 7.5$...".

This was taken into account in the Manual of Seismological Observatory Practice edited by Willmore [87]. It includes the recommendation to extend the measurement time window for *P*-wave magnitudes up to 60 s for very large earthquakes. But still, this has not yet become common practice (see Sect. "Introduction to Common Magnitude Scales: Potential and Limitations") although even

a limit of 60 s may not be sufficient for extreme events such as the Sumatra M_w 9.3 earthquake when the first $P1_{\max}$ appeared around 80 s and a second $P2_{\max}$ of comparable amplitude at about 330 s after the first *P*-wave onset (cf. Fig. 8).

To allow a quick rough estimate of earthquake rupture duration τ_d as a function of magnitude we derived from extrapolation of data published in [66] the average relation

$$\log \tau_d \approx 0.6M - 2.8. \quad (13)$$

It yields for $M = 6, 7, 8$ and 9 $\tau_d \approx 6$ s, 25 s, 100 s and 400 s, respectively. Measurement time windows of 5 s, 25 s or 60 s may therefore underestimate the magnitude of earthquakes with $M_w > 6$, > 7 or > 8 , respectively. We call this effect the time-window component of magnitude saturation. It aggravates the pure spectral saturation component. To avoid this in future, the new IASPEI standards of amplitude measurements for m_b and m_B (cf. [42]) recommend to measure $(A/T)_{\max} = A_{v\max}/2\pi$ in the entire *P*-phase train (time span including *P*, *pP*, *sP*, and possibly *PcP* and their codas but ending preferably before *PP*).

In fact the pioneers of the magnitude scales, Richter and Gutenberg, knew this, because they were still very familiar with the daily analysis of real seismic records and their complexity. Regrettably, they never wrote this down, with respect to magnitude measurements, in detail for easy reference. In the current era of automation and scientific depreciation of alleged "routine processes" the younger generation of seismologists usually had no chance to gather this experience themselves and occasionally introduced technologically comfortable but seismologically questionable practices. In an interview given in 1980 [75] Prof. Richter remembered that Gutenberg favored the body-wave scale in preference to the surface-wave scale because it is theoretically better founded. However, he said:

"... it gives results comparable with Gutenberg's only if his procedure is closely followed. Experience has shown that misunderstanding and oversimplified misapplications can occur. For instance, magnitude is sometimes assigned on the first few waves of the *P* group rather than the largest *P* waves as Gutenberg did."

In order to avoid too-short measurement time windows when searching for the largest *P* amplitude one can estimate the rupture duration independently from the duration of large *P*-wave amplitudes in high-frequency filtered BB records because the generation of high-frequency waves is directly related to the propagation of the rupture

front. Thus one may find P_{\max} for great earthquakes even beyond the theoretically expected PP arrival (cf. Fig. 8).

Magnitude Biases Due to Neglecting Multiple Source Complexity

Realizing that strong earthquakes usually consist of multiple ruptures Bormann and Khalturin [7] also wrote:

“In such cases we should determine the onset times and magnitudes of all clear successive P-wave onsets separately, as they give a first rough impression of the temporal and energetic development of the complex rupture process. ... The magnitude $MP = \log \sum_n (A_i/T_i) + Q(D, h)$ (n is the number of successive P-wave onsets) could be considered as a more realistic measure of the P-wave energy released by such a multiple seismic event than the m_b -values from ... (single amplitude) $(A/T)_{\max}$ within the first five half cycles or within the whole P-wave group.”

This magnitude, which is based on summed amplitudes in broadband records, is now called m_{bc} [10], which stands for cumulative body-wave magnitude.

The 1985 $M_w = 8.1$ Mexico earthquake was a striking example for the development of such a multiple rupture process in space and time ([58], Fig. 6). A dense network of high-frequency strong-motion recordings near to the source revealed that the earthquake had a total rupture duration of about 60 s and consisted of two main sub-ruptures with significantly increased slip-velocities (12–32 cm/s). These two fault segments were separated in space by roughly 100 km in strike direction and ruptured between 10–22 s and 34–50 s after rupture start. Such a complicated rupture is not well represented by calculating the average slip and rupture velocity for a single point-source model. Also the *corner frequencies* related to these smaller sub-ruptures will be higher and not correspond to $(L \times W)^{-1/2}$ of the total rupture area.

Such multiple ruptures are not an exception but rather the rule for earthquakes with magnitudes above 7.5 (and often also for smaller ones, even down to events with magnitudes around 5.0). The detailed patterns of the respective moment-rate curves differ from event to event (Fig. 7). Often they can not be approximated by a single-source triangular moment-rate function, as commonly assumed in the standard procedure for moment tensor solutions practiced at Harvard [78] and other centers.

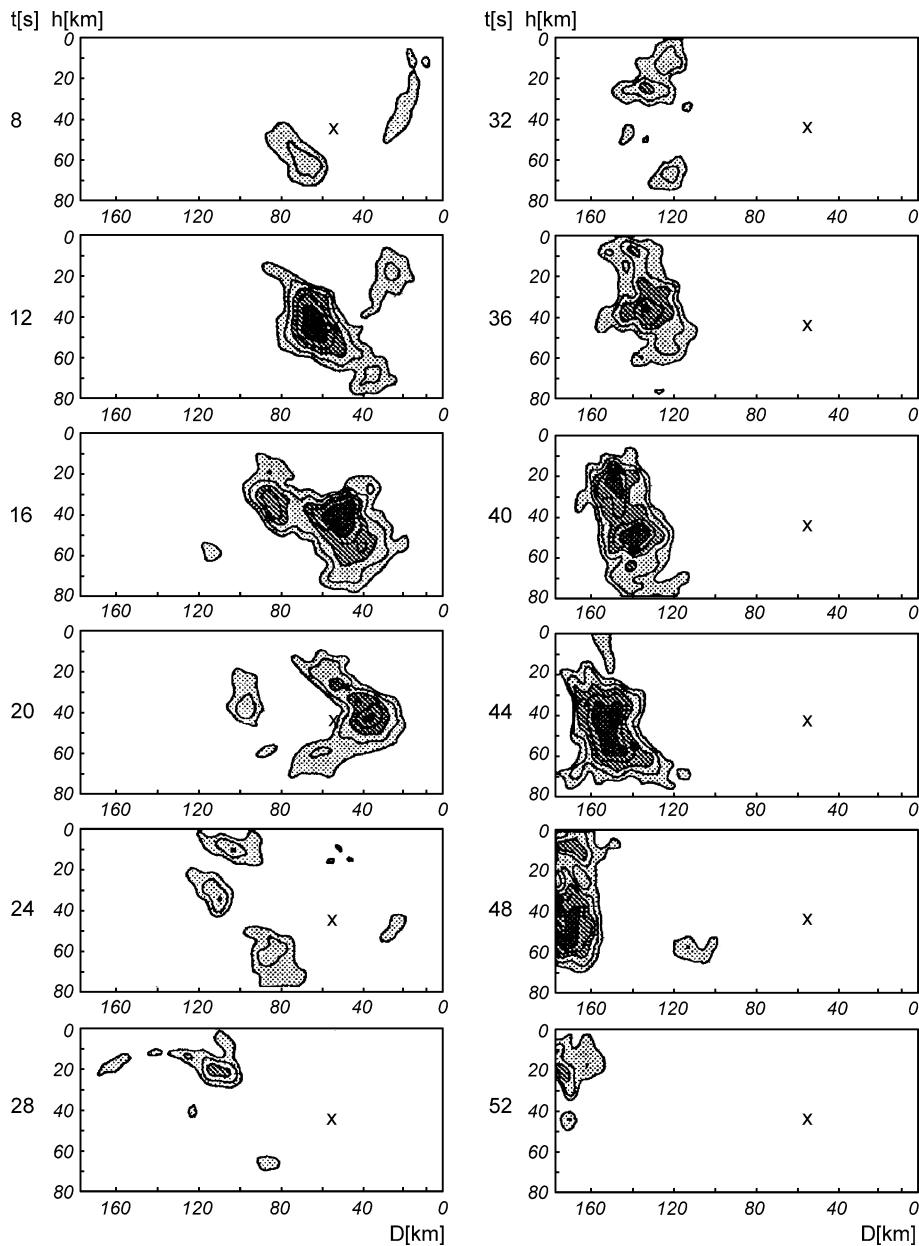
Therefore, the Harvard group [78] re-analyzed the data of the great Sumatra 2004 earthquake for which $M_w = 9.0$ had been calculated with the standard procedure.

Interactively fitting synthetic records for five successive point sources to the observed mantle surface-wave data in the 200–500 s period range yielded the same value of $M_w = 9.3$ as derived by [76] for a single-source model but using much longer periods between 20 min and 54 min. In fact, the multiple Centroid Moment Tensor (CMT) source analysis applied in [78] resembles the concept proposed in [7] for P -wave magnitudes of strong earthquakes, but applied to long-period surface waves. Presently, a multiple CMT source analysis still requires human interaction and takes too much time for early warning applications. Possible alternative procedures such as the automatic line source inversion [21] have been developed but demonstrated so far only for earthquakes with magnitudes < 7 for which classical m_b , M_s or M_w do not saturate due to source complexity.

Proposals for Faster Magnitude Estimates of Strong Earthquakes

Soon after the great Sumatra earthquake of 2004 several authors suggested improvements to obtain more reliable and faster magnitude estimates of strong earthquakes. Menke and Levin [59] proposed to use a representative selection of 25 globally distributed high quality stations of the IRIS (Incorporated Research Institutions for Seismology) Global Seismic Network as a reference data base of available strong long-period master-event records with known M_w . In case of a new strong earthquake, a search for the nearest (within a few hundred kilometers) reference event in the data base is performed and waveforms are compared for a time window of about 30 min. By adding the \log_{10} of the average amplitude ratio of the two events to the M_w of the master event, a moment magnitude estimate of the actual event is obtained. This procedure is based on the assumption of similarity of *source mechanisms* and radiation patterns, slip rates and stress drops, at least within the reference regions. The authors expect reasonably good magnitude estimates, with only small underestimation for events with $M_w > 8.6$. Thus warnings could be issued within about 40 min after OT (measurement time window plus travel-time to stations of a global network). This would still be relevant for distant coasts that might be affected by a tsunami. However, no data have been published until now that demonstrate the near-real-time operational capability of this procedure for a representative set of strong events.

Another approach by Lomax et al. [55,56,57] uses high-frequency seismograms ($f \geq 1$ Hz) that contain predominantly P signals radiated directly from the propagating rupture front and show little interference with later

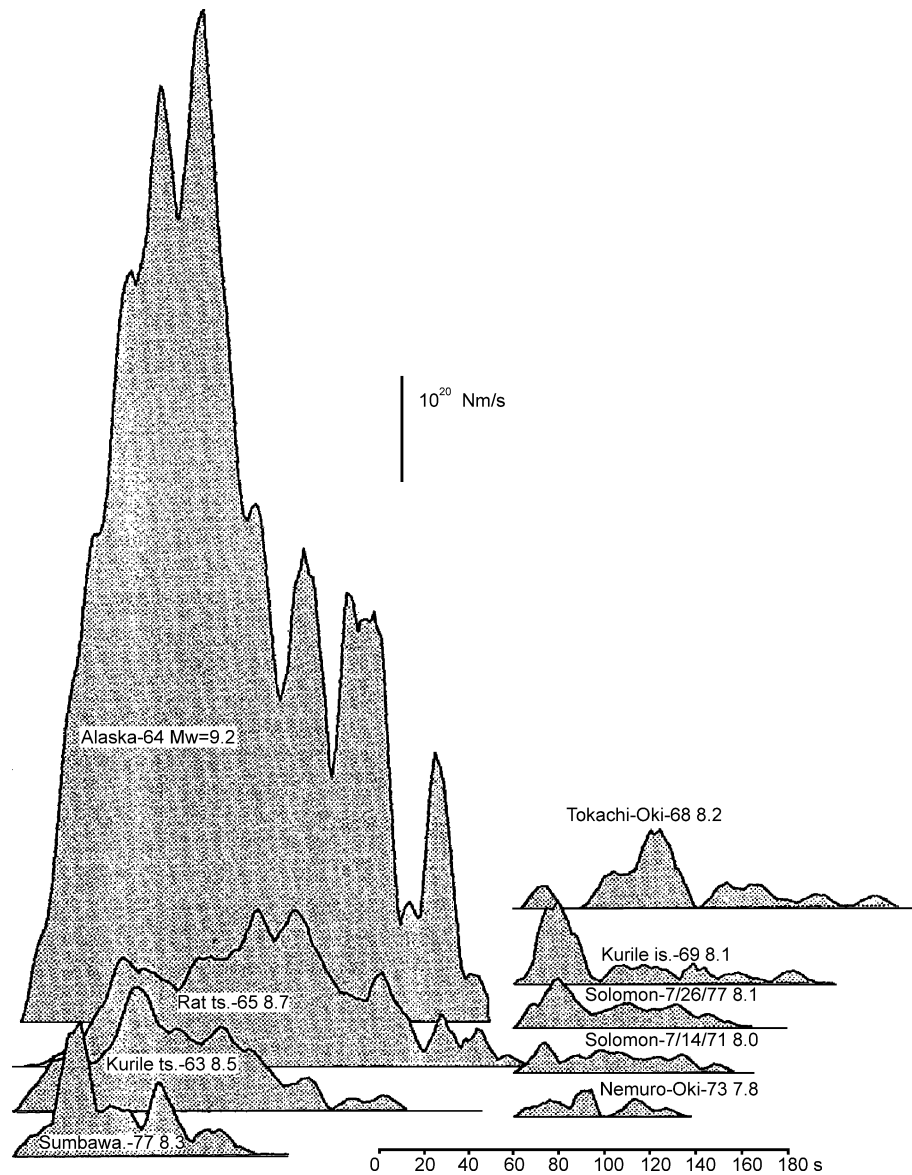


Earthquake Magnitude, Figure 6

Snapshots of the development in space and time of the inferred rupture process of the 1985 Michoacán, Mexico earthquake. The cross denotes the NEIC hypocenter position, the shading of the patches (from the outer part inwards *dotted*, *hatched* and *black*) relate to areas with velocities of dip slip (see *source mechanism*) in the ranges between 12 and 22 cm/s, 22 and 32 cm/s and greater than 32 cm/s. Redrawn and modified from Fig. 6 in [58]; taken from Fig. 3.8 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

secondary waves such as *PP* or *S*, thus providing a direct estimate of the rupture duration. Such recordings are available at teleseismic distances (30° – 90°) within about 20 min after OT, even after strong events with long durations and provide an early picture of the total rupture pro-

cess. When assuming constant rupture velocity and mean slip for stronger and weaker earthquakes, the seismic moment M_0 and thus moment magnitude M_w could be estimated by comparing the actual rupture duration (averaged from observations at several seismic stations) with that of



Earthquake Magnitude, Figure 7

Moment-rate functions for the largest earthquakes in the 1960 and 1970s (modified from Fig. 9, p. 1868 in [53]), taken from Fig. 3.7 in Vol. 1 of [6], © Seismological Society of America and IASPEI; with permission of the authors

a reference event with known M_0 and rupture duration. This is conceptually similar to the approach in [59] but with high-frequency observations and the ratio of rupture duration instead of amplitudes.

However, Hara [32] demonstrated with a large data set of strong earthquakes that it is difficult to estimate earthquake size reliably only from durations t of high-frequency radiation. Therefore, he measured duration t in combination with the maximum displacement amplitude A_{dmax}

within this time interval and derived the following empirical relation:

$$M = 0.79 \log A_{dmax} + 0.83 \log D + 0.69 \log t + 6.47 \quad (14)$$

with A_{dmax} , D and t in units of m, km and s, respectively. He applied Eq. (14) to 69 shallow earthquakes in the magnitude range $7.2 \leq M_w(\text{HRV}) \leq 9.0$ at distances between 30° and 85° and on average got a 1:1 relation between $M_w(\text{HRV})$ and his magnitude with a standard deviation

of 0.18 m.u. All event estimates were within ± 0.5 m.u. of $M_w(\text{HRV})$, with the exception of the heavily underestimated Denali/Alaska earthquake of 3 November 2002 (7.1 instead of 7.8). This is a promising and simple procedure.

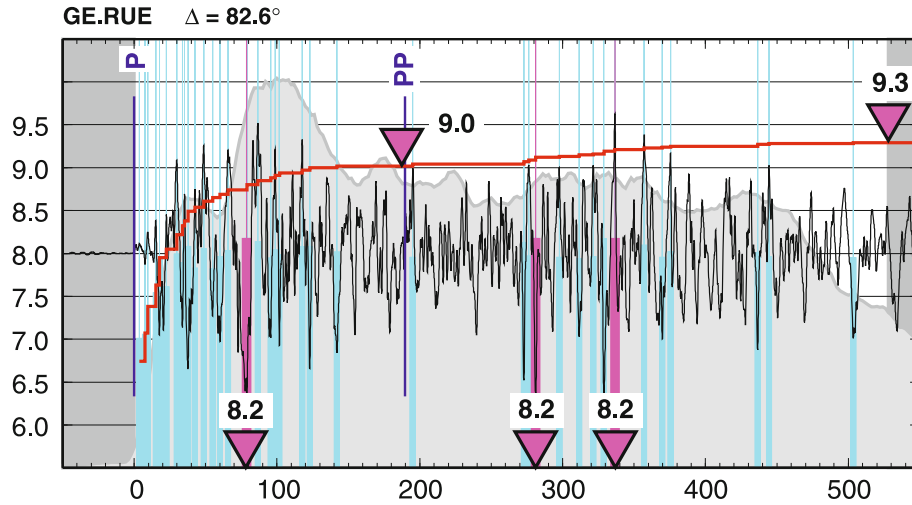
Bormann and Wylegalla [9] applied the earlier proposal in [7] to recordings with a velocity passband between 40 Hz and 125 s. They interactively summed up the maximum amplitudes of all visually discernible sub-ruptures in the recordings of several recent great earthquakes with $M_w \geq 8.3$, amongst them the tsunamigenic $M_w = 9.3$ Sumatra earthquake of 2004. For the latter they obtained a cumulative broadband body-wave magnitude $m_{\text{Bc}} = 9.3$ in records of just a single German station (RUE; $D = 82.5^\circ$) at the time of the second major amplitude maximum, some 330 s after the first P onset and 18 min after OT. For three more events with magnitudes M_w 8.3, 8.4 and 8.6 they calculated m_{Bc} values of 8.4, 8.4 and 8.6, respectively, i. e., excellent agreement. Subsequently, 50 more earthquakes in the magnitude range 6 to 9 were analyzed interactively [10] with the following results:

- Average difference $m_{\text{B}} - M_w(\text{HRV}) = 0.00 \pm 0.27$ in the range $6.0 \leq M_w(\text{HRV}) < 8$. For magnitudes > 7.8 –8, however, m_{B} tends to underestimate M_w , e. g., $m_{\text{B}} = 8.3$ for the Sumatra earthquake of 26 December 2004 based on the BB record of station RUE. Remarkably this m_{B} value is still very close to M_w , M_{wp} and M_e of the NEIC, which ranged between 8.2 and 8.5.
- The average difference $m_{\text{Bc}} - M_w(\text{HRV}) = +0.18 \pm 0.26$ in the range $6.0 \leq M_w(\text{HRV}) \leq 9.0$, i. e., m_{Bc} has a tendency to slightly overestimate $M_w(\text{HRV})$ on average, but not for $M_w > 8$ (see the four values above).

In [10] also first results of a fully automatic determination of m_{B} and m_{Bc} have been presented. The algorithm has been improved by incorporating automatic estimates of the rupture duration calculated from the envelope of the high-frequency P -wave radiation from filtered broadband records of globally distributed stations in a wide range of azimuths and source distances. In the case of strong earthquakes with long rupture duration this justifies the search for broadband P_{max} even beyond the onset of PP and to sum-up the amplitudes of major sub-ruptures over the whole rupture duration as defined above. Figure 8 gives an example for a BB record of the Sumatra earthquake of 26 December 2004. The largest P -wave amplitudes at about 80 s, 280 s and 330 s after the P onset each yield a single amplitude $m_{\text{B}} = 8.2$, whereas the cumulative magnitude $m_{\text{Bc}} = 9.3$ is in perfect agreement with the best moment magnitude estimates for this event.

The automatic algorithm for m_{B} and m_{Bc} determination has been in use since spring 2007 in the operational Indonesian prototype tsunami early warning system and yields online estimates of m_{B} . Before the implementation it had been tested whether the automatic procedure produces results that are comparable with those determined earlier interactively by two experienced seismogram analysts. Identical broadband records of 54 earthquakes in the magnitude range $6 \leq M_w(\text{HRV}) \leq 9$ were used for this comparison based on 138 m_{B} and 134 m_{Bc} values. The average difference between the interactively and automatically determined magnitudes was 0.03 and 0.02 m.u. with standard deviations of ± 0.13 and ± 0.19 m.u., respectively. This is in the range of other high-quality magnitude measurements. Even single station m_{B} and m_{Bc} estimates differed on average < 0.08 m.u. from average global network estimates based on up to hundreds of stations. Their standard deviations were $< \pm 0.25$ m.u. and decreased to ± 0.10 m.u. for m_{B} and ± 0.14 m.u. for m_{Bc} when just a few stations (between two and seven) were used to estimate the m_{B} and m_{Bc} event magnitudes. This documents both the reliability of the automatic procedure as well as the reliability of m_{B} and m_{Bc} estimates, even if derived from a few records of globally distributed high-fidelity stations. Thus, the automatic procedure is suitable for reproducibly determining the IASPEI recommended standard magnitude m_{B} and its proposed non-saturating extension m_{Bc} in near real-time. When using only observations in the distance range $21^\circ \leq D^\circ \leq 100^\circ$ saturation-free teleseismic magnitude estimates of earthquakes with potential for strong shaking damage and tsunami generation could be made available in near real-time within about 4 to 18 min after OT, depending on epicentral distance and rupture duration.

Compared to other more theoretically based methods such as M_{wp} and M_w , the empirical $m_{\text{B}} - m_{\text{Bc}}$ method is free of any hypothesis or model assumptions about the rupture process (single or multiple source), type of rupture mechanism, rupture velocity, average slip and/or stress drop, complexity or simplicity of the moment-release function, etc. It just measures velocity amplitudes on the unfiltered broadband record, complex or not, sums them up over the duration of the rupture process and calibrates them with the classical empirical broadband Q_{PV} function (Fig. 2 and [30]). However, one has to consider that – in contrast to all types of moment magnitudes – m_{B} and m_{Bc} are not determined from the maximum long-period displacement amplitudes, but from the maximum velocity amplitudes. Therefore, m_{B} (for earthquakes with $M_w < 8.0$) and m_{Bc} (for earthquakes with $M_w > 7.8$) are better estimators than M_w for the seismic energy released



Earthquake Magnitude, Figure 8

Velocity broadband record at the Berlin station RUE in $D^\circ = 82.6^\circ$ epicentral distance of the great $M_w 9.3$ tsunamigenic Sumatra earthquake of Dec. 26, 2004. The record is projected into a time-magnitude diagram as plotted by the automatic $m_B - m_{Bc}$ algorithm. The red inverted triangles mark the times and give the values of m_B for the three largest sub-ruptures. The red step curve shows the development of the cumulative magnitude m_{Bc} as a function of time. The inverted red triangles on this curve give the m_{Bc} before the onset of PP and at the end of the rupture process, about 530 s after the first P-wave onset, as estimated from the decay of the amplitude envelope of short-period filtered P-waves (see text)

by the earthquake and thus of its shaking-damage potential. Figure 9 compares m_B and m_{Bc} with $M_w(\text{HRV})$ for 76 earthquakes in the range $6 \leq M_w \leq 9.3$. The respective standard regression relations are:

$$M_w(\text{HRV}) = 1.22 m_B - 1.54 \pm 0.29 \quad (15)$$

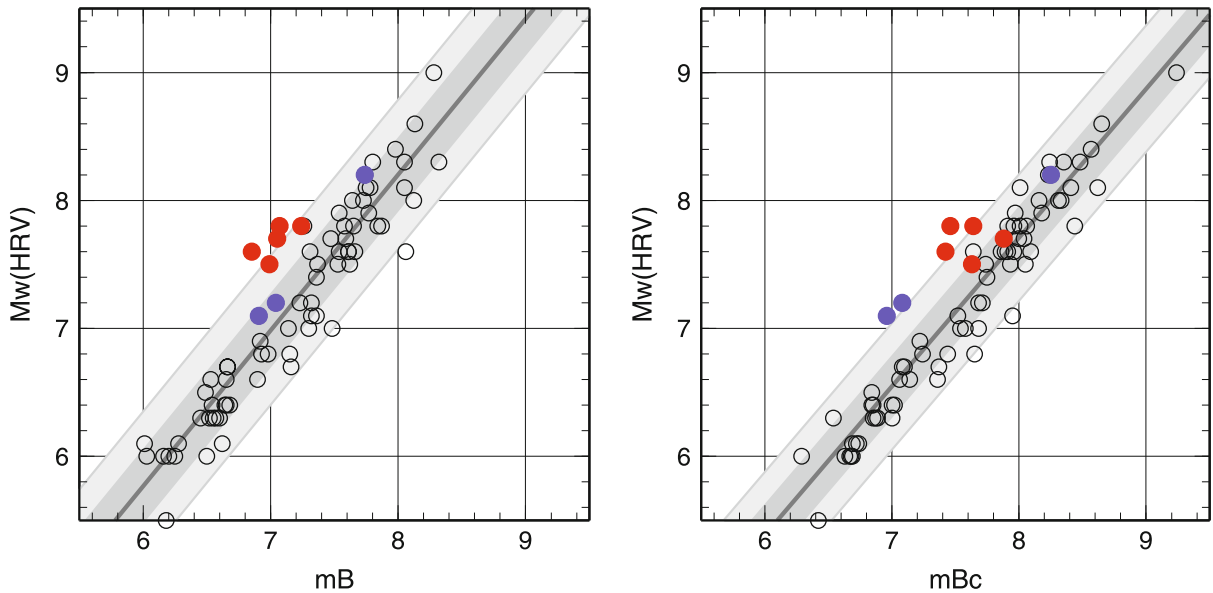
and

$$M_w(\text{HRV}) = 1.16 m_{Bc} - 1.59 \pm 0.25 \quad (16)$$

These scaling relations allow much faster estimates of M_w than current routine standard M_w procedures. The rough moment estimates derived from m_B and m_{Bc} data, $M_w(m_B)$ or $M_w(m_{Bc})$, are sufficiently reliable for initial earthquake and tsunami alarms with standard deviations of the M_w estimates of about ± 0.29 and ± 0.25 m.u., respectively. However, looking into details of the somewhat irregular data scatter one realizes how seismic source complexity may “spoil” such regression relations. The five data points marked red in Fig. 9(left and right) are distinct outliers in the left diagram, i.e., the respective m_B values are 0.5 to 0.75 m.u. smaller than $M_w(\text{HRV})$ although usually m_B scales rather well with $M_w(\text{HRV})$ between $6.5 < m_B < 8.0$. These points correspond to slow earthquakes, one in Peru (1996) and four are tsunami earthquakes as mentioned at the end of Sect. “Common Magnitude Estimates for the Sumatra 2004 $M_w 9.3$ Earthquake”.

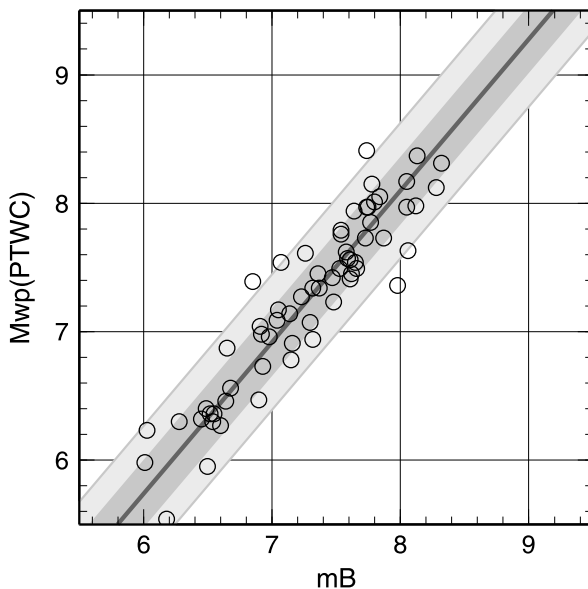
Their rupture durations ranged from about 100 s to 200 s, i.e., according to relationship (13) about 2–3 times longer than expected on average for their M_w value. Both m_B and M_e are usually much smaller than M_w for such events. In contrast, when calculating m_{Bc} , then these five data points all move close to the (not marked) 1:1 line in the $m_{Bc} - M_w(\text{HRV})$ diagram Fig. 9(right). Thus m_{Bc} becomes a good direct estimator of M_w for typical slow earthquakes, much better than via relation (16), which compensates for the usually too large m_{Bc} values of shallow depth and “normal” rupture earthquakes with $M_w < 8$. Thus, by determining rupture duration independently and treating very slow events separately, the standard deviation in relations (15) and (16) can be reduced. Moreover, the blue dots in Fig. 9 belong to very deep earthquakes with $h = 525$ km, 583 km and 631 km, respectively. Such deep earthquakes are “explosion-like” with comparably short rupture durations. Both m_B (for $M_w < 8$) and m_{Bc} yield values very close to M_w . In the $m_{Bc} - M_w$ diagram, which is dominated by shallow and normal rupture earthquakes, such deep events appear as outliers. However, rapid event locations with good depth estimates allow one to identify such events and m_{Bc} (or m_B) should then be taken directly as estimator of M_w and not via relation (16).

M_{wp} has so far been the fastest operationally determined estimator of M_w . Comparably fast automatic m_B determination is now implemented, complementary to



Earthquake Magnitude, Figure 9

Standard regression relationships of M_w (HRV) over m_B (left) and m_{Bc} (right). Red dots correspond to very slow earthquakes (Nicaragua 1992, Java 1994, New Britain Region 2000, Peru 2001 and Java 2006) and the blue dots belong to very deep earthquakes (Bolivia 1994, Philippines 2005 and Fiji Island 2006) with source depths $h = 631$ km, 525 km and 583 km, respectively. The gray band and the two white bands around the average straight line correspond to the width of one and two standard deviations in y -direction



Earthquake Magnitude, Figure 10

Standard regression of M_{wp} (PTWC) over m_B . The standard deviations in y -direction are marked as in Fig. 9. The M_{wp} data have been kindly provided by the PTWC (courtesy of B. Hirshorn)

M_{wp} , in the German Indonesian Tsunami Early Warning System (GITEWS). Figure 10 compares the relation be-

tween m_B and M_{wp} for our test data set. These two magnitudes scale almost 1:1, following the standard regression relation:

$$M_{wp} = 1.08 m_B - 0.638 \pm 0.24. \quad (17)$$

Future Requirements and Developments

Few national seismological data centers and stations report amplitude, period and/or magnitude data to the international data centers. The main reason is usually the lack of manpower to make competent measurements of these parameters interactively for the large amount of data recorded nowadays. Instrument responses of the seismographs used are sometimes not known accurately enough. There is, however, a growing practical and research need for such parameter data that have been determined according to international standards. Therefore, the most urgent requirements in the field of magnitudes are:

- Training of station and network operators to understand and practice proper magnitude measurements, instrument calibration and control;
- Implementation of the IASPEI magnitude standards [42];
- Making the tested and calibrated automatic algorithms available worldwide to data producers so that lack of

manpower is no longer a hindrance to mass-produce such data;

- Use of such standardized mass data with significantly reduced procedure-dependent errors for improved research into the attenuation properties of the Earth and deriving better magnitude calibration functions for all distance ranges;
- Comparison of magnitude data derived from identical record sets by applying both traditional and new standard measurement procedures and to derive standardized conversion relationships. This is a precondition for assuring long-term compatibility of magnitude data in national and international data catalogs and their usefulness for seismic hazard assessment and research;
- Improvement of current procedures for direct determination of seismic moment and energy in a wider magnitude range than currently possible, down to small magnitudes that are at present well covered only by M_L and m_b ;
- Development of regional calibration functions for m_b and m_B , which will permit more reliable and much faster body-wave magnitude estimates from records at distances down to about 5°
- Development and consequent use of standard procedures for M_0 and E_s measurements that assure non-saturating and globally compatible estimates of seismic moment and energy and of their related magnitude scales M_w and M_e ;
- Use of these data for in-depth studies in the regional variability of apparent stress conditions and their relevance for improving (time-variable) regional earthquake and tsunami hazard and risk assessment;
- Comprehensive testing of speed and reliability of the various methods recently proposed for more rapid (near) real-time magnitude estimates (e. g. [9,10,32,46,55,56,57,59,61,66]) under operational EWS conditions;
- Development of faster automated procedures for direct non-saturating M_w and M_e determination for improving quick and realistic disaster response;
- Development of alternative automatic (near) real-time procedures of magnitude determination such as the rapid finite-source analysis [21], their scaling to both seismic energy and moment and operational testing also for very large earthquakes.

Bibliography

Primary Literature

1. Abe K (1981) Magnitudes of large shallow earthquakes from 1904 to 1980. *Phys Earth Planet Int* 27:72–92
2. Abe K (1984) Complements to Magnitudes of large shallow earthquakes from 1904 to 1980. *Phys Earth Planet Int* 34:17–23
3. Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 72(4):1217–1231
4. Båth M (1981) Earthquake magnitude – recent research and current trends. *Earth Sci Rev* 17:315–398
5. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91(B2):2095–2112
6. Bormann P (ed) (2002) IASPEI New manual of seismological observatory practice, vol 1 and 2. GeoForschungsZentrum, Potsdam
7. Bormann P, Khalturin V (1975) Relations between different kinds of magnitude determinations and their regional variations. In: *Proceed XIVth General Ass European Seism. Comm Trieste Sept pp 16–22, 1974. Academy of Sciences of DDR, Berlin*, pp 27–39
8. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) IASPEI New manual seismological observatory practice. GeoForschungsZentrum Potsdam, chap 3, pp 1–94
9. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *EOS* 86(46):464
10. Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes m_B and m_{BC} for quick reliable estimation of the size of great earthquakes. <http://spring.msi.umn.edu/USGS/Posters/>
11. Bormann P, Liu R, Ren X, Gutdeutsch R, Kaiser D, Castellaro S (2007) Chinese national network magnitudes, their relation to NEIC magnitudes, and recommendations for new IASPEI magnitude standards. *Bull Seism Soc Am* 97(1B):114–127
12. Brune JN (1970) Tectonic stress and the spectra of shear waves from earthquakes. *J Geophys Res* 75:4997–5009
13. Brune JN, Engen GR (1969) Excitation of mantle Love waves and definition of mantle wave magnitude. *Bull Seism Soc Am* 49:349–353
14. Castellaro S, Mulargia F, Kagan YY (2006) Regression problems for magnitudes. *Geophys J Int* 165:913–930
15. Castellaro S, Bormann P (2007) Performance of different regression procedures on the magnitude conversion problem. *Bull Seism Soc Am* 97:1167–1175
16. Choy GL, Boatwright J (1995) Global patterns of radiated seismic energy and apparent stress. *J Geophys Res* 100, B9:18,205–18,228
17. Choy GL, Boatwright J, Kirby SH (2001) The radiated seismic energy and apparent stress of interplate and intraslab earthquakes at subduction zone environments: Implications for seismic hazard estimation. US Geological Survey Open-File Report 01–0005:18
18. Choy GL, Kirby S (2004) Apparent stress, fault maturity and seismic hazard for normal-fault earthquakes at subduction zones. *Geophys J Int* 159:991–1012
19. Choy GL, McGarr A, Kirby SH, Boatwright J (2006) An overview of the global variability in radiated energy and apparent stress. In: Abercrombie R, McGarr A, Kanamori H (eds) *Radiated energy and the physics of earthquake faulting*, AGU. *Geophys Monogr Ser* 170:43–57
20. Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra-Andaman earthquake estimated from 10-minute *P*-wave windows. *Bull Seism Soc Am* 97:18–24

21. Dregers DS, Gee L, Lombard P, Murray MH, Romanowicz B (2005) Rapid finite-source analysis and near-fault strong ground motions: Application to the 2003 M_w 6.5 San Simeon and 2004 M_w 6.0 Parkfield earthquakes. *Seism Res Lett* 76:40–48
22. Duda SJ (1965) Secular seismic energy release in the circum-Pacific belt. *Tectonophysics* 2:409–452
23. Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
24. Ekström G, Dziewonski AM (1988) Evidence of bias in estimations of earthquake size. *Nature* 332:319–323
25. Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. *Bull Seism Soc Am* 67:587–598
26. Gutenberg B (1945) Amplitudes of P, PP, and S and magnitude of shallow earthquakes. *Bull Seism Soc Am* 35:57–69
27. Gutenberg B (1945) Magnitude determination of deep-focus earthquakes. *Bull Seism Soc Am* 35:117–130
28. Gutenberg B (1945) Amplitude of surface waves and magnitude of shallow earthquakes. *Bull Seism Soc Am* 35(3):3–12
29. Gutenberg B, Richter CF (1954) *Seismicity of the earth and associated phenomena*, 2nd Edn. Princeton University Press, Princeton
30. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Ann Geofis* 9:1–15
31. Hanks TC, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84(B5):2348–2350
32. Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth Planet Space* 59:227–231
33. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 1. *Bull Seismol Soc Am* 54:1811–1842
34. Haskell N (1964) Total energy and energy spectral density of elastic wave radiation from propagating faults, vol 2. *Bull Seismol Soc Am* 56:125–140
35. Hatzidimitriou P, Papazachos C, Kiratzi A, Theodulidis N (1993) Estimation of attenuation structure and local earthquake magnitude based on acceleration records in Greece. *Tectonophysics* 217:243–253
36. Herak M, Herak D (1993) Distance dependence of M_S and calibrating function for 20 second Rayleigh waves. *Bull Seism Soc Am* 83:1881–1892
37. Herak M, Panza GF, Costa G (2001) Theoretical and observed depth corrections for M_S . *Pure Appl Geophys* 158:1517–1530
38. Hutton LK, Boore DM (1987) The M_L scale in Southern California. *Bull Seism Soc Am* 77:2074–2094
39. Hyvernaud O, Reymond D, Talandier J, Okal EA (1993) Four years of automated measurements of seismic moments at Papeete using the mantle magnitude M_m : 1987–1991. In: Duda SJ, Yanovskaya TB (eds) *Special section: Estimation of earthquake size*. *Tectonophysics* 217:175–193. Elsevier Science
40. <http://neic.usgs.gov/neis/sopar>
41. <http://www.globalcmt.org/CMTsearch.html>
42. http://www.iaspei.org/commissions/CSOI/Summary_of_WG_recommendations.pdf
43. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
44. Kanamori H (1983) Magnitude scale and quantification of earthquakes. *Tectonophysics* 93:185–199
45. Kanamori H (1988) The importance of historical seismograms for geophysical research. In: Lee WHK (ed) *Historical seismograms and earthquakes of the world*. Academic Press, New York, pp 16–33
46. Kanamori H (2005) Real-time seismology and earthquake damage prediction. *Ann Rev Earth Planet Sci* 33:195–214
47. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. *Nature* 390:461–464
48. Kanjo K, Furudate T, Tsuboi S (2006) Application of M_{wp} to the great December 26, 2004 Sumatra earthquake. *Earth Planet Space* 58:121–126
49. Katsumata M (1964) A method of determination of magnitude for near and deep-focus earthquakes (in Japanese with English abstract). *A J Seism* 22:173–177
50. Katsumata M (1996) Comparison of magnitudes estimated by the Japan Meteorological Agency with moment magnitudes for intermediate and deep earthquakes. *Bull Seism Soc Am* 86:832–842
51. Kaverina AN, Lander AV, Prozorov AG (1996) Global creep distribution and its relation to earthquake – source geometry and tectonic origin. *Geophys J Int* 135:249–265
52. Keilis-Borok VI (1959) On the estimation of displacement in an earthquake source and of source dimension. *Ann Geofis* 12:205–214
53. Kikuchi M, Ishida M (1993) Source retrieval for deep local earthquakes with broadband records. *Bull Seism Soc Am* 83:1855–1870
54. Lee V, Trifunac M, Herak M, Živčić M, Herak D (1990) M_L^{SM} computed from strong motion accelerograms recorded in Yugoslavia. *Earthq Eng Struct Dyn* 19:1167–1179
55. Lomax A (2005) Rapid estimation of rupture extent for large earthquakes: Application to the 2004, M9 Sumatra-Andaman mega-thrust. *Geophys Res Lett* 32:L10314
56. Lomax A, Michelini A (2005) Rapid determination of earthquake size for hazard warning. *EOS* 86(21):202
57. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid and accurate determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
58. Mendez AJ, Anderson JG (1991) The temporal and spatial evolution of the 19 September 1985 Michoacan earthquake as inferred from near-source ground-motion records. *Bull Seism Soc Am* 81:1655–1673
59. Menke W, Levin R (2005) A strategy to rapidly determine the magnitude of great earthquakes. *EOS* 86(19):185–189
60. Nakamura Y (1989) Earthquake alarm system for Japan railways. *Jpn Railw Eng* 109:1–7
61. Nakamura Y, Saita J (2007) UrEDAS, the earthquake warning system: today and tomorrow. In: Gasperini P, Manfredi G, Zschau J (eds) *Earthquake early warning systems*. Springer, Berlin, pp 249–281
62. Nuttli OW (1986) Yield estimates of Nevada test site explosions obtained from seismic Lg waves. *J Geophys Res* 91:2137–2151
63. Okal EA (1989) A theoretical discussion of time domain magnitudes: The Prague formula for M_S and the mantle magnitude M_m . *J Geophys Res* 94:4194–4204
64. Okal EA, Talandier J (1989) M_m : A variable-period mantle magnitude. *J Geophys Res* 94:4169–4193

65. Okal EA, Talandier J (1990) M_m : Extension to Love waves of the concept of a variable-period mantle magnitude. *Pure Appl Geophys* 134:355–384
66. Olson EL, Allen R (2005) The deterministic nature of earthquake rupture. *Nature* 438:212–215
67. Patton HJ (1998) Bias in the centroid moment tensor for central Asian earthquakes: Evidence from regional surface wave data. *J Geophys Res* 103(26):885–898
68. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys J Int* 142:684–702
69. Purcaru G, Berckhemer H (1978) A magnitude scale for very large earthquakes. *Tectonophysics* 49:189–198
70. Rezapour M, Pearce RG (1998) Bias in surface-wave magnitude M_S due to inadequate distance corrections. *Bull Seism Soc Am* 88:43–61
71. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seism Soc Am* 25:1–32
72. Richter CF (1958) *Elementary seismology*. W.H. Freeman, San Francisco
73. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? *Nature* 444:E5–E6
74. Soloviev SL (1955) Classification of earthquakes by energy value (in Russian). *Trudy Geophys Inst Acad Sci USSR* 39(157):3–31
75. Spall H (1980) Charles F. Richter – an interview. *Earthq Inf Bull* 12(1):5–8
76. Stein S, Okal E (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582
77. Talandier J, Okal EA (1992) One-station estimates of seismic moments from the mantle magnitude M_m : The case of the regional field ($1.5^\circ \leq \Delta \leq 15^\circ$). *Pure Appl Geophys* 138:43–60
78. Tsai VC, Nettles M, Ekström G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys Res Lett* 32(L17304):1–4
79. Tsuboi C (1954) Determination of the Gutenberg-Richter's magnitude of earthquakes occurring in and near Japan (in Japanese with English abstract). *Zisin Second Ser* 7:185–193
80. Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of M_w from broadband P waveforms. *Bull Seism Soc Am* 85:606–613
81. Tsuboi S, Whitmore PM, Sokolovski TJ (1999) Application of M_{wp} to deep and teleseismic earthquakes. *Bull Seism Soc Am* 89:1345–1351
82. Uhrhammer RA, Collins ER (1990) Synthesis of Wood-Anderson seismograms from broadband digital records. *Bull Seism Soc Am* 80:702–716
83. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of earthquake and engineering seismology, Part A*. Academic Press, Amsterdam, pp 733–746
84. Vanek J, Zátopek A, Kárník V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovév S, Shebalin N (1962) Standardization of magnitude scales. *Izv Acad Sci USSR, Geophys Ser*, pp 108–111 (English translation)
85. Weinstein S, Okal E (2005) The mantle magnitude M_m and the slowness parameter Θ : Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 95:779–799
86. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for M_{wp} . *Sci Tsunami Hazard* 20(4):187–192
87. Willmore PL (ed) (1979) *Manual of seismological observatory practice*, World data center A for solid earth geophysics. Report SE–20. Boulder, Colorado
88. Wu Z (2001) Scaling of apparent stress from broadband radiated energy catalogue and seismic moment catalogue and its focal mechanism dependence. *Earth Planets Space* 53:943–948
89. Wu KM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. *Bull Seism Soc Am* 95:347–353
90. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada regions. *J Geophys Res* 73:4681–4694

Books and Reviews

- Båth M (1979) *Introduction to seismology*. Birkhäuser, Basel
- Bolt BA (1999) *Earthquakes*, 4th edn. W.H. Freeman, San Francisco
- Duda S, Aki K (Eds) (1983) *Quantification of earthquakes*. *Tectonophysics* 93(3/4):183–356
- Gasperini P, Manfredi G, Zschau J (eds) (2007) *Earthquake early warning systems*. Springer, Berlin
- Kulháněk O (1990) *Anatomy of seismograms*. *Developments in solid earth geophysics*, vol 18. Elsevier, Amsterdam
- Lay T, Wallace TC (1995) *Modern global seismology*. Academic Press, San Diego
- Lee WHK (ed) (1988) *Historical seismograms and earthquakes of the world*. Academic Press, New York
- Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) (2002) *International handbook of earthquake and engineering seismology*, part A and B. Academic Press, London (an imprint of Elsevier Science)
- Scholz CH (2002) *The mechanics of earthquake faulting*, 2nd edn. Cambridge University Press, Cambridge
- Shearer PM (1999) *Introduction to seismology*. Cambridge University Press, Cambridge
- Stein S, Wysession M (2002) *Introduction to seismology, Earthquakes and earth structure*. Blackwell Publishing, Malden

Earthquake Monitoring and Early Warning Systems

WILLIAM H. K. LEE¹, YIH-MIN WU²

¹ US Geological Survey, Menlo Park, USA

² Department of Geosciences, National Taiwan University, Taipei, Taiwan

Article Outline

Glossary

Definition of the Subject

Introduction

Earthquake Monitoring: Instrumentation

[Earthquake Monitoring:](#)
[Regional and Local Networks](#)
[Seismograms and Derived Products](#)
[Earthquake Early Warning \(EEW\) Systems](#)
[Future Directions](#)
[Acknowledgments](#)
[Appendix: A Progress Report on Rotational Seismology](#)
[Bibliography](#)

Glossary

- Active fault** A *fault* (q.v.) that has moved in historic (e. g., past 10,000 years) or recent geological time (e. g., past 500,000 years).
- Body waves** Waves which propagate through the interior of a body. For the Earth, there are two types of seismic body waves: (1) compressional or longitudinal (*P* wave), and (2) shear or transverse (*S* wave).
- Coda waves** Waves which are recorded on a *seismogram* (q.v.) after the passage of *body waves* (q.v.) and *surface waves* (q.v.). They are thought to be back-scattered waves due to the Earth's inhomogeneities.
- Earthquake early warning system (EEWS)** An earthquake monitoring system that is capable of issuing warning message after an earthquake occurred and before strong ground shaking begins.
- Earthquake precursor** Anomalous phenomenon preceding an earthquake.
- Earthquake prediction** A statement, in advance of the event, of the time, location, and *magnitude* (q.v.) of a future earthquake.
- Epicenter** The point on the Earth's surface vertically above the *hypocenter* (q.v.).
- Far-field** Observations made at large distances from the *hypocenter* (q.v.), compared to the wave-length and/or the source dimension.
- Fault** A fracture or fracture zone in the Earth along which the two sides have been displaced relative to one another parallel to the fracture.
- Fault slip** The relative displacement of points on opposite sides of a *fault* (q.v.), measured on the fault surface.
- Focal mechanism** A description of the orientation and sense of slip on the causative fault plane derived from analysis of *seismic waves* (q.v.).
- Hypocenter** Point in the Earth where the rupture of the rocks originates during an earthquake and *seismic waves* (q.v.) begin to radiate. Its position is usually determined from arrival times of *seismic waves* (q.v.) recorded by *seismographs* (q.v.).
- Intensity, earthquake** Rating of the effects of earthquake vibrations at a specific place. Intensity can be estimated from instrumental measurements, however, it is formally a rating assigned by an observer of these effects using a descriptive scale. Intensity grades are commonly given in Roman numerals (in the case of the Modified Mercalli Intensity Scale, from I for "not perceptible" to XII for "total destruction").
- Magnitude, earthquake** Quantity intended to measure the size of earthquake at its source, independent of the place of observation. *Richter magnitude* (M_L) was originally defined in 1935 as the logarithm of the maximum amplitude of seismic waves in a seismogram written by a Wood–Anderson seismograph (corrected to) a distance of 100 km from the epicenter. Many types of magnitudes exist, such as *body-wave magnitude* (m_b), *surface-wave magnitude* (M_S), and *moment magnitude* (M_W).
- Moment tensor** A symmetric second-order tensor that characterizes an internal seismic point source completely. For a finite source, it represents a point source approximation and can be determined from the analysis of *seismic waves* (q.v.) whose wavelengths are much greater than the source dimensions.
- Near-field** A term for the area near the causative rupture of an earthquake, often taken as extending a distance from the rupture equal to its length. It is also used to specify a distance to a seismic source comparable or shorter than the wavelength concerned. In engineering applications, near-field is often defined as the area within 25 km of the fault rupture.
- Plate tectonics** A theory of global *tectonics* (q.v.) in which the Earth's lithosphere is divided into a number of essentially rigid plates. These plates are in relative motion, causing earthquakes and deformation along the plate boundaries and adjacent regions.
- Probabilistic seismic hazard analysis** Available information on earthquake sources in a given region is combined with theoretical and empirical relations among earthquake *magnitude* (q.v.), distance from the source, and local site conditions to evaluate the exceedance probability of a certain ground motion parameter, such as the peak ground acceleration, at a given site during a prescribed time period.
- Seismic hazard** Any physical phenomena associated with an earthquake (e. g., ground motion, ground failure, liquefaction, and tsunami) and their effects on land use, man-made structure, and socio-economic systems that have the potential to produce a loss.
- Seismic hazard analysis** The calculation of the *seismic hazard* (q.v.), expressed in probabilistic terms (See *probabilistic seismic hazard analysis*, q.v.). The result is usually displayed in a *seismic hazard map* (q.v.).

Seismic hazard map A map showing contours of a specified ground-motion parameter or response spectrum ordinate for a given *probabilistic seismic hazard analysis* (q.v.) or return period.

Seismic moment The magnitude of the component couple of the double couple that is the point force system equivalent to a *fault slip* (q.v.) in an isotropic elastic body. It is equal to rigidity times the fault slip integrated over the fault plane. It can be estimated from the far-field seismic spectrum at wave lengths much longer than the source size. It can also be estimated from the near-field seismic, geologic and geodetic data. Also called “scalar seismic moment” to distinguish it from *moment tensor* (q.v.).

Seismic risk The risk to life and property from earthquakes.

Seismic wave A general term for waves generated by earthquakes or explosions. There are many types of seismic waves. The principle ones are *body waves* (q.v.), *surface waves* (q.v.), and *coda waves* (q.v.).

Seismograph Instrument which detects and records ground motion (and especially vibrations due to earthquakes) along with timing information. It consists of a *seismometer* (q.v.) a precise timing device, and a recording unit (often including telemetry).

Seismogram Record of ground motions made by a *seismograph* (q.v.).

Seismometer Inertial sensor which responds to ground motions and produces a signal that can be recorded.

Source parameters of an earthquake The parameters specified for an earthquake source depends on the assumed earthquake model. They are origin time, *hypocenter* (q.v.), *magnitude* (q.v.), *focal mechanism* (q.v.), and *moment tensor* (q.v.) for a point source model. They include fault geometry, rupture velocity, stress drop, slip distribution, etc. for a finite fault model.

Surface waves Waves which propagate along the surface of a body or along a subsurface interface. For the Earth, there are two common types of seismic surface waves: Rayleigh waves and Love waves (both named after their discoverers).

Tectonics Branch of Earth science which deals with the structure, evolution, and relative motion of the outer part of the Earth, the lithosphere. The lithosphere includes the Earth’s crust and part of the Earth’s upper mantle and averages about 100 km thick. See *plate tectonics* (q.v.).

Teleseism An earthquake at an epicentral distance greater than about 20° or 2000 km from the place of observation.

Definition of the Subject

When a sudden rupture occurs in the Earth, elastic (seismic) waves are generated. When these waves reach the Earth’s surface, we may feel them as a series of vibrations, which we call an earthquake. *Seismology* is derived from the Greek word *σεισμός* (seismos or earthquake) and *λόγος* (logos or discourse); thus, it is the science of earthquakes and related phenomena. Seismic waves can be generated naturally by earthquakes or artificially by explosions or other means. We define earthquake monitoring as a branch of seismology, which systematically observes earthquakes with instruments over a long period of time.

Instrumental recordings of earthquakes have been made since the later part of the 19th century by seismographic stations and networks of various sizes from local to global scales. The observed data have been used, for example, (1) to compute the source parameters of earthquakes, (2) to determine the physical properties of the Earth’s interior, (3) to test the theory of plate tectonics, (4) to map active faults, (5) to infer the nature of damaging ground shaking, and (6) to carry out seismic hazard analyzes. Constructing a satisfactory theory of the complex earthquake process has not yet been achieved within the context of physical laws, e. g., realistic equations for modeling earthquakes do not exist at present. Good progress, however, has been made in building a physical foundation for the earthquake source process [62], partly as a result of research directed toward earthquake prediction.

Earthquakes release large amounts of energy that potentially can cause significant damage and human deaths. During an earthquake, potential energy (mainly elastic strain energy and some gravitational energy) that has accumulated in the hypocentral region over decades to centuries or longer is released suddenly [63]. This energy is partitioned into (1) radiated energy in the form of propagating seismic waves, (2) energy consumed in overcoming fault friction, (3) the energy which expands the rupture surface area or changes its properties (e. g., by pulverizing rock), and (4) heat. The radiated seismic energy is a small fraction (about 7%) of the total energy budget, and it can be estimated using the recorded seismograms. Take, for example, the 1971 San Fernando earthquake ($M_W = 6.6$) in southern California. Its radiated energy was about 5×10^{21} ergs, or about 120 kilotons of TNT explosives, or the energy released by six atomic bombs of the size used in World War II. The largest earthquake recorded instrumentally (so far) is the 1960 Chilean earthquake ($M_W = 9.5$). Its radiated energy was about 1.1×10^{26} ergs, an equivalent of about 2,600 megatons of TNT explosives, the energy released by about 130,000

atomic bombs. It is, therefore, no surprise that an earthquake can cause up to hundreds of thousands of human deaths, and produce economic losses of up to hundreds of billions of dollars.

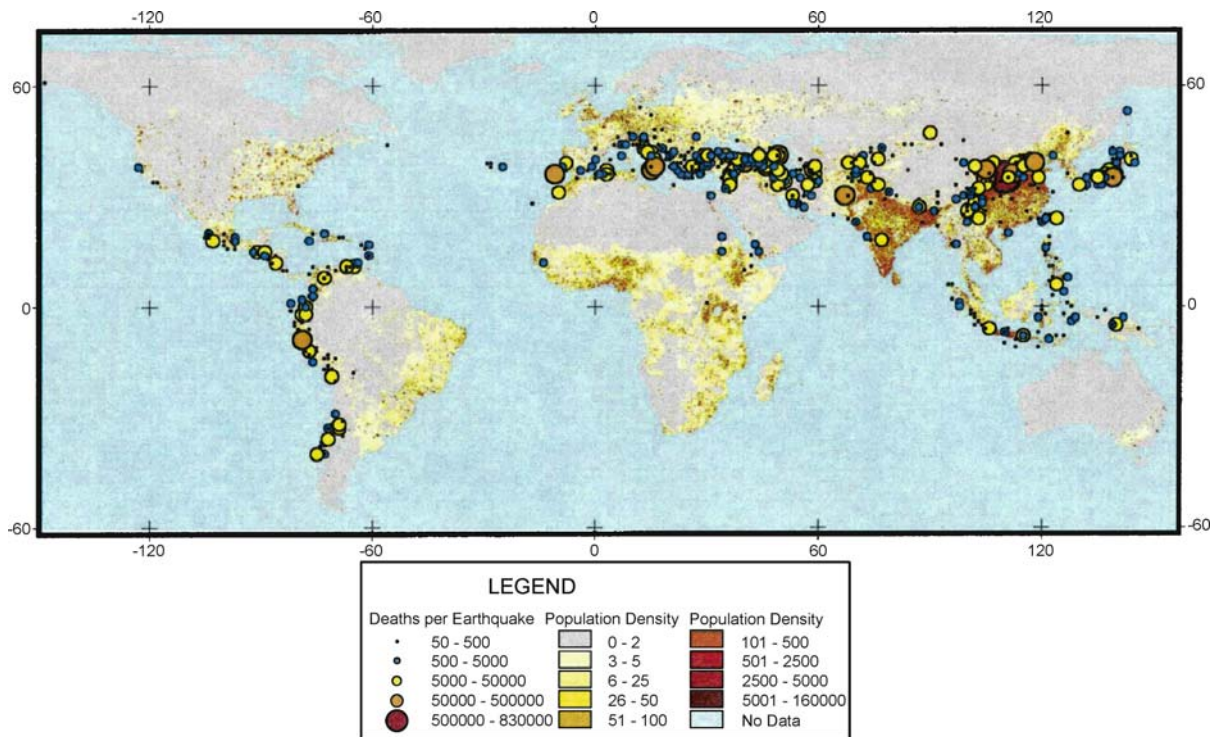
Monitoring earthquakes is essential for providing scientific data to investigate complex earthquake phenomena, and to mitigate seismic hazards. The present article is a brief overview of earthquake monitoring and early warning systems; it is intended for a general scientific audience, and technical details can be found in the cited references. Earthquakes are complex natural phenomena and their monitoring requires an interdisciplinary approach, including using tools from computer science, electrical and electronic engineering, mathematics, physics, and others. Earthquake early warning systems (which are based on earthquake monitoring) offer practical information for reducing seismic hazards in earthquake-prone regions.

After the “Introduction”, we will present a summary of earthquake monitoring, a description of the products derived from the analysis of seismograms, and a discussion of the limitations of these products. Earthquake early warning systems are then presented briefly, and we conclude with a section on future directions, including a progress

report on rotational seismology ([Appendix](#)). We present overviews of most topics in earthquake monitoring, and an extensive bibliography is provided for additional reading and technical details.

Introduction

Earthquakes, both directly and indirectly, have caused much suffering to mankind. During the 20th century alone about two million people were killed as a result of earthquakes. A list of deadly earthquakes (death tolls ≥ 25) of the world during the past five centuries was compiled by Utsu [115]. It shows that earthquakes of magnitude ≥ 6 (~ 150 per year worldwide) can be damaging and deadly if they occur in populated areas, and if their focal depths are shallow (< 50 km). Seismic risk can be illustrated by plotting the most deadly earthquakes of the past five centuries (1500–2000) over a map of current population density. This approach was used by Utsu [115], and his result is shown in Fig. 1. Most of these deadly earthquakes are concentrated (1) along the coasts of Central America, the Caribbean, western South America, and Indonesia, and (2) along a belt that extends from southern Europe,



Earthquake Monitoring and Early Warning Systems, Figure 1

Location of deadly earthquakes around the world, 1500–2000. Population density is shown by the *background colors*. See [115] for details

Earthquake Monitoring and Early Warning Systems, Table 1

Deadly Earthquakes/Tsunamis from 1896–2005 ([115] and recent sources)

Origin Time Year MM/DD Hr:Min (UTC, except L=local)	Hypocenter			Magnitude	Location	Deaths (Approximate)
	Lat. (deg)	Lon. (deg)	Depth (km)			
2005 10/08 3:50	34.432	73.573	10	7.6	Pakistan, Kashmir	80,361+
2004 12/26 0:58	3.298	95.778	7	9.2	Indonesia, Sumatra	283,106+
2003 12/26 1:56	29.004	58.337	15	6.6	Iran, Bam	26,000
2001 01/26 3:16	23.420	70.230	16	7.7	India, Gujarat, Bhuj	20,000+
1990 06/20 21:00	37.008	49.213	18	7.4	Iran, western	~40,000
1988 12/07 7:41	40.919	44.119	7	6.8	Armenia, Spitak	~40,000
1976 07/27 19:42	39.605	117.889	17	7.6	China, Tangshan	~242,000
1976 02/04 9:01	15.298	−89.145	13	7.5	Guatemala	23,000
1970 05/31 20:23	−9.248	−78.842	73	7.5	Peru	67,000
1948 10/05 20:12	37.500	58.000	0	7.2	USSR, Ashgabat	~65,000
1939 12/26 23:57	39.770	39.533	35	7.7	Turkey, Erzincan	33,000
1939 01/25 3:32	−36.200	−72.200	0	7.7	Chile, Chillian	28,000
1935 05/30 21:32	28.894	66.176	35	8.1	Pakistan, Quetta	60,000
1932 12/25 2:04	39.771	96.690	25	7.6	China, Gansu	~70,000
1927 05/22 22:32	37.386	102.311	25	7.7	China, Tsinghai	~100,000
1923 09/01 2:58	35.405	139.084	35	7.9	Japan, Kanto	143,000
1920 12/16 12:05	36.601	105.317	25	8.6	China, Gansu	~240,000
1915 01/13 6:52	42.000	13.500	0	6.9	Italy, Avezzano	33,000
1908 12/28 4:20	38.000	15.500	0	7.0	Italy, Messina	~82,000
1906 08/17 0:40	−33.000	−72.000	0	8.2	Chile, Valparaiso	20,000
1905 04/04 0:50	33.000	76.000	0	8.1	India, Kangra	20,000
1896 06/15 19:32L	39.500	144.000	0	8.2	Japan, Sanriku-oki	22,000

“~” denotes large uncertainties because a range of deaths had been reported.

“+” denotes a minimum value.

the Middle East, Iran, Pakistan and India, to China and Japan.

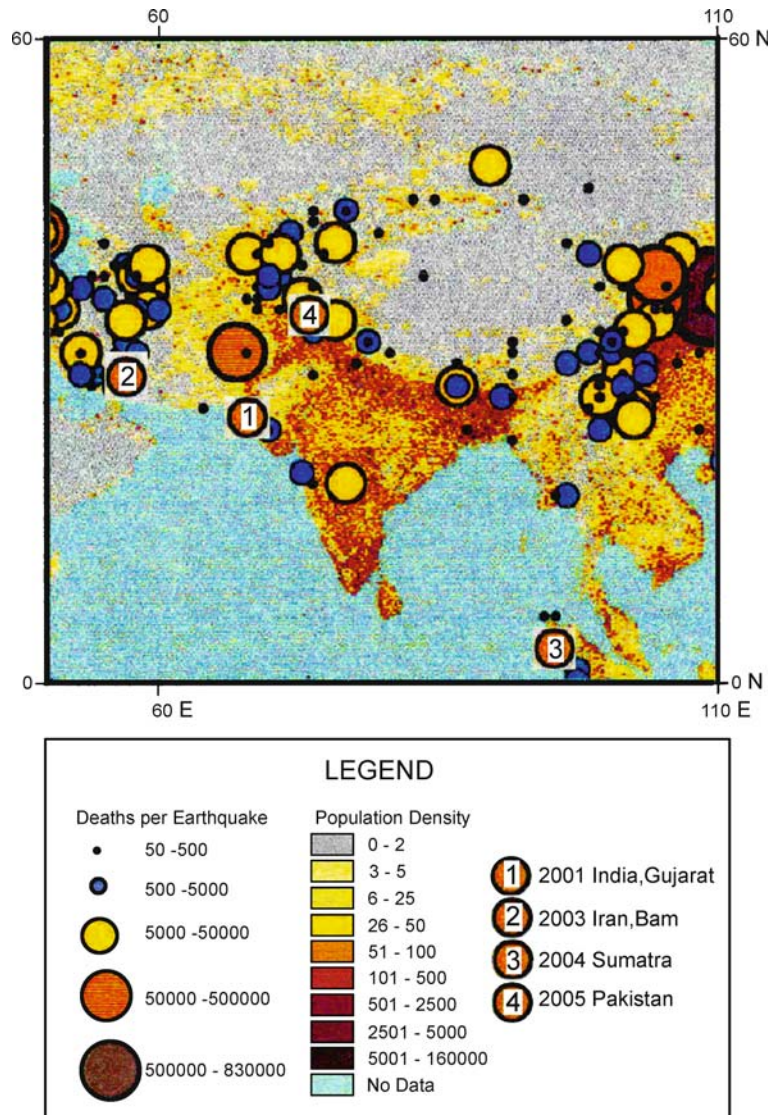
Table 1 lists the most deadly earthquakes (death toll > 20,000) of the past 110 years based on official estimates (often under-estimated for political reasons, or lack of accurate census data in many areas of the world). In the first 5 years of the 21st century, four disastrous earthquakes occurred in India, Indonesia, Iran, and Pakistan. In the 20th century, the average death toll caused by earthquakes (and tsunamis they triggered) was about 16,000 per year. For the past seven years the yearly death toll was about 60,000 – four times higher than the average in the previous century. In Fig. 2 we extracted a portion of Fig. 1 to illustrate the relationship between past earthquakes and population in India, Pakistan, northern Indonesia, and adjoining regions. We numbered the four most recent disastrous earthquakes in Fig. 2. It is obvious that the large populations in India, Indonesia, Iran, Pakistan, and their adjoining regions (over 1.5 billion people) has been and will continue to be adversely affected by earthquakes. Fatalities depend largely on resistance of building construction to

shaking, in addition to population density and earthquake occurrence.

In recent decades, population increases, accelerated urbanization, and population concentration along coastal areas prone to earthquakes suggest that many more earthquake-related fatalities will occur unless effective steps are taken to minimize earthquake and tsunami hazards.

Earthquake Monitoring: Instrumentation

Besides geodetic data [28], the primary instrumental data for the quantitative study of earthquakes are *seismograms*, records of ground motion caused by the passage of seismic waves. Seismograms are written by *seismographs*, instruments which detect and record ground motion along with timing information. A seismograph consists of three basic components: (1) a seismometer, which responds to ground motion and produces a signal proportional to acceleration, velocity, or displacement over a range of amplitudes and frequencies; (2) a timing device; (3) either a local recording unit which writes seismograms on paper, film, or elec-



Earthquake Monitoring and Early Warning Systems, Figure 2

Location of the 4 most deadly earthquakes of the 21st century (up to the end of 2007) on a map showing the location of the deadly earthquakes from 16th to 20th centuries (after [115] and Table 1)

tronic storage media, or more recently, a telemetry system for delivering the seismograms to a central laboratory for recording. Technical discussions of seismometry may be found, for example, in Wielandt [122], and of seismic instruments in Havskov and Alguacil [48]. An overview of challenges in observational earthquake seismology is given by Lee [71], and a useful manual of seismological observatory practice is provided by Bormann [12].

An *accelerograph* is a seismograph designed to record, on scale, the acceleration time history of strong ground

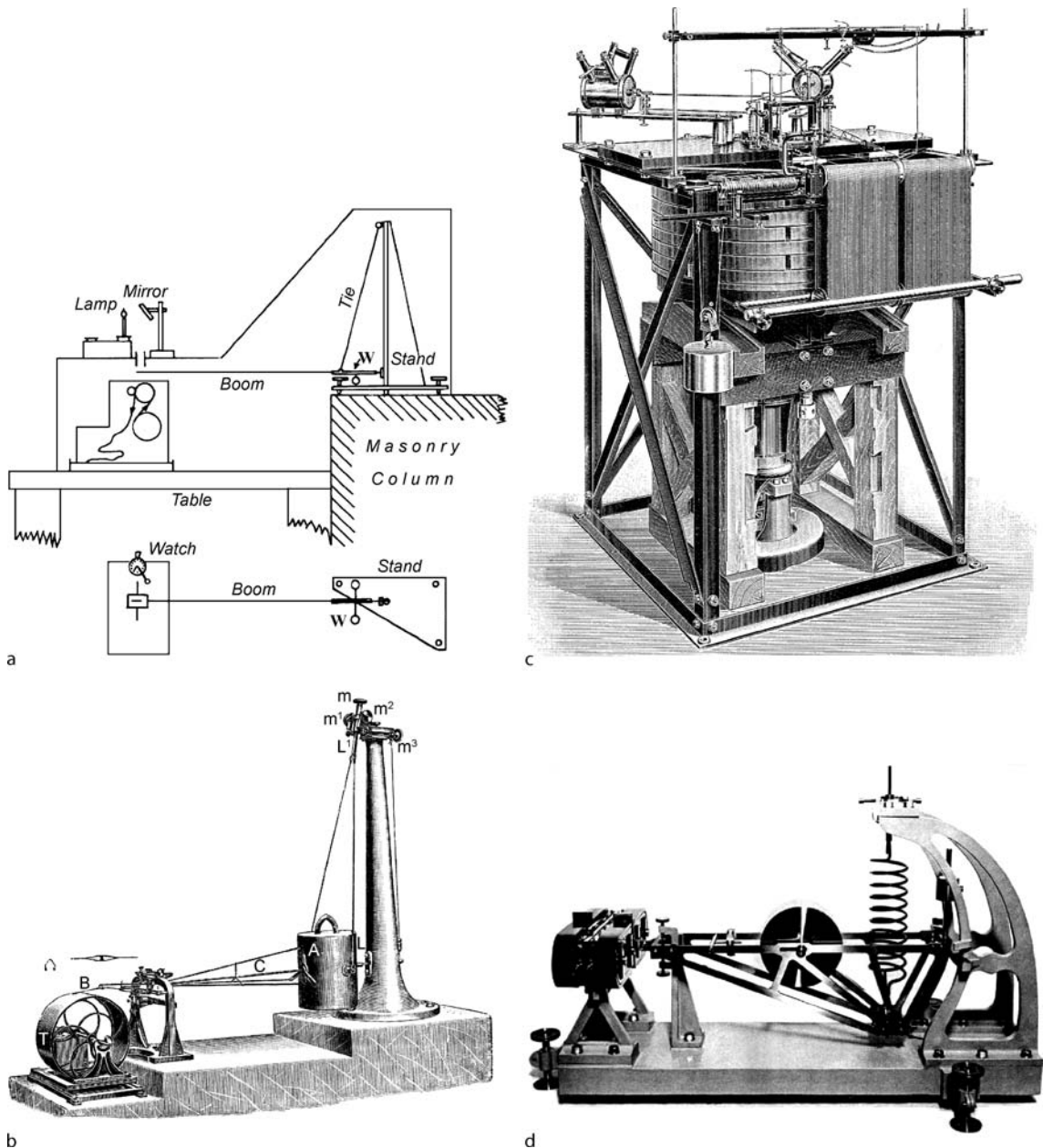
motions. Measuring acceleration is important for studying response of buildings to strong ground motions close to earthquakes. Many modern sensitive seismographs are *velocigraphs* recording the time history of ground velocity. They are designed to measure seismic waves of small amplitudes (because seismic waves attenuate quickly from their sources) either from small earthquakes nearby, or from large earthquakes that are far away.

A seismic network (or an “array”) is a group of seismographs “linked” to a central headquarters. Nowadays the

link is by various methods of telemetry, but in early days the links were by mail or telegrams, or simply by manual collection of the records. When we speak of a seismic *station*, we may mean an observatory with multiple instruments in special vaults or a small instrument package at a remote site.

Seismographs were first developed in the late 19th century, and individual seismographic observatories (often

a part of astronomical or meteorological observatories) began earthquake monitoring by issuing earthquake information in their station bulletins and other publications. However, in order to accurately locate an earthquake, data from several seismographic stations are necessary. It was then natural for many governments to assume responsibility for monitoring earthquakes within their territories. However, because seismic waves from earthquakes do not



Earthquake Monitoring and Early Warning Systems, Figure 3

Some classical seismographs: a Milne, b Bosch-Omori, c Wiechert, and d Galitzin (after [101])

recognize national boundaries, the need for international cooperation became clear. In the following subsections, we present an overview of the history and results of earthquake monitoring.

Historical Developments

In 1897, John Milne designed the first inexpensive seismograph, which was capable of recording very large earthquakes anywhere in the world. With a small grant from the British Association for the Advancement of Science (BAAS), a few other donations, and his own money, Milne managed to deploy about 30 of his instruments around the world, forming the first worldwide seismographic network. At the same time, seismogram readings were reported voluntarily to Milne's observatory at Shide on the Isle of Wight, England. A global earthquake summary with these seismogram readings was issued by Milne beginning in 1899. These summaries are now known as the "Shide Circulars". Milne also published progress and results in the "Reports of the BAAS Seismological Committee" from 1895 to 1913. A review of Milne's work and a reproduction of his publications as computer readable files were given by Schweitzer and Lee [101] and its attached CD-ROM. After Milne's death in 1913, Herbert H. Turner continued Milne's efforts, and in 1918 established publication of the International Seismological Summary (ISS).

The shortcomings of the Milne seismograph (low magnification, no damping, and poor time resolution) were soon recognized. Several improved seismographs (notably the Omori, Bosch–Omori, Wiechert, Galitzin, and Milne–Shaw) were developed and deployed in the first three decades of the 20th century. Figure 3 shows several of these classical seismographs (see Schweitzer and Lee [101] for further explanation). Although the ISS provided an authoritative compilation arrival-time data of seismic waves and determinations of earthquake hypocenters beginning in 1918, its shortcomings were also evident. These include difficulties in collecting the available arrival-time data around the world (which were submitted on a voluntary basis), and in the processing and analysis of data from many different types of seismographs. Revolutions and wars during the first half of the 20th century frequently disrupted progress, particularly impacting collection and distribution earthquake information.

In the late 1950s, attempts to negotiate a comprehensive nuclear test ban treaty failed, in part because of perceptions that seismic methods were inadequate for monitoring underground nuclear tests [95]. The influential Berkner report of 1959 therefore advocated major support for seismology [66]. As a result, the Worldwide Standard-

ized Seismograph Network (WWSSN) was created in the early 1960s with about 120 continuously recording stations located across much of the world, except China and the USSR [91]. Each WWSSN station was equipped with *identical* sets of short-period and long-period three-component seismographs and accurate chronometers. Figure 4 shows some of the equipment at a WWSSN station, including three-components of long-period seismometers, long-period recording and test instruments, and the time and power console. A similar set of three-component short-period seismometers and recording and test instruments, nearly identical in appearance, was also deployed at each station. Seismograms from the WWSSN were sent to the United States to be photographed on 70 mm film chips for distribution (about US\$ 1 per chip as then sold to any interested person).

The WWSSN network is credited with making possible rapid progress in global seismology, and with helping to spark the plate tectonics revolution of the late 1960s [117]. At about the same time, the Unified System of Seismic Observations (ESSN) of the former USSR and its allied countries was established, consisting of almost 100 stations equipped with Kirnos short-period, 1–20 s displacement sensors, and long-period seismographs.

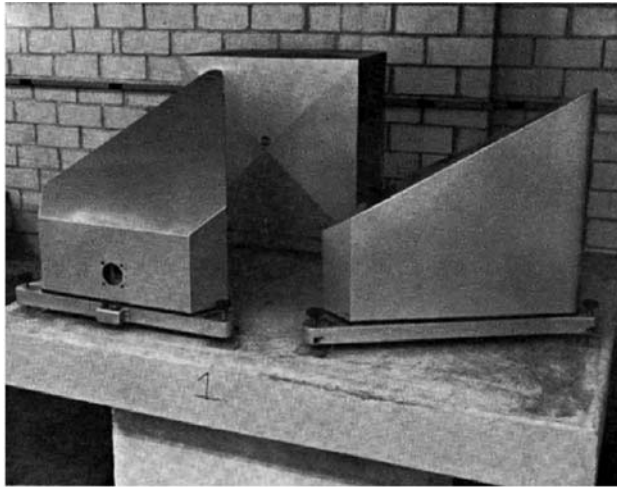
Samples of seismograms recorded on smoked paper and photographic paper or film by analog seismographs are shown in Figs. 5 and 6. Two efforts to preserve and make such records available online are now underway: the *SeismoArchives* (www.iris.edu/seismo/ [72]), and *Sismos* (sismos.rm.ingv.it [82]).

With the establishment of the WWSSN, the United States also assumed the task of monitoring earthquakes on a global scale beginning in the early 1960s. The mission of the US National Earthquake Information Center (NEIC, now part of the US Geological Survey) is "to determine rapidly the location and size of all destructive earthquakes worldwide and to immediately disseminate this information to concerned national and international agencies, scientists, and the general public" (<http://earthquake.usgs.gov/regional/neic/>).

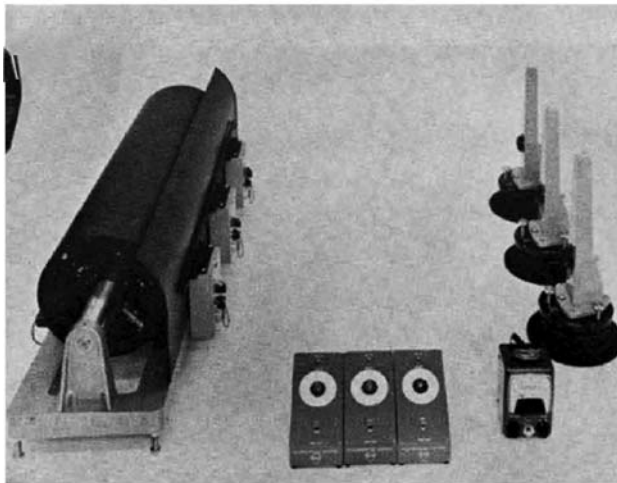
In 1964, the ISS was reorganized as the International Seismological Centre (ISC). Since then, the ISC (<http://www.isc.ac.uk/>) has issued annual global earthquake catalogs with a time lag of about two years [123].

Technical Considerations

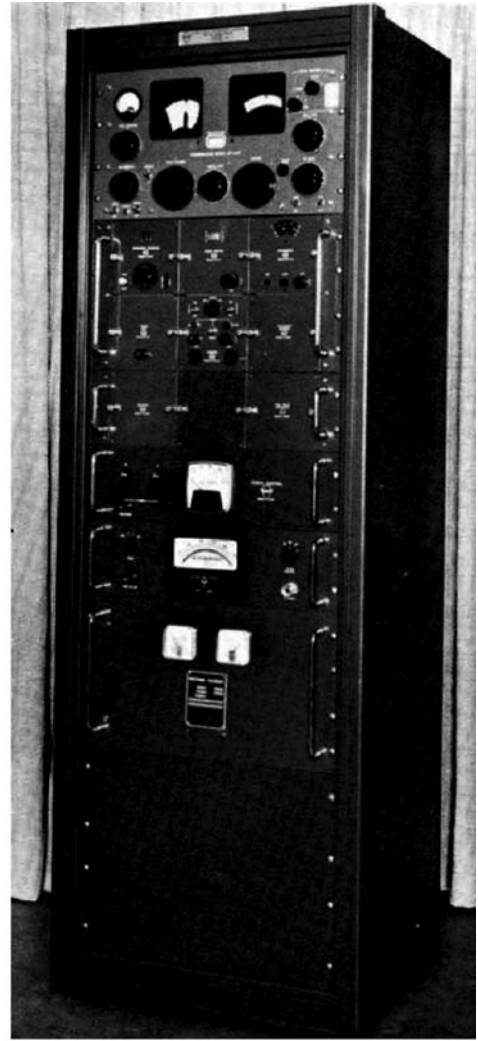
To record seismic waves, we must consider both the available technology for designing seismographs, and the nature of the Earth's background noise [121]. The Earth is constantly in motion. This "background" noise is usu-



a LONG-PERIOD SEISMOMETERS INSTALLED ON A PIER



b LONG-PERIOD AND TEST INSTRUMENTS



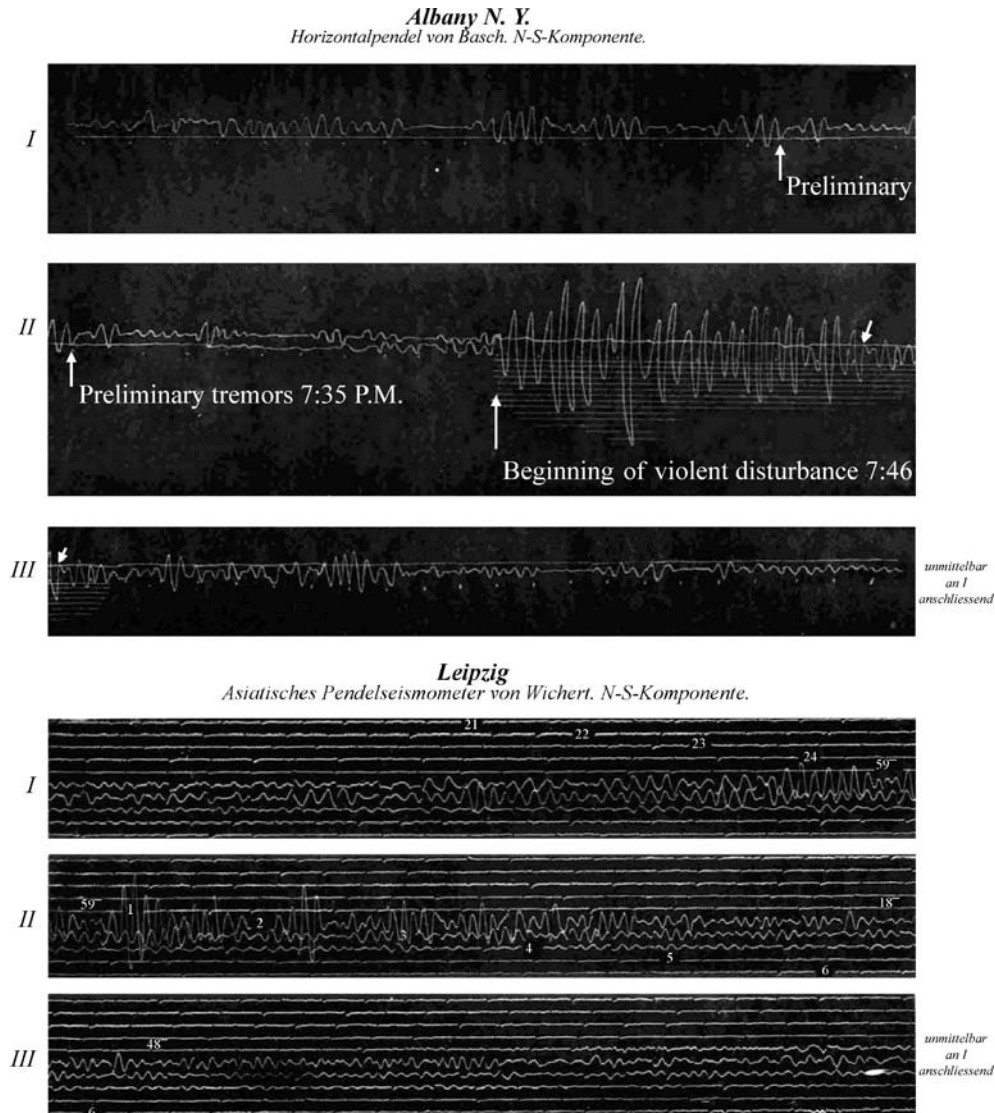
c TIME AND POWER CONSOLE

Earthquake Monitoring and Early Warning Systems, Figure 4

Some WWSSN station equipment: **a** Three-component, long-period seismometers installed on a seismic pier, **b** Long-period recording and test instruments, and **c** Time and power console. A similar set of three-component, short-period seismometers and recording/test instruments is not shown

ally classified as either (1) *microseisms*, which typically have frequencies below about 1 Hz, are often the largest background signals, and are usually caused by natural disturbances (largely caused by ocean waves near shorelines); or (2) *microtremors*, which have frequencies higher than about 1 Hz, and are due to human activities (such as traffic and machinery) and local natural sources (such as wind and moving vegetation). Ground motions from earthquakes vary more than ten orders of magnitude in amplitude and six orders of magnitude in frequency, depending on the size of the earthquake and the distance at

which it is recorded. Figure 7 illustrates the relative dynamic range of some common seismometers for global earthquake monitoring. A “low Earth noise” model [10,92] is the lower limit of Earth’s natural noise in its quietest locations – it is desirable to have instruments that are sensitive enough to detect this minimal background Earth signal. In the analog instrument era (i.e., prior to about 1980), short-period and long-period seismometers were designed separately to avoid microseisms, which have predominant periods of about 6 s. Short-period seismometers were designed to detect tiny ground motions from



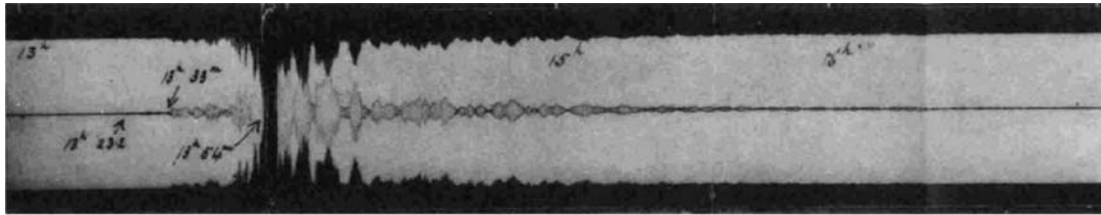
Earthquake Monitoring and Early Warning Systems, Figure 5
Some sample analog seismograms recorded on smoked paper

smaller, nearby earthquakes, while long-period instruments were designed to recover the motions of distant, larger earthquakes (“teleseisms”). Additionally, strong-motion accelerometers, generally recording directly onto 70 mm-wide film strips, were used to measure large motions from nearby earthquakes. In today’s much more capable digital instrumentation, two major types of instruments are deployed: (1) “broadband” seismometers, which replace and improve upon both short-period and long-period seismometers, and (2) strong-motion accelerometers for high-amplitude, high-frequency, seismic waves from local earthquakes, which often drive broadband seis-

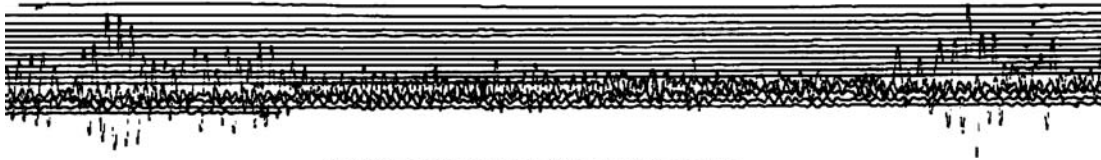
mometers off scale. While rare examples of the old analog instruments are still in use, the vast majority of instruments presently operating are digital.

In addition to having large variations in amplitudes and frequencies, seismic waves from earthquakes also attenuate rapidly with distance, that is, they lose energy as they travel, particularly at higher frequencies. We must consider these effects in order to monitor seismic waves effectively.

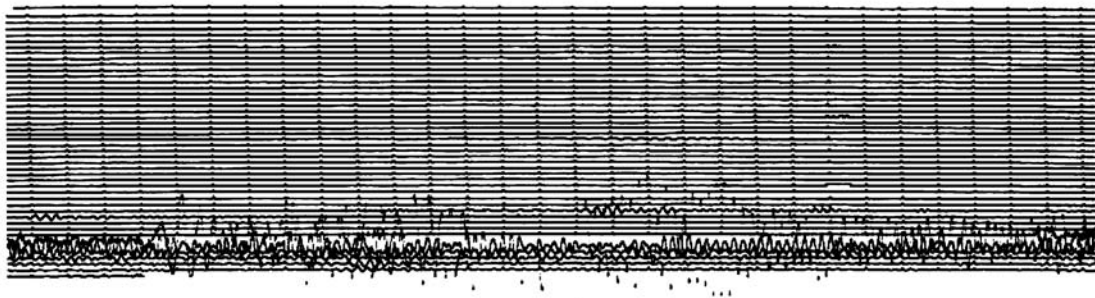
In 1935, C.F. Richter introduced the concept of *magnitude* to classify local earthquakes by their “size”, effectively the amount of energy radiated at the actual rup-



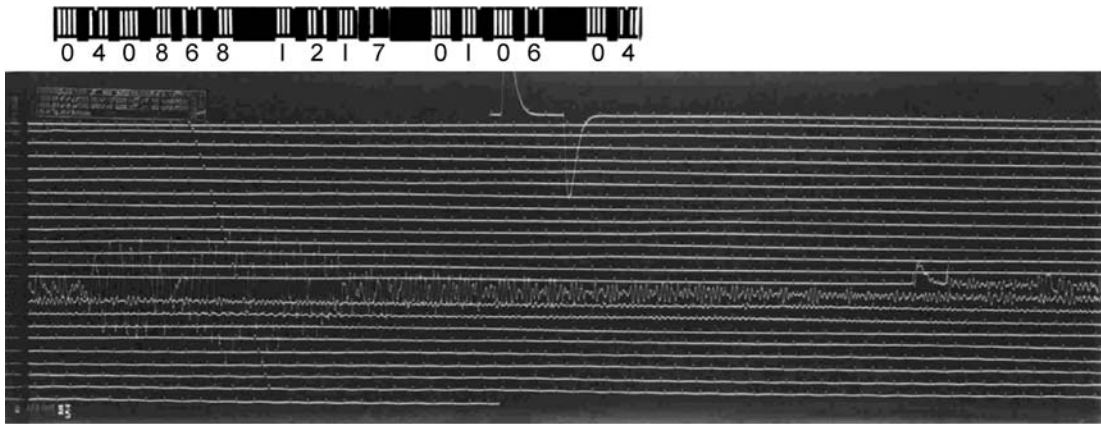
PAISLEY, SCOTLAND. Milne Seismograph. (From photographic copy.)



De Bilt, the Netherlands. Galitzin Seismograph.



Weston Observatory, USA. Benioff Seismograph



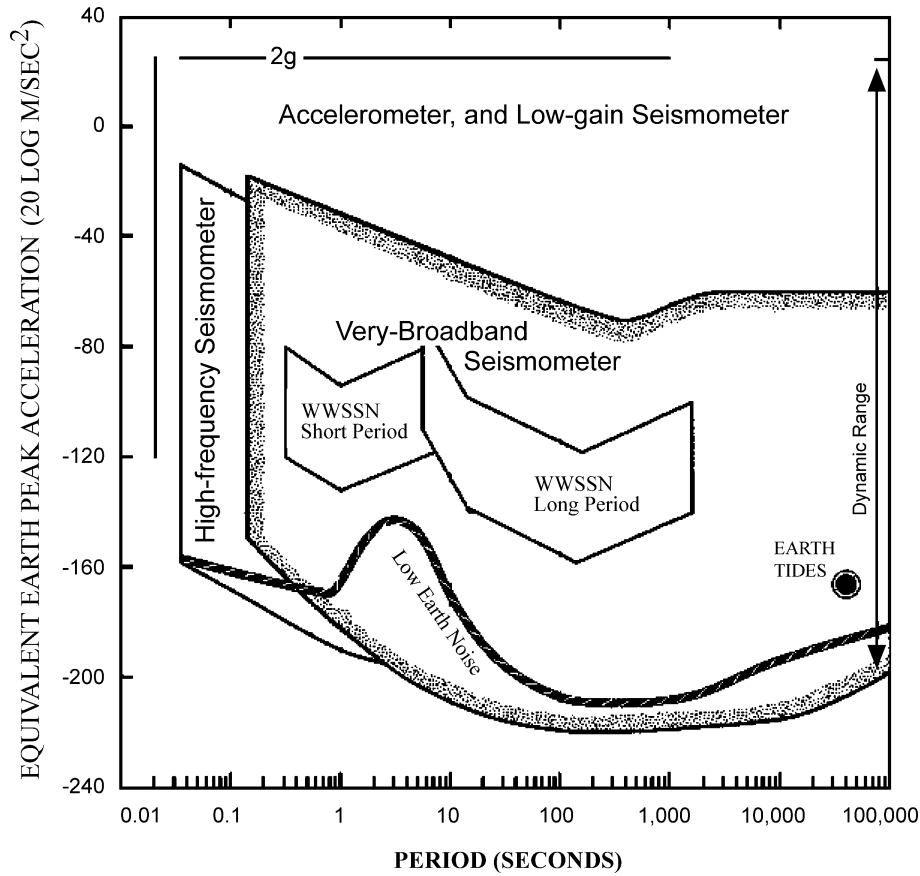
San Juan, Puerto Rico. WWSSN Long-Period Seismograph.

Earthquake Monitoring and Early Warning Systems, Figure 6

Some sample analog seismograms recorded on photographic paper or film

ture surface within the Earth. See the entry by Bormann and Saul ► [Earthquake Magnitude](#) for a discussion of the various magnitude scales in use. While every effort is made to make these different scales overlap cleanly, each has strengths and weaknesses that make one or another preferable in a given situation. Probably the most general

and robust of these methods is called a “moment magnitude”, symbolized as M_W . Existing instruments and environments are such that the smallest natural earthquakes we routinely observe close by are about magnitude = 1. The largest earthquake so far recorded by instrumentals is the $M_W = 9.5$ Chilean earthquake in 1960. In 1941, B.



Earthquake Monitoring and Early Warning Systems, Figure 7

Relative dynamic range of some common seismometers for global earthquake monitoring (modified from Fig. 1 in [54]). The Y-axis is marked in decibel (dB) where $\text{dB} = 20 \log(A/A_0)$; A is the signal amplitude, and A_0 is the reference signal amplitude

Gutenberg and C.F. Richter discovered that over large geographic regions the rate of earthquake occurrence is empirically related to their magnitudes by:

$$\log N = a - bM \quad (1)$$

where N is the number of earthquakes of magnitude M or greater, and a and b are numerical constants. It turns out that b is usually about 1, which implies that $M = 6$ earthquakes are about ten times more frequent than $M = 7$ earthquakes. Engdahl and Villasenor [24] show that there has been an *average* of about 15 major ($M \geq 7$) earthquakes per year over the past 100 years, and about 150 large ($M \geq 6$) earthquakes per year during this same time interval. Strong ground motions (above 0.1 g in acceleration) over sizeable areas are generated by $M \geq 6$ earthquakes; these are potentially damaging levels of ground shaking.

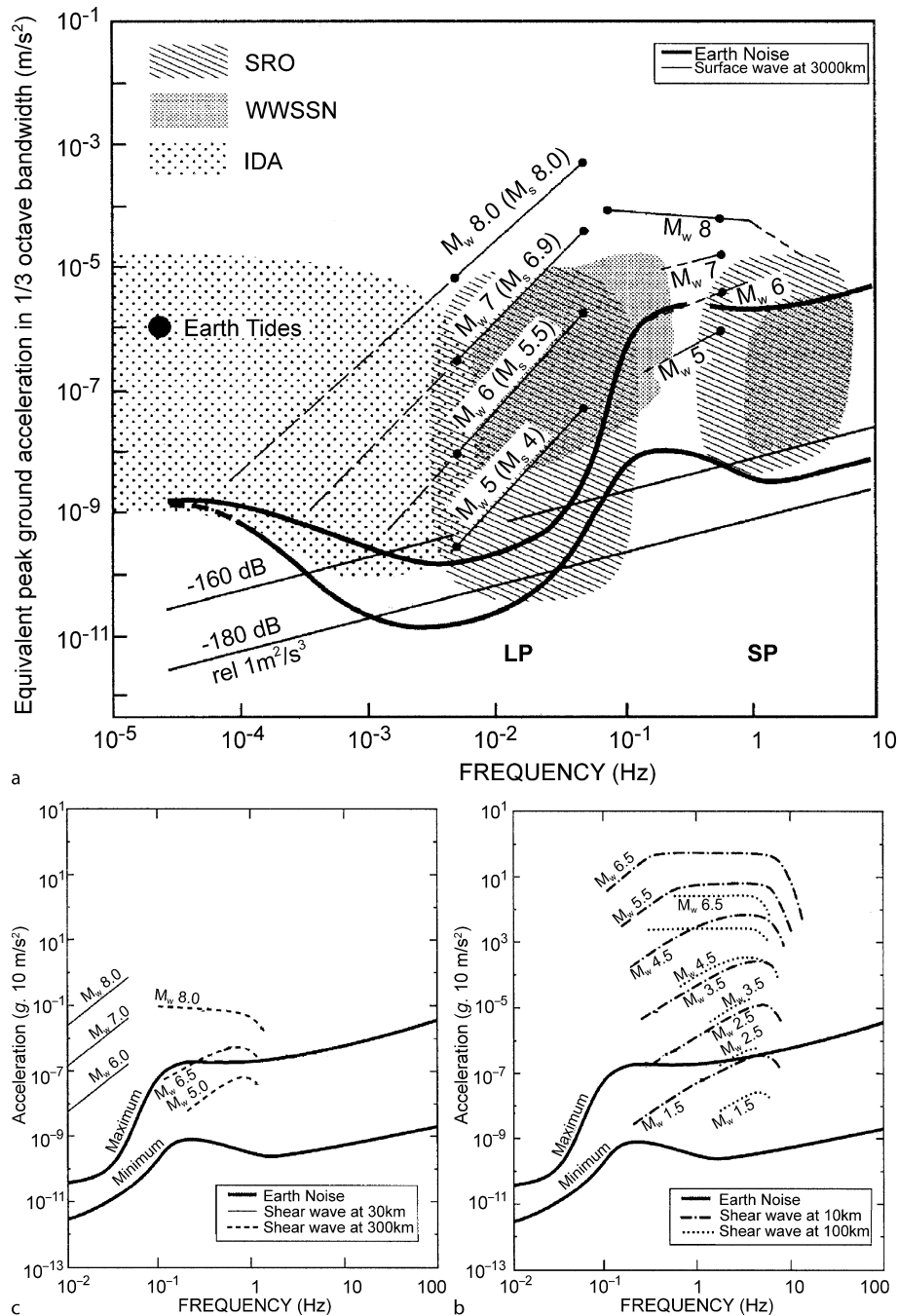
Earthquakes are classified by magnitude (M) as *major* if $M \geq 7$, as *moderate* to *large* if M ranges from 5 to 7,

as *small* if M ranges from 3 to 5, as *micro* if $M < 3$, and as *nano* if $M < 0$. An earthquake with $M \geq 7 \frac{3}{4}$ is often called *great*, and if $M \geq 9$, *mega*.

Earthquake Monitoring in the Digital Era

Figure 8 shows the expected amplitudes of seismic waves by earthquake magnitude. The top frame is a plot of the equivalent peak ground acceleration versus frequency. The two heavy curves denote the “minimum Earth noise”, and the “maximum Earth noise” (i.e., for seismographic station located in the continental interior versus near the coast).

The two domains of the WWSSN equipment, short-period long-period seismometers are shown as gray shading. The domains for two other instruments, SRO (Seismic Research Observatories Seismograph) and IDA (International Deployment of Accelerometers), are also shown; these were the early models of the current instruments

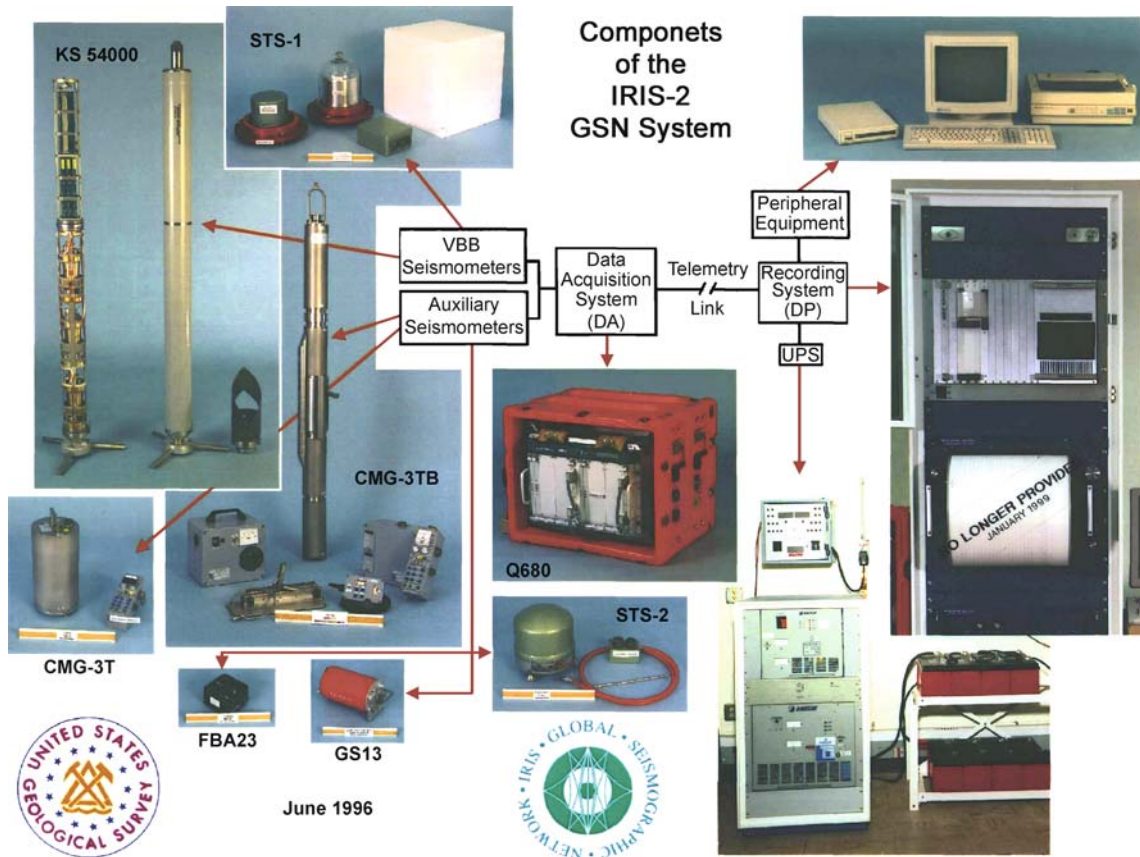


Earthquake Monitoring and Early Warning Systems, Figure 8

Expected amplitudes of seismic waves by earthquake magnitude. See text for explanations

now in operation in the *Global Seismographic Network* (GSN). The bottom two frames indicate expected amplitudes of seismic waves from earthquakes of a range of magnitudes (we use the moment magnitude, M_W). For

simplicity, we consider two cases: (bottom left) global earthquakes recorded at a large distance with a seismographic network spaced at intervals of about 1000 km, and (bottom right) local earthquakes recorded at short dis-



Earthquake Monitoring and Early Warning Systems, Figure 9

Components of the IRIS-2 GSN System: broadband seismometers, accelerometers and recording equipment

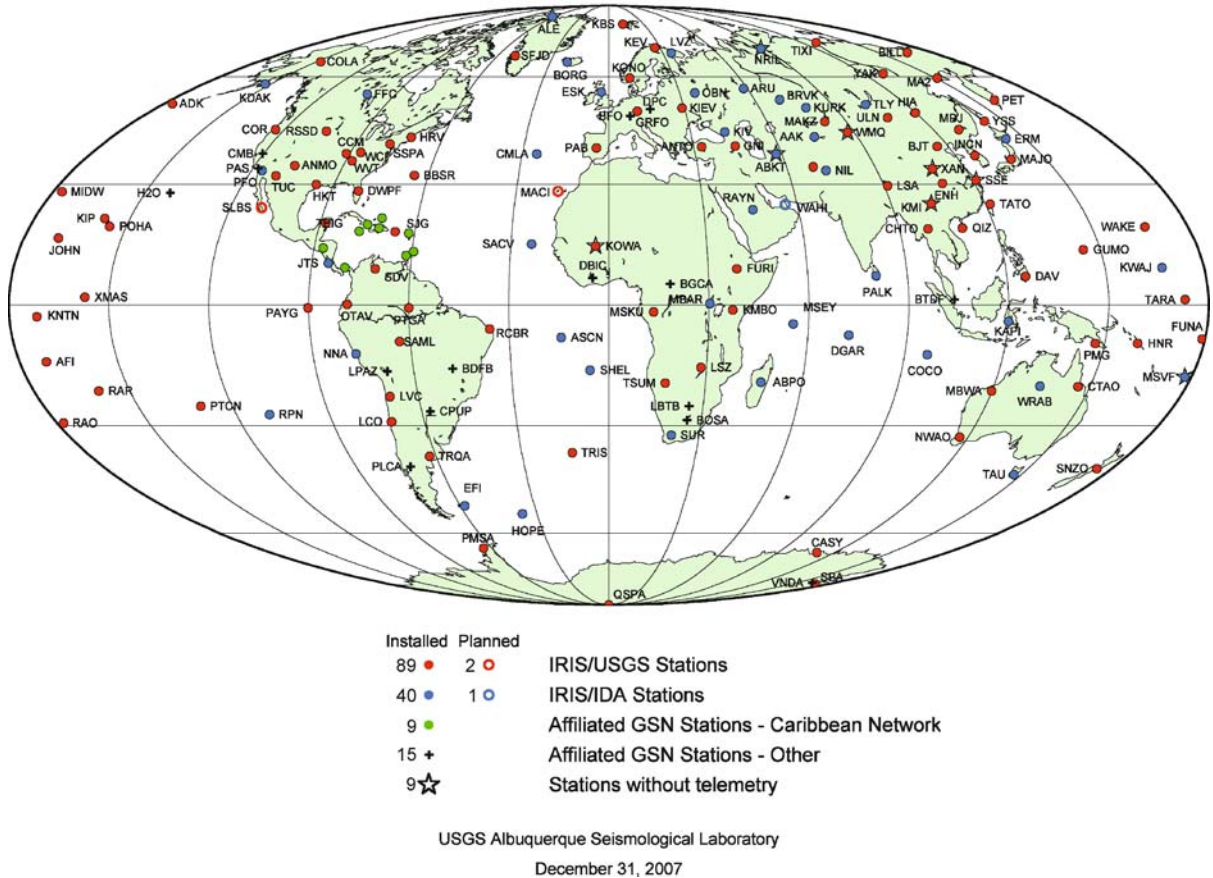
tances with a seismic array spaced at intervals of about 50 km. In the bottom left plot, the global-scale network, the expected amplitudes of *P*-wave and surface wave at 3000 km from the earthquake source are shown; for the bottom right plot, a local seismic array, the expected amplitudes of *S*-wave at 10 km and 100 km from the earthquake source are shown. Seismologists use this and similar figures in planning seismographic networks. Local noise surveys are usually conducted as well when designing specific seismographic networks.

With advances in digital technology, earthquake monitoring entered the digital era in the 1980s. Older analog equipment was gradually phased out as modern digital equipment replaced it [54]. The WWSSN was replaced by the *Global Seismographic Network* (GSN), a collaboration of several institutions under the IRIS consortium (<http://www.iris.edu/>). The goal of the GSN (<http://www.iris.edu/about/GSN/index.htm>) is “to deploy over 128 permanent seismic recording stations uniformly over the Earth’s sur-

face”. The GSN project provides funding for two network operators: (1) the IRIS/ASL Network Operations Center, in Albuquerque, New Mexico (operated by the US Geological Survey), and (2) the IRIS/IDA Network Operations Center in La Jolla, California (operated by personnel from the Scripps Institution of Oceanography). Components of a modern IRIS GSN seismograph system, which include broadband seismometers, accelerometers, and recording equipment, are shown in Fig. 9.

Figure 10 shows the station map of the Global Seismographic Network as of 2007. IRIS GSN stations continuously record seismic data from very broad band seismometers at 20 samples per second (sps), and also include high-frequency (40 sps) and strong-motion (1 and 100 sps) sensors where scientifically warranted. It is the goal of the GSN project to provide real-time access to its data via Internet or satellite. Since 1991, the IRIS Data Management Center has been providing easy access to comprehensive seismic data from the GSN and elsewhere [1].

Global Seismographic Network



Earthquake Monitoring and Early Warning Systems, Figure 10
Station map of the Global Seismographic Network (GSN) as of 2007

Earthquake Monitoring: Regional and Local Networks

A major development in earthquake monitoring was the establishment of seismographic networks optimized to record the many *frequent* but *smaller* regional and local earthquakes occurring in many locations. To observe as many of these nearby earthquakes as possible, inexpensive seismographs with high magnifications and low dynamic-range telemetry are used to record the smallest earthquakes feasible with current technology and local background noise. As a result, the recorded amplitudes often overdrive the instruments for earthquakes with $M \gtrsim 3$ within about 50 km of such seismographs. This is not a serious defect, since the emphasis for these networks is to obtain as many first arrival times as possible, and to detect and to locate the maximum number of earthquakes.

Because seismic waves from small earthquakes are quickly attenuated with increasing distance, it is also necessary to deploy many instruments at small station spacing (generally from a few to a few tens of kilometers), and to cover as large a territory as possible in order to record at least a few earthquakes every week. Since funding often is limited, these local and regional seismic networks are commonly optimized for the largest number of stations rather than for the highest quality data.

A Brief History

In the 1910s, the Carnegie Institution of Washington D.C. (CIW) was spending a great deal of money building the world's then largest telescope (100 inch) at Mount Wilson Observatory, southern California [38]. Since astronomers were concerned about earthquakes that might disturb their

telescopes, Harry O. Wood was able to persuade CIW to support earthquake investigations, and as a result, a regional network of about a dozen Wood–Anderson seismographs was established in southern California in the 1920s. See Goodstein [38] for the early history leading to the establishment of the California Institute of Technology (Caltech) and its Seismological Laboratory. Astronomers played important roles in getting seismic monitoring established in various other regions of the world as well.

Regional networks using different types of seismographs were established in many countries about this time, such as in Japan, New Zealand, and the USSR and its allies. In the 1960s, high-gain, short-period, telemetered networks were developed to study microearthquakes. To support detailed studies of local earthquakes and especially for the purpose of earthquake prediction, over 100 microearthquake networks were established in various parts of the world by the end of the 1970s [74]. These microearthquake networks comprised from tens to hundreds of short-period seismometers, generally with their signals telemetered into central recording sites for processing and analysis. High magnification was achieved through electronic amplification, permitting recording of very small earthquakes (down to $M = 0$), though this came at the expense of saturated records for earthquakes of $M \gtrsim 3$ within about 50 km. Unfortunately, the hope of discovering some sort of earthquake precursor from the data obtained by these microearthquake networks did not work out. For a review of the earthquake prediction efforts, please read Kanamori [60].

Some Recent Advances

Because of recent advances in electronics, communications, and microcomputers, it is now possible to deploy sophisticated digital seismograph stations at global, national, regional, and local scales for *real-time* seismology [64]. Many such networks, including temporary portable networks, have been implemented in many countries. In particular, various real-time and near real-time seismic systems began operation in the 1990s: for example, in Mexico [25], California [32,47], and Taiwan [110]. The Real-Time Data (RTD) system operated by the Central Weather Bureau (CWB) of Taiwan is based on a network of telemetered digital accelerographs [102]; since 1995, this system has used pagers, e-mail, and other techniques to automatically and rapidly disseminate information about the hypocenter, magnitude, and shaking amplitude of felt earthquakes ($M \gtrsim 4$) in the Taiwan region. The disastrous Chi-Chi earthquake ($M_W = 7.6$) of 20 September 1999 caused 2,471 deaths and total economic

losses of US\$ 11.5 billion. For this earthquake sequence, the RTD system delivered accurate information to government officials 102 seconds after the origin time of the main shock (about 50 seconds for most aftershocks), and proved to be useful in the emergency response of the Taiwan government [37,131].

Recording Damaging Ground Shaking

Observing teleseisms on a global scale with station spacing of several hundreds of kilometers does not yield critical information about near-source strong ground shaking required for earthquake structural engineering purposes and seismic hazard reduction. Broadband seismometers, which are optimized to record earthquakes at great distances, do not perform well in the near-field of a major earthquake. For example, during the 1999 Chi-Chi earthquake the nearest broadband station in Taiwan (epicentral distance of about 20 km) was badly overdriven, recorded no useful data beyond the arrival time of the initial *P*-wave, and finally failed about one minute into the shock.

A regional seismic network with station spacing of a few tens of kilometers cannot do the job either: the station spacing is still too large and the records are typically overdriven for earthquakes of $M \gtrsim 3$ (any large earthquake would certainly overdrive these sensitive instruments in the entire network). In his account of early earthquake engineering, Housner [51] credited John R. Freeman, an eminent engineer, with persuading the then US Secretary of Commerce to authorize a strong-motion program, and, in 1930, the design of an accelerograph for engineering purposes. In a letter to R.R. Martel, Housner's professor at Caltech, Freeman wrote:

I stated that the data which had been given to structural engineers on acceleration and limits of motion in earthquakes as a basis for their designs were all based on guesswork, that there had never yet been a precise measurement of acceleration made. That of the five seismographs around San Francisco Bay which tried to record the earthquake of 1906 not one was able to tell the truth.

Strong-motion recordings useful to engineers must be on-scale for damaging earthquakes and, in particular, from instruments located on or near built structures in densely urbanized environments, within about 25 km of the earthquake-rupture zone for sites on rock, or within about 100 km for sites on soft soils. Recordings of motions sufficient to cause damage at sites at greater distances are also of interest for earthquake engineering in areas likely to be affected by major subduction-zone earth-

quakes and in areas with exceptionally low attenuation rates [11]. In addition, densely-spaced networks of strong-motion recorders are needed to study the large variations in these motions over short distances [26,29].

Although several interesting accelerograms were recorded in southern California in the 1930s and 1940s, most seismologists did not pursue strong-motion monitoring until much later. The 1971 San Fernando earthquake emphatically demonstrated the need for more strong-motion data [9]. Two important programs emerged in the United States – the National Strong-Motion Program (<http://nsmp.wr.usgs.gov/>), and the California Strong Motion Instrumentation Program (<http://docinet3.consrv.ca.gov/csmip/>). However, the budgets for these programs were and continue to be small in comparison to other earthquake programs. High levels of funding for strong-motion monitoring, comparable to that of the GSN and the regional seismic networks, became available in Taiwan in the early 1990s, and in Japan in the mid-1990s. The Consortium of Organizations for Strong-Motion Observation Systems (<http://www.cosmos-eq.org/>) was established recently to promote the acquisition and application of strong-motion data.

Seismograms and Derived Products

Even before instruments were developed to record seismic waves from earthquakes, many scholars compiled catalogs of earthquake events noted in historical and other documents. Robert Mallet in 1852–1854 published the first extensive earthquake catalog of the world (1606 B.C.–A.D. 1842) totaling 6831 events [79]. Based on this compilation, Mallet prepared the first significant seismicity map of the Earth in 1858. Mallet's map is remarkable in that it correctly identifies the major earthquake zones of the Earth excepting for parts of the oceans. Although Mallet's earthquake catalog and similar compilations contain a wealth of information about earthquakes, they were made without the aid of instruments, and thus were subject to the biases of the observers as well as to population distributions. These non-instrumental earthquake catalogs also contain errors because the source materials were commonly incomplete and inconsistent regarding date, time, place names, and reported damage. Ambraseys et al. [8] discusses these difficulties for a regional case and Guidoboni [42] addresses the matter in general.

Today, seismograms are the fundamental data produced by earthquake monitoring. An analyst's first task is to find out when and where the earthquakes occurred, its size, and other characteristics. The accuracy of deter-

mining earthquake parameters, as well as the number of parameters used to characterize an earthquake, has progressed along with the availability of seismograms and computers, as well as advances in seismology. In the analog era, earthquake parameters were primarily the origin time, geographical location (epicenter), focal depth, and magnitude. A list of these parameters for earthquakes occurring over some time interval is called an *earthquake catalog*. A useful and common illustration of such results is a map showing the locations of earthquakes by magnitude (a seismicity map). Figure 11 is such a seismicity map for 1900–1999 as prepared by Engdahl and Villaseñor [24]. The map shows that moderate and large earthquakes are concentrated in tectonic active areas while most areas of the Earth are aseismic.

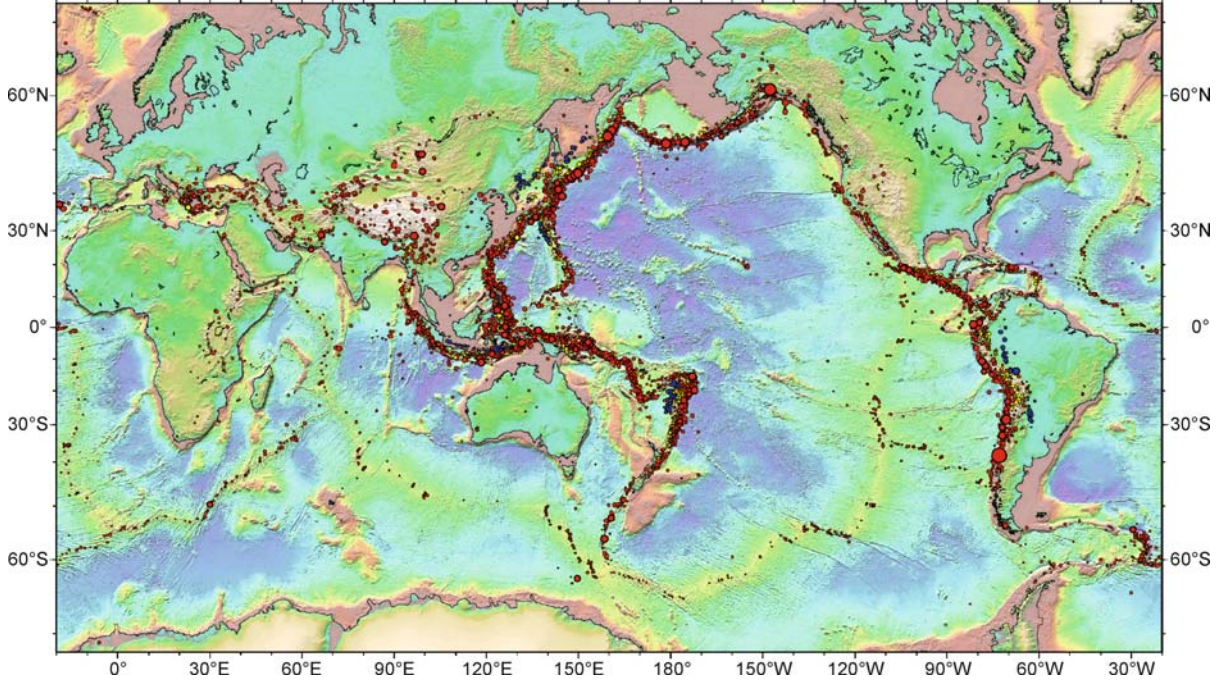
Earthquake Location

Several methods have been developed to locate earthquakes (i.e., determine origin time, latitude and longitude of the epicenter, and focal depth). Common to most of these methods is the use of arrival times of initial *P*- and *S*-waves. In particular, Geiger [33] applied the Gauss–Newton method to solve for earthquake location, which is a nonlinear problem, by formulating it as an inverse problem. However, since Geiger's method is computational intensive, it was not practical to apply it for the routine determinations of earthquake hypocenters until the advance of modern computers in the early 1960s.

Before computers became widely available starting in the 1960s, earthquakes were usually located by a manual, graphical method. In any location method, we assume that an earthquake is a point source and its sole parameters are origin time (time of occurrence, t_0) and hypocenter position (x_0, y_0, z_0). If both *P*- and *S*-arrival times are available, one may use the time intervals between *P*- and *S*-waves at each station (*S*-*P* times) and estimates of seismic wave velocities in the Earth to obtain a rough estimate of the epicentral distance, D , from that station:

$$D = [V_P V_S / (V_P - V_S)](T_S - T_P) \quad (2)$$

where V_P is the *P*-wave velocity, V_S the *S*-wave velocity, T_S the *S*-wave arrival time, and T_P the *P*-wave arrival time. For a typical crustal *P*-wave velocity of 6 km/s, and $V_P/V_S \approx 1.8$, the distance D in kilometers is about 7.5 times the *S*-*P* interval measured in seconds. If three or more epicentral distances are available, the epicenter may be placed at the intersection of circles with the stations as centers and the appropriate D as radii. The intersection



Earthquake Monitoring and Early Warning Systems, Figure 11
Seismicity of the Earth: 1900–1999 (see [24] for details)

will seldom be a point, and its areal extent gives a rough estimate of the uncertainty of the epicenter and hypocentral (focal) depth. In the early days, the focal depth was usually assumed or occasionally determined using a “depth phase” (generally, a ray that travels upward from the hypocenter and reflects back from the Earth’s surface, then arcs through the Earth to reach a distant seismograph).

Although Geiger [33] presented a method for determining the origin time and epicenter, the method can be extended easily to include focal depth. To locate an earthquake using a set of arrival times, τ_k , from stations at positions (x_k, y_k, z_k) , $k = 1, 2, \dots, m$, we must assume a model of seismic velocities from which theoretical travel times, T_k for a trial hypocenter at (x^*, y^*, z^*) to the stations can be computed. Let us consider a given trial origin time and hypocenter as the trial vector χ^* in a four-dimensional Euclidean space:

$$\chi^* = (t^*, x^*, y^*, z^*)^T \quad (3)$$

where the superscript T denotes the vector transpose. Theoretical arrival time, t_k , from χ^* to the k -th station is the theoretical travel time, T_k , plus the trial origin time, t^* . We now define the arrival time residual at the k -th station, r_k , as the difference between the observed and the theo-

retical arrival times. We may consider this set of station residuals as a vector in an m -dimensional Euclidean space and write:

$$\mathbf{r} = (r_1(\chi^*), r_2(\chi^*), \dots, r_m(\chi^*))^T. \quad (4)$$

We now apply the least squares method to obtain a set of linear equations solving for an adjustment vector, $\delta\chi$:

$$\mathbf{A}^T \mathbf{A} \delta\chi = -\mathbf{A}^T \mathbf{r}, \quad (5)$$

where A is the Jacobian matrix consisting of partial derivatives of travel time with respect to t , x , y , and z . A detailed derivation of the Geiger method is given by Lee and Stewart (see, pp 132–134 in [74]). There are many practical difficulties in implementing Geiger’s method for locating earthquakes, as discussed by Lee and Stewart (see, pp 134–139 in [74]). Although standard errors for these earthquake locations can be computed, they are often not meaningful because errors in the measurement of arrival times usually do not obey a Gaussian probability distribution. In recent years, many authors applied various nonlinear methods to locate earthquakes; a review of these methods is given by Lomax et al. ► [Earthquake Location, Direct, Global-Search Methods](#).

Earthquake Magnitude

After an earthquake is located, the next question is: how big was it? The Richter magnitude scale was originally devised to measure the “size” of an earthquake in southern California. Richter [96] defined the local (earthquake) magnitude, M_L , of an earthquake observed at any particular station to be:

$$M_L = \log A - \log A_0(\Delta) \quad (6)$$

where A is the maximum amplitude in millimeters as recorded by a Wood–Anderson seismograph for an earthquake at epicentral distance of Δ km. The correction factor, $\log A_0(\Delta)$, is the maximum amplitude at Δ km for a “standard” earthquake. Thus, three arbitrary choices enter into the definition of local magnitude: (1) the use of the Wood–Anderson seismographs, (2) the use of the common logarithm (i. e., the logarithm to the base 10), and (3) the selection of the standard earthquake, whose amplitudes as a function of distance Δ are represented by $A_0(\Delta)$.

In the 1940s, B. Gutenberg and C.F. Richter extended the local magnitude scale to include more distant earthquakes. Gutenberg [43] defined the surface-wave magnitude, M_S , as

$$M_S = \log(A/T) - \log A_0(\Delta^\circ) \quad (7)$$

where A is the maximum combined horizontal ground displacement in micrometers (μm) for surface waves with a period of 20 s, and $-\log A_0$ is tabulated as a function of epicentral distance Δ in degrees, in a similar manner to that for the local magnitude's $A_0(\Delta)$. Specifically, surface-wave magnitude is calculated from

$$M_S = \log A + 1.656 \log \Delta + 1.818 \quad (8)$$

using the prominent 20 s period surface waves commonly observed on the two horizontal-component seismograms from earthquakes of shallow focal depth.

Both magnitude scales were derived empirically and have scale-saturation problems, e. g., for very large earthquakes above a certain size the computed magnitudes of a particular type are all about the same. After the pioneering work of Charles F. Richter and Beno Gutenberg, numerous authors have developed alternative magnitude scales, as reviewed recently by Utsu [116] and by Bormann and Saul ▶ [Earthquake Magnitude](#). A current magnitude scale widely accepted as “best” (as having the least saturation problem and being a close match to an earthquake's total release of stress and strain) is the “moment magnitude”, M_W , computed from an earthquake's “moment tensor”.

Quantification of the Earthquake Source

As pointed out by Kanamori [59], it is not a simple matter to find a single measure of the “size” of an earthquake, simply because earthquakes result from complex physical processes. The elastic rebound theory of Harry F. Reid suggests that earthquakes originate from spontaneous slippage on active faults after a long period of elastic strain accumulation [94]. Faults may be considered the slip surfaces across which discontinuous displacement occurs in the Earth, while the faulting process may be modeled mathematically as a shear dislocation in an elastic medium (see [100], for a review). A shear dislocation (or slip) is equivalent to a double-couple body force [15,81]. The scaling parameter of each component couple of a double-couple body force is its *moment*. Using the equivalence between slip and body forces, Aki [2] introduced the *seismic moment*, M_0 , as:

$$M_0 = \mu \int D(A) dA = \mu s A \quad (9)$$

where μ is the shear modulus of the medium, A is the area of the slipped surface or source area, and s is the slip $D(A)$ averaged over the area A . If an earthquake produces surface faulting, we may estimate its rupture length, L , and its average slip, s , from measurement of that faulting. The area A may be approximated by Lh , where h is the focal depth (it is often, but not always, found that the hypocenter is near the bottom of the rupture surface). A reasonable estimate for μ is 3×10^{11} dynes/cm². With these quantities, we can estimate the seismic moment from Eq. (9).

Seismic moment also can be estimated independently from seismograms. From dislocation theory, the seismic moment can be related to the far-field seismic displacement recorded by seismographs. For example, Hanks and Wyss [46] showed that

$$M_0 = (\Omega_0/\psi_{\theta\phi}) 4\pi\rho R v^3 \quad (10)$$

where Ω_0 is the long-period limit of the displacement spectrum of either P or S waves, $\psi_{\theta\phi}$ is a function accounting for the body-wave radiation pattern, ρ is the density of the medium, R is a function accounting for the geometric spreading of body waves, and v is the body-wave velocity. Similarly, seismic moment can be determined from surface waves or coda waves [2,3].

In 1977, Hiroo Kanamori recognized that a new magnitude scale can be developed using seismic moment (M_0) by comparing the earthquake energy and seismic moment relation

$$E_S = (\Delta\sigma/2\mu) M_0, \quad (11)$$

where $\Delta\sigma$ is the stress drop and μ is the shear modulus, with the surface-wave magnitude and energy relation [45]

$$\log E_S = 1.5M_S + 11.8, \quad (12)$$

where E_S and M_0 are expressed in ergs and dyne-cm, respectively. The average value of $(\Delta\sigma/2\mu)$ is approximately equal to 1.0×10^{-4} . If we use this value in Eq. (11), we obtain

$$\log M_0 = 1.5M_S + 16.1. \quad (13)$$

It is known that M_S values saturate for great earthquakes (M_0 about 10^{29} dyne-cm or more) and, therefore, that Eqs. (12) and (13) do not hold for such great earthquakes. If a new moment-magnitude scale using the notation M_W is defined by

$$\log M_0 = 1.5M_W + 16.1 \quad (14)$$

then M_W is equivalent to M_S below saturation and provides a reasonable estimate for great earthquakes without the saturation problem [58]. The subscript letter W stands for the work at an earthquake fault, but soon M_W became known as the *moment magnitude*. Determining earthquake magnitude using seismic moment is clearly a better approach because it has a physical basis.

The concept of seismic moment led to the development of moment tensor solutions for quantifying the earthquake source, including its focal mechanism [35,36]; the seismic moment is just the scalar value of the moment tensor. Since the 1980s, Centroid-Moment-Tensor (CMT) solutions have been produced routinely for events with moment magnitudes (M_W) greater than about 5.5. The CMT methodology is described by Dziewonski et al. [22] and Dziewonski and Woodhouse [20]; a comprehensive review is given in Dziewonski and Woodhouse [21]. These CMT solutions are published yearly in the journal *Physics of the Earth and Planetary Interiors*, and the entire database is accessible online. This useful service is now performed by the Global CMT Project (<http://www.globalcmt.org/>), and more than 25,000 moment tensors have been determined for large earthquakes from 1976 to 2007. In the most recent decade, Quick CMT solutions [23] determined in near-real time have been added and are distributed widely via e-mail (<http://www.seismology.harvard.edu/projects/CMT/QuickCMTs/>).

Limitations of Earthquake Catalogs

In addition to international efforts to catalog earthquakes on a global scale, observatories and government agencies issue more-detailed earthquake catalogs at local, regional,

and national scales. However, earthquake catalogs from local to global scales vary greatly in spatial and temporal coverage and in quality, with respect to completeness and accuracy, because of the ongoing evolution of instrumentation, data processing procedures, and agency staff. An earthquake catalog, to be used for research, should have at least the following source parameters: origin time, epicenter (latitude and longitude), focal depth, and magnitude.

The International Seismological Summary and its predecessors provided compilations of arrival times and locations of earthquakes determined manually from about 1900 to 1963. Despite their limitations (notably the lack of magnitude estimates), these materials remain valuable. The first global earthquake catalog that contains both locations and magnitudes was published by Gutenberg and Richter in 1949, and was followed by a second edition in 1954 [44]. This catalog contains over 4,000 earthquakes from 1904 to 1951. Unfortunately, its temporal and spatial coverage is uneven as a result of rapid changes in seismic instrumentation, and of the interference of both World Wars. Nevertheless, the procedures used for earthquake location and magnitude estimation were the same throughout, using the arrival-time and amplitude data available to Gutenberg and Richter during the 1940s and early 1950s.

Since 1964, the International Seismological Centre has performed systematic cataloging of earthquakes worldwide by using computers and more modern seismograph networks. The spatial coverage of this catalog is not complete for some areas of the Earth (especially the oceans) because of the paucity of seismographic stations in such areas. By plotting the cumulative numbers of earthquakes above a certain magnitude versus magnitude, and using Eq. (1), the lower limit of completeness of an earthquake catalog may be estimated – it is the magnitude below which the data deviate below a linear fit to Eq. (1).

A *Centennial Earthquake Catalog* covering ISS- and ISC-reported global earthquakes from 1900–1999 was generated using an improved Earth model that takes into account regional variations in seismic wave velocities in the Earth's crust and upper mantle [24,118]. Engdahl and Villasenor [24] also compiled existing magnitude data from various authors and suggested preferred values. However, these “preferred magnitudes” were not determined by the same procedures. At present, the Global CMT Project (<http://www.globalcmt.org/>) provides the most complete online source parameters for global earthquakes (with $M_W > 5.5$), including Centroid-Moment-Tensor solutions. Although the CMT catalog starts in 1976, the improved global coverage of modern broadband digital seismographs began only in about 1990.

In summary, earthquake catalogs have been used extensively for earthquake prediction research and seismic hazard assessment since the first such catalog was produced. Reservations have been expressed about the reliability of the results and interpretations from these studies because the catalogs cover too little time and have limitations in completeness and accuracy (both random and systematic). Nevertheless, advances have been made in using earthquake catalogs to (1) study the nature of seismicity (e.g., ► [Seismicity, Critical States of: From Models to Practical Seismic Hazard Estimates Space](#)), (2) investigate earthquake statistics (e.g., ► [Earthquake Occurrence and Mechanisms, Stochastic Models for](#)), (3) forecast earthquakes (e.g., ► [Earthquake Forecasting and Verification](#)), (4) predict earthquakes (e.g., ► [Geo-complexity and Earthquake Prediction](#)), (5) assess seismic hazards and risk, and so forth.

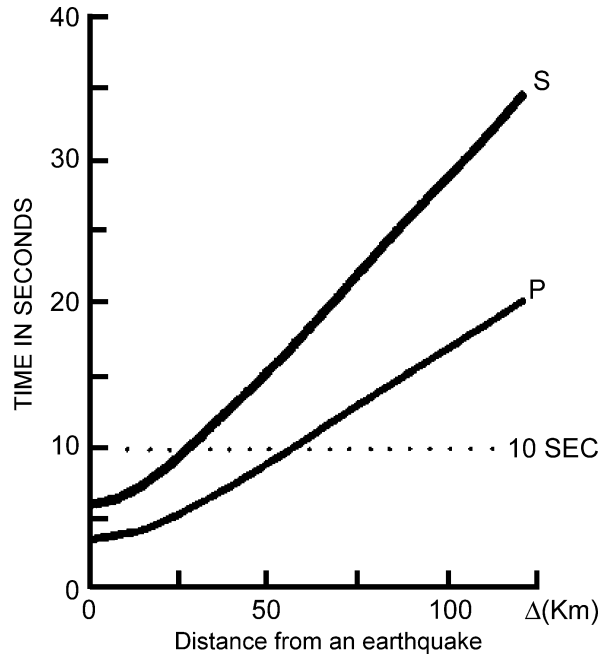
Earthquake Early Warning (EEW) Systems

With increasing urbanization worldwide, earthquake hazards pose ever greater threats to lives, property, and livelihoods in populated areas near major active faults on land or near offshore subduction zones. Earthquake early-warning systems can be useful tools for reducing the impact of earthquakes, provided that cities are favorably located with respect to earthquake sources and their citizens are properly trained to respond to the warning messages. Recent reviews of earthquake early warning systems may be found in Lee and Espinosa-Aranda [73], Kanamori [61], and Allen [6], as well as a monograph on the subject by Gasparini et al. [31].

Under favorable conditions, an EEW system can forewarn an urban area of impending strong shaking with lead times that range from a few seconds to a few tens of seconds. A lead time is the time interval between issuing a warning and the arrival of the *S*-waves, which are the most destructive seismic waves. Even a few seconds of advanced warning is useful for pre-programmed emergency measures at various critical facilities, such as the deceleration of rapid-transit vehicles and high-speed trains, the orderly shutoff of gas pipelines, the controlled shutdown of some high-technological manufacturing operations, the safe-guarding of computer facilities (e.g., disk-head parking), and bringing elevators to a stop at the nearest floor.

Physical Basis and Limitations of EEW Systems

The physical basis for earthquake early warning is simple: damaging strong ground shaking is caused primarily by shear (*S*) and subsequent surface waves, both of which travel more slowly than the primary (*P*) waves, and



Earthquake Monitoring and Early Warning Systems, Figure 12
Travel time of *P*-waves and of *S*-waves versus distance for a typical earthquake

seismic waves travel much more slowly than electromagnetic signals transmitted by telephone or radio. However, certain physical limitations must be considered, as shown by Fig. 12.

Figure 12 is a plot of the travel time for the *P*-wave and *S*-wave as a function of distance from an earthquake. We make the following assumptions about a typical destructive earthquake: (1) focal depth at ~ 20 km, (2) *P*-wave velocity ~ 8 km/s, and (3) *S*-wave velocity ~ 4.5 km/s. If an earthquake is located 100 km from a city, the *P*-wave arrives at the city after about 13 s, and the *S*-waves in about 22 s (Fig. 12). If we deploy a dense seismic network near the earthquake source area (capable of locating and determining the size of the event in about 10 s), we will have about 3 s to issue the warning before the *P*-wave arrives, and about 12 s before the more destructive *S*-waves and surface waves arrive at the city. We have assumed that it takes negligible time to send a signal from the seismic network to the city via electromagnetic waves, which travel at about one-third the velocity of light or faster (between about 100,000 and 300,000 km/s depending on the method of transmission).

From Fig. 12 it is clear that this strategy may work for earthquakes located at least about 60 km from the urban area. For earthquakes at shorter distances (~ 20 to

~ 60 km), we must reduce the time needed to detect the event and issue a warning to well under 10 s. This requirement implies that we must deploy a very dense seismic network very close to the fault and estimate the necessary parameters very fast. However, such dense networks are not economical to deploy using existing seismic instruments.

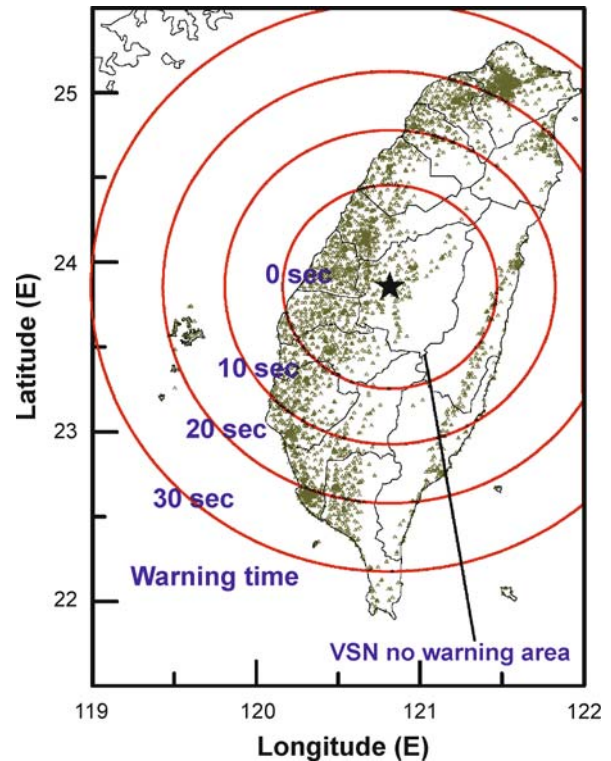
For earthquakes within 20 km of a city, there is little one can do other than installing motion-sensitive automatic shut-off devices at critical facilities (natural gas, for example) and hope that they are either very quick when responding to *S*-waves or are triggered by the onset of the *P*-wave. Normally an earthquake rupture more than ~ 100 km from an urban area does not commonly pose a large threat (seismic waves would be attenuated and spread out farther). There are exceptions caused either by unusual local site conditions, such as Mexico City, or by earthquakes with large rupture zones which therefore radiate efficiently to greater distances.

Design Considerations for EEW Systems

In the above discussion, we have assumed that one implements an earthquake early warning system with a traditional seismic network. Such EEW systems have limitation as illustrated by Fig. 13, which shows the expected early warning times for a repeat of the 1999 Chi–Chi earthquake. However, Nakamura and his colleagues have been successful in applying a single-station approach [84,99], where seismic signals are recorded and processed locally by the seismograph and an earthquake warning is issued whenever ground motions there exceed some trigger threshold. We will next discuss these two basic approaches, regional versus on-site in designing an earthquake early warning system.

Earthquake early warning capability can be implemented through a rapid reporting system (RRS) from a traditional network, assuming real-time telemetry into the network's central laboratory. This type of system provides, to populated areas and other sensitive locations, primary event information (hypocenter, magnitude, ground shaking intensities, and potential damage) about one minute after the earthquake begins. The RRS transmits this critical information electronically to emergency response agencies and other interested organizations and to individuals. Each recipient can then take action (some of which may be pre-programmed) shortly after the earthquake begins. Response measures can include the timely dispatch of rescue equipment and emergency supplies to the likely areas of damage.

California's ShakeMap [119,120], Taiwan's CWB, and Japan's JMA systems are typical examples of RSS. In



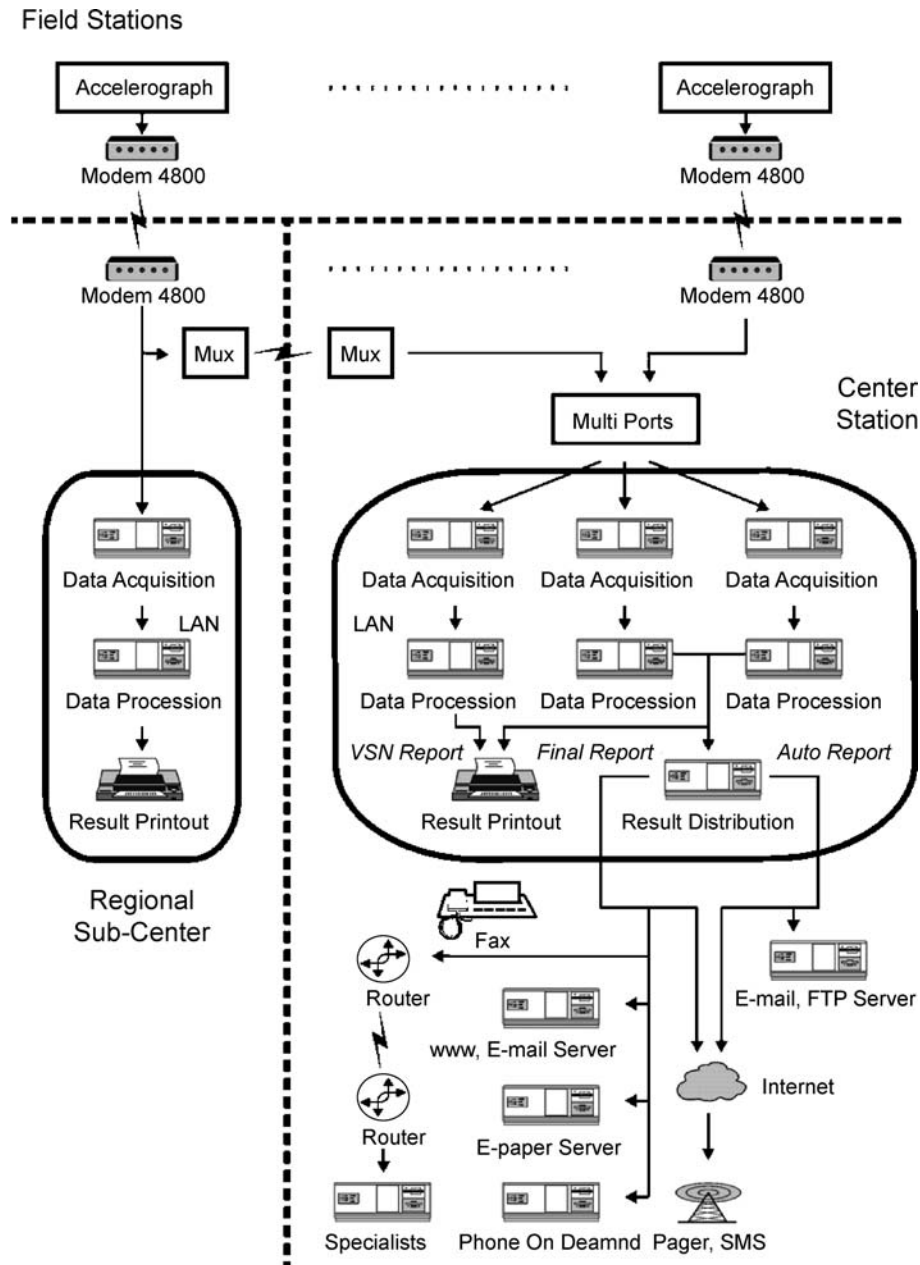
Earthquake Monitoring and Early Warning Systems, Figure 13 Expected EWS early warning times (indicated by circles) in Taiwan with respect to the occurrence of an event similar to the Chi–Chi earthquake of 20 September 1999. Triangles are locations of elementary schools, which can be regarded as a good indicator for the population density of Taiwan

the case of the Taiwan RRS, the CWB has, since 1995, provided intensity maps, hypocenters, and magnitudes within one minute of the occurrence of $M > 4$ earthquakes [110,128]. This system's reliability, documented by electronic messages to government agencies and scientists, has been close to perfect, particularly for large, damaging earthquakes. Figure 14 shows a block diagram of the Taiwan RRS, and details may be found in [128].

Using a set of empirical relationships derived from the large data set collected during the 1999 Chi–Chi earthquake, CWB now releases, within a few minutes of an event, the estimated distributions of PGA and PGV, refined magnitudes, and damage estimates [129]. This near-real-time damage assessment is useful for rapid post-disaster emergency response and rescue missions.

Regional Warning Versus Onsite Warning

Two approaches have been adopted for earthquake early warning systems: (1) regional warning, and (2) on-site



Earthquake Monitoring and Early Warning Systems, Figure 14

A block diagram showing the hardware of the Taiwan Earthquake Rapid Reporting System

warning. The first approach relies on traditional seismological methods in which data from a seismic network are used to locate an earthquake, determine the magnitude, and estimate the ground motion in the region involved. In the second approach, the initial ground motions (mainly *P* wave) observed at a site are used to predict the ensuing ground motions (mainly *S* and surface waves) at the same site.

The regional approach is more comprehensive, but takes a longer time to issue an earthquake warning. An advantage of this approach is that estimates of the timing of expected strong motions throughout the affected region can be predicted more reliably. The early warning system in Taiwan is a typical example and it uses a regional warning system called virtual sub-network approach (VSN) that requires an average of 22 s to determine earthquake

parameters with magnitude uncertainties of ± 0.25 . It provides a warning for areas beyond about 70 km from the epicenter (Fig. 13). This system has been in operation since 2002 with almost no false alarms [129]. With the advancement of new methodology and more dense seismic networks, regional systems are beginning to be able to provide early warnings to areas closer to the earthquake epicenter.

The regional approach has also been used in other areas. The method used in Mexico [25] is slightly different from the traditional seismological method. It is a special case of EEW system due to the relatively large distance (about 300 km in this case) between the earthquake source region (west coast of Central America) and the warning site (Mexico City). However, the warning is conceptually “regional”.

In Japan, various EEW techniques have been developed and deployed by the National Research Institute for Earth Science and Disaster Prevention (NIED) and Japan Meteorological Agency (JMA) since 2000 [49,57,89], ▶ **Tsunami Forecasting and Warning**. In particular, JMA has started sending early warning messages to potential users responsible for emergency responses [50]. The potential users include railway systems, construction companies, and others; and they are familiar with the implications of early warning messages, as well as the technical limitations of EEW [57].

Some Recent EEW Advances

Allen and Kanamori [7] proposed the Earthquake Alarm System (ElarmS) to issue an earthquake warning based on information determined from the *P*-wave arrival only. Kanamori [61] extended the method of Nakamura [84] and Allen and Kanamori [7] to determine a period parameter, τ_c , from the initial 3 s of the *P* wave. τ_c is defined as

$$\tau_c = 2\pi / \sqrt{r} \quad (15)$$

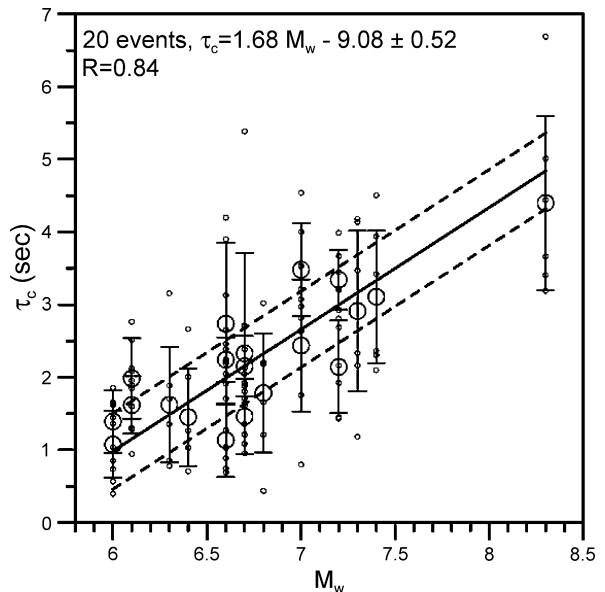
where

$$r = \frac{\int_0^{\tau_0} \dot{u}^2(t) dt}{\int_0^{\tau_0} u^2(t) dt} \quad (16)$$

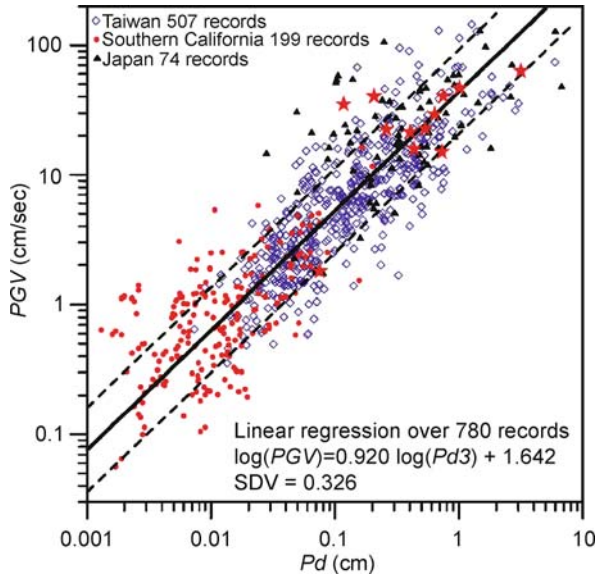
$u(t)$ is the ground-motion displacement; τ_0 is the duration of record used (usually 3 s), and τ_c , which represents the size of an earthquake, can be computed from the incoming data sequentially.

The τ_c method was used for earthquake early warning in southern California, Taiwan, and Japan by Wu and Kanamori [124,125,126] and Wu et al. [130]. At a given site, the magnitude of an event is estimated from τ_c and the peak ground-motion velocity (PGV) from P_d (the peak amplitude of displacement in the first 3 s after the arrival

of the *P* wave). The incoming 3-component signals are recursively converted to ground acceleration, velocity and displacement. The displacements are recursively filtered using an accusal Butterworth high-pass filter with a cut-off frequency of 0.075 Hz, and a *P*-wave threshold trigger is constantly monitored. When a trigger occurs, τ_c and P_d are computed. The relationships between τ_c and magnitude (M), and P_d and peak ground velocity (PGV) for southern California, Taiwan, and Japan were investigated. Figure 15 shows a good correlation between τ_c and M_w from the K-NET records in Japan, and Fig. 16 shows the P_d versus PGV plot for southern California, Taiwan, and Japan. These relationships may be used to detect the occurrence of a large earthquake and provide onsite warning in the area immediately around the station where the onset of strong ground motion is expected within a few seconds after the arrival of the *P*-wave. When the station density is high, the onsite warning methods may be applied to data from multiple stations to increase the robustness of an onsite early warning, and to complement the regional warning approach. In an ideal situation, such warnings would be available within 10 s of the origin time of a large earthquake whose subsequent ground motion may last for tens of seconds.



Earthquake Monitoring and Early Warning Systems, Figure 15
 τ_c estimates from 20 events using the nearest 6 stations of the K-NET. Small open circles show single-record results, and large circles show event-average values with one standard deviation bars. Solid line shows the least squares fit to the event-average values, and the two dashed lines show the range of one standard deviation



Earthquake Monitoring and Early Warning Systems, Figure 16
Relationship between peak initial displacement amplitude (P_d) measurements and peak ground velocity (PGV) for the records with epicentral distances less than 30 km from the epicenter in Southern California (red solid circles), Taiwan (blue diamonds) and Japan (black solid triangles). Solid line shows the least squares fit and the two dashed lines show the range of one standard deviation

Wu and Zhao [127] investigated the attenuation of P_d with the hypocentral distance R in southern California as a function of magnitude M , and obtained the following relationships:

$$M_{P_d} = 4.748 + 1.371 \times \log(P_d) + 1.883 \times \log(R) \quad (17)$$

and

$$\log(P_d) = -3.463 + 0.729 \times M - 1.374 \times \log(R). \quad (18)$$

For the regional warning approach, when an earthquake location is determined by the P -wave arrival times at stations close to the epicenter, this relationship can be used to estimate the earthquake magnitude. Their result shows that for earthquakes in southern California the P_d magnitudes agree with the catalog magnitudes with a standard deviation of 0.18 for events less than magnitude 6.5. They concluded that P_d is a robust measurement for estimating the magnitudes of earthquakes for regional early warning purposes in southern California. This method has also applied to Italian region by Zollo et al. [132] with a very good performance.

Because the on-site approach is faster than the regional approach, it can provide useful early warning to sites at

short distances from the earthquake epicenter where early warning is most needed. Onsite early warning can be generated by either a single station or by a dense array. For a single station operation, signals from P -waves are used for magnitude and hypocenter determination to predict strong ground shaking. Nakamura [83] first proposed this concept, developed the Urgent Earthquake Detection and Alarm System or UrEDAS [86], and introduced a simple strong-motion index for onsite EEW [85]. However, the reliability of on-site earthquake information is generally less than that obtained with the regional warning system. There currently is a trade-off between warning time and the reliability of the earthquake information. Generally, an information updating procedure is necessary for any EEW system. On-site warning methods can be especially useful in regions where a dense seismic network is deployed.

The Japan Meteorological Agency (JMA) began distribution of earthquake early warning information to the public in October 1, 2007 through several means, such as TV and radio [50] (<http://www.jma.go.jp/jma/en/Activities/eww.html>). The JMA system was successfully activated during the recent Noto Hanto and Niigata Chuetsu–Oki earthquakes in 2007, and provided accurate information of hypocenter, magnitude, and intensity about 3.8 s after the arrival of P -waves at nearby stations. The warning message reached sites further than about 30 km from the epicenter as an early warning alert (i. e., information arrived before shaking started at the site). This is a remarkable performance of the system for damaging earthquakes and gives promise of an early warning system as a practical means for earthquake damage mitigation. Although warning alert is most needed within 30 km of the epicenter, it is not feasible with the current density and configuration of the JMA network.

Lawrence and Cochran [68] proposed a collaborative project for rapid earthquake response and early warning by using the accelerometers that are already installed inside many laptop computers. Their Quake-Catcher Network (QCN) will employ existing laptops, which have accelerometers already installed, and desktops outfitted with inexpensive (under \$ 50) USB accelerometers to form the world's largest high-density, distributed computing seismic network for monitoring strong ground motions (<http://qcn.stanford.edu/>). By freely distributing the necessary software, anyone having a computer with an Internet connection can join the project as a collaborative member. The Quake-Catcher Network also has the potential to provide better understanding of earthquakes, and the client-based software is also intended to be educational, with instructive material displaying the current seismic signal and/or recent earthquakes in the region. It

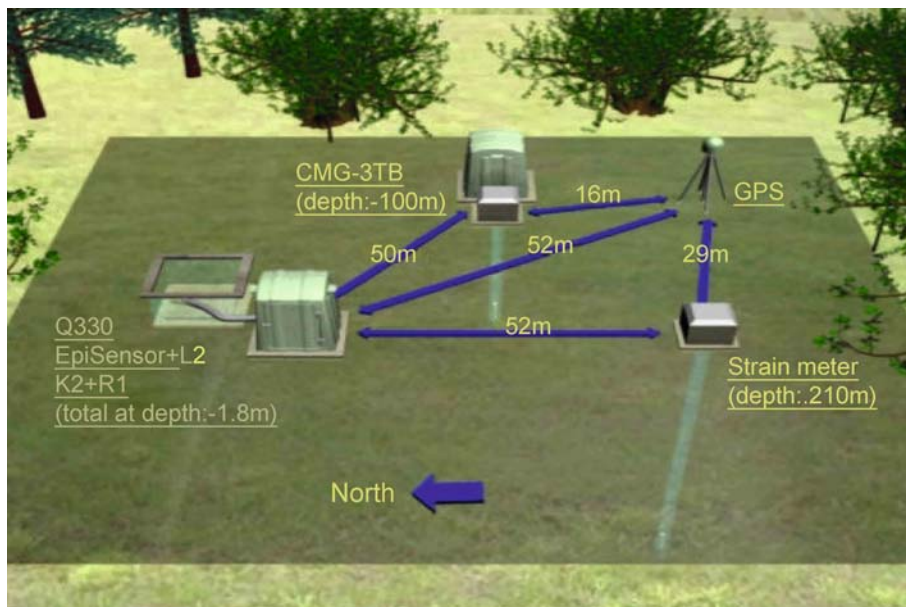
is an effective way to bring earthquake awareness to students and the general public.

Future Directions

To be successful, monitoring earthquakes requires large, stable funding over a long period of time. The most direct argument for governments to support long-term earthquake monitoring is to collect scientific data for hazard mitigation. In the past two decades about half a million of human lives have been lost due to earthquakes, and economic losses from earthquake damage total about \$200 billion. Future losses will be even greater as rapid urbanization is taking place worldwide. For example, the recent Japanese Fundamental Seismic Survey and Observation Plan (costing several hundred million US dollars) is a direct response to the economic losses of about \$100 billion due to the 1995 Kobe earthquake. In addition to scientific and technological challenges in monitoring earthquakes, seismologists must pay attention to achieve (1) stable long-term funding, (2) effective management and execution, and (3) delivery of useful products to the users.

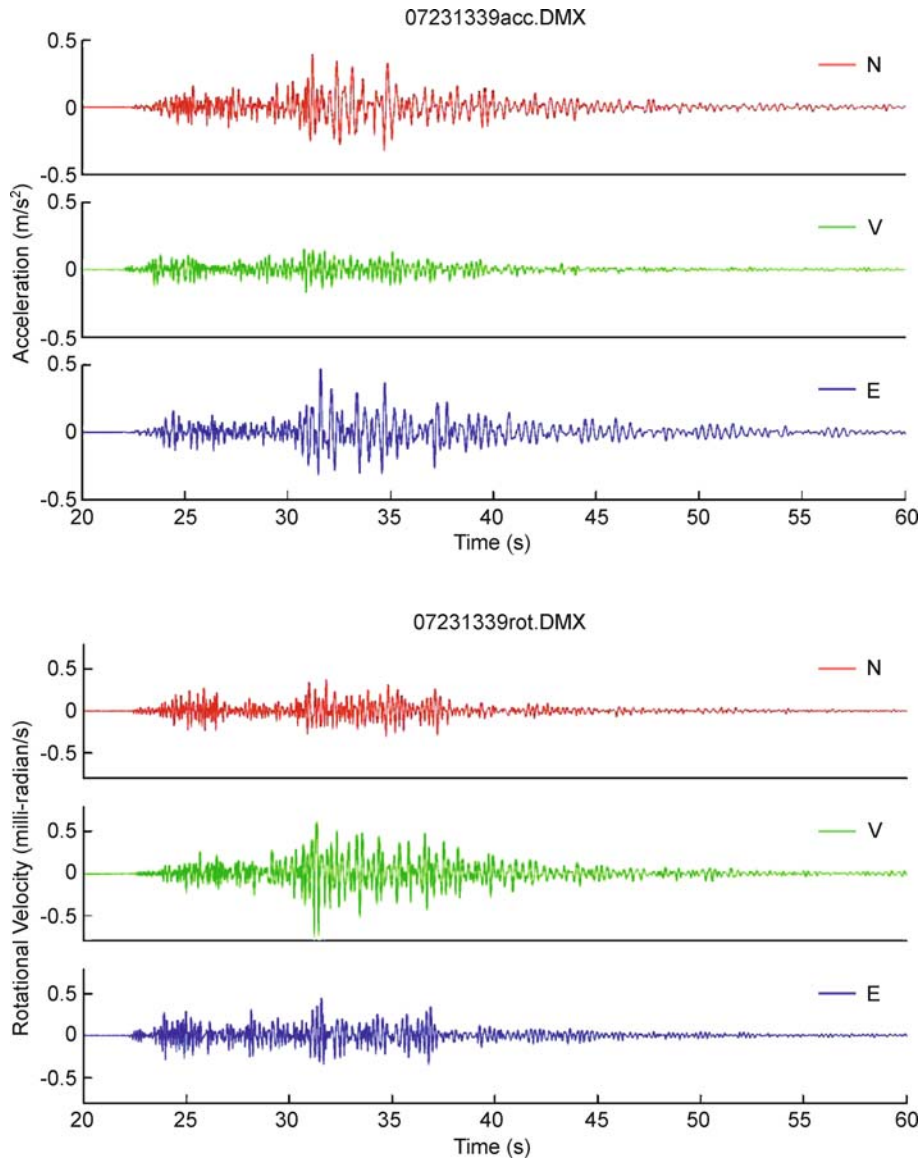
Seismologists benefit greatly from scientific and technological advances in other fields. For example, Global Positioning Systems (GPS) open a new window for mon-

itoring crustal deformation which is important to understand the driving forces that generate earthquakes (► [GPS: Applications in Crustal Deformation Monitoring](#), ► [Crustal Deformation During the Seismic Cycle, Interpreting Geodetic Observations of](#)). Under the US Earth Scope Program (<http://www.earthscope.org/>) the Plate Boundary Observatory (PBO) is covering the western Northern America and Alaska with a network of high precision GPS and strain-meter stations in order to measure deformation across the active boundary between the Pacific and North America plates (<http://www.earthscope.org/observatories/pbo>). As the sampling rate of GPS data increases, they can provide time histories of displacement during an earthquake. Monitoring earthquakes with multiple types of instruments and sensors is now increasingly popular, and “integrated” or “super” stations are increasingly common. Figure 17 shows an example of an integrated station (HGSD) in eastern Taiwan. Instruments deployed at the HGSD station in eastern Taiwan include: a broadband seismometer, a continuous GPS instrument, a strain-meter, and a 6-channel accelerograph (Model K2 by Kinemetrics) with an internal accelerometer and a rotational sensor (Model R-1 by eentec). A digital seismogram recorded at the HGSD station from an earthquake ($M_W = 5.1$) of July 23, 2007 at a distance of 34 km is



Earthquake Monitoring and Early Warning Systems, Figure 17

Instruments deployed at the HGSD station in eastern Taiwan. Clockwise from the top: (1) A broadband seismometer (Model CMG-3TB) installed at a depth of 100 m, (2) A continuous GPS instrument, (3) A strain-meter installed at a depth of 210 m, (4) A Model Q330 6-channel recorder with an accelerometer (Model EpiSensor) and a short-period seismometer (Model L2), and (5) A Model K2 6-channel accelerograph with an internal accelerometer and a rotational sensor (Model R-1)



Earthquake Monitoring and Early Warning Systems, Figure 18

A digital seismogram recorded at the HGSD station from an earthquake ($M_W = 5.1$) of July 23, 2007 at a distance of 34 km. *Top frame:* 3-component translational accelerations. *Bottom frame:* 3-component rotation velocity motions. N = North-South; V = Vertical, and E = East-West

shown in Fig. 18. The importance of rotational seismology and its current status are given in the Appendix.

A radically different design of seismographic networks (and earthquake early warning system in particular) is now possible using the “Sensor Network” developed by Intel Research. Intel is working with the academic community and industry collaborators to actively explore the potential of wireless sensor networks. This research is already demonstrating the potential for this new technology to

enhance public safety, reduce the cost of doing business, and bring a host of other benefits to business and society (http://www.intel.com/research/exploratory/wireless_sensors.htm).

It has been very difficult historically to obtain adequate and stable funding for long-term earthquake monitoring, largely because disastrous earthquakes occur infrequently. Since there are many pressing problems facing modern societies, almost all governments react to earthquake (and

tsunami) disasters only after the fact, and even then for relatively short periods of time. To advance earthquake prediction research and to develop effective earthquake warning systems will require continuous earthquake monitoring with extensive instrumentations in the near-field for decades and even centuries. Therefore, innovative approaches must be developed and perseverance is needed.

Acknowledgments

We thank John Evans, Fred Klein, Woody Savage, and Chris Stephens for reviewing the manuscript, their comments and suggestions greatly improved it. We are grateful to Lind Gee and Bob Hutt for information about the Global Seismographic Network (GSN) and for providing a high-resolution graphic file of an up-to-date GSN station map.

Appendix: A Progress Report on Rotational Seismology

Seismology is based primarily on the observation and modeling of three orthogonal components of translational ground motions. Although effects of rotational motions due to earthquakes have long been observed (e.g., [80]), Richter (see, p. 213 in [97]) stated that:

Perfectly general motion would also involve rotations about three perpendicular axes, and three more instruments for these. Theory indicates, and observation confirms, that such rotations are negligible.

However, Richter provided no references for this claim, and the available instruments at that time did not have the sensitivity to measure the very small rotation motions that the classical elasticity theory predicts.

Some theoretical seismologists (e.g., [4,5]) and earthquake engineers have argued for decades that the rotational part of ground motions should also be recorded. It is well known that standard seismometers and accelerometers are profoundly sensitive to rotations, particularly tilt, and therefore subject to rotation-induced errors (see e.g., [39,40,41,93]). The paucity of instrumental observations of rotational ground motions is mainly the result of the fact that, until recently, the rotational sensors did not have sufficient resolution to measure small rotational motions due to earthquakes.

Measurement of rotational motions has implications for: (1) recovering the complete ground-displacement history from seismometer recordings; (2) further constraining earthquake rupture properties; (3) extracting information about subsurface properties; and (4) providing addi-

tional ground motion information to engineers for seismic design.

In this Appendix, we will first briefly review elastic wave propagation that is based on the linear elasticity theory of simple homogeneous materials under infinitesimal strain. This theory was developed mostly in the early nineteenth century: the differential equations of the linear elastic theory were first derived by Louis Navier in 1821, and Augustin Cauchy gave his formulation in 1822 that remains virtually unchanged to the present day [103]. From this theory, Simeon Poisson demonstrated in 1828 the existence of longitudinal and transverse elastic waves, and in 1885, Lord Rayleigh confirmed the existence of elastic surface waves. George Green put this theory on a physical basis by introducing the concept of strain energy, and, in 1837, derived the basic equations of elasticity from the principle of energy conservation. In 1897, Richard Oldham first identified these three types of waves in seismograms, and linear elasticity theory has been embedded in seismology ever since.

In the following we summarize recent progress in rotational seismology and the need to include measurements of rotational ground motions in earthquake monitoring. The monograph by Teisseyre et al. [109] provides a useful summary of rotational seismology.

Elastic Wave Propagation

The equations of motion for a homogeneous, isotropic, and initially unstressed elastic body may be obtained using the conservation principles of continuum mechanics (e.g., [30]) as

$$\rho \frac{\partial^2 u_i}{\partial t^2} = (\lambda + \mu) \frac{\partial \theta}{\partial x_i} + \mu \nabla^2 u_i, \quad i = 1, 2, 3 \quad (\text{A1})$$

and

$$\theta = \sum_j \partial u_j / \partial x_j \quad (\text{A2})$$

where θ is the dilatation, ρ is the density, u_i is the i th component of the displacement vector \vec{u} , t is the time, and λ and μ are the elastic constants of the media. Eq. (A1) may be rewritten in vector form as

$$\rho (\partial^2 \vec{u} / \partial t^2) = (\lambda + \mu) \nabla (\nabla \cdot \vec{u}) + \mu \nabla^2 \vec{u}. \quad (\text{A3})$$

If we differentiate both sides of Eq. (A1) with respect to x_i , sum over the three components, and bring ρ to the right-hand side, we obtain

$$\partial^2 \theta / \partial t^2 = [(\lambda + 2\mu) / \rho] \nabla^2 \theta. \quad (\text{A4})$$

If we apply the curl operator ($\nabla \times$) to both sides of Eq. (A3), and note that

$$\nabla \bullet (\nabla \times \vec{u}) = 0 \quad (\text{A5})$$

we obtain

$$\partial^2 (\nabla \times \vec{u}) / \partial t^2 = (\mu/\rho) \nabla^2 (\nabla \times \vec{u}). \quad (\text{A6})$$

Now Eqs. (A4) and (A6) are in the form of the classical wave equation

$$\partial^2 \Psi / \partial t^2 = v^2 \nabla^2 \Psi, \quad (\text{A7})$$

where Ψ is the wave potential, and v is the wave-propagation velocity (a pseudovector; wave slowness is a proper vector). Thus a dilatational disturbance θ (or a compressional wave) may be transmitted through a homogenous elastic body with a velocity V_p where

$$V_p = \sqrt{[(\lambda + 2\mu)/\rho]} \quad (\text{A8})$$

according to Eq. (A4), and a rotational disturbance $\nabla \times \vec{u}$ (or a shear wave) may be transmitted with a wave velocity V_s where

$$V_s = \sqrt{\mu/\rho} \quad (\text{A9})$$

according to Eq. (A6). In seismology, and for historical reasons, these two types of waves are called the primary (P) and the secondary (S) waves, respectively.

For a heterogeneous, isotropic, and elastic medium, the equation of motion is more complex than Eq. (A3), and is given by Karal and Keller [65] as

$$\begin{aligned} \rho(\partial^2 \vec{u} / \partial t^2) = & (\lambda + \mu) \nabla (\nabla \bullet \vec{u}) + \mu \nabla^2 \vec{u} \\ & + \nabla \lambda (\nabla \bullet \vec{u}) + \nabla \mu \times (\nabla \times \vec{u}) + 2(\nabla \mu \bullet \nabla) \vec{u}. \end{aligned} \quad (\text{A10})$$

Furthermore, the compressional wave motion is no longer purely longitudinal, and the shear wave motion is no longer purely transverse. A review of seismic wave propagation and imaging in complex media may be found in the entry by Igel et al. [► Seismic Wave Propagation in Media with Complex Geometries, Simulation of.](#)

A significant portion of seismological research is based on the solution of the elastic wave equations with the appropriate initial and boundary conditions. However, explicit and unique solutions are rare, except for a few simple problems. One approach is to transform the wave equation to the eikonal equation and seek solutions in terms of wave fronts and rays that are valid at high frequencies. Another approach is to develop through specific boundary conditions a solution in terms of normal modes [77].

Although ray theory is only an approximation [17], the classic work of Jeffreys and Bullen, and Gutenberg used it to determine Earth structure and locate earthquakes that occurred in the first half of the 20th century. It remains a principal tool used by seismologists even today. Impressive developments in normal mode and surface wave studies (in both theory and observation) started in the second half of the 20th century, leading to realistic quantification of earthquakes using moment tensor methodology [21].

Rotational Ground Motions

Rotations in ground motion and in structural responses have been deduced indirectly from accelerometer arrays, but such estimates are valid only for long wavelengths compared to the distances between sensors (e.g., [16,34,52,88,90,104]). The rotational components of ground motion have also been estimated theoretically using kinematic source models and linear elastodynamic theory of wave propagation in elastic solids [14,69,70,111].

In the past decade, rotational motions from teleseismic and small local earthquakes were also successfully recorded by sensitive rotational sensors, in Japan, Poland, Germany, New Zealand, and Taiwan (e.g., [53,55,56,105,106,107,108]). The observations in Japan and Taiwan show that the amplitudes of rotations can be *one to two orders of magnitude greater than expected* from the classical linear theory. Theoretical work has also suggested that, in granular materials or cracked continua, asymmetries of the stress and strain fields can create rotations in addition to those predicted by the classical elastodynamic theory for a perfect continuum ([► Earthquake Source: Asymmetry and Rotation Effects](#)).

Because of lack of instrumentation, rotational motions have not yet been recorded in the near-field (within ~ 25 km of fault ruptures) of strong earthquakes (magnitude > 6.5), where the discrepancy between observations and theoretical predictions may be the largest. Recording such ground motions will require extensive seismic instrumentation along some well-chosen active faults and luck. To this end, several seismologists have been advocating such measurements, and a current deployment in southwestern Taiwan by its Central Weather Bureau is designed to “capture” a repeat of the 1906 Meishan earthquake (magnitude 7.1) with both translational and rotational instruments.

Rotations in structural response, and the contributions to the response from the rotational components of the ground motion, have also been of interest for many decades (e.g., [78,87,98]). Recent reviews on rotational motions in seismology and on the effects of the rota-

tional components of ground motion on structures can be found, for examples, in Cochard et al. [18] and Pillet and Virieux [93], and Trifunac [112], respectively.

Growing Interest – The IWGoRS

Various factors have led to spontaneous organization within the scientific and engineering communities interested in rotational motions. Such factors include: the growing number of successful direct measurements of rotational ground motions (e.g., by ring laser gyros, fiber optic gyros, and sensors based on electro-chemical technology); increasing awareness about the usefulness of the information they provide (e.g., in constraining the earthquake rupture properties, extracting information about subsurface properties, and about deformation of structures during seismic and other excitation); and a greater appreciation for the limitations on information that can be extracted from the translational sensors due to their sensitivity to rotational motions e.g., computation of permanent displacements from accelerograms (e.g., [13,39,40,41,93,113]).

A small workshop on Rotational Seismology was organized by W.H.K. Lee, K. Hudnut, and J.R. Evans of the USGS on 16 February 2006 in response to grassroots interest. It was held at the USGS offices in Menlo Park and in Pasadena, California, with about 30 participants from about a dozen institutions participating via teleconferencing and telephone [27]. This event led to the formation of the *International Working Group on Rotational Seismology* in 2006, inaugurated at a luncheon during the AGU 2006 Fall Meeting in San Francisco.

The *International Working Group on Rotational Seismology* (IWGoRS) aims to promote investigations of rotational motions and their implications, and the sharing of experience, data, software and results in an open web-based environment (<http://www.rotational-seismology.org>). It consists of volunteers and has no official status. H. Igel and W.H.K. Lee currently serve as “co-organizers”. Its charter is accessible on the IWGoRS web site. The Working Group has a number of active members leading task groups that focus on the organization of workshops and scientific projects, including: testing and verifying rotational sensors, broadband observations with ring laser systems, and developing a field laboratory for rotational motions. The IWGoRS web site also contains the presentations and posters from related meetings, and eventually will provide access to rotational data from many sources.

The IWGoRS organized a special session on *Rotational Motions in Seismology*, convened by H. Igel, W.H.K. Lee, and M. Todorovska during the 2006 AGU Fall Meet-

ing [76]. The goal of that session was to discuss rotational sensors, observations, modeling, theoretical aspects, and potential applications of rotational ground motions. A total of 21 papers were submitted for this session, and over 100 individuals attended the oral session.

The large attendance at this session reflected common interests in rotational motions from a wide range of geophysical disciplines, including strong-motion seismology, exploration geophysics, broadband seismology, earthquake engineering, earthquake physics, seismic instrumentation, seismic hazards, geodesy, and astrophysics, thus confirming the timeliness of IWGoRS. It became apparent that to establish an effective international collaboration within the IWGoRS, a larger workshop was needed to allow sufficient time to discuss the many issues of interest, and to draft research plans for rotational seismology and engineering applications.

First International Workshop

The *First International Workshop on Rotational Seismology and Engineering Applications* was held in Menlo Park, California, on 18–19 September 2007. This workshop was hosted by the US Geological Survey (USGS), which recognized this topic as a new research frontier for enabling a better understanding of the earthquake process and for the reduction of seismic hazards. The technical program consisted of three presentation sessions: plenary (4 papers) and oral (6 papers) held during the first day, and poster (30 papers) held during the morning of the second day. A post-workshop session was held on the morning of September 20, in which scientists of the Laser Interferometer Gravitational-wave Observatory (LIGO) presented their work on seismic isolation of their ultra-high precision facility, which requires very accurate recording of translational and rotational components of ground motions (3 papers). Proceedings of this Workshop were released in Lee et al. [75] with a DVD disc that contains all the presentation files and supplementary information.

One afternoon of the workshop was devoted to in-depth discussions on the key outstanding issues and future directions. The participants could join one of five panels on the following topics: (1) theoretical studies of rotational motions (chaired by L. Knopoff), (2) measuring far-field rotational motions (chaired by H. Igel), (3) measuring near-field rotational motions (chaired by T.L. Teng), (4) engineering applications of rotational motions (chaired by M.D. Trifunac), and (5) instrument design and testing (chaired by J.R. Evans). The panel reports on key issues and unsolved problems, and on research strategies and plans, can be found in Appendices

2.1 through 2.5 in Lee et al. [75]. Following the in-depth group discussions, the panel chairs reported on the group discussions in a common session, with further discussions among all the participants.

Discussions

Since rotational ground motions may play a significant role in the near-field of earthquakes, rotational seismology has emerged as a new frontier of research. During the Workshop discussions, L. Knopoff asked: Is there a quadratic rotation-energy relation, in the spirit of Green's strain-energy relation, coupled to it or independent of it? Can we write a rotation-torque formula analogous to Hooke's law for linear elasticity in the form

$$L_{ij} = d_{ijkl}\omega_{kl} \quad (\text{A11})$$

where ω_{kl} is the rotation,

$$\omega_{kl} = \frac{1}{2}(u_{k,l} - u_{l,k}). \quad (\text{A12})$$

L_{ij} is the torque density; and d_{ijkl} are the coefficients of rotational elasticity? How are the d's related to the usual c's of elasticity? If we define the rotation vector as

$$\vec{\Omega} = \frac{1}{2}(\nabla \times \vec{u}) \quad (\text{A13})$$

we obtain

$$-V_s^2 \nabla \times (\nabla \times \vec{\Omega}) = \partial^2 \vec{\Omega} / \partial t^2 - \frac{1}{2} \rho^{-1} (\nabla \times \vec{f}) \quad (\text{A14})$$

where the torque density is $\nabla \times \vec{f}$, \vec{f} is the body force density, and ρ is density of the medium. This shows that rotational waves propagate with S-wave velocity and that it may be possible to store torques. Eq. (15) is essentially an extension using the classical elasticity theory.

Lakes [67] pointed out that the behavior of solids can be represented by a variety of continuum theories. In particular, the elasticity theory of the Cosserat brothers [19] incorporates (1) a local rotation of points as well as the translation motion assumed in the classical theory, and (2) a couple stress (a torque per unit area) as well as the force stress (force per unit area). In the constitutive equation for the classical elasticity theory, there are two independent elastic constants, whereas for the Cosserat elastic theory there are six. Lakes (personal communication, 2007) advocates that there is substantial potential for using generalized continuum theories in geo-mechanics, and any theory must have a strong link with experiment (to determine the constants in the constitutive equation) and with physical reality.

Indeed some steps towards better understandings of rotational motions have taken place. For example, Twiss et al. [114] argued that brittle deformation of the Earth's crust ([► Brittle Tectonics: A Non-linear Dynamical System](#)) involving block rotations is comparable to the deformation of a granular material, with fault blocks acting like the grains. They realized the inadequacy of classical continuum mechanics and applied the Cosserat or micropolar continuum theory to take into account two separate scales of motions: macro-motion (large-scale average motion composed of macrostrain rate and macrospin), and micro-motion (local motion composed of microspin). A theoretical link is then established between the kinematics of crustal deformation involving block rotations and the effects on the seismic moment tensor and focal mechanism solutions.

Recognizing that rotational seismology is an emerging field, the *Bulletin of Seismological Society of America* will be publishing in 2009 a special issue under the guest editorship of W.H.K. Lee, M. Çelebi, M.I. Todorovska, and H. Igel.

Bibliography

Primary Literature

1. Ahern TK (2003) The FDSN and IRIS Data Management System: providing easy access to terabytes of information. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam pp 1645–1655
2. Aki K (1966) Generation and propagation of G waves from the Niigata Earthquake of June 16, 1964: Part 1. A statistical analysis. *Bull Earthq Res Inst* 44:23–72; Part 2. Estimation of earthquake moment, released energy, and stress-strain drop from the G wave spectrum. *Bull Earthq Res Inst* 44:73–88
3. Aki K (1969) Analysis of the seismic coda of local earthquakes as scattered waves. *J Geophys Res* 74:6215–6231
4. Aki K, Richards PG (1980) *Quantitative Seismology*. W.H. Freeman, San Francisco
5. Aki K, Richards PG (2002) *Quantitative Seismology: Theory and Methods*, 2nd edn. University Science Books, Sausalito
6. Allen RM (2007) Earthquake hazard mitigation: New directions and opportunities. In: Kanamori H (ed) *Earthquake Seismology. Treatise on Geophysics*, vol 4. Elsevier, Amsterdam, pp 607–648
7. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300:786–789
8. Ambraseys NN, Jackson JA, Melville CP (2002) Historical seismicity and tectonics: The case of the Eastern Mediterranean and the Middle East. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 747–763
9. Anderson JG (2003) Strong-motion seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International*

- Handbook of Earthquake and Engineering Seismology, Part B. Academic Press, Amsterdam pp 937–965
10. Berger J, Davis P, Ekström G (2004) Ambient earth noise: A survey of the Global Seismographic Network. *J Geophys Res* 109:B11307
 11. Borchardt RD (ed) (1997) Vision for the future of the US National Strong-Motion Program, The committee for the future of the US National Strong Motion Program. US Geol Surv Open-File Rept B97530
 12. Bormann P (ed) (2002) New Manual of Seismological Observatory Practice. GeoForschungsZentrum Potsdam http://www.gfz-potsdam.de/bib/nmsop_formular.html
 13. Boroschek R, Legrand D (2006) Tilt motion effects on the double-time integration of linear accelerometers: an experimental approach. *Bull Seism Soc Am* 96:2072–2089
 14. Bouchon M, Aki K (1982) Strain, tilt, and rotation associated with strong ground motion in the vicinity of earthquake faults. *Bull Seism Soc Am* 72:1717–1738
 15. Burridge R, Knopoff L (1964) Body force equivalents for seismic dislocations. *Bull Seism Soc Am* 54:1875–1888
 16. Castellani A, Boffi G (1986) Rotational components of the surface ground motion during an earthquake. *Earthq Eng Struct Dyn* 14:751–767
 17. Chapman CH (2002) Seismic ray theory and finite frequency extensions. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 103–123
 18. Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski M (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg, pp 391–411
 19. Cosserat E, Cosserat F (1909) *Theorie des Corps Deformables*. Hermann, Paris
 20. Dziewonski AM, Woodhouse JH (1983) An experiment in the systematic study of global seismicity: centroid-moment tensor solutions for 201 moderate and large earthquakes of 1981. *J Geophys Res* 88:3247–3271
 21. Dziewonski AM, Woodhouse JH (1983) Studies of the seismic source using normal-mode theory. In: Kanamori H, Boschi B (eds) *Earthquakes: Observation, Theory, and Interpretation*. North-Holland, Amsterdam, pp 45–137
 22. Dziewonski AM, Chou TA, Woodhouse JH (1981) Determination of earthquake source parameters from waveform data for studies of global and regional seismicity. *J Geophys Res* 86:2825–2852
 23. Ekström G (1994) Rapid earthquake analysis utilizes the Internet. *Comput Phys* 8:632–638
 24. Engdahl ER, Villasenor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 665–690
 25. Espinosa-Aranda JM, Jimenez A, Ibarrola G, Alcantar F, Aguilar A, Inostroza M, Maldonado S (1995) Mexico City Seismic Alert System. *Seism Res Lett* 66(6):42–53
 26. Evans JR, Hamstra RH, Kündig C, Camina P, Rogers JA (2005) TREMOR: A wireless MEMS accelerograph for dense arrays. *Earthq Spectr* 21(1):91–124
 27. Evans JR, Cochard A, Graizer V, Huang B-S, Hudnut KW, Hutt CR, Igel H, Lee WHK, Liu C-C, Majewski E, Nigbor R, Safak E, Savage WU, Schreiber U, Teisseyre R, Trifunac M, Wassermann J, Wu C-F (2007) Report of a workshop on rotational ground motion. US Geol Surv Open File Rep 20:2007–1145 <http://pubs.usgs.gov/of/2007/1145/>
 28. Feigl KL (2002) Estimating earthquake source parameters from geodetic measurements. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 607–620
 29. Field EH, Hough SE (1997) The variability of PSV response spectra across a dense array deployed during the Northridge aftershock sequence. *Earthq Spectr* 13:243–257
 30. Fung YC (1965) *Foundations of Solid Mechanics*. Prentice-Hall, Englewood Cliffs
 31. Gasparini P, Manfredi G, Zschau J (eds) (2007) *Seismic Early Warning Systems*. Springer, Berlin
 32. Gee L, Neuhauser D, Dreger D, Pasyanos M, Uhrhammer R, Romanowicz B (2003) The Rapid Earthquake Data Integration Project. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1261–1273
 33. Geiger LC (1912) Probability method for the determination of earthquake epicenters from the arrival time only. *Bull St Louis Univ* 8:60–71
 34. Ghayamghamian MR, Nouri GR (2007) On the characteristics of ground motion rotational components using Chiba dense array data. *Earthq Eng Struct Dyn* 36(10):1407–1429
 35. Gilbert F (1971) Excitation of the normal modes of the Earth by earthquake sources. *Geophys J R Astron Soc* 22:223–226
 36. Gilbert F, Dziewonski AM (1975) Application of normal mode theory to the retrieval of structural parameters and source mechanisms from seismic spectra. *Phil Trans Roy Soc Lond A* 278:187–269
 37. Goltz JD, Flores PJ, Chang SE, Atsumi T (2001) Emergency response and early recovery. In: 1999 Chi-Chi, Taiwan, Earthquake Reconnaissance Report. *Earthq Spectra Suppl A* 17:173–183
 38. Goodstein JR (1991) *Millikan's School: A History of the California Institute of Technology*. Norton, New York
 39. Graizer VM (1991) Inertial seismometry methods. *Izv Earth Phys Akad Nauk SSSR* 27(1):51–61
 40. Graizer VM (2005) Effect of tilt on strong motion data processing. *Soil Dyn Earthq Eng* 25:197–204
 41. Graizer VM (2006) Tilts in strong ground motion. *Bull Seis Soc Am* 96:2090–2106
 42. Guidoboni E (2002) Historical seismology: the long memory of the inhabited world. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 775–790
 43. Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:3–12
 44. Gutenberg B, Richter CF (1954) *Seismicity of the Earth*, 2nd edn. Princeton University Press, Princeton
 45. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Ann Geofis* 9:1–15

46. Hanks TC, Wyss M (1972) The use of body wave spectra in the determination of seismic source parameters. *Bull Seism Soc Am* 62:561–589
47. Hauksson E, Jones LM, Shakal AF (2003) TriNet: a modern ground motion seismic network. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1275–1284
48. Havskov J, Alguacil G (2004) *Instrumentation in Earthquake Seismology*. Springer, Berlin
49. Horiuchi S, Negishi H, Abe K, Kamimura A, Fujinawa Y (2005) An automatic processing system for broadcasting earthquake alarms. *Bull Seism Soc Am* 95:708–718
50. Hoshiba M, Kamigaichi O, Saito M, Tsukada S, Hamada N (2008) Earthquake early warning starts nationwide in Japan, EOS. *Trans Am Geophys Un* 89(8):73–74
51. Housner GW (2002) Historical view of earthquake engineering. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 13–18
52. Huang BS (2003) Ground rotational motions of the 1991 Chi-Chi, Taiwan earthquake as inferred from dense array observations. *Geophys Res Lett* 30(6):1307–1310
53. Huang BS, Liu CC, Lin CR, Wu CF, Lee WHK (2006) Measuring mid- and near-field rotational ground motions in Taiwan. Poster, presented at 2006 Fall AGU Meeting, San Francisco
54. Hutt CR, Bolton HF, Holcomb LG (2002) US contribution to digital global seismograph networks. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 319–322
55. Igel H, Cochard A, Wassermann J, Schreiber U, Velikoseltsev A, Dinh NP (2007) Broadband observations of rotational ground motions. *Geophys J Int* 168(1):182–197
56. Igel H, Schreiber U, Flaws A, Schuberth B, Velikoseltsev A, Cochard A (2005) Rotational motions induced by the M8.1 Tokachi-oki earthquake, September 25, 2003. *Geophys Res Lett* 32:L08309. doi:10.1029/2004GL022336
57. Kamigaichi O (2004) JMA Earthquake Early Warning. *J Japan Assoc Earthq Eng* 4:134–137
58. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2921–2987
59. Kanamori H (1978) Quantification of earthquakes. *Nature* 271:411–414
60. Kanamori H (2003) Earthquake prediction: an overview. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1205–1216
61. Kanamori H (2005) Real-time seismology and earthquake damage mitigation. *Annual Rev Earth Planet Sci* 33:195–214
62. Kanamori H, Brodsky EE (2000) *The physics of earthquakes*. *Phys Today* 54(6):34–40
63. Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) *Earthquakes: Radiated Energy and the Physics of Faulting*. Geophysical Monograph, vol 170. Am Geophys Union, Washington DC, pp 3–13
64. Kanamori H, Hauksson E, Heaton T (1997) Real-time seismology and earthquake hazard mitigation. *Nature* 390:461–464
65. Karal FC, Keller JB (1959) Elastic wave propagation in homogeneous and inhomogeneous media. *J Acoust Soc Am* 31:694–705
66. Kisslinger C, Howell BF (2003) Seismology and physics of the Earth's interior in the USA, 1900–1960. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, p 1453
67. Lakes RS (1995) Experimental methods for study of Cosserat elastic solids and other generalized continua. In: Mühlhaus H (ed) *Continuum Models for Materials with Micro-structure*. Wiley, New York, pp 1–22
68. Lawrence JF, Cochran ES (2007) The Quake Catcher Network: Cyberinfrastructure bringing seismology into schools and homes. American Geophysical Union, Fall Meeting 2007, abstract #ED11C-0633
69. Lee VW, Trifunac MD (1985) Torsional accelerograms. *Int J Soil Dyn Earthq Eng* 4(3):132–139
70. Lee VW, Trifunac MD (1987) Rocking strong earthquake accelerations. *Int J Soil Dyn Earthq Eng* 6(2):75–89
71. Lee WHK (2002) Challenges in observational seismology. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, 269–281
72. Lee WHK, Benson RB (2008) Making non-digitally-recorded seismograms accessible online for studying earthquakes. In: Fréchet J, Meghraoui M, Stucchi M (eds) *Modern Approach in Historical Seismology: Interdisciplinary studies of past and recent earthquakes*. Springer, Berlin, pp 403–427
73. Lee WHK, Espinosa-Aranda JM (2003) Earthquake early warning systems: Current status and perspectives. In: Zschau J, Koppers AN (eds) *Early Warning Systems for Natural Disaster, Reduction*. Springer, Berlin, pp 409–423
74. Lee WHK, Stewart SW (1981) *Principles and Applications of Microearthquake Networks*. Academic Press, New York
75. Lee WHK, Celebi M, Todorovska MI, Diggles MF (2007) Rotational seismology and engineering applications: Proceedings for the First International Workshop, Menlo Park, California, USA, 18–19 September. US Geol Surv Open File Rep 2007–1144. <http://pubs.usgs.gov/of/2007/1144/>
76. Lee WHK, Igel H, Todorovska MI, Evans JR (2007) Rotational Seismology: AGU Session, Working Group, and Website. US Geol Surv Open File Rep 2007–1263. <http://pubs.usgs.gov/of/2007/1263/>
77. Lognonne P, Clevede E (2002) Normal modes of the Earth and planets. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 125–147
78. Luco JE (1976) Torsional response of structures to obliquely incident seismic SH waves. *Earthq Eng Struct Dyn* 4:207–219
79. Mallet R (1858) Fourth report on the facts of earthquake phenomena. *Ann Rep Brit Assn Adv Sci* 28:1–136
80. Mallet R (1862) *Great Neapolitan Earthquake of 1857, vol I, II*. Chapman and Hall, London
81. Maruyama T (1963) On the force equivalents of dynamical elastic dislocations with reference to the earthquake mechanism. *Bull Earthq Res Inst* 41:467–486
82. Michelini A, De Simoni B, Amato A, Boschi E (2005) Collecting, digitizing and distributing historical seismological data. EOS 86(28) 12 July 2005

83. Nakamura Y (1984) Development of the earthquake early-warning system for the Shinkansen, some recent earthquake engineering research and practical in Japan. The Japanese National Committee of the International Association for Earthquake Engineering, pp 224–238
84. Nakamura Y (1988) On the urgent earthquake detection and alarm system, UrEDAS. *Proc Ninth World Conf Earthq Eng* 7:673–678
85. Nakamura Y (2004) On a rational strong motion index compared with other various indices. 13th World Conf Earthq Eng, Paper No 910
86. Nakamura Y, Saita J (2007) UrEDAS, The Earthquake Warning System: today and tomorrow. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake Early Warning Systems*. Springer, Berlin, pp 249–281
87. Newmark NM (1969) Torsion in symmetrical buildings. *Proc. Fourth World Conference on Earthquake Eng*, vol II. pp A3/19–A3/32
88. Niazi M (1987) Inferred displacements, velocities and rotations of a long rigid foundation located at El-Centro differential array site during the 1979 Imperial Valley, California, earthquake. *Earthq Eng Struct Dyn* 14:531–542
89. Odaka T, Ashiya K, Tsukada S, Sato S, Ohtake K, Nozaka D (2003) A new method of quickly estimating epicentral distance and magnitude from a single seismic record. *Bull Seism Soc Am* 93:526–532
90. Oliveira CS, Bolt BA (1989) Rotational components of surface strong ground motion. *Earthq Eng Struct Dyn* 18:517–526
91. Oliver J, Murphy L (1971) WWNSS: Seismology's global network of observing stations. *Science* 174:254–261
92. Peterson J (1993) Observations and modeling of seismic background noise. *US Geol Surv Open File Rep*, 93–322
93. Pillet R, Virieux J (2007) The effects of seismic rotations on inertial sensors. *Geophys J Int.* doi:10.1111/j.1365-246X.2007.03617.x
94. Reid HF (1910) The California Earthquake of 18 April 1906, vol 2. *The Mechanics of the Earthquake*. Carnegie Inst, Washington DC
95. Richards PG (2002) Seismological methods of monitoring compliance with the Comprehensive Nuclear Test-Ban Treaty. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 369–382
96. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seis Soc Am* 25:1–32
97. Richter CF (1958) *Elementary Seismology*. Freeman, San Francisco
98. Rutenberg A, Heidebrecht AC (1985) Rotational ground motion and seismic codes. *Can J Civ Eng* 12(3):583–592
99. Saita J, Nakamura Y (2003) UrEDAS: the early warning system for mitigation of disasters caused by earthquakes and tsunamis. In: Zschau J, Koppers AN (eds) *Early Warning Systems for Natural Disaster, Reduction*. Springer, Berlin, pp 453–460
100. Savage JC (1978) Dislocation in seismology. In: Nabarro FRN (ed) *Dislocation in Solids*. North-Holland, Amsterdam, 251–339
101. Schweitzer J, Lee WHK (2003) Old seismic bulletins: a collective heritage from early seismologists. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1665–1717 (with CD-ROM)
102. Shin TC, Tsai YB, Yeh YT, Liu CC, Wu YM (2003) Strong-motion instrumentation programs in Taiwan. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, Amsterdam, pp 1057–1062
103. Sokolnikoff IS (1956) *Mathematical Theory of Elasticity*, 2nd edn. McGraw-Hill, New York
104. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Park-field, California, produced by the m 7.4 Landers earthquake of 28 June 1992: Observations from the UPSAR dense seismograph array. *J Geophys Res* 100:675–690
105. Suryanto W, Igel H, Wassermann J, Cochard A, Schubert B, Vollmer D, Scherbaum F (2006) Comparison of seismic array-derived rotational motions with direct ring laser measurements. *Bull Seism Soc Am* 96(6):2059–2071
106. Takeo M (1998) Ground rotational motions recorded in near-source region. *Geophys Res Lett* 25(6):789–792
107. Takeo M, Ito HM (1997) What can be learned from rotational motions excited by earthquakes? *Geophys J Int* 129:319–329
108. Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of theory and recording. *Annali Geofis* 46:671–685
109. Teisseyre R, Takeo M, Majewski E (eds) (2006) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Berlin
110. Teng TL, Wu L, Shin TC, Tsai YB, Lee WHK (1997) One minute after: strong motion map, effective epicenter, and effective magnitude. *Bull Seism Soc Am* 87:1209–1219
111. Trifunac MD (1982) A note on rotational components of earthquake motions on ground surface for incident body waves. *Soil Dyn Earthq Eng* 1:11–19
112. Trifunac MD (2006) Effects of torsional and rocking excitations on the response of structures. In: Teisseyre R, Takeo M, Majewski M (eds) *Earthquake Source Asymmetry, Structural Media and Rotation Effects*. Springer, Heidelberg, pp 569–582
113. Trifunac MD, Todorovska MI (2001) A note on the useable dynamic range of accelerographs recording translation. *Soil Dyn Earthq Eng* 21(4):275–286
114. Twiss R, Souter B, Unruh J (1993) The effect of block rotations on the global seismic moment tensor and the patterns of seismic P and T axes. *J Geophys Res* 98(B1):645–674
115. Utsu T (2002) A list of deadly earthquakes in the world (1500–2000). In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 691–717
116. Utsu T (2002) Relationships between magnitude scales. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 733–746
117. Uyeda S (2002) Continental drift, sea-floor spreading, and plate/plume tectonics. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 51–67
118. Villaseñor A, Engdahl ER (2007) Systematic relocation of early instrumental seismicity: Earthquakes in the International Seismological Summary for 1960–1963. *Bull Seism Soc Am* 97:1820–1832

119. Wald DJ, Quitoriano V, Heaton TH, Kanamori H (1999) Relationships between peak ground acceleration, peak ground velocity, and modified Mercalli intensity in California. *Earthq Spectr* 15:557–564
120. Wald DJ, Quitoriano V, Heaton TH, Kanamori H, Scrivner CW, Worden CB (1999) TriNet “ShakeMaps”: Rapid generation of peak ground motion and intensity maps for earthquakes in Southern California. *Earthq Spectr* 15:537–555
121. Webb SC (2002) Seismic noise on land and on the seafloor. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 305–318
122. Wielandt E (2002) Seismometry. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, Amsterdam, pp 283–304
123. Willemann RJ, Storchak DA (2001) Data collection at the International Seismological Centre. *Seism Res Lett* 72:440–453
124. Wu YM, Kanamori H (2005) Experiment on an onsite early warning method for the Taiwan early warning system. *Bull Seism Soc Am* 95:347–353
125. Wu YM, Kanamori H (2005) Rapid assessment of damaging potential of earthquakes in Taiwan from the beginning of P Waves. *Bull Seism Soc Am* 95:1181–1185
126. Wu YM, Kanamori H (2008) Exploring the feasibility of on-site earthquake early warning using close-in records of the 2007 Noto Hanto earthquake. *Earth Planets Space* 60:155–160
127. Wu YM, Zhao L (2006) Magnitude estimation using the first three seconds P-wave amplitude in earthquake early warning. *Geophys Res Lett* 33:L16312. doi:10.1029/2006GL026871
128. Wu YM, Chen CC, Shin TC, Tsai YB, Lee WHK, Teng TL (1997) Taiwan Rapid Earthquake Information Release System. *Seism Res Lett* 68:931–943
129. Wu YM, Hsiao NC, Lee WHK, Teng TL, Shin TC (2007) State of the art and progresses of early warning system in Taiwan. In: Gasparini P, Manfredi G, Zschau J (eds) *Earthquake Early Warning Systems*. Springer, Berlin, pp 283–306
130. Wu YM, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters, τ_c and P_d , for southern California. *Geophys J Int* 169:667–674
131. Wu YM, Lee WHK, Chen CC, Shin TC, Teng TL, Tsai YB (2000) Performance of the Taiwan Rapid Earthquake Information Release System (RTD) during the 1999 Chi-Chi (Taiwan) earthquake. *Seism Res Lett* 71:338–343
132. Zollo A, Lancieri M, Nielsen S (2006) Earthquake magnitude estimation from peak amplitudes of very early seismic signals on strong motion records. *Geophys Res Lett* 33:L23312. doi:10.1029/2006GL027795
- Kanamori H, Boschi E (1983) *Earthquakes: Observation, Theory and Interpretation*. North-Holland, Amsterdam
- Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction*. Springer, Berlin
- Lee WHK, Meyers H, Shimazaki K (eds) (1988) *Historical Seismograms and Earthquakes of the World*. Academic Press, San Diego
- Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2002) *International Handbook of Earthquake and Engineering Seismology, Part A*. Academic Press, San Diego, pp 933 (and 1 CD-ROM)
- Lee WHK, Kanamori H, Jennings JC, Kisslinger C (eds) (2003) *International Handbook of Earthquake and Engineering Seismology, Part B*. Academic Press, San Diego, pp 1009 (and 2 CD-ROMs)
- Pujol J (2003) *Elastic Wave Propagation and Generation in Seismology*. Cambridge Univ Press, Cambridge
- Zschau J, Koppers AN (eds) (2003) *Early Warning Systems for Natural Disaster Reduction*. Springer, Berlin

Earthquake Networks, Complex

SUMIYOSHI ABE^{1,2}, NORIKAZU SUZUKI³

¹ Department of Physical Engineering, Mie University, Tsu, Japan

² Institut Supérieur des Matériaux et Mécaniques, Le Mans, France

³ College of Science and Technology, Nihon University, Chiba, Japan

Article Outline

Glossary

Definition of the Subject

Introduction

Construction of an Earthquake Network

Scale-free Nature of Earthquake Network

Small-World Nature of Earthquake Network

Hierarchical Structure

Mixing Property

Period Distribution

Future Directions

Addendum

Bibliography

Books and Reviews

- Abercrombie R, McGarr A, Kanamori H, Di Toro G (2006) *Earthquakes: Radiated Energy and the Physics of Faulting*. Geophysical Monograph, vol 170. American Geophysical Union, Washington DC
- Bolt BA (1993) *Earthquakes*. W.H. Freeman, New York
- Chen YT, Panza GF, Wu ZL (2004) *Earthquake Hazard, Risk, and Strong Ground Motion*. Seismological Press, Beijing
- Kanamori H (ed) (2007) *Earthquake Seismology, Treatise on Geophysics*, vol 4. Elsevier, Amsterdam

Glossary

Network or graph A network (or a graph) [28] consists of vertices (or nodes) and edges (or links) connecting them. In general, a network contains loops (i. e., edges with both ends attached to the same vertices) and multiple edges (i. e., edges more than one that connect two different vertices). If edges have their directions,

such a network is called directed. A simple graph is a network, in which loops are removed and each multiple edge is replaced by a single edge. In a stochastic network, each connection is inherently probabilistic. A classical random graph is a simple example, in which each two vertices are connected by an edge with probability p and unconnected with probability $1 - p$ ($0 < p < 1$).

Connectivity distribution or degree distribution The connectivity distribution (or the degree distribution), $P(k)$, is the probability of finding vertices with k edges in a stochastic network. In a directed network, the number of incoming/outgoing edges is called the in-degree/out-degree. Connectivity of a classical random graph obeys the Poissonian distribution in the limit of the large number of vertex [11,14,20], $P(k) = e^{-\lambda} \lambda^k / k!$ (λ : a positive parameter, $k = 0, 1, 2, \dots$), whereas a scale-free network [11,12,14,20] has a power-law shape, $P(k) \sim k^{-\gamma}$ (γ : a positive exponent), for large k .

Preferential attachment rule This is a concept relevant to a growing network, in which the number of vertices increases. Preferential attachment [11,12,14,20] implies that a newly created vertex tends to link to pre-existing vertices with the probability $\Pi(k_i) = k_i / \sum_j k_j$, where k_i stands for the connectivity of the i th vertex. That is, the larger the connectivity of a vertex is, the higher the probability of getting linked to a new vertex is.

Clustering coefficient The clustering coefficient [27] is a quantity characterizing an undirected simple graph. It quantifies the adjacency of two neighboring vertices of a given vertex, i. e., the tendency of two neighboring vertices of a given vertex to be connected to each other. Mathematically, it is defined as follows. Assume the i th vertex to have k_i neighboring vertices. There can exist at most $k_i(k_i - 1)/2$ edges between the neighbors. Define c_i as the ratio

$$c_i = \frac{\text{actual number of edges between the neighbors of the } i\text{th vertex}}{k_i(k_i - 1)/2} . \quad (1)$$

Then, the clustering coefficient is given by the average of this quantity over the network:

$$C = \frac{1}{N} \sum_{i=1}^N c_i , \quad (2)$$

where N is the total number of vertices contained in the network. The value of the clustering coefficient of

a random graph, C_{random} , is much smaller than unity, whereas a small-world network has a large value of C which is much larger than C_{random} .

Hierarchical organization Many complex networks are structurally modular, that is, they are composed of groups of vertices that are highly interconnected to each other but weakly connected to outside groups. This hierarchical structure [22] can conveniently be characterized by the clustering coefficient at each value of connectivity, $c(k)$, which is defined by

$$c(k) = \frac{1}{NP_{\text{SG}}(k)} \sum_{i=1}^N c_i \delta_{k_i, k} , \quad (3)$$

where c_i is given by (1), N the total number of vertices, and $P_{\text{SG}}(k)$ the connectivity distribution of an undirected simple graph. Its average is the clustering coefficient in (2): $C = \sum_k c(k) P_{\text{SG}}(k)$. A network is said to be hierarchically organized if $c(k)$ varies with respect to k , typically due to a power law, $c(k) \sim k^{-\beta}$, with a positive exponent β .

Assortative mixing and disassortative mixing Consider the conditional probability, $P(k'|k)$, of finding a vertex with connectivity k' linked to a given vertex with connectivity k . Then, the nearest-neighbor average connectivity of vertices with connectivity k is defined by [20,21,26]

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k) . \quad (4)$$

If $\bar{k}_{nn}(k)$ increases/decreases with respect to k , mixing is termed assortative/disassortative. A simple model of growth with preferential attachment is known to possess no mixing. That is, $\bar{k}_{nn}(k)$ does not depend on k . The above-mentioned linking tendency can be quantified by the correlation coefficient [17] defined as follows. Let $e_{kl}(= e_{lk})$ be the joint probability distribution for an edge to link with a vertex with connectivity k at one end and a vertex with connectivity l at the other. Calculate its marginal, $q_k = \sum_l e_{kl}$. Then, the correlation coefficient is given by

$$r = \frac{1}{\sigma_q^2} \sum_{k,l} kl (e_{kl} - q_k q_l) , \quad (5)$$

where $\sigma_q^2 = \sum_k k^2 q_k - (\sum_k k q_k)^2$ stands for the variance of q_k . $r \in [-1, 1]$, and if r is positive/negative, mixing is assortative/disassortative [17,20].

Definition of the Subject

Complexity is an emergent collective property, which is hardly understood by the traditional approach in natural

science based on reductionism. Correlation between elements in a complex system is strong, no matter how largely they are separated both spatially and temporally, therefore it is essential to treat such a system in a holistic manner, in general.

Although it is generally assumed that seismicity is an example of complex phenomena, it is actually nontrivial to see how and in what sense it is complex. This point may also be related to the question of primary importance of why it is so difficult to predict earthquakes.

Development of the theory of complex networks turns out to offer a peculiar perspective on this point. Construction of a complex earthquake network proposed here consists of mapping seismic data to a growing stochastic graph. This graph, or network, turns out to exhibit a number of remarkable behaviors both physically and mathematically, which are in common with many other complex systems. The scale-free and small-world natures are typical examples. In this way, one will be able to obtain a novel viewpoint of seismicity.

Introduction

Seismicity is a field-theoretical phenomenon. Released energy of each earthquake may be regarded as a field amplitude defined at a discrete spacetime point. However, in contrast to a familiar field theory such as the electromagnetic theory, both amplitudes and locations are intrinsically probabilistic. The fault distribution may geometrically be fractal [18], and the stress distribution superposed upon it often has a complex landscape. Accordingly, seismicity is characterized by extremely rich phenomenology, which attracts physicists' attention from the viewpoint of science of complex systems.

There are at least two celebrated empirical laws known in seismology. One is the Gutenberg–Richter law [16], which states that the frequency of earthquakes obeys a power law with respect to released energy. This power-law nature makes it difficult or even meaningless to statistically distinguish earthquakes by their values of magnitude because of the absence of typical energy scales. The other is the Omori law [19], which states that the rate of the frequency of aftershocks following a main shock algebraically decays with respect to time elapsed from the main shock. This slow relaxation reminds one of complex glassy dynamics [13]. Such a viewpoint is supported by the discovery of the aging phenomenon and the scaling law for aftershocks [2].

Another point, which seems less noticed, is that correlation of two successive events is strong, no matter how large their spatial separation is. There is, in fact, an obser-

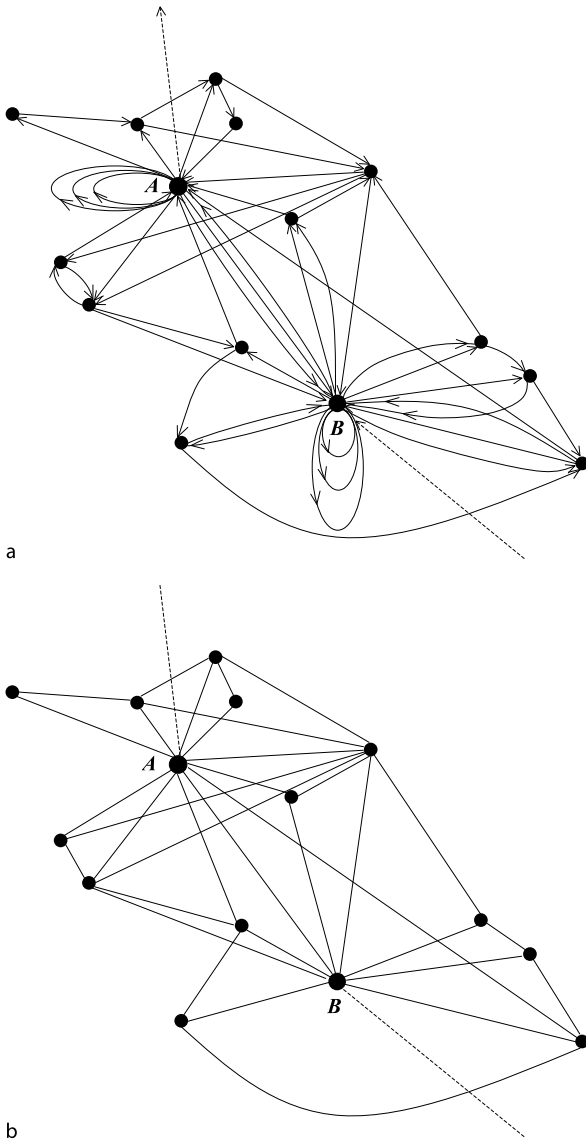
vation [23] that an earthquake can be triggered by a foregoing one more than 1000 km away. The reason why two successive events are indivisibly related can also be found in another observation [3,6] that both spatial distance and time interval between two successive events obey the q -exponential distributions in nonextensive statistics [1,15,24], which offers a statistical-mechanical framework for describing complex systems. Thus, the correlation length can be enormously large and long-wave-length modes of seismic waves play an important role. This has a strong similarity to phase transitions and critical phenomena. Accordingly, it may not be appropriate to use spatial windows in analysis of seismicity. Furthermore, all of the data in a relevant area (ideally the whole globe, though still not satisfactorily available) should be treated based on the nonreductionistic standpoint.

The network approach is a powerful tool for analyzing kinematical and dynamical structures of complex systems in a holistic manner. Such a concept was introduced to seismology by the present authors in 2004 [4] in order to represent complexity of seismicity. The procedure described in Sect. “Construction of an Earthquake Network” allows one to map a seismic time series to a growing stochastic network in an unambiguous way. Vertices and edges of such a network correspond to coarse-grained events and event-event correlations, respectively. Yet unknown microscopic dynamics governing event-event correlations and fault-fault interactions are replaced by these edges. Global physical properties of seismicity can then be explored by examining its geometric (e.g., topological etc.) and dynamical features. It turns out that earthquake networks have a number of intriguing properties, some of which are shared by many other natural as well as artificial systems including metabolic networks, food webs, the Internet, the world-wide web, and so on [11,14,20]. This, in turn, enables seismologists to study seismicity in analogy with such relatively better understood complex systems. Thus, the network approach offers a novel way of analyzing seismic time series and casts fresh light on the physics of earthquakes.

In this article, only the data taken from California is utilized. However, it has been ascertained that the laws and trends discussed here are universal and hold also in other geographical regions including Japan.

Construction of an Earthquake Network

An earthquake network is constructed as follows [4]. A geographical region under consideration is divided into small cubic cells. A cell is regarded as a vertex if earthquakes with any values of magnitude above a certain de-



Earthquake Networks, Complex, Figure 1

a A schematic description of earthquake network. The *dashed lines* correspond to the initial and final events. The vertices, *A* and *B*, contain main shocks and play roles of hubs of the network. **b** The undirected simple graph reduced from the network in **a**.

tection threshold occurred therein. Two successive events define an edge between two vertices. If they occur in the same cell, a loop is attached to that vertex. This procedure enables one to map a given interval of the seismic data to a growing probabilistic graph, which is referred to as an earthquake network (see Fig. 1a).

Several comments are in order. Firstly, this construction contains a single parameter: cell size, which is a scale of coarse graining. Once cell size is fixed, an earthquake

network is unambiguously defined. However, since there exist no a priori operational rule to determine cell size, it is important to notice how the properties of an earthquake network depend on this parameter. Secondly, as mentioned in Sect. “[Introduction](#)”, edges and loops efficiently represent event-event correlation. Thirdly, an earthquake network is a directed graph in its nature. Directedness does not bring any difficulties to statistical analysis of connectivity (degree, i.e., the number of edges attached to the vertex under consideration) since, by construction, the in-degree and out-degree are identical for each vertex except the initial and final vertices in analysis. Therefore, the in-degree and out-degree are not distinguished from each other in the analysis of the connectivity distribution (see Sections “[Scale-free Nature of Earthquake Network](#)” and “[Mixing Property](#)”). However, directedness becomes essential when the path length (i.e., the number of edges) between a pair of connected vertices, i.e., the degree of separation between the pair, is considered. This point is explicitly discussed in the analysis of the period distribution in Sect. “[Period Distribution](#)”. Finally, a full directed earthquake network has to be reduced to a simple undirected graph, when its small-worldness and hierarchical structure are examined (see Sections “[Small-World Nature of Earthquake Network](#)” and “[Hierarchical Structure](#)”). There, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). The path length in this case is the smallest value among the possible numbers of edges connecting a pair of vertices.

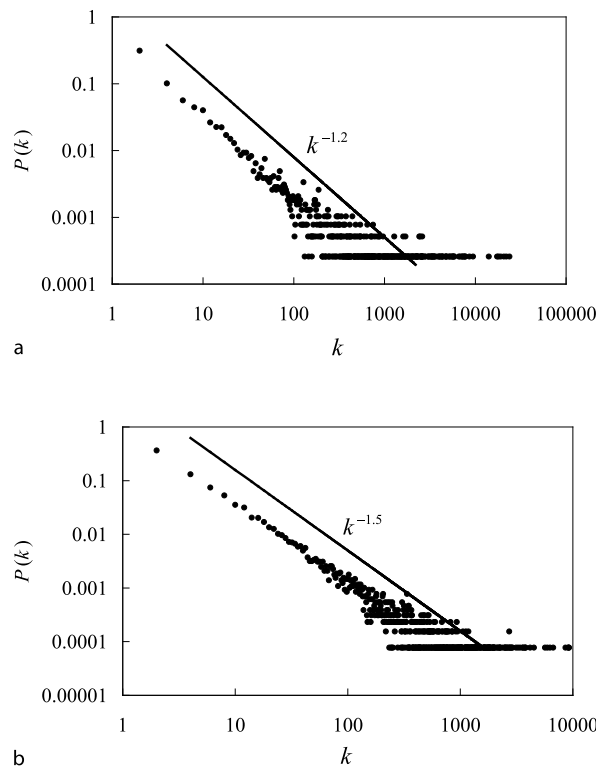
Scale-free Nature of Earthquake Network

An earthquake network contains some special vertices which have large values of connectivity. Such “hubs” turn out to correspond to cells with main shocks. This is due to a striking fact discovered from real data analysis that aftershocks associated with a main shock tend to return to the locus of the main shock, geographically. This is the primary reason why a vertex containing a main shock becomes a hub. The situation is analogous to the preferential attachment rule for a growing network [11,14,20]. According to this rule, a newly created vertex tends to be connected to the (already existing) i th vertex with connectivity k_i with probability, $\Pi(k_i) = k_i / \sum_j k_j$. It can generate a scale-free network characterized by the power-law connectivity distribution [11,12]:

$$P(k) \sim k^{-\gamma}, \quad (6)$$

where γ is a positive exponent.

In Fig. 2, the connectivity distribution of the full earthquake network with loops and multiple edges is pre-



Earthquake Networks, Complex, Figure 2
The log-log plots of the connectivity distributions of the earthquake network constructed from the seismic data taken in California [the Southern California Earthquake Data Center (<http://www.data.scec.org/>)]. The time interval analyzed is between 00:25:58 on January 1, 1984 and 22:21:52.09 on December 31, 2003. The region covered is $29^{\circ}06.00'N$ – $38^{\circ}59.76'N$ latitude and $113^{\circ}06.00'W$ – $122^{\circ}55.59'W$ longitude with the maximal depth 175.99 km. The total number of events is 367 613. The data contains no threshold for magnitude (but “quarry blasts” are excluded from the analysis). Two different values of cell size are examined: a $10 \text{ km} \times 10 \text{ km} \times 10 \text{ km}$ and b $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$. All quantities are dimensionless.

sented [4]. From it, one appreciates that the earthquake network in fact possesses connectivity of the form in (6) and is therefore scale-free. The smaller the cell size, the larger the exponent, γ , is, since the number of vertices with large values of connectivity decreases as cell size becomes smaller. The scale-free nature may be interpreted as follows. As mentioned above, aftershocks associated with a main shock tend to be connected to the vertex of the main shock, satisfying the preferential attachment rule. On the other hand, the Gutenberg–Richter law states that frequency of earthquakes decays slowly as a power law with respect to released energy. This implies that there appear to be quite a few giant components, and accordingly the network becomes highly inhomogeneous.

Earthquake Networks, Complex, Table 1
The small-world properties of the undirected simple earthquake network. The values of the number of vertices, N , the clustering coefficient, C , (compared with those of the classical random graphs, C_{random}) and the average path length, L are presented. The data employed is the same as that in Fig. 2.

Cell size	$10 \text{ km} \times 10 \text{ km} \times 10 \text{ km}$	$5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$
Number of vertices	$N = 3869$	$N = 12913$
Clustering coefficient	$C = 0.630$ ($C_{\text{random}} = 0.014$)	$C = 0.317$ ($C_{\text{random}} = 0.003$)
Average path length	$L = 2.526$	$L = 2.905$

Small-World Nature of Earthquake Network

The small-world nature is an important aspect of complex networks. It shows how a complex network is different from both regular and classical random graphs [27]. A small-world network resides in-between regularity and randomness, analogous to the edge of chaos in nonlinear dynamics.

To study the small-world nature of an earthquake network, a full network has to be reduced to a simple undirected graph: that is, loops are removed and each multiple edge is replaced by a single edge (see Fig. 1b). This is because in the small-world picture one is concerned only with simple linking pattern of vertices.

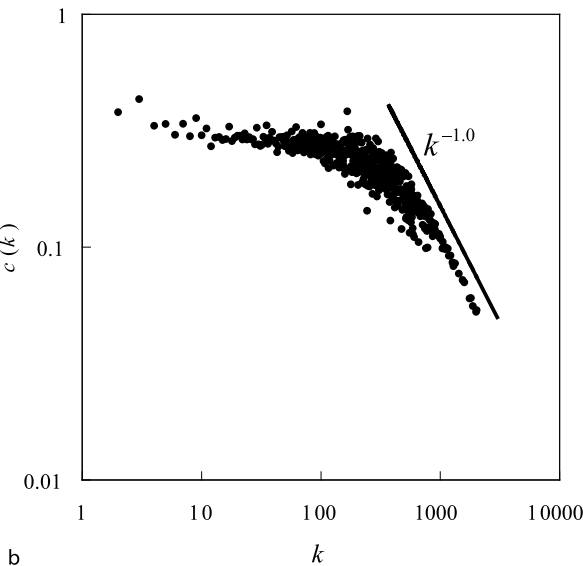
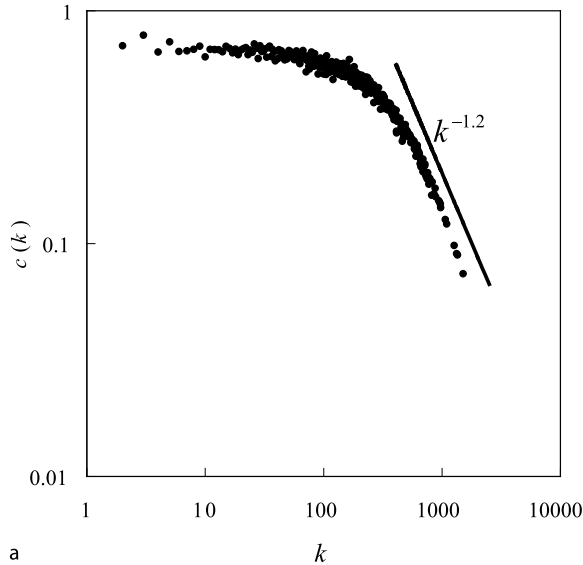
A small-world network is characterized by a large value of the clustering coefficient in (2) and a small value of the average path length [27]. The clustering coefficient quantifies the tendency of two neighboring vertices of a given vertex to be connected to each other. A small-world network has a large value of the clustering coefficient, whereas the value for the classical random graph is very small [11,14,20,27]: $C_{\text{random}} = \langle k \rangle / N \ll 1$, where N and $\langle k \rangle$ are the total number of vertices and the average value of connectivity, respectively.

In Table 1, the results are presented for the clustering coefficient and the average path length [5,9]. One finds that the values of the clustering coefficient are in fact much larger than those of the classical random graphs and the average path length is short. Thus, the earthquake network is an important feature of small-world network.

Hierarchical Structure

As seen above, seismicity generates a scale-free network of an earthquake network further, one may examine if it is hierarchically organized [8]. The hierarchical structure can be revealed by analyzing the clustering coefficient as a function of connectivity. The connectivity-dependent cluster-

ing coefficient, $c(k)$, is defined in (3). This quantifies the adjacency of two vertices connected to a vertex with connectivity, k , and gives information on hierarchical organization of a network.



Earthquake Networks, Complex, Figure 3

The log-log plots of the connectivity-dependent clustering coefficient for two different values of cell size: **a** 10 km × 10 km × 10 km and **b** 5 km × 5 km × 5 km. The analyzed period is between 00:25:8.58 on January 1, 1984 and 22:50:49.29 on December 31, 2004, which is taken from the same catalog as in Fig. 2. The region covered is 28°36.00'N–38°59.76'N latitude and 112°42.00'W–123°37.41'W longitude with the maximal depth 175.99 km. The total number of the events is 379728. All quantities are dimensionless.

In Fig. 3, the plots of $c(k)$ are presented [8]. As can be clearly seen, the clustering coefficient of the undirected simple earthquake network asymptotically follows the scaling law

$$c(k) \sim k^{-\beta} \quad (7)$$

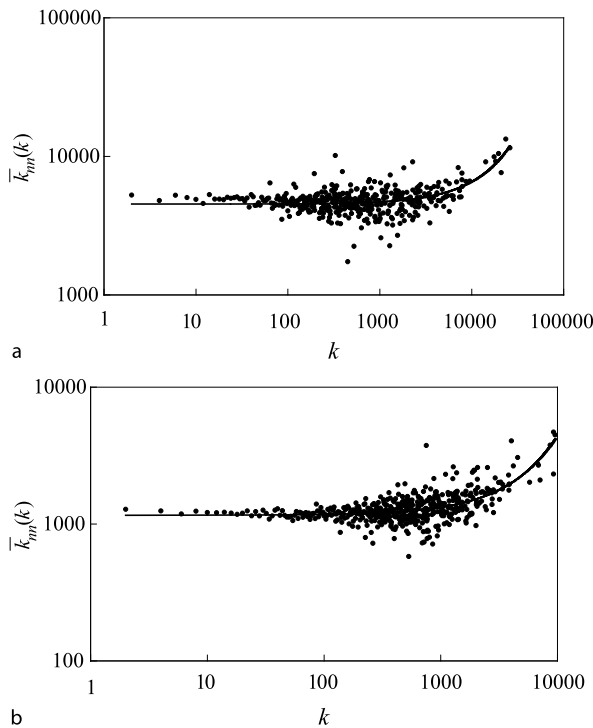
with a positive exponent β . This highlights hierarchical organization of the earthquake network.

Existence of the hierarchical structure is of physical importance. The earthquake network [11,12,14,20]. However, the standard preferential-attachment-model is known to fail at generating hierarchical organization [22]. To mediate between growth with preferential attachment and the presence of hierarchical organization, the concept of vertex deactivation has been introduced in the literature [25]. According to this concept, in the process of network growth, some vertices deactivate and cannot acquire new edges any more. This has a natural physical implication for an earthquake network: active faults may be deactivated through the process of stress release. In addition, the fitness model [26] is also known to generate hierarchical organization. This model generalizes the preferential attachment rule in such a way that not only connectivity but also “charm” of vertices (i. e., attracting a lot of edges) are taken into account. Seismologically, fitness is considered to describe intrinsic properties of faults such as geometric configuration and stiffness. Both of these two mechanisms can explain a possible origin of the complex hierarchical structure, by which relatively new vertices have chances to become hubs of the network. In the case of an earthquake network, it seems plausible to suppose that the hierarchical structure may be due to both deactivation and fitness.

A point of particular interest is that the hierarchical structure disappears if weak earthquakes are removed. For example, setting a lower threshold for earthquake magnitude, say $M_{th} = 3$, makes it difficult to observe the power-law decay of the clustering coefficient hierarchical structure of an earthquake network is largely supported by weak shocks.

Mixing Property

The scale-free nature, small-worldness, growth with preferential attachment, and hierarchical organization all indicate that earthquake networks are very similar to other known networks, for example, the Internet. However, there is at least one point which shows an essential difference between the two. It is concerned with the mixing property, which is relevant to the concept of the nearest-



Earthquake Networks, Complex, Figure 4
The log-log plots of the nearest-neighbor average connectivity for two different values of cell size: a $10 \text{ km} \times 10 \text{ km} \times 10 \text{ km}$ and b $5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$. The data employed is the same as that in Fig. 3. The solid lines show the trends depicted by the exponentially increasing functions. All quantities are dimensionless.

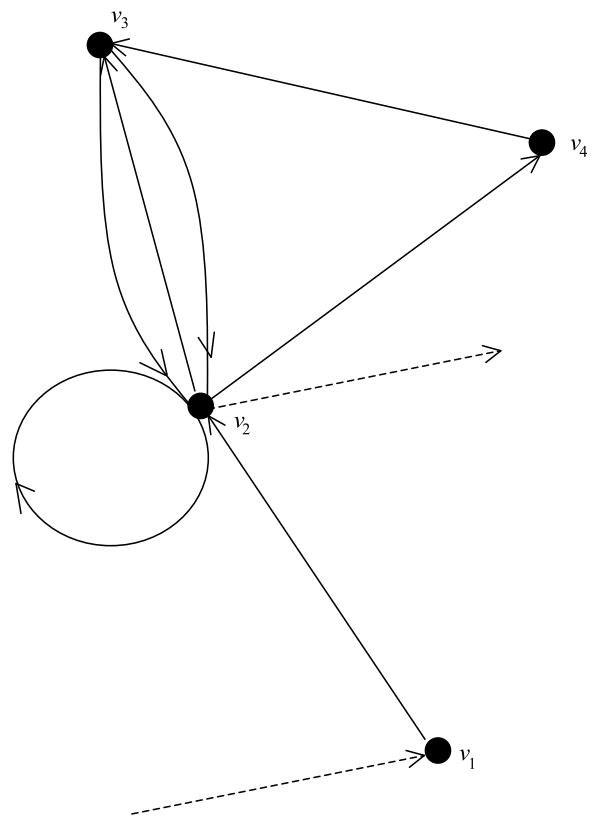
Earthquake Networks, Complex, Table 2
The values of the dimensionless correlation coefficient. The data employed is the same as that in Fig. 3. Positivity of the values implies that mixing is assortative.

$10 \text{ km} \times 10 \text{ km} \times 10 \text{ km}$	$5 \text{ km} \times 5 \text{ km} \times 5 \text{ km}$
$r = 0.285$	$r = 0.268$

neighbor average connectivity $\bar{k}_{nn}(k)$ [in (4)], of a full network with loops and multiple edges.

The plots of this quantity are presented in Fig. 4. There, the feature of assortative mixing [8] is observed, since $\bar{k}_{nn}(k)$ increases with respect to connectivity k . Therefore, vertices with large values of connectivity tend to be linked to each other. That is, vertices containing stronger shocks tend to be connected among themselves with higher probabilities.

To quantify this property, the correlation coefficient in (5) is evaluated [8]. The result is summarized in Table 2. The value of the correlation coefficient is in fact positive, confirming that the earthquake network has assorta-

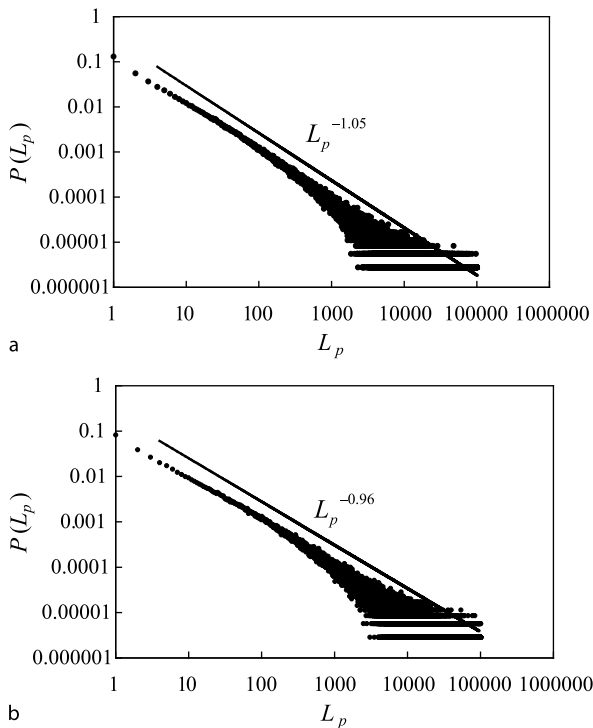


Earthquake Networks, Complex, Figure 5
A full directed network: $\dots \rightarrow v_1 \rightarrow v_2 \rightarrow v_3 \rightarrow v_2 \rightarrow v_2 \rightarrow v_4 \rightarrow v_3 \rightarrow v_2 \rightarrow \dots$. The period associated with v_3 is 4, whereas v_2 has 1, 2 and 3.

tive mixing. On the other hand, the Internet is of disassortative mixing [17,20,21,26]. That is, the mixing properties of the earthquake network and the Internet are opposite to each other. It is noticed however that the loops and multiple edges play essential roles for the assortative mixing: an undirected simple graph obtained by reducing a full earthquake network turns out to exhibit disassortative mixing. These are purely the phenomenological results, and their physical origins still have yet to be clarified.

Period Distribution

So far, directedness of an earthquake network has been ignored. The full directed network picture is radically different from the small-world picture for a simple undirected graph. It enables one to consider interesting dynamical features of an earthquake network. As an example, here the concept of period [7] is discussed. This is relevant to the question “after how many earthquakes does an event return to the initial cell, statistically?” It is therefore of obvious interest for earthquake prediction.



Earthquake Networks, Complex, Figure 6

The log-log plots of the period distribution for two different values of cell size: **a** $10\text{ km} \times 10\text{ km}$ and **b** $5\text{ km} \times 5\text{ km} \times 5\text{ km}$. The data employed is the same as that in Fig. 2. All quantities are dimensionless.

Period in a directed network is defined as follow. Given a vertex of a network, there are various closed routes starting from and ending at this vertex. The period, L_p , of a chosen closed route is simply the number of edges forming the route (see Fig. 5).

The period distribution, $P(L_p)$, is defined as the number of closed routes. The result is presented in Fig. 6 [7]. As can be seen there, $P(L_p)$ obeys a power law

$$P(L_p) \sim (L_p)^{-\alpha}, \quad (8)$$

where α is a positive exponent. This implies that there exist a number of closed routes with significantly long periods in the network. This fact makes it highly nontrivial to statistically estimate the value of period.

Future Directions

In the above, the long-time statistical properties of an earthquake network have mainly been considered. On the other hand, given the cell size, an earthquake network represents all the dynamical information contained in a seismic time series, and therefore the study of its time evolu-

tion may give a new insight into seismicity. This, in turn, implies that it may offer a novel way of monitoring seismicity.

For example, it is of interest to investigate how the clustering coefficient changes in time as earthquake network dynamically evolves. According to the work in [10], the clustering coefficient remains stationary before a main shock, suddenly jumps up at the main shock, and then slowly decays to become stationary again following the power-law relaxation. In this way, the clustering coefficient successfully characterizes aftershocks in association with main shocks.

A question of extreme importance is if precursors of a main shock can be detected through monitoring dynamical evolution of earthquake network. Clearly, further developments are needed in science of complex networks to address to this question.

Addendum

Some authors (e.g., Sornette and Werner) raised questions about the applicability of complex earthquake network using online accessible earthquake catalog data and on the validity of our results. A preprint of an article by the authors discussing these questions can be found in the e-print available at <http://arxiv.org/abs/0708.2203>.

Bibliography

1. Abe S, Okamoto Y (eds) (2001) Nonextensive statistical mechanics and its applications. Springer, Heidelberg
2. Abe S, Suzuki N (2003) Aging and scaling of earthquake aftershocks. *Physica A* 332:533–538
3. Abe S, Suzuki N (2003) Law for the distance between successive earthquakes. *J Geophys Res* 108(B2):2113 ESE 19-1-4
4. Abe S, Suzuki N (2004) Scale-free network of earthquakes. *Europhys Lett* 65:581–586
5. Abe S, Suzuki N (2004) Small-world structure of earthquake network. *Physica A* 337:357–362
6. Abe S, Suzuki N (2005) Scale-free statistics of time interval between successive earthquakes. *Physica A* 350:588–596
7. Abe S, Suzuki N (2005) Scale-invariant statistics of period in directed earthquake network. *Eur Phys J B* 44:115–117
8. Abe S, Suzuki N (2006) Complex earthquake networks: Hierarchical organization and assortative mixing. *Phys Rev E* 74:026113-1-5
9. Abe S, Suzuki N (2006) Complex-network description of seismicity. *Nonlin Processes Geophys* 13:145–150
10. Abe S, Suzuki N (2007) Dynamical evolution of clustering in complex network of earthquakes. *Eur Phys J B* 59:93–97
11. Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
12. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
13. Debenedetti PG, Stillinger FH (2001) Supercooled liquids and the glass transition. *Nature* 410:259–267

14. Dorogovtsev SN, Mendes JFF (2003) Evolution of networks: from biological nets to the Internet and WWW. Oxford University Press, Oxford
15. Gell-Mann M, Tsallis C (eds) (2004) Nonextensive Entropy: Interdisciplinary Applications. Oxford University Press, Oxford
16. Gutenberg B, Richter CF (1944) Frequency of earthquakes in California. *Bull Seismol Soc Am* 34:185–188
17. Newman MEJ (2002) Assortative mixing in networks. *Phys Rev Lett* 89:208701-1-4
18. Okubo PG, Aki K (1987) Fractal geometry in the San Andreas fault system. *J Geophys Res* 92(B1):345–355
19. Omori F (1894) On the after-shocks of earthquakes. *J Coll Sci Imper Univ Tokyo* 7:111–200
20. Pastor-Satorras R, Vespignani A (2004) Evolution and structure of the Internet: a statistical physics approach. Cambridge University Press, Cambridge
21. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the Internet. *Phys Rev Lett* 87:258701-1-4
22. Ravasz E, Barabási A-L (2003) Hierarchical organization in complex networks. *Phys Rev E* 67:026112-1-7
23. Steeples DW, Steeples DD (1996) Far-field aftershocks of the 1906 earthquake. *Bull Seismol Soc Am* 86:921–924
24. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. *J Stat Phys* 52:479–487
25. Vázquez A, Boguña M, Moreno Y, Pastor-Satorras R, Vespignani A (2003) Topology and correlations in structured scale-free networks. *Phys Rev E* 67:046111-1-10
26. Vázquez A, Pastor-Satorras R, Vespignani A (2002) Large-scale topological and dynamical properties of the Internet. *Phys Rev E* 65:066130-1-12
27. Watts DJ, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393:440–442
28. Wilson RJ (1996) Introduction to graph theory, 4th edn. Prentice Hall, Harlow

Earthquake Nucleation Process

YOSHIHISA IIO
Disaster Prevention Research Institute,
Kyoto University, Kyoto, Japan

Article Outline

Glossary
Definition of the Subject
Introduction
Contribution from the Development
of Earthquake Early-Warning Systems
Observations of Initial Rupture Processes
Discussion
Future Directions
Acknowledgments
Bibliography

Glossary

Nucleation process The process in which rupture velocity accelerates from quasi-static to dynamic. The dynamic rupture velocity almost equals the shear wave velocity.

Nucleation zone The portion of the fault where rupture velocity accelerates from quasi-static to dynamic.

Initial rupture process The rupture process that precedes the largest slip. This term is used when the acceleration of rupture velocity is not clear. This is a wider concept that includes the earthquake nucleation process. The area where the initial rupture process occurs is called the initial rupture fault.

Slip velocity The dislocation velocity at a point on the fault. The rupture velocity is the velocity at which the rupture front is expanding.

Preslip model An earthquake source model having a detectable size nucleation zone.

Cascade model An earthquake source model in which smaller sub-events successively trigger larger sub-events. A sub-event is the same as a small earthquake if it does not trigger a successive sub-event.

Stress drop (static stress drop) The amount of shear stress change at a point on the fault before and after an earthquake. It is proportional to the strain released on the fault.

Dynamic stress drop The difference between the initial shear stress and the minimum frictional stress at a point on the fault during fault slip.

Fault strength The shear stress level necessary to initiate slip at a point on the fault.

M Magnitude. Earthquake size computed basically from waveform amplitudes and focal distances.

Seismic moment The most reliable measure of earthquake size which is determined from the products of the rigidity near the fault, the amount of slip, and the area of the fault surface.

M_w Moment magnitude. Earthquake magnitude derived from the seismic moment.

Definition of the Subject

Earthquake prediction in the long, intermediate, and short terms is essential for the reduction of earthquake disasters. However, it is not practical at present, in particular, for the intermediate and short time scales of a few days to years. This is mainly because we do not know exactly how and why earthquakes begin and grow larger or stop. Theoretical and laboratory studies have confirmed that earthquakes do not begin abruptly with dynamic rupture propagation. They show that a quasi-static rupture

growth precedes dynamic rupture. Thus, if we can detect the quasi-static rupture growth, we could forecast the following dynamic rupture. A key issue is how natural earthquakes initiate. To solve this issue, a first approach would be to investigate the very beginning parts of observed waveforms of earthquakes, since they can reflect the earthquake nucleation process from a quasi-static to a dynamic rupture. This paper reviews the studies analyzing the beginning parts of observed waveforms, and shows what we presently understand about the earthquake nucleation process.

Introduction

Earthquakes initiate at a small portion on a fault. Then, their rupture fronts expand outward until they stop. Some large earthquakes have a rupture extent greater than 1000 km (e.g., the 2004 Sumatra Earthquake), while fault lengths of very small microearthquakes ($M = 0$) are estimated to be a few meters [13]. Fault lengths of earthquakes range at least over 6 orders of magnitude. Surprisingly, the concept that earthquakes are self-similar is widely accepted in spite of the difference in fault length (e.g., [40]). One example of the similarity is that the average fault slip is proportional to the fault length. In other words, the static stress drop is constant independent of earthquake size.

The similarity law raises fundamental questions: Why do large earthquakes grow larger? What is the difference between large and small earthquakes? An end member model represents earthquakes as ruptures that grow randomly, and terminate in an earlier stage for smaller earthquakes while continuing longer for larger earthquakes. This type of model has been proposed mainly to explain the frequency-size distribution of earthquakes (e.g., [8]) and implies that it is impossible to forecast the final size of earthquakes at the time of the rupture initiations. However, another end member model predicts that larger earthquakes have a larger “seed” than smaller earthquakes and that large and small earthquakes are different even at their beginnings (e.g., [66]).

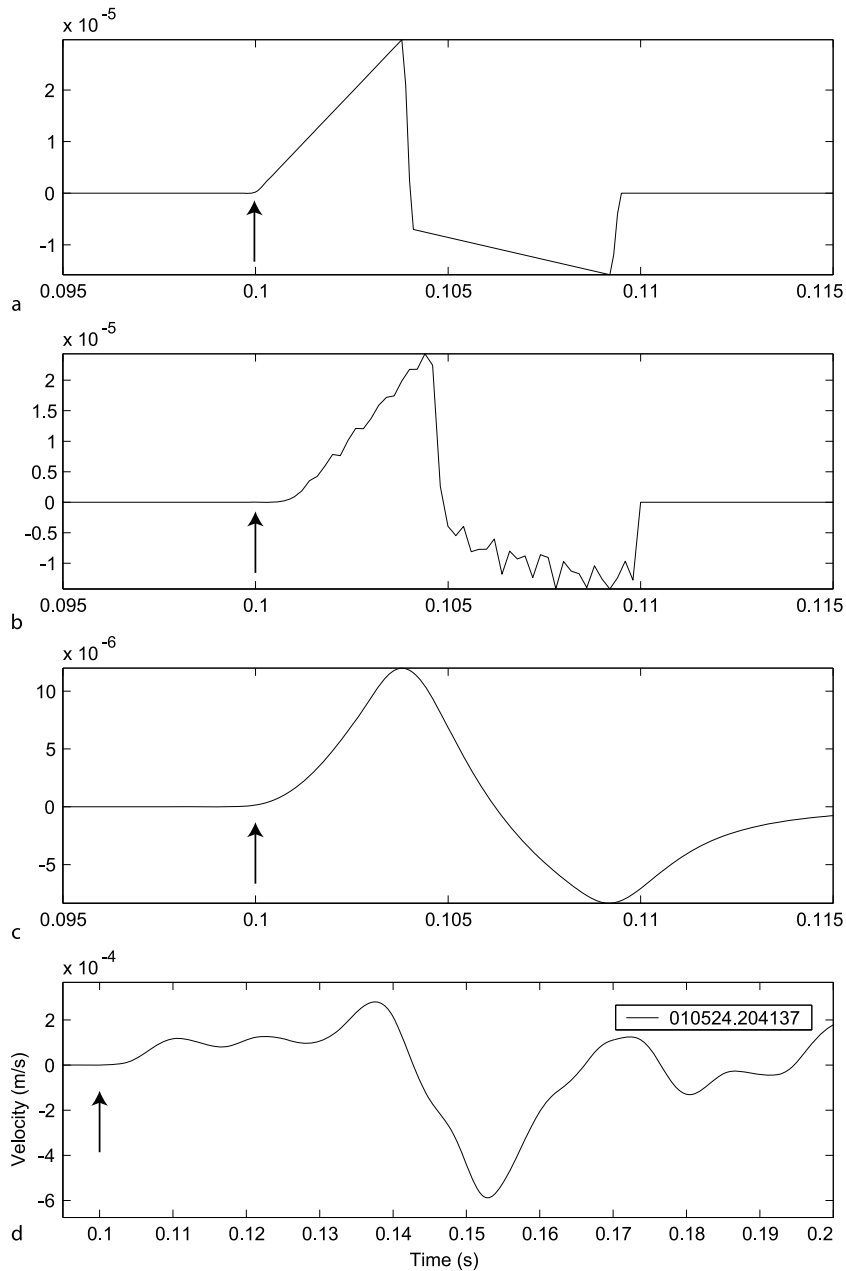
The key issue is how earthquakes initiate and grow larger.

All the theoretical models of earthquake sources predict that earthquake shear failures begin with a quasi-static rupture growth on a small portion on the fault (e.g., [5]). The shear failures become unstable and the rupture growth begins to accelerate after the energy released by the rupture growth equals to or becomes larger than the work necessary for producing new fracture surfaces. Finally, the rupture velocity, the velocity at which

the rupture front is expanding, reaches a constant value comparable to the shear wave velocity. We call the process from the rupture initiation to rupture growth acceleration the earthquake nucleation process, and the portion of the fault where the rupture growth accelerates the nucleation zone. The important point is that the rupture velocity (and slip velocity) accelerates during the nucleation process, since it is a transient process from quasi-static to dynamic. One of the major differences between various models is whether the size of the nucleation zone is much smaller than the final fault length and is not detected by observations, and whether the size of the nucleation zone is different between large and small earthquakes.

If the size of the nucleation zone is extremely small and can be approximated as a point, the waveforms radiated from the earthquake are shown in Fig. 1a. Figure 1a displays the initial rise of the P-wave velocity pulse at a far-field (distant relative to the fault length) station, when the waveform propagates in a purely elastic homogenous medium with no attenuation and scattering under the following assumption: the rupture initiates at a point, the circular rupture front expands at a constant velocity, the stress on the fault abruptly drops from an initial level to a final level, and the stress drop is constant over the fault, as modeled by Sato and Hirasawa [60]. In this case, the initial portion of the P-wave velocity pulse is characterized by a linear rise (and the initial rise of the displacement pulse is quadratic with time, due to increasing circular fault area with time). The tangent of the linear rise is proportional to the stress drop and rupture velocity.

On the other hand, if the size of the nucleation zone is relatively large, the velocity pulse can be approximated as shown in Fig. 1b. It is seen that the linear rise is delayed and the slope of the initial portion gradually increases. Theoretical models explain such a slow rise as follows. Shibazaki and Matsu'ura [64] demonstrated that the slow rise could be explained by gradually accelerating rupture and slip velocities in a relatively large nucleation zone. The size of the nucleation zone is controlled by the critical slip distance, D_c , that is the amount of slip necessary to drop the peak shear strength down to the dynamic frictional level (e.g., [19]). This frictional behavior is called slip-weakening, which is theoretically predicted (e.g., [15]) and is observed in laboratory experiments (e.g., [52,56]). According to their theory, since large slips are necessary to decrease the friction on faults with a large D_c , the rupture and slip velocities are not accelerated in the beginning on the faults. Sato and Kanamori [61] modeled the slow rise by Griffith's fracture criterion based on the energy balance indicated above, as the rupture velocity grad-



Earthquake Nucleation Process, Figure 1

Examples of P-wave velocity pulses. Arrows indicate onsets of P-waves. **a** Source pulse from the circular fault model [60]. Stress drop: 1 MPa, fault radius: 17 m, rupture velocity: 0.8 V_s , take-off angle: 61 degrees. **b** Source pulse from a circular fault model with a rupture velocity accelerating with time for first 20% of the total duration time. The other parameters are the same as **a**. **c** Q convolved velocity pulse at a distance of 3.58 km from the source. Q is set as 300. The other parameters are the same as **a**. **d** Velocity waveform of a complicated shape observed in the Western Nagano Prefecture region

ually increases under the assumptions of a large pre-existing fault and small trigger factor (instantaneous small stress increment) on the fault. The pre-existing fault is often called the initial crack. In this model, large initial

cracks result from large fracture energies on the fault and the rupture and slip velocities are not accelerated right after the rupture onset owing to the large fracture energies. In this paper, the model having a detectable size of the nu-

cleation zone is called the preslip model, following Beroza and Ellsworth [9].

It is important to examine whether initial rises of observed velocity pulses are linear or gradually increasing, since the above two models predict different initial rises. Although the examination seems to be very easy at first glance, it is actually very difficult. The path effect is an inevitable obstacle to the examination, in particular for small earthquakes. The effects of anelastic attenuation and/or scattering contaminate observed waveforms and can reproduce a slow rise of observed waveforms even though the waveforms show a linear rise at the source. Figure 1c shows the initial rise of velocity pulses calculated by considering anelastic attenuation by a convolution of a Q operator [7]. It is found that the linear rise is delayed as shown in Fig. 1b.

Another serious obstacle is the complexity of observed waveforms. They are not often as simple and smooth as those shown in Figs. 1b and c, but complicated as shown in Fig. 1d. If several small patches in a relatively large nucleation zone break during the nucleation process [9,65] as shown in Fig. 2a, it is difficult to detect an acceleration of the rupture velocity from the complicated waveforms that include several small phases radiated from the breaks of the small patches. In other words, it is difficult to infer the nucleation process from the complicated waveforms. On the other hand, the model with small nucleation zones claims that the complicated waveforms are radiated only from successive breaks of small sub-events and a delayed break of the largest sub-event, as shown in Fig. 2b. This concept is known as the cascade model (e.g., [1,9,23,78]). In this model, a former sub-event triggers successive larger sub-events and the final event is the largest among detectable sub-events. Although it is not always clear how larger sub-events are delayed from former smaller sub-

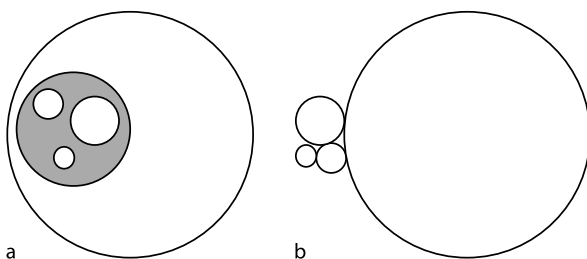
events, the growth of the rupture can raise the possibility that larger sub-events are generated. For small earthquakes also, it is possible that very small sub-events trigger successive sub-events. In this case, their waveforms produced by the source are complicated, but the observed waveforms are likely to show a smooth initial rise as shown in Figs. 2b and c, due to the path effects. Thus, in this paper, both smooth initial rises of small earthquakes and complicated initial portions of large earthquakes are discussed in the same manner.

By the way, if the size of the nucleation zone is relatively large, it could be possible to detect a quasi-static rupture growth at the very beginning of the earthquake nucleation process by near-field broadband instruments, such as strain meters. However, it has not, to present, succeeded except for a few very large earthquakes, such as the 1964 Chile earthquake (e.g., [41]). This fact suggests the possibility that the size of the nucleation zone is too small to be recorded by strain meters or the duration of the quasi-static rupture process is much longer than a practical frequency range of strain meters. Furthermore, the observations for very large earthquakes were not explained by the nucleation process that occurs near the hypocenter but by slow slips on the downward extension of the seismogenic faults (e.g., [33,41]). Thus, we will examine data obtained by seismometers.

It is very important to carefully analyze observed waveforms recorded at a short focal distance by a wide-dynamic range and frequency range. This paper reviews various studies related to the earthquake nucleation process, indicates the problems about these studies, and summarizes the current observations. First, in Sect. “Introduction”, contributions from earthquake early-warning systems are reviewed, since they are directly influenced by the problems cited above. In Sect. “Contribution from the Development of Earthquake Early-Warning Systems”, important studies about the initial rupture process are reviewed basically in the order in which they were published. In Sect. “Observations of Initial Rupture Processes”, a probable model for the process will be proposed based on the reviews in the former sections.

Contribution from the Development of Earthquake Early-Warning Systems

To solve the problem of how earthquakes initiate and grow larger, the most straightforward approach is a comparison between initial rises of observed waveforms of large and small earthquakes. Important studies were done for the development of earthquake early-warning systems which hope to forecast, as early as possible, the final size of earth-



Earthquake Nucleation Process, Figure 2

Schematic source models in which small amplitude phases are generated by breaking of small fault patches within the nucleation zone (shaded portion) a, and as successive small subevents b. Large and small circles are the mainshock fault and small subevent fault patches, respectively

quakes from observed waveforms, from the very beginning parts of seismograms. Olson and Allen [57] applied the method by Allen and Kanamori [3] to a new dataset including many large earthquakes ($M_w > 6$) and claimed that the final size of earthquakes ($3 < M_w < 8$) can be estimated from waveforms of the first several seconds. The method is based on the scaling relationship between the predominant period of waveforms and the final size of earthquakes, which was first derived by Aki [2]. To implement the scaling relationship in earthquake early-warning systems, Allen and Kanamori [3] adopted an algorithm to estimate the period from a short waveform, computing regressively the ratio of velocity and displacement amplitudes (modified after Nakamura [49]).

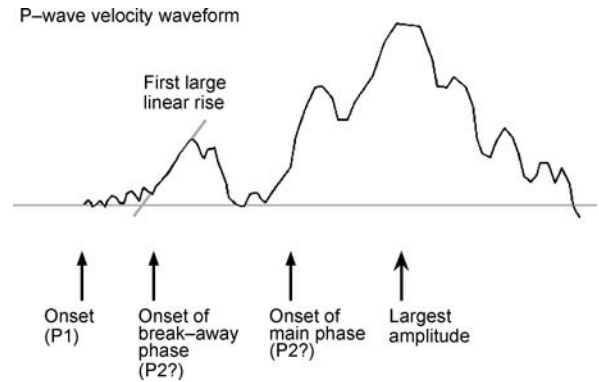
For smaller earthquakes, it is quite natural to be able to estimate the final size from the first several seconds of the waveforms, since the waveforms cover the entire source duration. The problem is estimating large earthquakes ($M_w > 5.5$). Rydelek and Horiuchi [58] evaluated the statistical significance of the results by Olson and Allen [57] and reported that the results are not clear. Rydelek and Horiuchi [58] analyzed waveforms observed by Hi-net, a nation-wide high gain seismometer network operated by NIED in Japan, and found no trend between the period and earthquake size for larger earthquakes. On the other hand, Wu et al. [77] analyzed many waveforms recorded at the southern California Seismic Network stations and showed that the period increases with magnitude ($4 < M < 7.5$). For earthquake early-warning system applications, where a quick response is essential, a precise waveform analysis is not required. A more careful analysis is needed to answer the question raised by Rydelek and Horiuchi [58].

Observations of Initial Rupture Processes

Measurements of Initial Parts of Observed Seismograms

How are Initial Portions of the Waveforms Measured? Observed waveforms of large to small earthquakes are measured in various manners in the studies reviewed in the following sections. Sometimes special names are given to a characteristic phase of the waveforms. First, the problems concerning the measurement of observed waveforms will be discussed.

Many studies try to find which phase is radiated from the source process with a constant rupture velocity and slip velocity, since the nucleation process is basically an accelerating rupture process. Since the nucleation process is a beginning rupture process, it is likely that the rupture front just after the nucleation process expands nearly cir-



Earthquake Nucleation Process, Figure 3

Schematic illustration of large and small earthquakes to show the first linear rise and the main phase

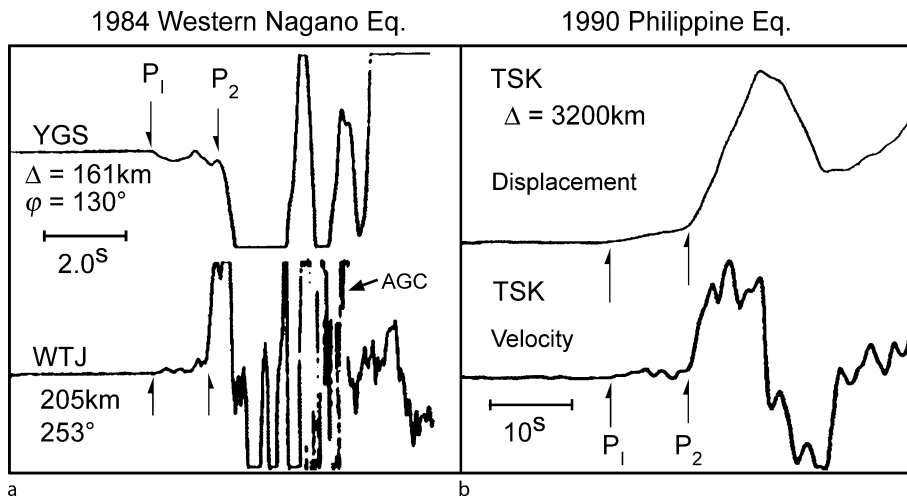
cularly. Consequently, the phase following the nucleation phase is thought to show a linear rise in velocity waveforms. For this reason, the first linear rise in velocity waveforms is extensively investigated in various studies.

Various studies also measure a pulse of the largest amplitude or a group of pulses including the largest amplitude, which is called the main phase in this paper. Amplitudes of waveforms are thought to show the maximum value after they reach terminal velocity, which is the maximum rupture velocity determined by the shear wave velocity. However, the largest amplitude does not necessarily occur just after rupture velocities reach the terminal velocity; all the waveforms preceding the maximum phase do not necessarily reflect the nucleation phase.

The problem is determining which phase is radiated just after rupture velocities reach terminal velocity. Furthermore, since slip velocities can cover a wide range of magnitude, it is important to investigate whether slip velocities are similar to the average values of earthquakes. In the following, when it is necessary to explicitly indicate the phase generated by faulting with a nearly constant rupture velocity and dynamic stress drop that are representative of average values for earthquakes, the phase is called the ordinary phase.

As schematically shown in Fig. 3, for large earthquakes, the first linear rise is not always a part of the main phase, which has the largest amplitude. For small earthquakes, the main phase often shows a linear rise. However, it is possible that the main phase is not the ordinary phase, as we discuss in a later section.

Large Earthquakes One of the first studies that indicated the importance of small amplitude waveforms preceding the main pulse is Furumoto and Nakanishi [24],



Earthquake Nucleation Process, Figure 4

Initial parts of observed waveforms of large earthquakes [73]. P_1 and P_2 indicate the onsets of P-waves and the large amplitude phases, respectively. Velocity waveforms are shown in a and the lower panel of b. Figure 7 in [73], copyright 1992 by Elsevier Science Publishers B.V.

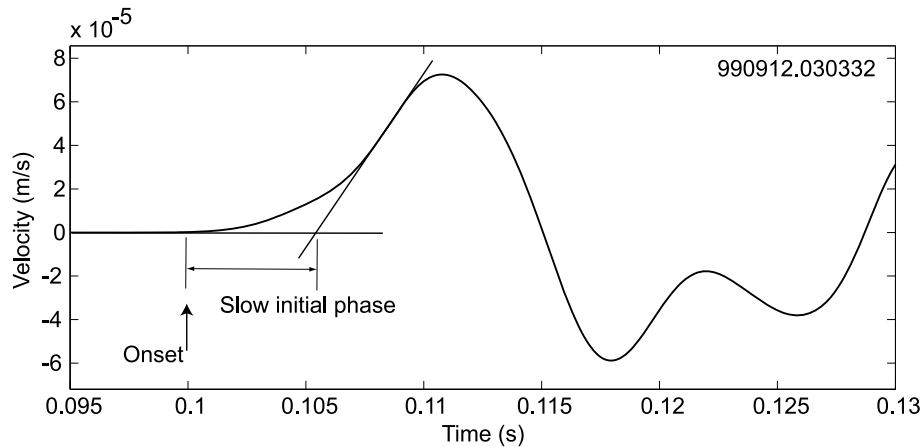
which analyzed long-period waveforms from a world-wide network; this was followed by many similar studies (e.g., [4,14,68]). Umeda [72,73] analyzed broadband seismograms at local, regional and teleseismic distances and found that small amplitude waveforms preceded large amplitude waveforms, as shown in Fig. 3. Umeda [72] called the onsets of the P-waves and large amplitude waveforms P_1 and P_2 , respectively. He originally thought that the large amplitude waveforms were generated by somewhat special processes that characterized large earthquakes, e.g., breaks of many small faults triggered by both static and dynamic stress concentrations due to the preceding fault motion and/or a dynamic connection of separated faults [74], while the smaller waveforms were radiated from ordinary smooth rupture propagation. His important result is the relationship between the duration of P_2-P_1 and magnitude ($3 < M < 7$), suggesting earthquake sizes are scaled with P_2-P_1 , the duration of small amplitude waveforms.

The determinations of P_2 seem to be reasonable from his figures. However, the criterion of P_2 determination is not necessarily clear and quantitative compared to the studies cited in the following sections. In many cases it seems that P_2 is the onset of the main phase that includes the largest amplitude, in particular for smaller earthquakes, but this point is not clearly shown in his papers. Furthermore, it is seen that absolute velocity amplitudes after P_1 do not show an accelerating increase with time, although both the preslip model and the cascade

model basically indicate that velocity amplitudes should have a growth that is faster than linear with time.

One of the important problems raised by Umeda [72, 73] is identification of the ordinary phase, the main phase or preceding smaller amplitude waveforms. The ordinary phase is the phase generated by faulting with a nearly constant rupture velocity and a dynamic stress drop, which are representative of average values for earthquakes. Furumoto and Nakanishi [24] regarded the main phase as the ordinary phase with a special rupture process occurring before the main phase, while in Umeda [72,73] the opposite is the case.

Small Earthquakes Iio [30,31] analyzed waveforms of microearthquakes recorded at very short focal distances (down to a few hundreds of meters) using instruments with a wide-frequency response and found that the initial rises of velocity pulses did not show a linear increase but, rather, a convex downward (or upward) shape as approximated by t^n (t is the time measured from the onset and $2 < n < 4$). He termed the initial rise "the slow initial phase", and measured the duration of the slow initial phase relative to the main phase that shows a linear increase, as shown in Fig. 5. He found that the duration of the slow initial phase is proportional to the earthquake size. Although the slow initial phase of microearthquakes could be the effect of anelastic attenuation as shown in Sect. "Introduction", Iio et al. [32] analyzed velocity pulses recorded at a 10 kHz sampling frequency at numerous stations at



Earthquake Nucleation Process, Figure 5

Slow initial phase of the velocity pulse observed in the Western Nagano Prefecture region and the measurement of the duration of the slow initial phase relative to the main phase that shows a linear increase of velocity amplitudes. The vertical arrow indicates the onset of P-waves, determined by their amplitudes. The portion indicated by the horizontal arrow is defined as the slow initial phase. The inclined line is the tangent at the maximum slope. The same pulse is shown in Fig. 7 by the black line

short focal distances using an instrument with a wide dynamic range and frequency range, and concluded that the slow initial phase mainly reflects the source process for M2 events.

These observations are unique and important, but their interpretations and implications contain several fundamental problems and might cause some misunderstandings. First, it was intuitively regarded that the main phase showing the linear rise was the ordinary phase radiated from a circular fault with constant rupture and slip velocities of ordinary magnitudes. This was derived from the interpretation that the slow initial phase reflected the nucleation process in which the rupture and/or slip velocities gradually accelerated. However, the interpretation has not been thoroughly examined. As emphasized by Ellsworth and Beroza [21] and Beroza and Ellsworth [9], waveforms recorded by seismometers basically reflect slips on the fault that have accelerated above a certain level.

The second problem is the meaning of the scaling relationship indicating that larger earthquakes have a longer slow initial phase. Although the possibility was suggested from this relationship that larger earthquakes had a larger nucleation zone, this is true only if the main phase is the ordinary phase and the slow initial phase reflects the nucleation process. The relationship could be also misunderstood as larger earthquakes begin more slowly. The observational results mean only that the portion of velocity pulse approximated by t^n ($2 < n < 4$) is longer for larger earthquakes. Furthermore, Iio [30,31] mentioned nothing about the difference between the slopes right after the on-

sets of larger and smaller earthquakes. Thus, another possible explanation is that for larger earthquakes, the slope of the velocity pulse increases with time for a longer period and the maximum slope becomes larger, even though the initial slope is the same as that for smaller earthquakes. This possibility implies that the final size of earthquakes is not estimated only from the initial part of velocity pulses, as pointed by Mori and Kanamori [48], although larger earthquakes have a greater dynamic stress drop. Several studies have indicated that the initial rises of large earthquakes ($M > 6$) are similar to small events occurring in the vicinity of their hypocenters (e.g., [12]), while other studies have indicated that larger earthquakes display a larger initial rise (e.g., [35,50]). However, it is possible that the initial parts of these waveforms mainly reflect the initial rise radiated from a small patch within the initial rupture fault, as shown in Sect. "Introduction". In that case, the results did not negate the existence of an observable earthquake nucleation process for large events. Further, these results were obtained only for a few earthquake sequences and need to be systematically examined for a greater data set.

The third problem is the propagation effects from the source to the observation stations. Iio [30,31] thought that observed waveforms basically reflect the characteristics of the source process in which the rupture and slip velocities gradually increase during the slow initial phase, assuming that source pulses of small earthquakes are simple and consist of a single event. However, it is possible that anelastic attenuation modifies two connected linear trends

of small and large slopes from two subevents, creating a smoothed waveform of convex downward shape. This point is qualitatively examined in a later section. Furthermore, it might be possible that waveforms near the source are complex; several very short duration small pulses may precede the main phase, since short duration pulses can be smoothed by the path effect to produce a slow initial phase.

Estimate of Source Time Functions

Ellsworth and Beroza [21] and Beroza and Ellsworth [9] focused on initial parts of seismograms of a very wide range of earthquake sizes ($1 < M_w < 8$) and estimated source time functions. They paid attention to the first large linear rise of velocity pulses and deduced that the signal before the linear rise reflected the earthquake nucleation process, since a linear rise of velocity pulses is characteristic of the waveforms radiated from an ordinary circular fault model. They termed the linear rise “the breakaway phase” and the portion before “the seismic nucleation phase”. The term “seismic” is added, since waveforms detected by seismometers are radiated from dynamic slip on the fault. The criterion for the detection of the breakaway phase is not necessarily quantitative, but it seems from their figures that their classifications are reasonable. As shown above, they regarded the breakaway phase as the ordinary phase, since they roughly estimated that dynamic stress drops of the breakaway phases were several or several tens of MPa, that is comparable to average values of earthquakes. However, it is possible that the seismic nucleation phase is the ordinary phase, since the rupture velocity and slip velocity during the seismic nucleation phase is not quantitatively estimated.

Kilb and Gombert [43] analyzed initial portions of the waveforms of the Northridge earthquake, which were also used in Ellsworth and Beroza [21] as a typical example, and claimed that the initial portions are very similar to those from nearby small earthquakes. They inferred from these results that the cascade model is plausible. However, their results did not exclude the preslip model in which the rupture and slip velocities are accelerating in the nucleation zone, since they analyzed only the very beginning portions and not the following portions that might result from slow slips in the nucleation zone. Shibasaki et al. [67] discussed the possibility that large earthquakes begin by a breaking of a small patch in the nucleation zone. As discussed in the Introduction, the breaking of small patches can mask the nucleation process.

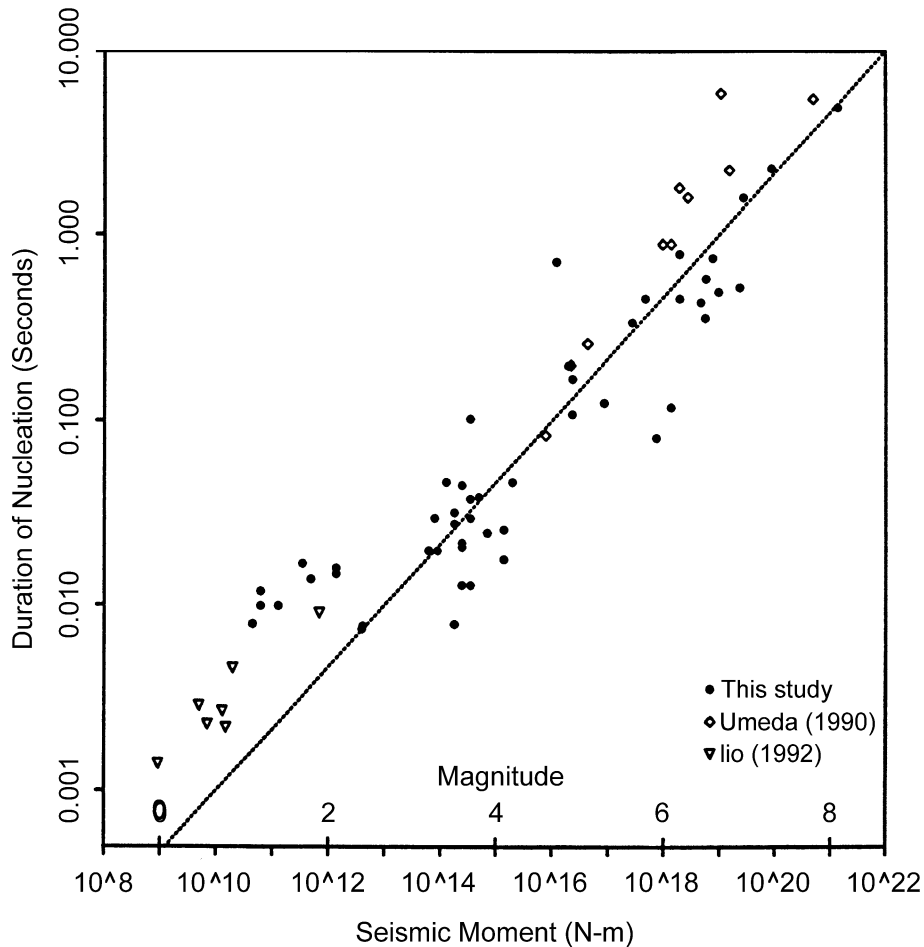
Ellsworth and Beroza [21] demonstrated the relationship between the duration of the seismic nucleation phase

and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73], as shown in Fig. 6, and concluded that the duration of the seismic nucleation phase scaled with the seismic moment. However, it is seen in Fig. 6 that the data of Iio [30,31] and Umeda [72,73] are shifted upward from the regression line. It seems that their data for small earthquakes also shifted upward from the regression line. The offsets of these shifts are about 1 order of magnitude. For larger earthquakes, this may be because Umeda [72,73] determined P2 as the onset of the main phase, while Ellsworth and Beroza [21] determined onsets of the first large linear rises. For smaller earthquakes, Iio [30,31] also detected the onsets of the main phases. They regarded the phases as the first large linear rises, but it is possible that the onsets of the first large linear rises are earlier than the main phases, as discussed in the following section.

Estimate of Source Processes

Small Earthquakes In order to solve the problems cited in the previous sections, several studies tried to estimate the source process that produces the initial phases. Hiramatsu et al. [26] analyzed rising parts of seismograms using the model of Sato and Kanamori [61], in which larger initial cracks generate a slower initial phase under a small triggering factor.

It was inferred from deep borehole (1800 m) seismograms by Hiramatsu et al. [26] that initial rises of velocity pulses of 5 events were explained by the ordinary circular fault model [60], namely the initial crack is too small to be detected, while the other 7 events needed a large initial crack. They first estimated source processes of the beginning of such small earthquakes; however, some basic problems remained. The first is that they modeled the first half cycle of velocity pulses. Since even portions of the first half cycle of the waveform can be affected by rupture arrest, it is not reasonable to fit the first half cycle of the waveform by using a model that does not include a reasonable rupture stopping process. Furthermore, a closer inspection of their results (Fig. 5 of [26]) reveals that some of the waveforms modeled with an initial crack do not display a smooth increase of the rising slope but appear to consist of two linear phases with different slopes. It is possible that these earthquakes result from a cascade rupture of a first and second sub-events with smaller and larger dynamic stress drops, respectively. For the data analyzed by Iio [30,31], waveforms with a predominantly longer slow initial phase appear to show a similar feature. Furthermore, it is not clear which is the ordinary phase, the first or second linear phase.



Earthquake Nucleation Process, Figure 6

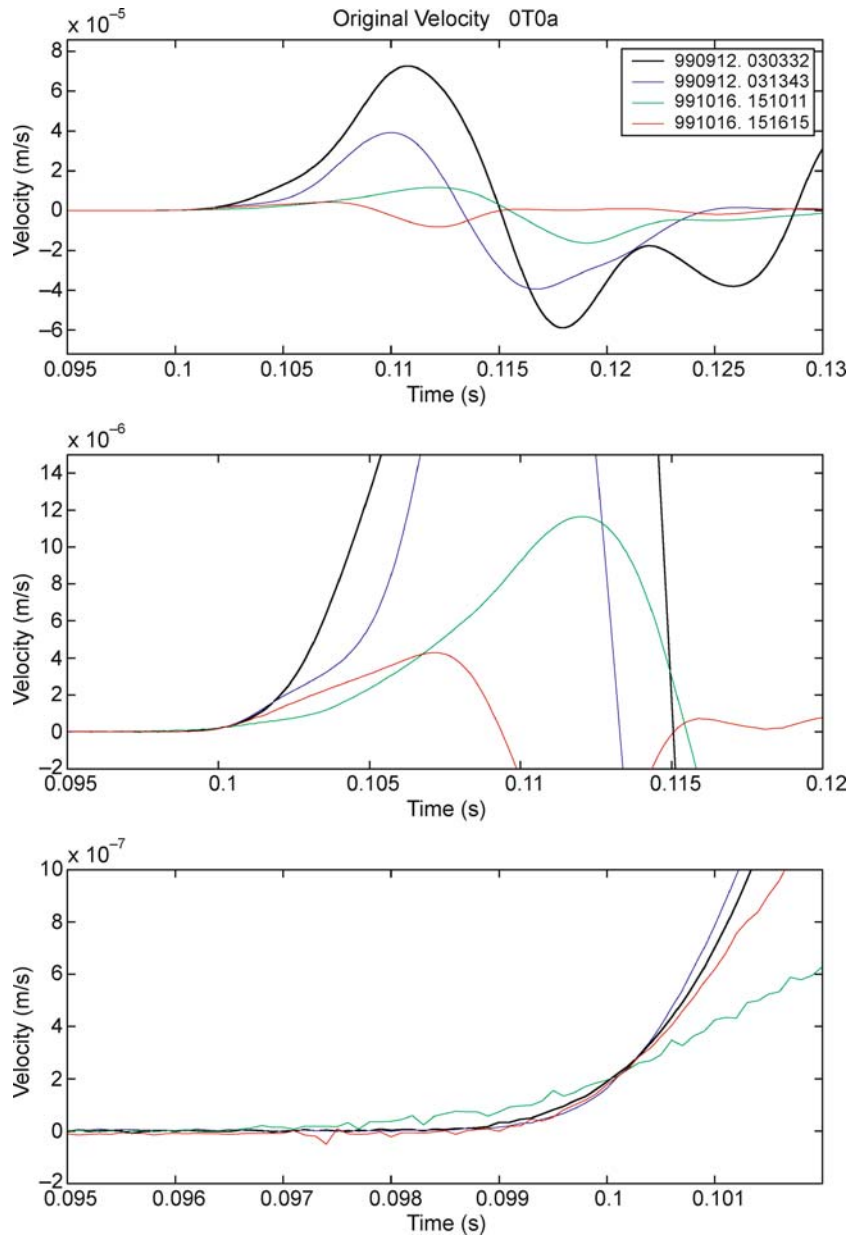
Relationship between the duration of the seismic nucleation phase and the seismic moment, together with the data of the slow initial phases by Iio [30,31] and P2-P1 by Umeda [72,73](Beroza and Ellsworth [9])

The problem of identifying the “ordinary phase”, the initial or main phase, is again pointed out. Recently, Iio et al. [34] analyzed the 10 kHz sampling data [32,36] and obtained results that suggest the initial rise is the ordinary phase.

Figure 7 shows velocity pulses of microearthquakes ($0.5 < M < 2.0$) for which relative locations are estimated within 100 m and fault plane solutions are similar [34]. These pulses were recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [37,76]. The focal distances are about 3.6 km. It is seen that the smallest event (shown in red) shows a linear rise except for the first 1 ms, while slopes of the other events increase with time. Thus, if the linear rise is explained by a circular fault model of the Sato and Hirasawa type with an ordinary rupture velocity and stress drop, the increasing slopes re-

flect increasing slip velocities accelerated beyond ordinary values.

The linear rise of the waveforms was modeled by various kinematic fault models. Figure 8 displays the comparison of synthesized and observed velocity pulses for three borehole stations at depths of 800, 150, and 100 m [37]. The fault plane was determined from focal mechanisms and hypocentral distributions around these events. Since all three waveforms were not explained simultaneously by a Sato and Hirasawa type of circular fault model, they used a fan fault of various angles. Further, they did not assume that the slip terminated simultaneously over the fault, but that it begins to stop at a point on the fault edge and the stopping phase propagates circularly at a constant velocity. The stress drop, rupture and stopping phase expanding velocities, fault radius, fan angle and fan fault geometry were



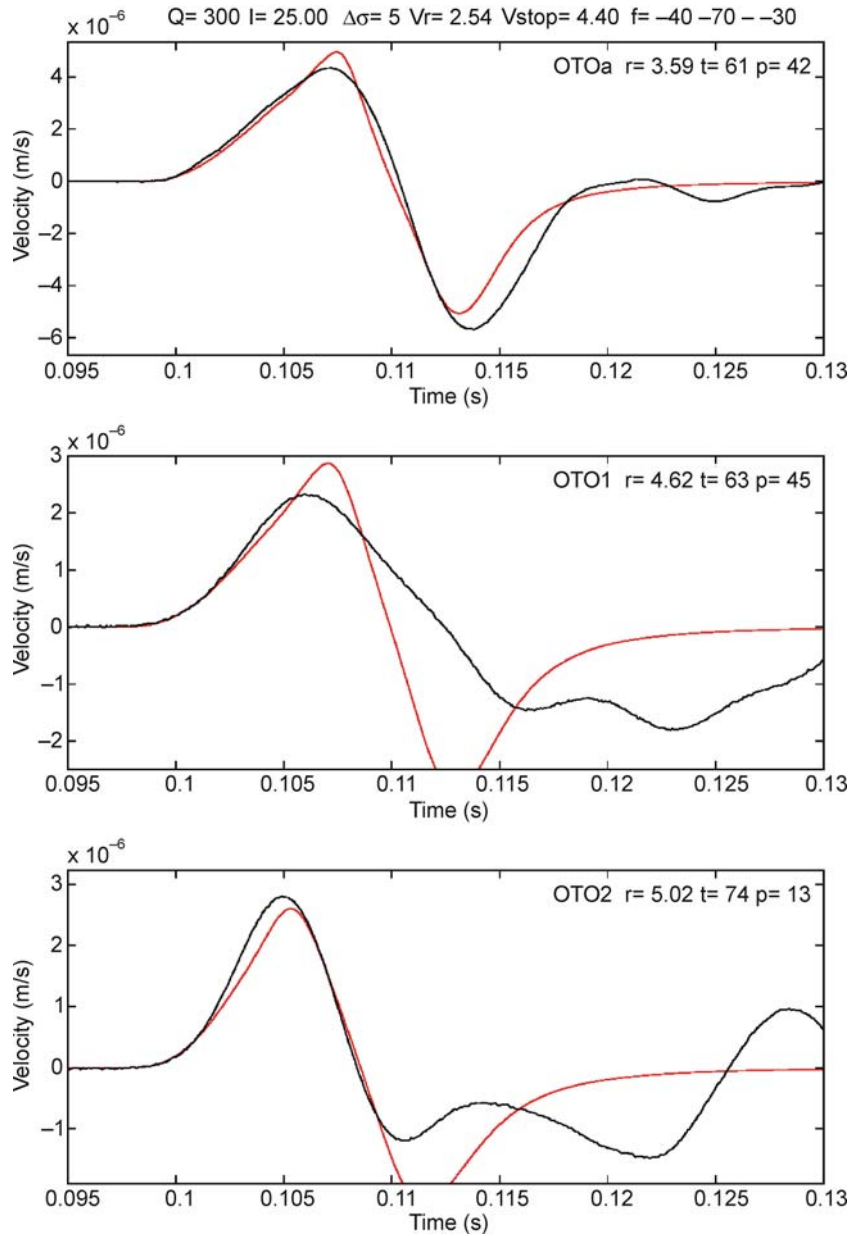
Earthquake Nucleation Process, Figure 7

Comparison of velocity pulses of microearthquakes ($0.5 < M < 2.0$) of which relative locations are estimated within 100 m and fault plane solutions are similar, recorded at a borehole station at a depth of 800 m in the Western Nagano prefecture region [34]. Focal distances are about 3.6 km

determined by a grid search technique. Q values are set as 300 by trial and error, to fit the very beginning initial rises with duration of about 1 ms.

It is seen that the observed waveforms in the first half cycle are well explained by the fan fault model, except for the middle trace, which might be contaminated by

a surface reflection. The rupture velocity was estimated as $0.8 V_s$ (the shear wave velocity), which is similar to that for large earthquakes. The stress drop was estimated as about 5 MPa by the formula for the circular fault of the equivalent fault area. Although they modeled only two events, they found that a few percent of events that oc-



Earthquake Nucleation Process, Figure 8

Comparison of observed and synthesized waveforms for an event that displays a linear rise, shown in Fig. 7 by the red line [34]. Waveforms observed at three borehole stations of depths of 800, 150, and 100 m are shown in the *top*, *middle* and *bottom* panels, respectively. The synthesized waveforms are calculated by a fan fault model in the homogenous half space using the following parameters: The stress drop is estimated as about 5 MPa. The rupture and stopping phase expanding velocities are $0.8 V_s$ and $0.8 V_p$ (the P-wave velocity), respectively. Fault radius: 25 m, fan angle: 40. A Q value is 300

occurred within about 5 km from the 800 m borehole station also showed such a linear rise. Furthermore, the other events generally show a steeper initial rise than the linear rise events. Consequently, it is likely that small earth-

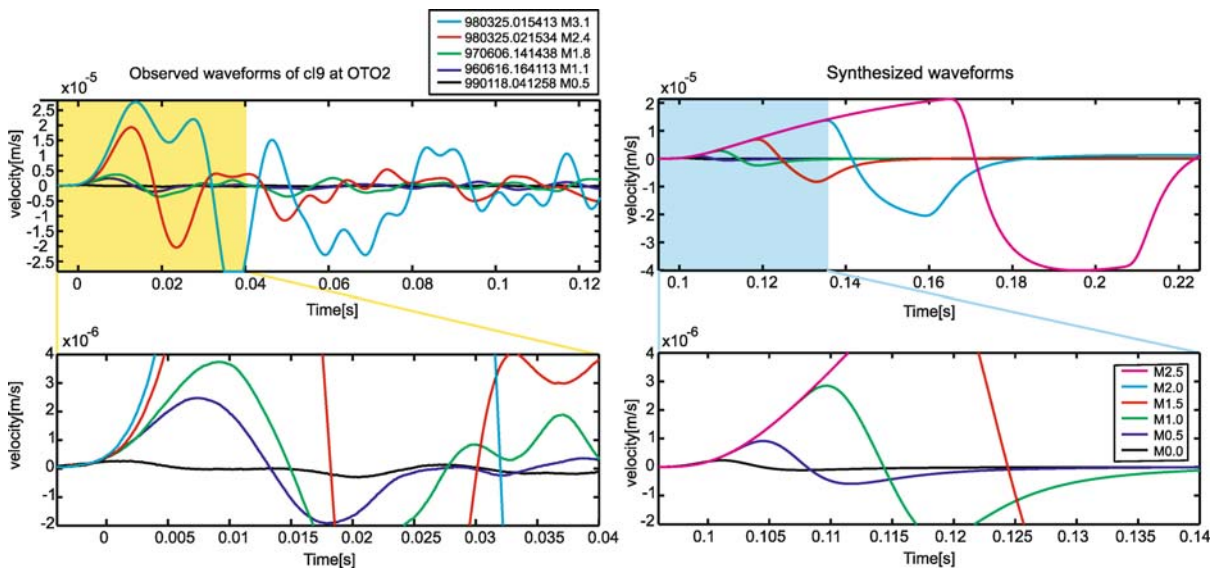
quakes occurring in the Western Nagano prefecture region have a rupture velocity equivalent to those of large earthquakes even in the initial rupture process. This suggests that the slow initial phase of small earthquakes does

not reflect the nucleation process that is characterized by accelerating rupture velocity, since the rupture velocity has already accelerated to a shear wave velocity in an early stage of rupture growth.

What do steeper main phases of larger earthquakes in the Western Nagano region reflect? In order to clarify this problem, Miura et al. [46] analyzed waveforms of earthquakes ($0.0 < M < 4.0$) occurring from 1996 to 2003 in the Western Nagano prefecture region. They selected M3 events ($3.0 < M < 4.0$) and identified 21 earthquake clusters that consist of M3 events and earthquakes occurring within 500 m of the hypocenters of M3 events. They investigated P-wave velocity pulses observed at three borehole stations for each cluster, in particular the difference in pulse shapes of large and small earthquakes. They found that waveforms of half of the clusters displayed complicated shapes, as shown in Fig. 1d, which are characterized by more inflection points than simple pulses [63]. The other half showed simple waveforms that enabled them to clarify the difference between large and small earthquakes. One example of waveforms is shown in Fig. 9. The observed P-wave velocity pulses of $0.5 < M < 3.1$ are displayed in different magnifications. It is seen that initial rises are similar for the first 1 ms but that slopes of larger events increase with time. The theoretical pulses from the circular fault model [60] are synthesized at the same focal

distances for a similar magnitude range in Fig. 9b, assuming a constant stress drop independent of earthquake size, and a Q value of 230, which was obtained by modeling the waveform of the smallest event. Responses of seismometers are also included. The synthesized pulse shapes of larger events do not display a linear rise but a slight convex upward shape. It is found that a distinct difference between theoretical and observed waveforms is seen in the maximum slope of larger events. This figure clearly demonstrates that larger earthquakes have larger dynamic stress drops than smaller events. Similar features are also seen in the other clusters that show simple waveforms. These results imply that the similarity law does not hold for a small range of magnitudes in the Western Nagano prefecture region, as indicated by Venkataraman et al. [58]. The similarity law predicts pulses as shown in the right hand side of the figure.

Large Earthquakes Sato and Mori [62] used the same method as Hiramatsu et al. [26] and analyzed the very beginnings of waveforms for a very wide range of magnitudes ($3 < M < 8$) recorded by high gain short-period seismometers at local distances. They showed that large initial cracks (a few tens of meters) are necessary to explain the initial rises and that initial crack lengths are almost constant, independent of the eventual earthquake size. Al-



Earthquake Nucleation Process, Figure 9

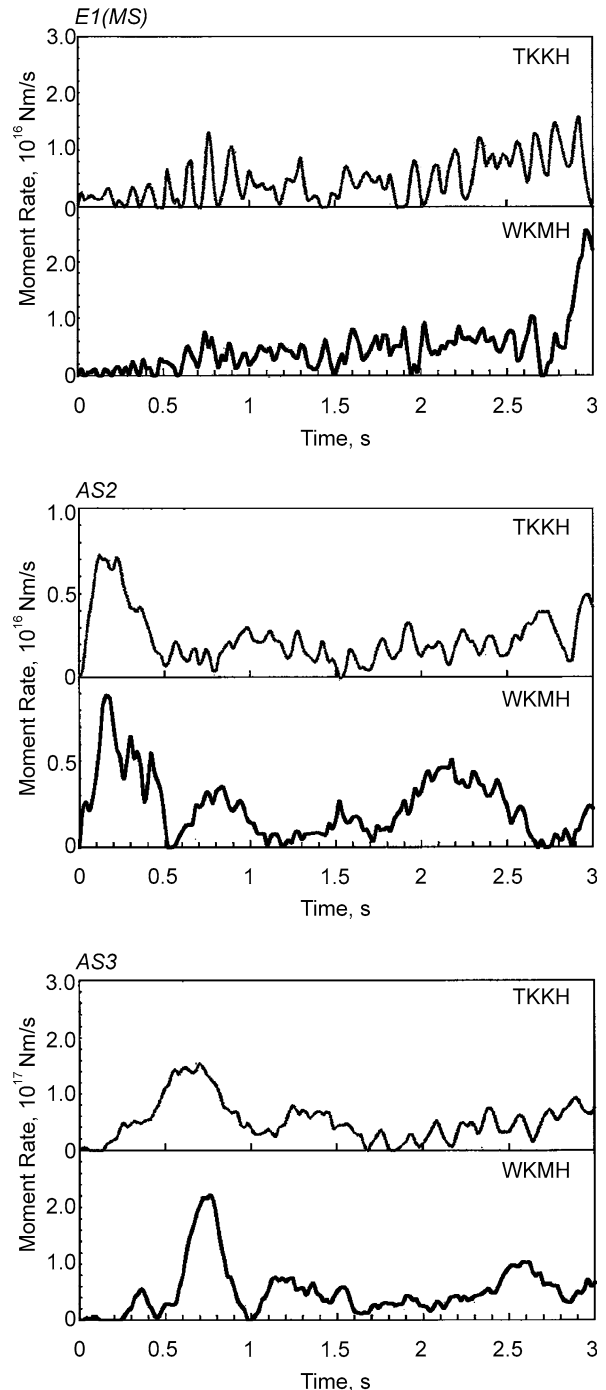
Comparison of observed (left) and synthesized (right) waveforms for events of an earthquake cluster that shows smooth waveform shapes [46]. The cluster consists of the M3 class events and earthquakes occurring within 500 m from their hypocenters. The synthesized waveforms are calculated by an ordinary circular fault model [60] at the same focal distance. A Q value of 230 was obtained by modeling the waveform of the smallest event. Responses of the instruments are also convolved

though they might have analyzed only the beginning of the first sub-event in the nucleation zone, it is important that they found that very beginning portions of large earthquakes show an accelerating rise over the first 0.1 s.

For large earthquakes, initial rupture processes can be estimated by near-source broadband seismograms. Shibazaki et al. [67] determined a P2 phase 0.5 s after the P-wave onset (P1) for the Kobe earthquake and performed a waveform inversion using the initial portion of seismograms. They estimated that the slip velocity slowly accelerated for 0.5 s. The average rupture velocity within 0.5 s after the onset of dynamic rupture was about 2 km/s, which is not very low. However, their results possibly detect the earthquake nucleation process, since it is difficult to estimate accelerating rupture velocity in the small nucleation zone of about 500 m, which was inferred from the waveform inversion. It is noted that their estimate of P2-P1 of 0.5 s is consistent with the scaling relationship of Ellsworth and Beroza [21], but is not consistent with that of Umeda [72,73]. The initial phase of the Kobe earthquake shows an accelerated increase of velocity amplitude with time [67], while a significant part of the data analyzed by the other studies do not show the accelerated increase. These results suggest the possibility that a major part of the initial rises analyzed in the other studies do not reflect the earthquake nucleation process.

Recently, longer durations of P2-P1 were observed for two intraplate earthquakes in Japan, the 2000 Western Tottori Prefecture earthquake (M7.3) and the 2004 West-off Fukuoka earthquake (M7.0). The initial phases of long durations are suitable for analyzing the rupture processes. For the Western Tottori Prefecture earthquake, Hirata [27] measured a duration of 2.5 s for P2-P1, determined the location of the P2 source and estimated the initial rupture process by a waveform inversion. They estimated that during the first 2.5 s, a small slip occurred within the limited area around the hypocenter and later, a large slip began near the location of P2 about 5 km southeast of the hypocenter. The average rupture velocity during the first 2.5 s was estimated to be 1.8 km/s from the waveform inversion of the initial phase. The stress drop in the limited portion was estimated to be small compared with the large slip area.

Since the duration of P2-P1 is long, the above features are also obtained in ordinary waveform inversions. Iwata and Sekiguchi [39] estimated the slip distribution using strong ground motion seismograms and geodetic data, and found a large slip area located about 5 km southeast of the hypocenter. They regarded the large slip area as an “asperity” on the fault. It has been well known that slip distributions of large earthquakes are heterogeneous



Earthquake Nucleation Process, Figure 10

Source time functions for the initial part of the 2004 West-Off Fukuoka earthquake (M7.0) The waveforms after the onsets are shown. The mainshock (top), a M4.5 aftershock (middle), and a M5.4 aftershock (bottom) estimated by an empirical green function method [75]. Zisin, 59:250, Figure 11, copyright 2007 by the Seismological Society of Japan

and large slips occur within a limited portion on the fault (e. g., [44]). Although the origin of large slip areas has not yet been clarified, one interpretation is that fault strength is higher in the areas of large slip than in the surrounding areas. This concept is called the asperity model. The observations about P1 and P2 shown above can be explained by the asperity model: The earthquake rupture begins at the weakest portion on the fault, then propagates to stronger portions where large slips occur.

For the 2004 West-off Fukuoka earthquake (M7.0), similar results were obtained. A P2 phase was detected by Yamaguchi et al. [75] and the duration of P2-P1 was long, estimated as 3.38 s. They determined the location of the P2 source as 3.44 km southeast of the hypocenter. The average rupture velocity during the 3.38 s period was computed to be about 1.02 km/s from the distance between the locations of P1 and P2. These results are consistent with those from ordinary waveform inversion studies (e. g., [6]). They also estimated the source time functions of the mainshock and a few large aftershocks by an empirical Green function method, as shown in Fig. 10. It was found that the source time function of the mainshock is not impulsive but has a gradual onset and long duration, while those of the aftershocks are impulsive. Only one aftershock shows a small initial phase, as shown in Fig. 10c. The seismic moment released before the P2 phase is comparable for those of M4 aftershocks. These results clearly show that the initial rise of the West-off Fukuoka Prefecture earthquake is different from those of the small aftershocks and is characterized by a small stress drop. It is suggested that large earthquakes do not accidentally grow larger.

For these two earthquakes with long initial phases, the average rupture velocities and stress drops on the initial rupture faults are small; however, the initial phases do not show an accelerating increase of velocity amplitude with time. These facts suggest the possibility that long initial phases do not reflect the nucleation process.

Discussion

Summary of the Observations

The reviews in the previous sections have revealed several important characteristics concerning the initial phases that are commonly seen in many studies. In this section, first, these characteristics are summarized, and then, an inferred initial rupture process will be discussed.

The almost linear initial rises were observed in the Western Nagano region and were well explained by a fan fault model with a constant rupture velocity comparable to the shear wave velocity. On the other hand, nearby larger earthquakes showed an initial rise in which the slope is of

the same order of magnitude as that of the linear rise, plus increasing slopes with times of a few to several ms after the onset [34]. Thus, it is thought that the rupture velocity of these larger earthquakes is also comparable to the shear wave velocity during the initial phase. Furthermore, observed waveforms do not necessarily display a gradual increase of rising slope but sometimes show a discrete change in the slope, suggesting that the initial rise of the velocity pulse at the source is not smooth. Consequently, successive sub-events with larger dynamic stress drops possibly produce larger earthquakes for these events occurring in the Western Nagano Prefecture. These inferences are consistent with the results obtained by Hiramatsu et al. [26] that about half of their data are explained by an ordinary circular fault model and longer initial phases are not always smooth. For small earthquakes ($M < 4$), the slow initial phase probably does not reflect the earthquake nucleation process. The nucleation size of small earthquakes is probably small and thus, the very beginning of observed waveforms should be analyzed very carefully considering path effects. This matter is beyond the scope of this paper and is left to future studies.

Larger earthquakes showed a variety of initial phases, as pointed out by Ellsworth and Beroza [21]. It is likely that the observed initial phase of the Kobe earthquakes reflects the earthquake nucleation process [67]. Also, that of the Northridge earthquake probably reflects the earthquake nucleation process, since the very weak initial phase shows an accelerated increase of velocity amplitudes with time [21], although the very beginning part of the initial phase is similar to the waveform of nearby small aftershocks [43]. Further, several data of Ellsworth and Beroza [21], in particular, those shifted downward from the regression line shown in Fig. 6 possibly reflect the earthquake nucleation process. However, it is likely that those of the other large earthquakes do not reflect the earthquake nucleation process, since they do not show an accelerated increase with time but rather are roughly flat. These facts are clear for the two intraplate earthquakes that have a relatively long initial phase (e. g., [27,75]). For these two earthquakes, the average rupture velocity during the initial phases are estimated as 1.02 to 1.8 km/s, slightly smaller than ordinary values, and the stress drops on these large initial rupture faults are estimated to be small (e. g., [27,75]). Ordinary waveform inversions for these earthquakes showed that the main phases were generated by the breaking of asperities (e. g., [6,28,39]).

As summarized above, a major part of the initial phases of large earthquakes do not show an accelerating increase of velocity amplitudes with time and it is thought that the rupture and slip velocities of these earthquakes are

not accelerating during this time period. Consequently, it is thought that the initial portions of the observed waveforms of these large earthquakes do not reflect the nucleation process that is characterized by the transition from a quasi-static to a dynamic state; instead they represent a part of the dynamic rupture process characterized by a smaller stress drop, before the breaking of large asperities. Small earthquakes possibly show similar characteristics with an initial rupture process that is characterized by an ordinary rupture velocity and much smaller dynamic stress drop than the main phase. Although the nucleation process is probably seen in the very beginning stage of the large earthquakes analyzed by Sato and Mori [62] and Shibazaki et al. [67], it is likely that the scaling relationships between the duration of the initial phases and the rupture size shown by Umeda [72,73], Iio [30,31], and a major part of the data Ellsworth and Beroza [21] do not reflect the nucleation process but a part of the dynamic rupture process before the breaks of relatively large asperities.

A Possible Model

As summarized in the previous section, since a major part of the observed data do not reflect the earthquake nucleation process, in particular the data from small earthquakes, we cannot examine the preslip model here.

The observed data do not match the cascade model, since the cascade model basically predicts self-similar fault breaks for successive events, not increasing dynamic stress drop events. A variation on the cascade model, the hierarchy fault model explains increasing slopes of initial rises by the abrupt increase of the moment rate function due to a longer fault edge at a higher hierarchy level [23]. However, as shown in Fig. 6, the difference between the slopes of larger and smaller earthquakes can be greater than one order of magnitude, so it is difficult to explain the observed data only by the change in rupture front geometry. It may be necessary to consider differences in dynamic fault parameters.

So the questions to be answered are, why do larger earthquakes have a longer slow initial rupture process? In the first place, why do earthquakes need an initial rupture process to break a stronger portion on the fault? Why don't earthquakes initiate as a breaking of an asperity without the initial phase?

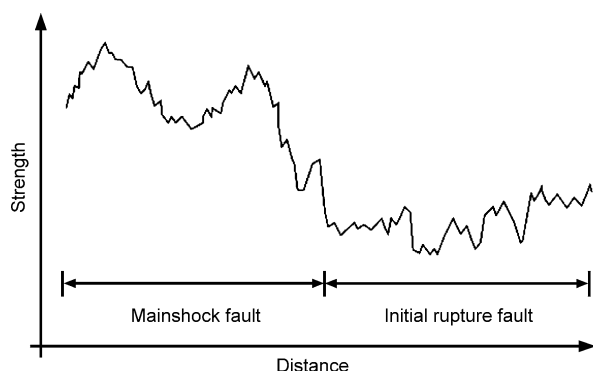
The key to solving these problems may lie in the studies of the two intraplate earthquakes with longer initial phases. The important observations about these earthquakes are the geometries of the initial rupture fault and mainshock fault. For the 2000 Tottori earthquake, it is

found that azimuths of the initial rupture fault and mainshock fault are estimated as N135°E and N150°E, respectively [51]. Since the direction of the maximum compressional stress is estimated as N90°–100°E [42], it is found that the initial rupture fault is favorably oriented, while the mainshock fault is unfavorably oriented. For the 2004 West-off Fukuoka prefecture earthquake, the azimuth of the initial rupture fault is different from that of the mainshock fault by 20 degrees (e. g., [38,69,71]). A more obvious example is obtained for the Landers earthquake, where the Emerson fault with the largest slip is unfavorably oriented, compared with the faults that had broken before (e. g., [25]). These observations indicate that a larger slip occurred on an unfavorably oriented fault. It is possible that the mainshock fault, which generates the main phase, generally has a higher strength than the initial rupture fault. This may be the reason why larger dynamic stress drops occurred on faults of successive larger events.

If this proposed strength profile on the initial and mainshock faults holds for all the earthquakes, the above questions can be re-written: Why is a larger weak initial rupture fault necessary to break a larger strong mainshock fault?

A similar problem has been discussed by Ohnaka [55]. He derived the relationship between the critical slip distance D_c and the asperity size from laboratory experiments and the physics of contacts on faults. Since D_c is related to the size of the nucleation zone (e. g., [52]) and it is inferred from the results of waveform inversions that the asperity size is proportional to the total fault length (e. g., [47]), his relationship might be regarded as the relationship between the lengths of initial rupture fault and the mainshock fault. However, the D_c used in his relationship is that of the asperity, not of the initial rupture fault. Further, it is inferred from the above discussion that a major part of observed initial phases probably do not reflect the nucleation process (namely, D_c).

It is important to clarify strength profiles along faults; however, we do not presently have enough information. In the following, we will assume the fault strength is controlled by the geometry of the fault surface, as estimated for the Tottori earthquake. More concretely, it is assumed that the angle between the tangent of the local fault surface and the direction of the uniform principal stresses controls the fault strength at each point. In this case, it is deduced that the fault strength shows a fractal-like distribution along the fault as shown in Fig. 11, since the geometry of fault surface is thought to be fractal [45]). The fault strength profile should include various wavelengths, but here we will consider the longest wavelength to investigate the interaction between the asperity and initial



Earthquake Nucleation Process, Figure 11

Schematic illustration of the strength profile along the fault deduced from a fractal geometry of fault surface. The rupture initiates at the weakest portion on the initial rupture fault and then propagates to a portion of higher strengths

rupture fault, since amplitudes of a longer wavelength are thought to be larger on fractal fault surfaces (e. g., [11]).

Under the above assumptions, the asperity and initial rupture fault are attributed to the portions of higher and lower strengths on the strength profile, respectively. Strength profiles of faults of larger earthquakes are thought to have a longer wavelength. Actually, it is empirically derived that the asperity size is proportional to the total fault length [47]. Consequently, it is inferred from these results that larger earthquakes have larger initial rupture faults. On initial rupture faults, the fault strength increases with distance from the hypocenter, since the hypocenter is the weakest point on the fault. In other words, the strength at a rupture front increases with rupture growth. Thus, it is thought that the initial rupture does not expand smoothly and the slip velocity does not significantly accelerate, but a large asperity can break after a breaking of the initial rupture fault. This is only a possible qualitative model for the initial rupture process and it suggests a possibility for explaining the observed data. It should be examined by extensive studies.

We do not presently have any clear answers to the subtitle question, 'Does the initiation of earthquake rupture knows about its termination?'. Even if the above model is correct, we also have to know the geometry of the fault surface and the time, slip and slip velocity dependent stress on the fault for large and small earthquakes, in order to simulate the rupture propagation and must understand the factors that control the earthquake size.

Future Directions

This paper reviewed studies analyzing the very beginning portions of observed waveforms of earthquakes and

showed what we have presently clarified about the earthquake nucleation process. To make further progress, the most straightforward path is to investigate the initial rupture process by precise waveform inversions for large earthquakes. In this case, new inversion methods, as proposed by Uchide and Ide [70], are useful. Furthermore, it is necessary to use broadband near field data, since it is possible that very slow slips occur on the initial rupture fault in association with the initial rupture process or for a very long time before the initial rupture process. For smaller earthquakes, high resolution data, as those in the Western Nagano Prefecture region, are necessary for investigating their rupture processes. These extensive studies about the initial rupture process can clarify the true feature of the earthquake nucleation process.

Acknowledgments

The project in the Western Nagano Prefecture is a co-operative study with Shigeki Horiuchi, Shiro Ohmi, Hisao Ito, Yasuto Kuwahara, Eiji Yamamoto, Kentaro Omura, Koichi Miura, Bun'ichiro Shibasaki, and Haruo Sato. We thank James Mori and Masumi Yamada for their critical reviews of the manuscript. This work is partly supported by JSPS.KAKENHI (19204043), Japan. We are grateful for two anonymous reviewers for their critical and thoughtful comments.

Bibliography

1. Abercrombie R, Mori J (1994) Local observations of the onset of a large earthquake, 28 June 1992. Landers, California. *Bull Seismol Soc Am* 84:725–734
2. Aki (1967) Scaling law of seismic spectrum. *J Geophys Res* 72:1217–1231
3. Allen RM, Kanamori H (2003) The potential for earthquake early warning in Southern California. *Science* 300(5620):786–789
4. Anderson JG, Bodin P, Brune JN, Pince J, Singh SK, Quaas R, Ohnate M (1986) Strong ground motion from the Michoacan, Mexico, earthquake. *Science* 233:1043–1049
5. Andrews DJ (1976) Rupture velocity of plane strain shear cracks. *J Geophys Res* 81:5679–5687
6. Asano K, Iwata T (2006) Source process and near-source ground motions of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planet Space* 58:93–98
7. Azimi SA, Kalinin AV, Kalinin VV, Pivovarov BL (1968) Impulse and transient characteristics of media with linear quadratic absorption laws, *Izv. Earth Phys* 1968(2):88–93
8. Bak P, Teng C (1989) Earthquakes as self-organized critical phenomenon. *J Geophys Res* 94:635–15–637–15
9. Beroza GC, Ellsworth WL (1996) Properties of the seismic nucleation phase. *Tectonophysics* 261:209–227
10. Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. *Bull Seismol Soc Am* 68:1117–1131

11. Brown SR, Scholz CH (1985) Broad bandwidth study of the topography of natural rock surfaces. *J Geophys Res* 90:12575–12582
12. Brune JN (1979) Implications of earthquake triggering and rupture propagation for earthquake prediction based on premonitory phenomena. *J Geophys Res* 84:2195–2198
13. Cheng X, Fenglin Niu, Silver PG, Horiuchi S, Takai K, Ito H, Iio Y Similar microearthquakes observed in western Nagano, Japan and implications to rupture mechanics. *J Geophys Res* 112:B04306. doi:10.1029/2006JB004416
14. Christensen DH, Ruff LJ (1986) Rupture process of the Chilean earthquake, 3 March 1985. *Geophys Res Lett* 13:721–724
15. Das S, Scholz CH (1982) Theory of time-dependent rupture in the Earth. *J Geophys Res* 86:6039–6051
16. Deichman N (1997) Far-field pulse shapes from circular sources with variable rupture velocities. *Bull Seismol Soc Am* 87:1288–1296
17. Dieterich JH (1978) Preseismic fault slip and earthquake prediction. *J Geophys Res* 83:3940–3948
18. Dieterich JH (1979) Modelling of rock friction: 1 Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
19. Dieterich JH (1986) A model for the nucleation of earthquake slip. In: *Earthquake source mechanics*. Geophysical Monograph. Maurice Ewing Series, vol 6. Am Geophys Union, Washington DC, pp 37,36–47
20. Dodge DA, Beroza GC, Ellsworth WL (1996) Detailed observations of California foreshock sequences: Implications for the earthquake initiation process. *J Geophys Res* 101(22):371–392
21. Ellsworth WL, Beroza GC (1995) Seismic evidence for an earthquake nucleation phase. *Science* 268:851–855
22. Ellsworth WL, Beroza GC (1998) Observation of the seismic nucleation phase in the 1995 Ridgecrest, California sequence. *Geophys Res Lett* 25:401–404
23. Fukao Y, Furumoto M (1985) Hierarchy in earthquake size distribution. *Phys Earth Planet Inter* 37:149–168
24. Furumoto M, Nakanishi I (1983) Source times and scaling relations of large earthquakes. *J Geophys Res* 88:2191–2198
25. Hardebeck JL, Hauksson E (2001) The crustal stress field in southern California and its implications for fault mechanics. *J Geophys Res* 106(21):859–882
26. Hiramatsu Y, Furumoto M, Nishigami K, Ohmi S (2002) Initial rupture process of microearthquakes recorded by high sampling borehole seismographs at the Nojima fault, central Japan. *Phys Earth Planet Inter* 132:269–279
27. Hirata M (2003) The initial rupture process of the 2000 Western Tottori Earthquake. Master Thesis, Kyoto University
28. Horikawa H (2006) Rupture process of the 2005 West Off Fukuoka Prefecture, Japan, earthquake. *Earth Planet Space* 58:87–92
29. Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent break in earthquake scaling due to path and site effects on deep borehole recordings. *J Geophys Res* 108(B5):2271; doi:10.1029/2001JB001617
30. Iio Y (1992) Slow initial phase of the P-wave velocity pulse generated by microearthquakes. *Geophys Res Lett* 19:477–480
31. Iio Y (1995) Observation of the slow initial phase generated by microearthquakes: Implications for earthquake nucleation and propagation. *J Geophys Res* 100:15333–15349
32. Iio Y, Ohmi S, Ikeda R, Yamamoto E, Ito H, Sato H, Kuwahara Y, Ohminato T, Shibasaki B, Ando M (1999) Slow initial phase generated by microearthquakes occurred in the Western Nagano prefecture, Japan -the source effect-. *Geophys Res Lett* 26(13):1969–1972
33. Iio Y, Kobayashi Y, Tada T (2002) Large earthquakes initiate by the acceleration of slips on the downward extensions of seismogenic faults. *Earth Planet. Sci Lett* 202:337–343
34. Iio Y, Horiuchi S, Ohmi S, Ito H, Kuwahara Y, Yamamoto E, Omura K, Miura K, Shibasaki B, Sato H (2006) Slow initial phase of microearthquakes. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, A48 (in Japanese)
35. Ishihara Y, Fukao Y, Yamada I, Aoki H (1992) Rising slope of moment rate functions: the 1989 earthquakes off east coast of Honshu. *Geophys Res Lett* 19:873–876
36. Ito S (2003) Study for the initial rupture process of microearthquakes in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. Ph.D. Thesis, Tohoku University (in Japanese)
37. Ito S, Ito H, Horiuchi S, Iio Y (2004) Local attenuation in western Nagano, central Japan, estimated from seismograms recorded in three boreholes. *Geophys Res Lett* 31:L20604; doi:10.1029/2004GL020745
38. Ito Y, Obara K, Takeda T, Shiomi K, Matsumoto T, Sekiguchi S, Hori S (2006) Initial-rupture fault, main-shock fault, and after-shock faults: Fault geometry and bends inferred from centroid moment tensor inversion of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planet Space* 58:69–74
39. Iwata T, Sekiguchi H (2002) Source process and near-source ground motion during the 2000 Tottori-ken Seibu earthquake (M_w 6.8). Reports on Assessments of Seismic local-site effects at plural test sites. MEXT, pp 231–241
40. Kanamori H, Anderson DL (1975) Theoretical bases for some empirical relations in seismology. *Bull Seism Soc Am* 65:1073–1095
41. Kanamori H (1996) Initiation process of earthquakes and its implications for seismic hazard reduction strategy. *Proc Natl Acad Sci* 93:3726–3731
42. Kawanishi R, Iio Y, Yukutake Y, Katao H, Shibutani T (2006) Estimate of the stress field in the region of the 2000 Western Tottori earthquake. Program and abstracts of 2006 fall meeting of the Seismological Society of Japan, P099 (in Japanese)
43. Kilb D, Gomberg J (1999) The initial subevent of the 1994 Northridge, California, Earthquake – is earthquake size predictable? *J Seismol* 3:409–420
44. Lay T, Kanamori H, Ruff L (1982) The asperity model and the nature of large subduction zone earthquakes. *Earthq Predict Res* 1:3–71
45. Mandelbrot BB (1982) *The fractal geometry of nature*. W.H. Freeman, New York
46. Miura K, Iio Y, Yukutake Y, Takai K, Horiuchi S (2005) The feature of initial motion for waveforms of microearthquakes in Western Nagano, Japan. Program and abstracts of 2005 fall meeting of the Seismological Society of Japan, P103. (in Japanese)
47. Miyake H, Iwata T, Irikura K (2003) Source characterization for broadband ground motion simulation: Kinematic heterogeneous source model and strong motion generation area. *Bull Seism Soc Am* 93:2531–2545
48. Mori J, Kanamori H (1996) Initial rupture of earthquake in the 1995 Ridgecrest, California sequence. *Geophys Res Lett* 23:2437–2440
49. Nakamura Y (1988) *Proc World Conference on Earthquake Engineering*, VII, 6763

50. Nakatani M, Kaneshima S, Fukao Y (2000) Size-dependent microearthquake initiation inferred from high-gain and low-noise observations at Nikko district, Japan. *J Geophys Res* 105(B12):28095–28110; doi:10.1029/2000JB900255
51. Ohmi S, Watanabe K, Shibutani T, Hirano N, Nakao S (2002) The 2000 Western Tottori Earthquake—Seismic activity revealed by the regional seismic networks. *Earth Planet Space* 54:819–830
52. Ohnaka M, Kuwahara Y, Yamamoto K, Hirasawa T (1986) Dynamic breakdown processes and the generating mechanism for high-frequency elastic radiation during stick-slip instabilities. In: Das S, Boatwright J, Scholz CH, AGU (eds) *Earthquake source mechanics*. Geophysical Monograph, vol 37. Maurice Ewing Series, vol 6. American Geophysical Union, Washington DC, pp 13–24
53. Ohnaka M, Kuwahara Y (1990) Characteristic features of local breakdown near a crack-tip in the transition zone from nucleation to unstable rupture during stick-slip shear failure. *Tectonophysics* 175:197–220
54. Ohnaka M, Shen L (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupture surfaces. *J Geophys Res* 104:817–844
55. Ohnaka M (2000) A physical scaling relation between the size of an earthquake and its nucleation zone size. *Pure Appl Geophys* 157:2259–2282
56. Okubo PG, Dieterich JH (1984) Effects of physical fault properties on frictional instabilities produced on a simulated faults. *J Geophys Res* 89:5817–5827
57. Olson EL, Allen RM (2006) Is earthquake rupture deterministic? *Nature* 442:E5–E6; doi:10.1038/nature04963
58. Rydelek P, Horiuchi S (2006) Is earthquake rupture deterministic? (Reply). *Nature* 442:E6; doi:10.1038/nature04964
59. Sato T (1994) Seismic radiation from circular cracks growing at variable rupture velocity. *Bull Seismol Soc Am* 84:1199–1215
60. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. *J Phys Earth* 21:415–431
61. Sato T, Kanamori H (1999) Beginning of earthquakes modeled with the Griffith's fracture criterion. *Bull Seismol Soc Am* 89:80–93
62. Sato K, Mori J (2006) Scaling relationship of initiations for moderate to large earthquakes. *J Geophys Res* 111:B05306; doi:10.1029/2005JB003613
63. Sato K, Mori J (2006) Relation between rupture complexity and earthquake size for two shallow earthquake sequences in Japan. *J Geophys Res* 10.1029/2005JB003613
64. Shibazaki B, Matsu'ura M (1992) Spontaneous processes for nucleation, dynamic propagation, and stop of earthquake rupture. *Geophys Res Lett* 19:1189–1192
65. Shibazaki B, Matsu'ura M (1995) Foreshocks and pre-events associated with the nucleation of large earthquakes. *Geophys Res Lett* 22(10):1305–1308; doi:10.1029/95GL01196
66. Shibazaki B, Matsu'ura M (1998) Transition process from nucleation to high-speed rupture propagation: Scaling from stick-slip experiments to natural earthquakes. *Geophys J Int* 132:14–30
67. Shibazaki B, Yoshida Y, Nakamura M, Nakamura M, Katao H (2002) Rupture nucleations in the 1995 Hyogo-ken Nanbu earthquake and its large aftershocks. *Geophys J Int* 149:572–588
68. Spudich P, Cranswick E (1984) Direct observation of rupture propagation during the 1979 Imperial Valley earthquake using a short-baseline accelerometer array. *Bull Seismol Soc Am* 74:2083–2114
69. Takenaka H, Nakamura T, Yamamoto Y, Toyokuni G, Kawase H (2006) Precise location of the fault plane and the onset of the main rupture of the 2005 West Off Fukuoka Prefecture earthquake. *Earth Planets Space* 58:75–80
70. Uchide T, Ide S (2007) Development of multiscale slip inversion method and its application to the 2004 Mid-Niigata Prefecture earthquake. *J Geophys Res* doi:10.1029/2006JB004528
71. Uehira K, Yamada T, Shinohara M, Nakahigashi K, Miyamachi H, Iio Y, Okada T, Takahashi H, Matsuwo N, Uchida K, Kanazawa T, Shimizu H (2006) Precise aftershock distribution of the 2005 West Off Fukuoka Prefecture Earthquake ($M_j = 7.0$) using a dense onshore and offshore seismic network. *Earth Planet Space* 58:1605–1610
72. Umeda Y (1990) High-amplitude seismic waves radiated from the bright spot of an earthquake. *Tectonophysics* 175:81–92
73. Umeda Y (1992) The bright spot of an earthquake. *Tectonophysics* 211:13–22
74. Umeda Y, Yamashita T, Tada T, Kame N (1996) Possible mechanisms of dynamic nucleation and arresting of shallow earthquake faulting. *Tectonophysics* 261:179–192
75. Yamaguchi S, H Kawakata, T Adachi, Y Umeda (2007) Features of initial process of rupture for the 2005 West off Fukuoka Prefecture Earthquake. *Zisin Ser 2*:241–252 (in Japanese)
76. Venkataraman A, Beroza GC, Ide S, Imanishi K, Ito H, Iio Y (2006) Measurements of spectral similarity for microearthquakes in western Nagano, Japan. *J Geophys Res* 111:B03303; doi:10.1029/2005JB003834
77. Wu Y, Kanamori H, Allen R, Hauksson E (2007) Determination of earthquake early warning parameters, τ_c and P_d , for southern California. *Geophys J Int* (OnlineEarly Articles) doi:10.1111/j.1365-246X.2007.03430.x
78. Wyss M, Brune J (1967) The Alaska earthquake of 28 March 1964—a complex multiple rupture. *Bull Seismol Soc Am* 57:1017–1023

Earthquake Occurrence and Mechanisms, Stochastic Models for

DAVID VERE-JONES

Statistical Research Associates and Victoria University, Wellington, New Zealand

Article Outline

Glossary

Definition of the Subject

Introduction

Historical Overview

Stochastic Models for Earthquake Mechanisms

Models for Paleoseismological

and Historical Earthquakes

Point Process Models for Regional Catalogues

Stochastic Models with Precursors

Further Topics

Future Directions

Acknowledgments

Bibliography

Glossary

Stochastic occurring by chance;

Stochastic process physical or other process evolving in time governed in part by chance.

Earthquake mechanism physical processes causing the occurrence of an earthquake.

Independent events events not affecting each other's probability of occurrence.

Branching process process of ancestors and offspring, as in the model of nuclear fission.

Point process stochastic process of point-events in time or space.

Probability forecast prediction of the probability distribution of the time and other features of some future event, as distinct from a forecast for the time (etc.) of the event itself.

Model test a statistical test for the extent to which a stochastic model is supported by the relevant data.

Precursory signal observed quantity which affects the occurrence probability of a future event (earthquake).

Definition of the Subject

Stochastic models for earthquake mechanism and occurrence combine a model for the physical processes generating the observable data (catalog data) with a model for the errors, or uncertainties, in our ability to predict those observables. Such models are essential to properly quantify the uncertainties in the model, and to develop probability forecasts. They also help to isolate those features of earthquake mechanism and occurrence which can be attributed to mass action effects of a statistical mechanical character. We do not consider in this paper applications of the models to earthquake engineering and insurance.

Introduction

The complexity of earthquake phenomena, the difficulty of understanding and monitoring the processes involved in their occurrence, and the consequent difficulty of accurately predicting them, are now widely accepted points of view. What are stochastic models, and what role do they play in aiding our understanding of such phenomena?

The present article is an attempt to address these questions. We start from the beginnings, the distinction between stochastic and deterministic models, and the first attempts to model earthquake phenomena in stochastic or statistical terms. We then follow through with a systematic account of some of the main classes of stochastic models that are currently in use, discussing in turn earthquake mechanisms, historical earthquakes, regional catalogs, descriptive patterns, and earthquake precursors.

The focus throughout is on the stochastic modeling aspects, rather than on statistical procedures or associated algorithms. As a result we have given only brief mention to pattern-recognition techniques, or to descriptive procedures such as the estimation of fractal dimensions or of second order properties, which do not lead to fully defined models. Again, although a primary use of stochastic models is in developing probability forecasts, we have limited ourselves to the briefest account of how such forecasts can be produced and assessed. Nor do we directly consider applications to engineering and insurance problems.

The fundamental difference between a physical model and a stochastic model, in broad terms, is that while the physical model attempts to fully describe and predict the process under study, the stochastic model treats some aspects of the physical process as out of range of exact modeling, at least for the time being, and replaces it by some unpredictable and hence random process. The resulting stochastic model should reproduce those aspects of the physical phenomenon which are important and accessible to measurement, but may relegate the rest to dice-tossing or one of its more contemporary avatars such as Brownian motion or the Poisson process.

Across their many different fields of application, two broad roles for stochastic models may be distinguished. The first is epitomized by statistical mechanics. Here the stochastic model plays an integral role in understanding the physical processes themselves. The macroscopic phenomena that we are able to observe directly – temperature, pressure and the like – are shown to be a consequence, not of the details of the collision processes at the microscopic level, but of their mass interactions, which are governed largely by laws of an essentially statistical character such as the law of large numbers or the central limit theorem. For predicting the macroscopic behavior, it is not necessary to know the details of the complex interactions between individual molecules; it is sufficient to replace them by a simple random process that nonetheless preserves the crucial physical aspects such as mean velocities and angular distributions.

Within seismology such a role is implicit when the fracture processes within the earth's crust are compared to

‘frozen turbulence’, or in applications of branching process or percolation theory to explain energy distributions, or in discussions of the fracture strength of materials as functions of the density and size distribution of microcracks, or in the use of cellular automata and similar models to explain the appearance of long-range correlations and power-law distributions in the approach to criticality of certain types of complex systems.

In the other, by far more common, type of application, the stochastic model is used as a basis for planning and prediction. In such situations it is vital to know, not just a forecast value, but also something about the reliability of that value. It is also vitally important that the models can be fully fitted to the observable data. Most branches of applied statistics have evolved in response to such requirements. Within seismology, applied models of this type are needed in discussions of earthquake risk for insurance or building codes, in many parts of engineering seismology, and in the development of decision rules for earthquake response and emergency planning. Probability forecasts of any kind, including all forecasts with some associated estimate of precision, necessarily rely on stochastic models of this kind.

Many decades ago, the famous geophysicist and seismologist Sir Harold Jeffreys, who is also regarded as a pioneer in inferential statistics, argued that, to be worthy of its name, every physical theory should contain within itself the means not only of predicting the relevant quantities, but also of predicting their uncertainties [45]. In our terminology, he was arguing that every physical theory should be based on a stochastic model. In the classical studies of physics and astronomy, the uncertainties in the model are assumed to be due to nothing deeper than observational errors. In a subject such as seismology, however, the uncertainties are much more fundamental.

While general patterns of earthquake behavior may be predicted from physical theories, the predictions do not extend to the times and locations of individual earthquakes. Moreover, the available observational data are rarely more than indirectly relevant to the physical processes controlling the details of earthquake occurrence, as these usually take place many kilometers beneath the surface of the earth, and out of range of direct observation.

Stochastic models of earthquake occurrence that can be used for earthquake prediction must somehow marry the limited physical theory to the limited available data. Attempts to grapple with this central problem have intensified in recent years. They form one factor in the emergence of ‘Statistical Seismology’ as a new sub-discipline. Another, perhaps dominating, factor, is the enormous improvement in both the quantity and quality of the data that

are available, whether from earthquake catalogs, from GPS measurements of ground deformation, or, less commonly, from data on auxiliary quantities such as well levels, electrical signals, ionospheric depression and others thought to have a potential bearing on earthquake occurrence. The high quality data demand a comparable quality in the statistical modeling and analysis.

Historical Overview

The forerunner of any serious statistical modeling is the availability of reliable and relevant data. For models of earthquake occurrence this means the availability of good quality earthquake catalogs. Broadly speaking, such catalogs had to wait, not only until around the turn of the 20th century, when the first instrumental records became available, but until the development of modern instrumentation and the extensive station networks which came into being after the second World War. Before then, the lack of any consistent measure of the size of an earthquake, and the general unevenness of network coverage, made the records of limited value for statistical purposes.

A turning point was the appearance of the first edition of the classic text [28] by Gutenberg and Richter (1949). For the first time it gave a comprehensive overview of the major features of the seismicity of the earth, and of the key empirical relations governing earthquake occurrence. From that time onwards, the way was open for serious statistical analysis, although recent data is far more comprehensive and detailed. Modern instrumental catalogs, prepared from digital records telemetered to a local center from a dense network of stations, may contain hundreds of thousands of events down to very small magnitudes. Typically such catalogs list for each event the origin time (initiation of rupture), epicenter (latitude and longitude of place of first motion), depth, magnitude or seismic moment (alternative measures of earthquake size), and often other parameters relating to the fault mechanism (orientation of the fault and direction of first motion).

The availability of these high-quality catalogs, alongside the increasing availability of data from GPS measurements and other earthquake-related phenomena, is a key reason for the recent upsurge of interest and research in statistical seismology. The broad aims of this emerging field may be described as finding statistical models to describe and make use of such data, and to marry it with the existing physical theory.

Even preinstrumental catalogs inspired the investigation of two statistical issues at least: does the occurrence

of major earthquakes exhibit some form of periodicity in time? do the numbers of events in time intervals of fixed length follow a Poisson distribution? We comment briefly on these two questions before looking at statistical models more generally.

Periodicity of Earthquakes

Periodicity of earthquakes, in some more or less regular sense, was the earliest issue to be investigated, and inspired many early studies, including one of Schuster's classic papers on the periodogram [112]. Until the 1930s, however, neither the data nor the statistical techniques were sufficiently developed to allow the question to be properly resolved. On the statistical side, for instance, the periodogram was a new concept, and statistical tests based on it were still in development. Schuster's paper, applied to a special case, contains within itself all the basic elements of point process spectral theory, starting from the finite Fourier transform, calculating the equivalent of rough significance levels using Rayleigh's random flights, and briefly treating the problems caused by binning the data and by clustering. Jeffreys [44] was one of the first to use a modern statistical approach to tackle the question, while Davison [21] reviewed many of the earlier studies and came to the conclusion that most of those studies were inconclusive.

The topic remains controversial, although it is now clear that no obvious periodicities exist. The most important current contenders for small-scale periodicities are in relation to earth tides. It is suggested that the small fluctuations in crustal stress due to the relative movements of the moon and sun around the earth may be large enough to trigger earthquake activity under favorable conditions, for example in regions already under high stress. A careful recent study of lunar tides on microearthquakes, with further references, is given in [40]. The possibility of using the response of small-scale seismicity to lunar tides as a possible indicator of regions in some near-critical state, and hence as a precursor for larger events, has been suggested in [139]; a statistical analysis is given in [142].

The possibility of long-term periodicities, of the order of decades or centuries, is unclear because of the shortage of data; substantial fluctuations certainly exist. The problem is still full of potential traps. In testing for periodic effects, for example, it is essential to take into account earthquake clustering, and whether or not the period being tested for is preassigned (as for the lunar cycle) or suggested by the data. Both of these issues are illustrated in the discussion in [134] of Kawasumi's historical data for large earthquakes in the Kanto region of Japan.

The Poisson Distribution and Process

The distribution

$$p_n = (\mu^n/n!)e^{-\mu}, \quad \mu > 0, \quad n \geq 0,$$

was introduced by Poisson as an approximation to the binomial distribution when the number of trials N becomes very large but the probability p of success becomes very small, the two balancing in such a way in such a way that the expected number $\mu = Np$ of successes remains moderate in size.

Earthquakes were included among the examples studied by von Bortkiewicz [136] in his 1898 compilation of phenomena to which he could apply 'the law of small numbers', the name he gave to the Poisson approximation to the binomial. The question was studied in greater depth by later writers, including Gutenberg and Richter [28], and several important qualifications were noted. In the first instance, the disturbing effect of aftershocks was pointed out, and so the Poisson distribution was supposed to apply just to main shocks. Then other disturbing effects, such as trends and longer-term fluctuations in activity, were noted. In fact almost no catalog fits the Poisson description exactly, and for research purposes its role as a base-line model for 'standard seismicity' has been replaced by the ETAS model (see Sect. "The ETAS Model"), which provides a much better approximation to the clustering properties of smaller earthquakes.

Nevertheless the simple Poisson form is still the principal basis for determining earthquake risk and for earthquake insurance practices. Underlying its continued relevance is the same idea underlying Poisson's original approximation to the binomial: when the data under examination consists of rare 'successes' from many different and essentially unrelated sources, the Poisson distribution generally emerges as a good approximation.

It is necessary to distinguish between the *Poisson distribution* and the *Poisson process*. The Poisson process refers to an evolutionary model for the occurrence of events in time or space or both. Its principal characteristics, at least in the stationary case, are

1. The number of events within any bounded region (interval in time; area or volume in space) follows a Poisson distribution with parameter μ proportional to the size (length, area etc) of the region selected for study;
2. The numbers of events in disjoint regions are independent random variables.

The second condition embodies the famous 'lack of memory' property of the Poisson process: the temporal version asserts that the occurrence of one or more events before

a certain time has no effect on the occurrence probabilities of subsequent events. It dictates the exponential form of the distribution of the time interval between events, and under simple conditions even dictates the form of the Poisson distribution itself; see, for example, the discussion in Chap. 2 of [20].

The Empirical Laws of Seismology

The advent of more complete and reliable catalogs saw the recognition of a number of statistical regularities in the occurrence of earthquakes. Two of these are central features of seismicity studies today.

Omori's Law Already by the end of the 19th century, the Japanese pioneer seismologist Omori had made detailed studies of some large Japanese aftershock sequences [92], and suggested that the frequency of aftershock occurrence, say $\lambda(\tau)$, decayed approximately hyperbolically with the time τ after the main event:

$$\lambda(\tau) \approx A/\tau$$

where A is a constant characteristic of the particular mainshock and associated with its size. His own and subsequent studies suggested the need for refinements, and the most widely accepted form today is the the Omori–Utsu formula

$$\lambda(\tau) = A/(c + \tau)^p \quad (1)$$

where the parameters A , c , p are again peculiar to the individual aftershock sequence, c is generally small (of the order of seconds to days) and p is close to 1. A detailed study of the history and other issues associated with the Omori law over the 100 years 1894–1994 is given in [124].

The simplest stochastic model for aftershocks is that suggested by Jeffreys in [44], namely an inhomogeneous Poisson process in which $\lambda(\tau)$ is interpreted as the current value of the time-varying Poisson intensity; the independence property (2) of the Poisson process of Sect. “[The Poisson Distribution and Process](#)” is retained, but the mean parameter for the number of events in (s, t) is now $\mu = \int_s^t \lambda(\tau) d\tau$.

The Gutenberg–Richter (GR) Law The law was formulated after the definition of earthquake magnitude gave an objective method of quantifying the size of an earthquake. It is a basic component of [28], although a similar relationship, based on the more qualitative maximum intensity concept, had been formulated somewhat earlier by Ishimoto and Iida for Japanese earthquakes [39].

The GR law provides a summary of the magnitude data in a catalog of earthquakes with magnitudes complete above a certain threshold, say M_0 . It is commonly written in the form

$$\begin{aligned} \{\text{Number of events above magnitude } M\} \\ \approx 10^{a-b(M-M_0)} \quad (2) \end{aligned}$$

or equivalently

$$\begin{aligned} \{\text{Proportion of events above Magnitude } M\} \\ = 10^{-b(M-M_0)} \quad (3) \end{aligned}$$

It is a pity in our view that the former rather than the latter of these two forms has become traditional. The danger then is that a becomes treated as a separate parameter instead of as a normalizing constant, $a = \log_{10} N$, where N is the total number of events under consideration. One reason for this tradition may have been the common (and incorrect) use of ordinary least squares methods to compute the line of best fit from a graph of the binned numbers. Such an approach will certainly produce a slope as one of the parameters, but the estimate is unstable and distorts the interpretation unless it is especially modified to fit the distribution function context.

The second form makes it clear that what we are looking at is an empirical probability distribution, and that the right hand side could equally and more appropriately be written in the form

$$10^{-b(M-M_0)} = e^{-\beta(M-M_0)}$$

where $\beta = b \log_e 10 \approx 2.3b$. Then it is clear that the G-R law asserts that, under suitable conditions, the empirical distribution of magnitudes is approximately exponential.

In principle, a could be regarded as a parameter in an extended model for the space-time-magnitude distribution of events in a given space-time window, but such an interpretation is rarely given.

How valid the exponential distribution remains when examined in greater detail, and whether, and if so by what, it should be replaced for general modeling purposes, is still a subject of debate. The main reservation relates to the possibility of extremely large events, which is physically unreasonable and can lead to misleading conclusions if used in simulation studies of long-term behavior.

Of the many alternatives offered, which include truncated and multi-parameter versions (see [123] for a listing and software), perhaps the most plausible is the ‘tapered Pareto distribution’, or ‘Kagan distribution’ (e.g. [48, 135]), with distribution function written out in terms of

seismic moments or energies as (4) below. This distribution arises in branching and similar models for crack propagation (see [123]), and is derived from maximum-entropy considerations in [73]; a further derivation from a critical phase transition in a finite elastic solid is given in [26].

The use of magnitude itself as a basic variable is also open to question. It is not a uniquely or tightly defined quantity. In terms of quantities such as energies or seismic moments with a more direct physical interpretation, the exponential distribution for magnitudes becomes a Pareto (inverse power-law) distribution,

$$\Pr\{E > x\} = (x/x_0)^{-\alpha}, \quad (x > x_0)$$

through a transformation of the form

$$\log_{10}(E) \approx 1.5M + \text{const } n,$$

where E is the energy. This illustrates the fact that the magnitude scale is essentially a decibel scale, ultimately a consequence of its initial definition in terms of the logarithm of the maximum amplitude of the trace on a seismograph. The tapered Pareto form mentioned earlier is

$$\Pr\{X > x\} \approx Cx^{-\alpha}e^{-\gamma x}, \quad (x \geq x_0), \quad (4)$$

where C is a normalizing constant, or the variant with a similar form for the density.

Båth's Law The so-called Båth's Law asserts, loosely, that in an aftershock sequence, the difference between the magnitude of the mainshock and that of the largest aftershock is around 1.2 magnitude units.

Although noted by Båth in 1965, and even earlier by Utsu, this regularity has never enjoyed quite the same status as the other two laws. The question, still an active topic of debate, is whether it represents a physical phenomenon in its own right, or is merely a consequence of the more general properties of earthquake clustering. It was suggested in [125] (see also the reviews and more extensive studies in [17,63]) that it might be simply a consequence of the statistical properties of the largest and second largest in a sequence of events following the G-R law. More recently, it has been linked to the 'productivity function' of earthquake clustering: the expected number of aftershocks increases typically as an exponential function $Ke^{\alpha M}$ of the magnitude of the main shock, with Båth's Law resulting when the exponent α equals the exponent β in the GR law (see [25]). This suggestion is supported by the appearance of a Båth's law phenomenon in the ETAS model, where there is certainly no explicit model feature relating to Båth's law, but there is an exponential productivity function [38].

Stochastic Models for Earthquake Mechanisms

General Considerations

The earliest model for earthquake mechanism is Reid's elastic rebound model [98]. It was inspired by studies of large-scale earthquakes, in particular the famous San Francisco earthquake of 1906. The upper part of the crust is deformed elastically by large scale tectonic motions, then ruptures and rebounds when its breaking strength is reached, resulting in an earthquake.

The many attempts to marry this physical picture with simple stochastic ideas lead typically to models based broadly on the renewal process, and will be picked up in the discussion in Sect. "Background and Data".

In this section we look rather at models which describe the behavior at the microscopic level, the evolution of the fracture itself, and can be used to explain the basic empirical laws, among other features.

There are strong links with theories on the strength of materials, starting from the classic studies of Griffiths (e.g. [27]) on crack extension in brittle materials, and the role of microfractures in controlling the fracture strength of glass. Griffiths' crack theory is basic to models of fracture in brittle materials, whether at the scale of rock fracture in laboratory specimens or fault propagation in the earth's crust (see e.g. [42,82,109,110]).

Griffiths' ideas were later developed by Weibull [136] into a model explaining the variations in strength of otherwise similar specimens of rock and many other substances. Weibull supposed that the underlying cause was the random distribution of microfracture lengths in the specimen, and used an argument based on the distribution of the length of the largest such microfracture to deduce a form for the distribution of strengths. Indeed it is from these studies that the 'Weibull distribution' takes its name.

The branching process, percolation, and cellular automata interpretations of the earthquake process start from similar general premises. The underlying idea is that, instead of progressing smoothly, as might a fault or fracture in a homogeneous elastic medium, the progress of the fault in a medium containing many weaknesses is controlled by the essentially random locations of these weaknesses. The various models which have been proposed differ mainly in the assumptions governing these random locations.

In Otsuka's original 'go-game' model [93], points were laid down on a lattice in much the same way as in the game of 'Go', but at random, using a simulation technique, with the interpretation that the enclosed pieces determined a rupture area. In [108] this was idealized into

a model linking weaknesses located on a Bethe lattice, where every node has one input link and the same fixed number of outward links, and each node may or may not be a point of weakness.

Otsuka's model has both branching process and percolation model interpretations, with a considerable literature surrounding extensions of both. There also links to other, apparently more deterministic, approaches to the generation of the empirical laws, for example through block slider models or in the general class of complex systems. Although the models differ in approach and detail, the size distributions and the like often turn out to be very similar to those derived from the branching models. As in statistical mechanics, the properties have their origin in the mass interactions of many small components, and are relatively insensitive to the details at the microscopic level. The simpler statistical models, such as the branching models, allow these distributions to be explored directly by analytical means. In complex system theory the aim is rather to show how similar results arise from approximating the deterministic equations governing large families of interacting bodies.

Branching Models

The conceptual framework here is that the crack initiates from an initial weakness (dislocation or microfracture) and spreads to one or more others, or terminates, the 'others' being interpreted as 'offspring' and the initial weakness as the 'ancestor'. Each 'other' then acts as an ancestor in its own right, and the process continues until either all branches have died out (subcritical and critical cases) or the process explodes (supercritical state). The behavior is controlled by a 'criticality parameter' ρ , effectively the mean number of offspring per ancestor. The subcritical, critical, and supercritical cases correspond respectively to $\rho < 1$, $\rho = 1$ and $\rho > 1$.

This model was developed in general form in [123], following [109] and the earlier work on the 'go-game' model in Japan. Related ideas occur in many places papers by Kagan and Kagan and Knopoff, see especially the extended branching model described in [54].

The distribution of the size or energy release of the rupture is then obtained by counting the total number of offspring before the process dies out (critical or subcritical cases). The remarkable feature here is that even when the individual offspring distributions are very regular, the total size distribution approaches a power-law (Pareto) form whose basic parameters are independent of the details of the offspring distribution. In the limiting critical case, the power-law distribution for sizes has $\Pr\{N > n\} \sim Cn^{-1/2}$,

corresponding roughly, assuming equal energies/event on average, to a G-R law with $b \approx 0.75$.

When the process is just subcritical the Pareto distribution becomes a tapered Pareto distribution, with power-law behavior for moderate to large events, and an exponential tail-off at high magnitudes which cuts in at a point determined by the distance from criticality, $\delta = 1 - \rho$. Again the behavior is otherwise largely independent of the details of the offspring distribution.

Many further developments and ramifications of this underlying model have been proposed. One of the deepest is the simulation model for earthquakes developed in [54], starting from the scale of dislocations or other defects in the rock fabric, and incorporating temporal, directional, and distance factors into the model evolution to develop an impressive array of properties akin to those of real earthquakes. The model still awaits a full analytical treatment.

It is also remarkable that a branching model underlies one of the most successful models for earthquake occurrence at the regional level, namely the ETAS model described in Sect. "The ETAS Model". The fact that the same mechanism seems implicated at both levels lends plausibility to Kagan's conjecture that the physical process is one and the same at all scales, and that our attempts to decompose it into elements at the fracture formation and inter-fracture stages are more a result of our perceptions and measuring instruments than they are of the underlying physical processes.

Percolation Models

The classical percolation model starts from a two- or three-dimensional lattice, the sites (or alternatively the bonds between lattice points) being randomly and independently labeled 'open' or 'closed' with a fixed probability p and its complement $1 - p$. A crack initiated at an open site links up all contiguous open sites until it can spread no further. In both cases, a critical regime, characterized by a critical value of the probability p , marks the transition between subcritical (small finite events only) and supercritical (infinite or explosive events) regimes. As with the branching models, it is assumed that the crust is generally in or just below the critical state.

An underlying difficulty is that the available observational data are insufficient to provide any easy control over the best interpretation. As with the branching process model again, the percolation models lead to forms of the G-R law, and with additional features can often be extended to cover aftershock phenomena. [15], and [65,66,67] are among the many papers which discuss

and develop these ideas. [5] highlights some of the difficulties of interpretation.

Percolation processes are extensively used in statistical physics to model phase transitions, and their appearance here invites an interpretation of fracture as a phenomenon analogous in some ways to a phase transition. Ideas from the phase transition context that have been transferred to both earthquakes and fracture mechanics include especially features characteristic of the approach to the critical conditions required for the occurrence of a phase transition: the development of long-range correlations, the appearance of power-law or fractal distributions, and approximate self-similarity. Many authors have sought to develop these analogies, often using analogue or simulation models, and attempted to use the appearance of different interaction ranges to identify the approach to near-critical stress conditions in the crust. See [9,10,26,121], as well as the papers cited above, for further references and discussion of such ideas.

Cellular Automata and Self-Organizing Criticality

A third type of model with a similar general role is the cellular automaton, with the distinction that the application here is not to a single faulting or fracture episode, but to a whole network of interacting faults. The simple basic form, first applied to the earthquake context by Bak and Tang [3], again relates to a two-dimensional lattice model. With each point of the lattice is associated a certain integer stress or force, say Z_{ij} for points on a 2-dimensional lattice $\{i, j\}$. The external force (the ‘immigrants’ in this context) is manifested through the addition of unit force to a site chosen at random through the lattice or on its boundary. When the force exceeds a certain critical value Z_c on a given site, a ‘microfracture’ occurs, and single units of force are transferred to each of the four directly adjacent sites, while four units are subtracted from the force at the original site. Such transfers may overload one or more of the adjacent sites, which then in turn transfer units of stress to their neighbors (including possibly the initial site), and so on until the system is at rest. Then another unit is added in and a further redistribution of stress takes place. The whole episode is interpreted as an earthquake, and the total number of steps in the episode is taken as proportional to the energy of the earthquake.

The process as a whole is said to exhibit ‘self-organizing (or ‘self-organized’) criticality’. Even if the process is started from a situation where the forces are set to zero at all sites, they will gradually build up, first to the stage where small individual episodes take place, then, as more and more sites approach the critical value of stress,

the episodes become larger, until a stochastically stationary state is reached where the input of stress units is just balanced by the loss of stress units from points on the boundary of the region. So long as the region under consideration is sufficiently large, a process reaching the critical regime exhibits many of the features already indicated as characteristic of the approach to a phase-change: a G-R relation, long-range correlation effects, and (with some elaborations) an Omori-type phenomenon for after-shock sequences.

Models for Paleoseismological and Historical Earthquakes

Background and Data

We move now to models designed for use with data on earthquake occurrences. These generally belong to the second class of stochastic models referred to in the introduction. They should be able to be fitted to real catalog data; simulations from them should mimic real catalogs; and they should be useful in real applications, capable in particular of generating probability forecasts.

We have grouped the models into two main types, those developed to model large earthquakes on historical or even geological time scales, and those developed for use with modern instrumental catalogs, where smaller events are included. The main difference between the two types of model is their treatment of clustering; this is largely ignored in models of the first type, but plays a central role for models of the second type. Models of the first type are considered in the present section, models of the second type in Sect. “Point Process Models for Regional Catalogs”.

The distinction between the groups is associated with one of the longest and still unresolved debates over earthquake mechanism, namely the validity of the *characteristic earthquake hypothesis*. Crudely stated, this asserts that, for any given fault or fault segment, there exists an earthquake of approximately fixed magnitude, which is determined by the physical attributes of the fault, and repeats itself after time intervals of approximately fixed length. Since faults occur over a very wide range of sizes (themselves having a power law or Pareto distribution), this does not contradict, but rather suggests a different origin for, the GR distribution.

The empirical evidence for such a hypothesis is equivocal. Its main support comes from the paleoseismological studies on repeated events along a single fault (e. g. [113]), but the data from such studies is usually so limited that it is hard to accept the evidence as conclusive. Other supporting evidence, again observed sometimes but not always, is the occurrence of a hump, corresponding to repeating

earthquakes with similar magnitudes, in the frequency-magnitude distribution for selected regions. For large regions (scale of major faults), this may occur around magnitudes 6–7, suggesting that the larger events from that region occur more regularly (with higher relative frequency) than would be expected from the GR model. One difficulty with the hypothesis is that several of the best-known sequences, such as the Parkfield earthquake sequence, ultimately deviate from the prescribed regularity. Studies of microearthquakes suggest that in some circumstances similar-sized small events may repeat themselves several times in almost identical locations.

Such results suggest that no simple, universal mode of behavior is likely to be found in earthquakes from particular fault structures. Indeed, recent studies by Ben-Zion and colleagues (see [7,8,9,10] and further references therein) have emphasized the possible role played by evolving heterogeneities and damage rheology in the occurrence patterns on a fault system, and have suggested that, according to their ages and past histories, some faults may exhibit characteristic earthquake behavior while others exhibit GR behavior and others again may alternate between the two.

For paleoseismic studies, each data point is extracted with effort from trenching along fault traces or similar exercises. Moreover, only the largest events leave traces identifiable over thousands of years or longer, while estimating magnitudes and other characteristics is at best informed guess-work. Even the dates, usually determined from some form of radio-carbon or other isotope-based method, can be subject to substantial errors.

Similarly in historical studies, such as Ambraseys' history of Persian earthquakes [1], only the largest events affecting a given region or territory are likely to appear sufficiently prominently in the historical records to allow the size and epicenter of the earthquake to be estimated even roughly. Periods of civil war, famine, foreign invasion and bureaucratic neglect leave gaps and further uncertainties which are difficult if not impossible to resolve.

Despite such difficulties, these data provide the only records we have of seismic activity over periods stretching backwards in time beyond the last hundred years or so, and are worthy of the most serious attempts to collect and interpret.

We consider a sequence of three models, starting from the simple renewal model, then considering variants more closely linked to the elastic rebound model.

Renewal Models

With paleological data in particular, attention is generally focused on major events along a single fault, where magni-

tudes are poorly constrained, and the stochastic elements are introduced primarily to describe, and if possible predict, the time intervals between events.

For a renewal process, magnitudes are neglected, and it is assumed that the successive intervals are independent, both of each other and of other processes, with a common distribution. The independence assumptions are questionable, but with no further information available, this is at least a reasonable starting model.

Let $f(x)$ denote the density and $F(x)$ the distribution function of the common interval distribution. If the observation record, over $(0, T)$, say, comprises an interval of length ℓ_0 to the first recorded event, then n complete intervals $\ell_1, \ell_2, \dots, \ell_n$, and finally an unfinished interval ℓ_{n+1} , the likelihood is given by

$$L(\ell_0, \ell_1, \dots, \ell_n, \ell_{n+1}) = a(\ell_0) \left[\prod_{i=1}^n f(\ell_i) \right] b(\ell_{n+1}), \quad (5)$$

where ℓ_0 and ℓ_{n+1} are the incomplete intervals from the commencement of study to the first event, and from the last event to the end of the study, respectively, $a(x) = [1 - F(x)]/\mu$, $b(x) = 1 - F(x)$, $\mu = \int_0^\infty uf(u)du$.

The term $a(x)$ at the beginning of the sequence is the appropriate form to use if the process can be supposed stationary, but nothing is known about events before the commencement of the observation period. The term $b(x) = 1 - F(x)$ at the end of the sequence merely acknowledges the fact that the final interval has begun but not yet concluded.

The main uses of the model are in estimating long-term average or static hazards, in which case the mean of the interevent times plays the crucial role and the form of the distribution is largely irrelevant (so that a Poisson approximation would generally be adequate). The other application is to estimating the residual time to the next event, which is governed by the extended hazard function

$$h(y|x) = f(x+y)/[1 - F(x)] \quad (y \geq 0; x > 0), \quad (6)$$

giving the density of the distribution of the time y from the present to the next event, given that time x has elapsed since the last event.

Distributions commonly used in these situations include the Weibull, gamma, log-normal and Brownian first passage time (inverse normal). Recent work has tended to favor the last of these: see [79]. Two further recent studies which look carefully at the statistical issues, including those relating to errors in the occurrence times, are in [64] and [121]. In long period studies, care needs to be taken that consistent procedures have been used over the

whole period, particularly in the determination of magnitude thresholds, and rules for the exclusion of aftershocks (which must be removed here since otherwise they would contradict the assumption of i.i.d. intervals).

Time- and Slip-Predictable Models

The time-predictable model, introduced by Shimazaki and Nakata in [116], is a widely used alternative for major events along a given fault, when magnitudes as well as interoccurrence times are available.

As in the elastic rebound model, it is supposed that stress along a particular fault builds up linearly until a critical value is reached, representing in some sense the strength of the fault. The size (magnitude) of the resulting event is not known beforehand, but is supposed to be selected randomly either from the standard G-R form or a variant suggested by the characteristic earthquake model.

Once it has occurred, the stress along the fault is instantaneously reduced by an amount determined by the magnitude of the event. Then stress build-up continues until the critical stress level is reached again. The time needed for this to occur is determined by the stress released by the previous event, and so is known, whence the 'time-predictable' title.

The major unknown in the model is the rate of stress build-up between events. If only observation times and magnitudes are available, this can be estimated, albeit crudely, by regressing the observed time intervals onto the magnitudes of the preceding events. If the magnitudes are determined up to a normal error term, with variance independent of the magnitude, this will result in a log-normal distribution for the length of the time interval following an event of given magnitude, and indeed this is commonly used. Another approach would be to use geological data to provide an initial ('prior') distribution for the slip rate, and put the further analysis into a Bayesian framework.

A simple model used for predictive purposes in some of the papers by the Working Group on Californian Earthquakes (see [137] for example), can be represented as

$$\log T_i = A + M_i + \epsilon_i \quad (7)$$

where the ϵ_i are independent, normally distributed errors with zero mean and constant variance, the M_i are the observed magnitudes of the events, and $A = -\log V$ is the logarithm of the slip rate, estimated from geological and GPS studies. Given the time x since the last major event on the fault, the distribution of the remaining time y until the next event is governed by the extended hazard function

of the log-normal distribution, as in the discussion (6) of the renewal model, but with the mean adjusted to take into account the extra information provided by the magnitude of the previous event.

An underlying but subtle logical difficulty with the model is that if applied on a long time basis, the assumption of i.i.d. lognormal errors leads to unbounded fluctuations in the accumulated sums

$$V \sum_{i=1}^n T_i - \sum_{i=1}^N S_i = V \sum_{i=1}^N S_i (\epsilon_i - 1),$$

where S_i is an estimate of the slip from an event with magnitude M_i . The cumulative sum on the right hand side can oscillate without bound, implying the unphysical possibility of indefinitely large fluctuations in the accumulated stress.

Shimazaki and Nakata also suggested a dual version, the *slip-predictable* model, characterized by a return after each event to a constant resting stress. The time at which the next event will occur is unknown, but given the time since the last event, the minimum expected size is determined by the stress accumulated since the previous event. [57] develops a more detailed version for use in earthquake engineering applications.

The Stress-Release Model

The stress-release model is an attempt to address similar issues from within a stochastic point process framework (see for example Chap. 7 of [20]), incorporating both occurrence times and magnitudes. As in the previous case, it is assumed that the rate of stress build-up is constant (say ρ), and that sizes of successive events are i.i.d. and independent of the stress level at the time of occurrence. Most commonly they are assumed to follow the exponential form associated with the GR law, but this is not inherent in the model.

The crucial difference with the time-predictable model is that, instead of assuming that the strength of the crust is fixed, it is assumed to be variable with distribution function say $\Phi(s)$ with density $\phi(s)$. The probability that the next earthquake occurs when the stress passes through $s, s + ds$, but not before, is then given by the hazard function $\Psi(s) = \phi(s)/[1 - \Phi(s)]$. This hazard function $\Psi(s)$ determines the pattern of occurrence probabilities. Most commonly, it is taken to have an exponential form $\Psi(s) = Ae^{\lambda s}$, corresponding to the double exponential distribution function $\Phi(S) = 1 - e^{-A[e^{\lambda S} - 1]}$ for the breaking strength itself. This has a well-marked mode at $(-\log A)/\lambda$ if A is rather small.

In stochastic point process terms, the quantity $\lambda^*(t) = \Psi[X(t)]$ can be interpreted as the *conditional intensity* of the model, meaning approximately the instantaneous occurrence rate, given the history of the process up to time t :

$$\lambda^*(t)dt \approx E[dN(t) | \mathcal{H}(t)] \approx \Pr\{dN(t) > 0 | \mathcal{H}(t)\}. \quad (8)$$

Roughly speaking, the process behaves locally like a Poisson process with instantaneous rate $\lambda^*(t)$, which in the stress-release model can be written more explicitly as

$$\lambda^*(t) = \Psi[X(t)] = \Psi \left[X(0) + \rho t - \sum_1^{N(t)} S_n \right]. \quad (9)$$

This model has several useful features. First, the fact that a simple explicit form exists for the conditional intensity means that it can be readily incorporated into standard procedures for maximum likelihood estimation, simulation, and prediction (see again Chap. 7 of [20]). In particular, the likelihood ratio for a set of observed events (t_i, M_i) over the interval $[0, T]$ can be written in the form

$$\log L/L_0 = \left[\sum \log[\lambda^*(t_i)/\lambda] - \int_0^T [\lambda^*(u) - \lambda] du \right] + \sum_1^{N(T)} \log[g(M_i)/g_0(M_i)] \quad (10)$$

where λ is the rate of the background (null) model, assumed constant rate Poisson, $g(x)$ and $g_0(x)$ are the densities of the proposed and background magnitude distributions, and the magnitudes are assumed independent.

Second, as in earlier, related work by Knopoff [58], the current stress level, say $X(t)$, is Markovian, for the current value of $X(t)$ determines the probability of the next jump occurring, while the remaining components (size of jump, rate of build-up between jumps) are independent of the past history of the process. Hence the extensive knowledge of Markov processes can be brought to bear on the properties of $X(t)$ (e. g. [11]).

A third point is that as the stress level increases, the rate of occurrence of new events will remain relatively high until a large enough event occurs to reduce the stress level to substantially lower values. The model therefore embodies a modest form of accelerated moment release [43].

The model assumes only a simple scalar concept for regional stress, much as in the early chapters of [83], and does not allow for stress interactions between regions. To address the latter point, the *coupled stress release model* was introduced by Shi Yaolin and students [68], to allow stress transfers between regions as well as simple

stress drops. Further discussions and examples are in [4] and [71].

Point Process Models for Regional Catalogues

Data Consistency and Declustering

Regional catalogs, based on instrumental data from the last century or so, present a very different picture, but one with its own problems also. Of these, the two most important are the maintenance of consistency and the problem of clustering (or declustering).

It is characteristic of such catalogs that the networks supplying the data undergo many changes with the passing of the years. Although it is just these changes that have made possible the more serious statistical studies of recent years, they create their own problems in terms of lack of data consistency. Using such data for any form of long-term study requires continual vigilance over questions such as improvements and other changes in the individual network stations and their instruments, shifts in magnitude definitions or thresholds, changes in the routines used in determining epicenter locations, policy decisions over the events to be listed in the catalog, etc. Unless such factors are carefully listed and properly allowed for, they can easily lead to misinterpretation of statistical features observed in the data. As just one illustration, [18] gives some vivid examples of features of apparent physical interest which in fact have their origins in catalog artefacts induced by changes in magnitude registration.

An even more vexed question is whether, and if so how, to remove major clusters from (i. e. ‘decluster’) the catalog. Large aftershock sequences look simple to identify and remove, but the process is considerably more difficult than it might appear.

The possible justifications for doing so are two-fold. If it is believed that the large events are different in kind from the smaller events, then declustering is simply a procedure to isolate the events of primary importance. This assumption was once standard, and in any case the large events appear to be responsible for the major part of the large-scale tectonic motion. But with data on small events becoming ever more plentiful and increasingly reliable, their role is undergoing reassessment.

The second justification for removing aftershocks and other clusters is that they are a nuisance. They negate the assumptions of independence which lie at the basis of most standard statistical tests (e. g. for trends or periodic effects), they greatly complicate analysis and interpretation, and they require elaborate and difficult techniques to deal with explicitly.

Nevertheless, most statisticians, myself included, would tend to look askance at throwing away a substantial portion of the data on the basis of what are inevitably somewhat ad-hoc rules. Many procedures for removing aftershocks have been proposed, and the fact that none has gained general acceptance is evidence of this underlying problem. Moreover, while declustering removes the most obvious earthquake clusters, it rarely removes the clustering completely. The interpretation of results based on the remaining data remains equivocal, partly physical and partly man-induced.

For such reasons we do not discuss declustering in detail in the present article, but concentrate rather on procedures for modeling the data without removing the aftershocks.

A general caution in handling clustered data is not to presume that standard statistical procedures, especially tests, can be applied without modification. In general, the presence of clustering severely affects significance levels. For example, attention is drawn in [81] to the dangers of assessing the significance of precursory effects without properly allowing for clusters. A similar point occurs in assessing the significance of periodic effects, as was pointed out by Schuster in [112], as well as more recently in [129] and no doubt in other places.

To allow for the effects of clustering, and to examine the structuring features themselves, some form of explicit modeling is generally desirable. For example, one possible approach to highly clustered data is to remove as much of the gross clustering as possible with a basic cluster model, and then examine the residuals from fitting the model. Ogata, Zhuang and colleagues have recently developed various techniques, described in [86] and [143] for example, for examining the residuals from catalog data initially fitted by the ETAS model. Alternatively the cluster model can be fitted locally (i. e. with parameters allowed to vary in time or in space and time), and the parameter variations examined to shed more light on the features of interest: [89] contains a compelling example of such an analysis.

We proceed to describe three types of cluster model, starting from the ETAS model itself. All three models are defined through the form of the conditional intensity function, as outlined in the discussion of the stress-release model. In all three models again, magnitudes are allocated independently and randomly, either according to the GR law, or some variant such as the tapered Pareto distribution for seismic moments. The final feature in common is that in all three models the main component in the conditional intensity is a linear combination of contributions from past events.

The ETAS Model

The ETAS model (the initials standing for Epidemic Type Aftershock Sequence) first appeared in Ogata's paper [86], but was preceded by a series of studies by Ogata and colleagues in Tokyo on processes which, like the ETAS model itself, have conditional intensities of the linear, Hawkes type, following [34,35]. Earlier cluster models included the Neyman-Scott process, reincarnated as a 'trigger model' in [133] and [126].

In its basic time-magnitude form, the ETAS model has conditional intensity

$$\lambda^*(t, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu + A \sum_{i:t_i < t} e^{\alpha(M_i-M_0)} f(t-t_i) \right\}, \quad (11)$$

where the first term on the RHS is the GR density for magnitudes, μ is an arrival rate for background (ancestor) events, the constant A is related to the criticality of the process, the 'productivity function' $e^{\alpha(M-M_0)}$ describes how the number of first-generating offspring increases with magnitude of the parent event, and $f(u) = p c^p / (c + u)^{1+p}$, $c > 0$, $p > 0$ is the density (here a Pareto form) for the distribution of the temporal lag between the arrival or birth of the parent and that of its offspring.

In the full space-time-magnitude version

$$\lambda^\dagger(t, x, M) = \beta e^{-\beta(M-M_0)} \cdot \left\{ \mu h(x) + A \sum_{i:t_i < t} e^{\alpha(M_i-M_0)} f(t-t_i) g(x-x_i) \right\}. \quad (12)$$

The new terms are the density h of new arrivals over the spatial region, and the density g in space for the location of an 'offspring' event about its parent. We suppose that f , g and h are all normalized to form probability densities.

One of the main attractions of the Hawkes' processes, including the ETAS model, is that they have a branching process interpretation, first pointed out in [36] and implicit already in the description of the conditional intensity. For example, the criticality parameter (mean number of offspring per ancestor, averaged over the magnitude distribution for the ancestor) is given by $A/(1 - \alpha/\beta)$ for both the above forms. Thus a stable version of the process can exist only if $\alpha < \beta$, and then only if A is small enough. Of course, branching process ideas appear in many earthquake occurrence models, notably in Kagan's work (e. g. [36]).

While the branching process interpretation gives much insight into the structure of the ETAS model, statistical analysis depends crucially on the representations in (11) and (12), since they lead to the relatively tractable form (10) for the likelihood.

For computational purposes, the likelihood of a general marked point process, of which the ETAS models are examples, is often written most conveniently in terms of the conditional intensity $\lambda_g^*(t)$ for the *ground process*, the overall occurrence of points, irrespective of location or mark, and the conditional mark (in our case space and magnitude coordinates) distribution $f^*(x, M|t)$, so that $\lambda^*(t, x, M) = \lambda_g^*(t) f^*(x, M|t)$. The star indicates that the quantities so labeled are in general conditional on (and hence functions of) the histories up to time t . Provided f^* is normalized to a probability density for any given past history, we can write the likelihood ratio in the form

$$\log L_1/L_0 = \left[\sum_{i=1}^{N(T)} \log[\lambda_g^*(t_i)/\lambda] - \int_0^T [\lambda_g(t) - \lambda] dt \right] + \sum_{i=1}^{N_g(T)} \log[f^*(x_i, M_i | t_i)/f(x_i, M_i)], \quad (13)$$

where the terms λ and $f(x, M)$ relate to the rate and mark distribution for the background process (null model), here taken to be a constant rate Poisson process with independent (and usually GR) magnitudes.

This form represents the likelihood ratio as the sum of two terms, the first involving the time points only, and the second involving the marks (spatial locations) given the time points. In many models, the parameters appearing in the two terms have no common variables, in which case optimization can be carried out for the two terms separately.

Because the ETAS model fits well to catalogue data over a wide range of scales and contexts, in recent years its properties have been examined in detail, with the aim of verifying its ability, or otherwise, to reproduce specific features of the real process, such as Bath's law or the occurrence of foreshocks; see, for example, [37,38].

Moreover, the procedures developed by Ogata and colleagues for fitting versions of the model in which the parameters can vary in both location and time, and for detecting local departures from a good fit of the model, have made the ETAS model a powerful diagnostic tool. In this way it has been used to estimate local variations in the stress field (e.g. [89,91]), or changes in seismicity due to the intrusion of ground water [30,58].

For long it was believed that an immigration component, coupled to a subcritical branching structure for the

offspring, was the only way to produce a stable process with branching structure. However, it was shown recently in [12] that when the temporal lag distribution $f(\cdot)$ of (11) is very long-tailed, a critical Hawkes process can sustain itself indefinitely as a 'process without ancestors'. Another somewhat unexpected extension, described in [132], is to a self-similar version over an infinite range of magnitudes.

Perhaps the one serious limitation of the ETAS model is its rather poor performance as an intermediate-term predictor. The reason for this is that its predictive power is basically dependent on its ability to fit aftershock sequences. Hence it does not show significant gains, even against the Poisson model, until the time intervals between forecasts are of similar order of magnitude to the time intervals between the larger events in an aftershock sequence.

The Kagan–Jackson Models

Kagan and Jackson have proposed a number of forms of which we refer to two, the long-term and short-term versions of [41]. The long-term version has its origins in [51], while the short-term version has its origins in [55].

In the long-term model, the current value of the conditional intensity, within a spatial region A and based on observations within $(0, t)$, has the form

$$\lambda^*(t, x, M) = f_t(x)g(M)h(t) \quad (14)$$

where $h(t)$ is the overall current risk (ground process intensity), $g(M)$ is a (fixed) magnitude distribution, commonly that corresponding to the tapered Pareto form for seismic moments, and

$$f_t(x) = \sum_{0 < t_i < t} k(x - x_i) / \sum_{0 < t_i < t} \int_A k(x' - x_i) dx'$$

k being a spatial kernel function. Thus f_t is a normalized sum of contributions from previous events within the observation period and spatial region. It is time-independent except insofar as the advent of additional events requires additional renormalization.

Although the model is well-defined by its conditional intensity and an initial condition at $t = 0$, and can be fitted by likelihood methods much as for the ETAS model, the renormalization introduces a non-linear component into the model which makes its properties more difficult to analyze than those of the ETAS model. Moreover, like other forms of moving average model, it is non-ergodic: there is no unique stationary form to which it will converge from different initial conditions. Nevertheless, it serves the principal purpose of providing a baseline comparison for other putative prediction models for the same region.

The short-term form is very close to the spatial ETAS model. As in the ETAS model, each past event is associated with both a spatial and a temporal decay function, the temporal decay following the Omori law. Again, however, it involves a renormalization rather than the introduction of an explicit immigration term into the model, although some versions of the model allow a small quota of ‘surprises’ in parts of the region with no previous earthquakes.

An important role for both models has been to focus attention on the need for systematic, long-term evaluation and comparison of forecasting models, and to provide a more relevant null model than the constant-rate Poisson process.

The EEPAS Model

This model grew out of several decades of experimentation with precursory swarm models by Evison and Rhoades; see [24] and [104] for more of the history and underlying concepts. The precursory swarm models identify groups of moderate-sized earthquakes as precursory swarms and use these as possible precursors of large earthquakes. Since both precursory phenomena and forecast phenomena belong ultimately to the same process of earthquake formation, a more satisfactory approach is to try and develop a joint model for both phenomena. This, in effect, is what the EEPAS model achieves. It has its own rationale, based on a theory of growth and development of crustal fractures outlined in the papers cited above, and has been successfully applied in several major seismic regions (e.g. [100] and [105]).

Much as in the ETAS model, the conditional intensity has the general form

$$\lambda^*(t, m, x) = \mu \lambda_0(t, m, x) + \sum_{t_i < t} w_i \eta(m_i) r(M | M_i) f(t - t_i | M_i) g(x - x_i | M_i), \quad (15)$$

but the details are significantly different.

First, the conditional intensity in (15) is not taken to apply to the whole catalogue from which the events on the right side are derived, but only for events above a higher threshold. Thus, the model might be used for modeling (and predicting) events over magnitude 5.8, but would take explanatory data from the catalogue of events with $M \geq 4$.

Second, the functions f and g are not based on Omori-type decay formula, but on logarithmic regressions for the time and space delays between an initiating event and the event it may precede, and the magnitudes of the two

events. Thus for example

$$f(u | M_i) = \frac{1}{u \sigma_T \sqrt{2\pi}} \exp \left[-\frac{(\log u - a_T - b_T M_i)^2}{2\sigma_T^2} \right],$$

with an analogous expression for $g(w)$, while $r(m)$ takes the form

$$r(m | M_i) = \frac{1}{\sigma_M \sqrt{2\pi}} \exp \left[-\frac{(m - a_M - b_M M_i)^2}{2\sigma_M^2} \right].$$

They differ considerably from the functional forms used in the ETAS model, but are similar to relations used in the precursory swarm models.

The weight factors w_i are commonly set to unity, but in more refined analyzes may be down-weighted when the triggering event has been identified as an aftershock. One way of finding suitable weights is to carry out an initial ETAS stochastic declustering, as in [143], and base the weights on the probability that a given event is independent.

The further normalizing factor $\eta(m_i)$ in (15) is introduced, much as in the Kagan–Jackson models, to offset the absence of any immigration term, and to compensate for the input from earthquakes below the magnitude threshold. It is adjusted for each magnitude class m_i so that the overall rates follow the G-R law.

In practice the contribution from the baseline rate density is often so small as to be negligible.

Our impression is that the EEPAS model is currently the best-performing of the general seismicity models in the sense of producing the highest average probability gains or entropy scores (see Sect. “Assessing Probability Forecasts”) for predicting moderate to large events on intermediate time scales.

Stochastic Models with Precursors

General Considerations

The search for reliable earthquake precursors has a long and troubled history. High hopes in the 1970s met many disappointments, some at least arising from an inadequate appreciation of the many statistical pitfalls. These difficulties are now much better appreciated, but even so there are relatively few studies based on a satisfactory statistical model, incorporating a proper assessment of the uncertainties, and showing a significant precursory effect.

From a modeling point of view it is important to distinguish between *complete* and *partial* models. In a complete model, both the earthquakes and the precursors are included as components of an overarching joint process, the earthquakes forming one marginal process and the precursors another. In principle such a complete model

should be the aim, but in most situations either the background physics is insufficiently understood, or observations on the precursors are inadequate, or the statistical analysis is too difficult, to allow such a joint analysis to proceed. In many situations also, the modeling and analysis lie outside the realms of conventional statistical models (in dealing with self-similarity, for example), raising further procedural problems.

In a partial model, no attempt is made to model the precursors as a stochastic process. They are treated as given, and their data used in regression-like procedures to modify the probabilities of earthquake occurrence. They can be used retrospectively to examine the performance of a suggested predictive relation, but their use in developing probability forecasts is limited to the short term because there is no model to forecast the future behavior of the precursors.

In the later parts of this section we illustrate some of the modeling approaches that can be used with precursors, using examples drawn mainly from our own experience.

One issue that commonly arises is that the precursor signal data are derived from observations taken at fixed sampling intervals. To match such data to the point process data for the earthquakes, it is generally easiest to switch the whole analysis to a discrete-time study. In this case the continuous time point process models of the previous sections need to be replaced by approximating discrete time models. The two most common of these are the logistic, or binary data, models, and the discrete Poisson process models, illustrated in the first two examples below. In these two classes of models, the probabilities p_n that an event occurs in the n th interval (respectively, the means μ_n of the Poisson distribution for that interval) play the role of the conditional intensity function $\lambda(t)$ of the continuous time models of the previous section.

Example 1: Logistic Regression Analysis of M8 Series

Logistic regressions are used to directly assess the effect of precursor observations on the event probabilities p_n . Suppose that at the n th interval, observations $(U_1^n, U_2^n, \dots, U_k^n)$ are available on k precursors. Dropping the n for brevity, the logistic regression takes the general form

$$\log \left[\frac{p}{1-p} \right] = \alpha_0 + \sum_{k=1}^K \alpha_k U_k \quad (16)$$

where the left side is the logit (log-odds) transform of the event probability, and the right side is the regression term. This representation corresponds to the canonical form for a binomial distribution as a member of the exponential

family. Standard routines exist for estimating the parameters α_k by maximum likelihood or closely related methods, and form part of the generalized linear model procedures.

As an example we consider the model used in [32,33] on the output from the M8 algorithm on New Zealand data.

The M8 algorithm itself is not a stochastic process model, but a decision procedure for calling an earthquake alert based on the analysis of 7 contributing series from earthquake data within a specified 'region (circle) of investigation'. It is the best known of a number of pattern-recognition algorithms developed by the Russian group headed by Keilis-Borok during the 1970s and subsequently. The basic form of the algorithm is described in [56] and [59], with recent reviews in [60,61].

The heart of the algorithm consists of a set of decision rules, based on the joint behavior of the 7 time series, for calling an alert, or more specifically the announcement of a 'TIP' (time of increased probability of an event above a given magnitude threshold) over the region of investigation. The time series are updated every six months, and a TIP extends for three years in the first instance.

A key feature of the analysis in [32,33] is that, in each six-month interval, the values of the seven series are combined by a non-linear formula (linear methods seem less effective) into the value U_1^n of a 'critical series' which is then used as the single precursor in a logistic regression model. The non-linear formula mimics the structure of the decision rules used in declaring a TIP.

The logistic regression analysis then provides the probability for the occurrence of an event over the specified magnitude threshold within the region of investigation for the current 6-monthly period. Much of the further discussion in [32,33] is concerned with combining the outputs from overlapping regions of investigation.

Note that the analysis is typical of that for a partial model; a complete model for M8 would model the joint distribution of the M8 series and the target events.

Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals

In a discrete-time Poisson-type model, the number of events Z_n in the n th time interval is modeled as a Poisson variable with mean μ_n^* that is treated as a function of the past history in much the same way as the conditional intensity in the continuous-time model. The likelihood ratio against a constant mean Poisson process takes the form

$$\log L/L_0 = \sum_1^N Z_n \log(\mu_n^*/\mu) - \sum_1^N (\mu_n^* - \mu). \quad (17)$$

The analogy with the point process form (10) is very obvious, particularly as the time intervals become small so that with high probability the Z_n are either 0 or 1.

The way in which μ_n^* depends on the past can be very general, subject only to the constraint $\mu_n^* \geq 0$. In particular, μ_n^* can depend on prior observations both of the process itself and of auxiliary (precursory) variables. If a linear conditioning of the type (16) is required, the multiplicative form

$$\log \frac{\mu_n^*}{\mu} = \sum_{k=1}^K \alpha_k^{(n)} U_k^{(n)} \quad (18)$$

can be used, ensuring that μ_n is positive, and slotting into the canonical form for the Poisson distribution as a further member of the exponential family, so that the generalized linear model procedures become available. However, the form of the likelihood (17) is usually simple enough to be maximized directly even in more general cases.

An example is the analysis of ULF (ultra low frequency) electric field data in [144].

The ULF signal referred to here is made up of small fluctuations in electric potential measured some meters below the ground surface by sensitive and well shielded electrodes. Its role as a precursor, and the physical explanation of the phenomenon, if it exists, are still unclear. The data analyzed in the paper cited come from some thirty years of recordings from stations around Beijing in China.

Here the base-line (reference) model is a self-exciting (Hawkes type) process in discrete time for the daily earthquake numbers in a wider region around Beijing. Such a model is needed as a reference model because it takes into the inherent clustering effects of the earthquakes themselves. Otherwise there is a temptation to interpret (wrongly) all improvements over the constant rate Poisson process as due to the signals and not to the inherent clustering of the earthquakes. The regressands were the daily readings of ULF anomalies at a set of some half-dozen recording stations around Beijing, simplified to form 0-1 series of observations above a threshold. In fact the daily numbers of earthquakes were small enough for the corresponding continuous and discrete time models to be essentially identical.

Two analyzes were carried out, first with a linear Hawkes-type representation incorporating the effects of past earthquakes alone (self-exciting model), none of the ULF data being used, and second with a double (mutually exciting) linear Hawkes-type representation for the effects of both past earthquakes and ULF signals on the current rate. Likelihood ratios were taken first with respect to a constant rate Poisson process, optimizing parameters in

both cases, then (as a ratio of ratios) for the first and second models against each other, to allow the improvement due to adding in the information from the ULF signals to be assessed.

In this study the model was also tested in reverse mode, to see whether the earthquakes improved the likelihood performance of a Hawkes-type model for the ULF signals alone. The results were positive in the direct mode and negative in the reverse mode. Either way, the models were partial, not complete, as no attempt was made to provide a joint model for the earthquakes and the ULF anomalies together.

Discrete-time Poisson-type models are also used as the basis for model testing within the RELM testing center in Southern California [111]. The modeler supplies the Poisson rates $\mu(n, r, m)$ not only for each time interval (n) but also for each spatial bin (r) and magnitude bin (m). The (approximative) assumption is then made that all the Poisson variables relating to a given time interval are conditionally independent given the current Poisson rates, so that a likelihood ratio of the form (17) can still be used, and made the basis of comparing different proposed models.

Example 3: Point Process Regression Models

As already hinted at, continuous time (point process) procedures can be developed along similar lines to those in the previous example, if the past history includes information on auxiliary (precursory) variables as well as the history of the point process itself.

The form of such dependences can be very general, but if a linear form is wanted it can be incorporated through expressions of the form

$$\log [\lambda^*(t)/\lambda_0^*(t)] = \sum_{k=1}^K \alpha_k U_k(t) \quad (19)$$

for the ratio of intensities, in a similar way to the multiplicative form for the discrete Poisson model. In the special case that $\lambda_0(t)$ corresponds to a renewal process, this is the well-known *Cox regression model* [18]. A comprehensive treatment of models of this kind, with mainly medical and social science applications, is in [2].

Few examples of this type are known to us in the seismological literature, and this may be one area with scope for further development. For example, it is possible that the time-predictable model could be reformulated as a Cox regression model, by building in the dependence on the size of the previous event as a regressand. The closest to a model of this kind that already considered in this article is perhaps the EEPAS model, where the events used in

developing the conditional intensities are mainly smaller than the events being modeled, and so could be described as precursors. Thus the conditional intensity for the larger events being modeled is regressed onto the magnitudes, times and locations of the smaller events in the catalogue. If the renormalization and immigration terms were omitted, the conditional intensity (15) of the EEPAS model would have a similar basic form to (19) above.

Example 4: Foreshocks

Foreshocks, with their potential for earthquake prediction, have been of interest to seismologists since the days of Omori and other earthquake pioneers. The hope that increased quantity and quality of catalogue data would lead to a more definitive picture of how, when and why foreshocks occur has not yet been realized, however. Features which would discriminate foreshocks from other earthquakes or earthquake clusters have proved hard to identify, although careful studies by Ogata and colleagues (e.g. [90]) do suggest that limited opportunities for discrimination may exist.

As precursors, foreshocks retain a specific but limited role. Some degree of forecasting power is available simply from the fact that any newly-observed event, outside those in clearly defined aftershock sequences, has the potentiality of being a foreshock. Recent studies of foreshock occurrence from this point of view are given (among others) in [46] for Southern California, [96] for large global events, [80] for New Zealand earthquakes.

The studies suggest that, leaving aside events in an obvious aftershock sequence, between 5 and 10% of earthquakes with magnitudes 4 and over are likely to be followed by a larger event within time and space windows of the order of 4–5 days and 20–30 km radius respectively. No very sophisticated stochastic model is required to describe such a feature: within the defined window, the probability that a larger event will occur is simply increased from its background value to about 5%, and multiplied by a standard GR factor to take into account magnitude variation; a further separate factor can be used to take into account sub-regional variations of foreshock probabilities.

In the last few decades foreshocks have been studied from a more general point of view, as evidence or otherwise of the self-similarity of earthquake occurrence. In a branching model such as the ETAS model, no foreshock feature is explicitly built into the model, but from among a given parent's offspring, one will occasionally appear with a magnitude larger than that of its parent – with frequencies again in the vicinity of 5%, approximately irrespective of the magnitude of the parent.

That just such an interpretation may apply also to real earthquakes is suggested in [25]. Foreshocks then are not a specific physical phenomenon, but just parent events which happen to have offspring larger than themselves. An interesting connection between the probability that an initial event is a foreshock (followed by a larger event as offspring), and the distribution of the Bath's law magnitude gap, is put forward in [25] and [38].

Further Topics

There are many further topics where stochastic modeling ideas are relevant, even if they do not necessarily involve the development and fitting of a full model. For example, the last few decades have seen considerable work on the development of procedures for producing and assessing probability forecasts, and for quantitatively describing spatial or space-time point patterns. In this section we list what seem to us to be some of the more important topics of this kind, space precluding more than very cursory accounts.

Generating Probability Forecasts

There is no longer any very hard and fast line between predictions and probability forecasts. It has long been recognized that any prediction of the time and place of a forthcoming event must be accompanied by some statement of the uncertainties in the prediction. But this, as discussed in the introduction, is precisely the motivation for introducing stochastic models. Using such a model, the uncertainties can be rephrased in terms of the probabilities of occurrence within specified time intervals and spatial regions, i.e. by probability forecasts. In our view such probability forecasts represent the most useful way of summarizing information about the uncertainties regarding future events. In statistical jargon, they represent the 'predictive distributions' for future events, and form the basis not only for the probability forecasts themselves, but also for any associated cost-benefit analyses.

Even within a probabilistic framework, the idea that precise, medium-term or even short-term forecasts may be possible has looked increasingly like a pipe dream over the last few decades. Nevertheless, it is not yet entirely ruled out. While the emphasis in the last few decades has been on increased surveillance, improved knowledge and understanding of long-term hazards, and the reduction of risks from earthquake hazards, medium term (months to years) probability forecasts are winning a new role as refinements of more traditional long-term, static hazards for both building and insurance purposes. In addition very short term forecasts play a useful role in connection with

the progression of aftershock sequences, and in developing real-time warnings for trains, gas supplies, and other facilities at high risk from a serious earthquake.

All of the models outlined in the previous sections can be used to develop probability forecasts. Typically, such forecasts are derived from simulations rather than from analytical studies. For example, simulation schemes for use with conditional intensity models are described, with further references, in Chap. 7 of [20].

For complete models, forecasting schemes making use of such procedures involve first fitting a model on all the data up to the current time, then simulating the model as far as is desired into the future, and using these simulations to estimate any required probabilities or expected values as in a Monte Carlo study.

Forecasts for partial models can proceed along similar general lines, although in the first instance the dependence on auxiliary variables restricts the probability forecasts to just the next forecasting period. To obtain forecasts beyond this first step, and in the absence of any updated precursor data, it is necessary to develop a further set of 2-step forecasts, using only the current precursor data, and so on successively, the forecasts gradually reducing in power.

Assessing Probability Forecasts

Two main approaches to assessing probability forecasts are through likelihoods and probability gains, and through their performance in decision schemes based on the forecast probability exceeding some threshold. Since most procedures are reduced in practice to forecasts for finite forecasting periods, we outline assessment procedures for this case only.

The *probability gain* for each forecast (i. e. for each time, time-space or other interval for which a probability is forecast) is the ratio of the forecast probability for the observed occurrence in that interval to the corresponding probability from a standard reference model, such as a simple or compound Poisson process, for that interval.

The sample averages of the log-probability-gains, or *entropy scores* in the terminology of [34], over all or certain classes of intervals or of observed outcomes, provide useful insights into the performance of the forecasts. From them one can quickly perceive the outcomes which the scheme is forecasting well, and those which it is forecasting badly.

The overall average, or more properly the expected value, of these entropy scores is called in [34] the *information gain* relative to the reference model. In a complete model, the information gain is a numerical characteristic of the model, giving an upper bound for the improvement in performance that can be expected for the proposed

model, relative to the reference model, when the proposed model is the true model. In many cases it reduces to the expected value of the mean likelihood ratio. For example, if successive forecasts, derived from the use of a particular model, give probabilities $p_1^*, p_2^*, \dots, p_N^*$ for the successive observed outcomes, and \bar{p} is a constant probability used as reference, then the sum

$$(1/N) \sum_{n=1}^N \log[p_n^*/\bar{p}]$$

is the empirical entropy score for that model as well as the mean log-likelihood ratio. It approximates the corresponding expected value (the information gain) if the model used is the true model. If the true model is unknown, as will usually be the case in practice, the above average approximates the difference in the Kullback–Leibler distances between the given model and the true model, and the reference model and the true model. For a set of models, that giving the largest entropy score should be that closest to the true model in the sense of Kullback–Leibler distance. In a partial model, expectations cannot be taken over values of the auxiliary variable, since its distribution is not included in the model specification, but at least empirical averages based on past observations can be developed. [34] gives further background and examples; the basic idea of using log-probabilities (i. e. loglikelihoods) as an indicator of forecasting performance goes back at least to [52].

We noted in Subsect. “[Example 2: Discrete-Time Poisson-Type Model for ULF Electric Signals](#)” that such comparisons of likelihood ratios form the basis of the assessment procedures used in the RELM testing center for probability forecasting schemes; see further [111].

The more traditional procedures for assessing probability forecasts suppose that the schemes are first turned into prediction schemes by predicting that an event will occur whenever the forecast probability of that event exceeds a certain threshold value. The results can then be put into a 2 x 2 table (occurrences or non-occurrences, versus predictions or non-predictions). If the entries in the table are labeled as

- (a) number of successful forecasts of occurrence,
- (b) number of failures to predict,
- (c) number of successful forecasts of non-occurrence,
- (d) number of false alarms,

the commonly used *R-score*, or *Hanssen–Kuiper skill score*, can be defined as

$$R = \frac{ac - bd}{(a + b)(c + d)} = \frac{a}{a + b} - \frac{d}{c + d} = \frac{c}{c + d} - \frac{b}{a + b}. \quad (20)$$

The R -score varies between -1 and $+1$, with 0 denoting forecasts independent of outcomes, and the two extreme values perfect non-prediction and perfect prediction, respectively. [91] illustrates the use of the R -score to evaluate Chinese yearly forecasts, even though these are based on expert opinion rather than any probability model.

The usefulness of this approach was greatly extended in [84,85], by allowing the threshold probability p_c to vary and calculating for each such value the ratios

$$\nu(p_c) = \frac{c}{(a+c)}$$

and

$$\tau(p_c) = \frac{(a+b)}{(a+b+c+d)}$$

The resulting $\nu - \tau$ diagram, obtained by plotting these quantities against each other, provides a comprehensive summary of the behavior of the probability forecasting scheme. The diagram typically consists of a convex (downwards) curve reducing to the diagonal joining the points $(0, 1)$ and $(1, 0)$ in the case of purely random forecasts, and to the two $(0, 1)$ segments of the axes in the case of perfect prediction. It is essentially a Q-Q (quantile-quantile) plot of the distribution of the proportion of time on trial, and the the proportion of failures to predict. See the papers quoted or [34] for further details.

It should be emphasized that the procedures described are concerned with the scientific issues of assessing the quality of proposed models rather than the practical issues of issuing and using probability forecasts. The latter raise just as many difficulties, if not more, than the former. Among these are the need to take into account errors in the model as well as the uncertainties described by the model itself, the need to develop decision-making frameworks that can take advantage of the information within probability forecasts, and the need to address the social, political and economic consequences of issuing forecasts.

Change-Point Models

Change-point models are used retrospectively to identify time-points at which a change occurs in quantities such as a mean value or rate. They are not themselves precursors, but are used rather to indicate the onset of a period which is anomalous in some sense, and may therefore have some precursory significance. In practice, one of the biggest difficulties in using such methods with seismicity data lies not in identifying the change point, but determining the region from which the data for the change-point analysis should taken.

The procedure consists essentially of dividing an observation time period, say $(0, t)$, into two segment $(0, t_1)$ and (t_1, t) , and finding the value of t_1 which maximizes the discrepancy in the values of the quantity being studied. The value of this maximized discrepancy can then be tested for significance using the null hypothesis that the values in both periods are equal.

Such a change-point technique is developed in [86] to detect the onset of precursory quiescence within a selected observation region. The data is first fitted to the ETAS model. The time axis is then transformed by the random time transformation

$$\tau = \int_0^t \lambda^*(u) du$$

which has the effect of transforming the original point process, with conditional intensity $\lambda^*(t)$, into a unit rate Poisson process (see, for example, Sect. 7.4 in [20]). If, however, there is a change in the parameters of the original ETAS model, this will show up as a change in the rate or other perturbation in the unit rate Poisson process, for which many tests are available.

This technique can be applied both to background events in a specified region (e.g. [140]), or to events within the course of an aftershock sequence, anticipating the occurrence of a large aftershock (e.g. [77]).

Ogata, Toda and colleagues make use of similar but more sophisticated procedures for identifying the regions and onset times of stress-shadowing through careful monitoring of activity of small events; see in particular [89,91].

A different type of change-point analysis has been developed by W. Smith for detecting changes in b -value. Although one of the earliest possibilities suggested, the precursory role of b -value changes as a precursor has never been unambiguously identified. Early papers such as [82] or [109] suggested a link to physical features such as heterogeneity of the crustal material, or changing levels of stress. The latter interpretation is at least partially supported by the branching and similar models for earthquake mechanism outlined in Sect. “Branching Models”.

Smith’s approach is based on modifications of the CUSUM procedure widely used in quality control contexts for detecting departures from normal behavior in a production process. They are essentially methods for detecting a change in slope of cumulative occurrences or other similar sums. For further details and reviews of earlier work on b -value changes, see [117,118].

Finally, there is the possibility of a Bayesian approach to change-point problems in seismology; some discussion and an example are given in [94].

Change point models are closely related to *hidden Markov models* in which the rates and other characteristics of the observed process change with the state of a Markov process which is hidden or at best only partly observable. Such models have been widely used in speech recognition and modeling of IT traffic (see, e. g. [72]). Their use in seismology is relatively new (see [23] for a recent example) but their potential seems worthy of further exploration.

Moment Measures and Correlation Functions

Moment structure plays an important role in most stochastic processes, and the same is true for point processes. The main attention is on second order or correlation properties. Indeed, where distributions are more or less Gaussian, the distinction between models and second order properties is largely nominal, since a Gaussian model is fully described by its means and covariances. The same is not true for count data, for which the second order properties form an important but not in general a definitive aspect of the overall model structure, since a range of different models can be developed to fit particular second order characteristics.

Early papers in the seismological context include [16, 53, 128]. There is also a considerable literature in astrophysics, relating especially to the distribution of galaxies and the role of 2-point, 3-point and higher order correlation functions (e. g. [76]).

Second-order properties for point processes mean properties of point-pairs. Their behavior is described through the second moment measure, whose density (when it exists) is given for $x_1 \neq x_2$ by

$$m_2(x_1, x_2)dx_1dx_2 = E[N(dx_1)N(dx_2)] .$$

Apart from a renormalization, this is also the 2-point correlation function of [53]. It is also the basis of the covariance measure, with density

$$c(x_1, x_2) = m_2(x_1, x_2) - m_1(x_1)m_1(x_2) \quad (x \neq y) .$$

This also exists in various renormalized forms, for example the radial correlation function defined, when the process is isotropic, by

$$\rho(r) = dK(r)/dA(r)$$

where $A(r) = \pi r^2$ is the area of a circle centered on an arbitrary point of the process, and $K(r)$ (often referred to as ‘Ripley’s K-function’) is the expected number of additional points (i. e. apart from the point at the center) in the same circle (see e. g. [22, 104]).

In a general treatment, the coordinates x_1, x_2 may combine both space and time components. The second moment densities then give information about the expected density of occurrence of additional points at given time or space intervals about a given point taken as origin. In this way they can display distance variations in the strength of the clustering tendency.

The second order techniques are particularly valuable when the process is stationary in time, so that spectral methods can be used. In this situation the covariance measure becomes a function of time through the difference $t_1 - t_2$ in the time coordinates of the two points being considered:

$$c_2(x_1, x_2) = c_2(t_1 - t_2; y_1, y_2) ,$$

where y_1, y_2 may represent locations or magnitudes. Taking Fourier transforms with respect to time leads to a spectral density, multivariate if the dependence on space or magnitude is retained, which can be used for the analysis of periodic effects in point processes. [133] is an early example of the spectral analysis of earthquake occurrence data. More general discussions can be found in [13] or Chap. 8 of [20].

Principal Component Analysis

Principal components are the names given to the eigenvectors in the orthogonal decomposition of a model-derived or empirical covariance matrix. Being symmetric and positive definite, the diagonal representation of such a matrix has non-negative eigenvalues, which measure the proportion of variation associated with the given eigenvector. Thus the principal components associated with the largest eigenvalues define those linear combinations of the observation vector components which explain the greatest amount of variability.

In geophysical applications, notably in meteorology, the observations typically arise as time sequences of observations from a set of recording stations. Each time point gives a vector observation. The principal components associated with the largest eigenvalues then often describe recognizable weather occurrence patterns, and are used as a means for identifying and classifying such patterns.

This idea admits many specializations and extensions. For stationary processes in time, for example, the covariance matrix has a special structure ($c_{i,j} = c_{i-j}$) which for infinite series, or a finite series with periodic structure, causes the eigenvectors to reduce to complex exponentials $e^{i\omega_r n}$. These periodic terms form the principal components of the sequence, while the corresponding eigenvalues are amplitudes associated with the given frequen-

cies. Thus standard spectral analysis can be interpreted as a special form of principal component analysis. The so-called Karhunen–Loeve theory (see e. g. [69]) is an extension to very general covariance operators admitting a diagonal decomposition.

Adaptations of these ideas for earthquake data have recently been considered by Rundle, Tiampo and colleagues (see e. g. [107,120]). Here the data is typically of point process form (0-1 functions) on a sequence of observation regions, centered either on a lattice of space or space-time points, or on the centers of a family of small fault segments over an interacting fault system. Their papers combine the Karhunen–Loeve theory with ideas originating in dynamical systems theory, and seek to extend the analysis to provide short or medium term probability forecasts.

Fractals and Fractal Dimensions

Many of the models so far considered show long-range dependence, power-law distributions, and some form of self-similarity (similar appearances on different scales). The term ‘fractal’ is commonly used in this context, with a variety of interrelated interpretations. A wide-ranging review of such properties for earthquakes is given in [121].

In the basic text [74] of Mandelbrot, fractals are introduced as a class of geometric objects (sets) with features which repeat on a family of reducing scales. The Cantor set in one dimension and the snowflake curve in 2 dimensions are well-known examples. Their characteristic feature is that the Hausdorff dimension of the fractal set may differ from the dimension of the space in which it is embedded. Thus the Cantor set has dimension $\log_3(2) < 1$ despite being embedded in a 1-dimensional space. As in this example, these aberrant dimensions often have a non-integer value, whence the name.

Various empirical procedures have been devised for determining these fractional dimensions. For example, for a set in 2 dimensions, we may count the proportion of cells from a 2-dimensional lattice which have a non-empty intersection with the set, and consider the behavior as the cell size in the lattice approaches zero. Typically this results in a log-log relation between number of affected cells and the cell dimension, corresponding in fact to power-law behavior. Such methods can be applied also to physical sets such as the traces of faults on the earth’s surface (see [121] and [45] for examples). If a suitably linear portion of the log-log plot can be found, the slope is used as an estimate of their fractal dimension.

Because power-laws arise inevitably in the study of fractal sets, the term *fractal behavior* has come to be used loosely to describe the behavior of any probability distri-

bution for which

$$1 - F(x) = \Pr(X > x) \sim cx^{-\alpha},$$

or even when power-law forms arise in features such as time or space correlation functions.

A different fractal concept derives from the work of Renyi [99]. Renyi’s dimension estimates (also called multifractal dimensions) reflect the irregularities of a measure (distribution) rather than a set. Thus, if a probability distribution on a square is used as an example, and the square is divided into a lattice of small squares Δ_i each of side Δ , we may consider the limit as $\Delta \rightarrow 0$ of the ratios

$$d_q(\Delta) = \left[\log \sum_i p(\Delta_i)^q \right] / \log \Delta \quad (q \neq 1) \quad (21)$$

with

$$d_1(\Delta) = \left[\sum_i p(\Delta_i) \log p(\Delta_i) \right] / \log \Delta. \quad (22)$$

For each q , d_q defines a *multifractal dimension* associated with the distribution, and for varying q the family of such dimensions defines a type of transform of the underlying probability distribution.

The term multifractal is used because the distribution may exhibit different forms of singularity, associated with different power-law behavior, at different points of the space. In applications, for example in turbulence and meteorology as well as seismology, the process is visualized as a descending sequence of random eddies (‘random cascades’) with dimensions shrinking to zero. It is in this sense that Kagan [49] coined the phrase ‘frozen turbulence’ to describe earthquake processes in the crust.

The computations in the equations above can be applied empirically to quantities such as the counts or the energy release from earthquakes and used to estimate the multifractal dimensions of some underlying spatial distribution. For general discussions of multifractals, see [95] and [31]; the latter gives an extended discussion of applications of the theory to earthquake data.

With count data, there is a link between the multifractal dimension d_q for integer q and the q -point correlation function. In particular d_2 describes the growth rate of the two-point correlation function for vanishingly small separations; see [53,131].

Self-Similarity

Self-similarity is used in a broad sense to describe processes in space or time where, much as in a fractal set, key structural features are preserved on a descending sequence

of scales. Plots of fault traces or epicenters do indeed suggest such scale invariance properties.

In stochastic process theory, self-similarity (auto-modeling) is used to denote a precisely defined property of random processes or random measures. In particular, a random measure $\xi(A)$ defined on sets in the plane (or other Euclidean space) is said to be *self-similar*, with similarity index H , if a change of scale in the space can be compensated for by a change of scale in the quantity being measured:

$$\xi(rA) =^D r^{-H} \xi(A), \quad r > 0 \quad (23)$$

where $=^D$ means the two sides have equal probability distributions.

Self-similar random measures of this form arise in turbulence theory, seismology, finance, and elsewhere. In one dimension, the concept can be reinterpreted as implying that a stochastic process has *self-similar increments*. Brownian and fractional Brownian motions both have increments of this type. However, the Brownian motions take both positive and negative values, whereas any random measures associated with earthquakes (for example, by way of energy release) should be non-negative.

Thus the self-similar processes arising in seismology have a rather special character, since they must be non-negative as well as self-similar. The only known examples are purely atomic random measures: $\xi(A)$ is obtained by summing the contributions (energies, for example) from point-events in A .

One such example is well-known: the *stable* random measures, which are characterized by a strong independence property, and a power law distribution for event sizes equivalent to a G-R relation for magnitudes. Superficially, this example matches the earthquake data rather well, but it fails to incorporate one key feature of earthquake occurrence: the lack of independence inherent in earthquake clustering. Recently, however, it was shown in [132] that it is possible to define a variant on the ETAS model which combines self-similarity with a non-trivial dependence (cluster) structure. It is not known at present how wide this class of models may be.

Block-Slider and Related Mechanical Models

The mechanical models attempt to illuminate the processes of earthquake occurrence by devising complex mechanical systems that will reproduce many observed features of earthquake catalogs. The reason for including a brief section on such models in the present article is that one of their underlying purposes is to demonstrate that deterministic models of sufficient complexity can exhibit just

that random-appearing behavior that characterizes the appearance of earthquake data. Two important issues then arise. The first is whether the mechanical models can be matched closely enough to any real seismic system to produce forecasts similar to or better than those produced by the stochastic models. The second is whether those stochastic models currently fitted to earthquake catalogs also provide good descriptions of the catalogs produced by the mechanical models. Such studies may then suggest ways in which the stochastic models themselves can be improved.

In the pioneering article [14], the mechanical system comprises a series of blocks, linearly connected by springs, which are pulled over a rough plane. The resultant movement is not smooth, but jerky, as the tension in one of the springs builds up to the point where it overcomes the frictional forces opposing motion and a slip occurs. Often movement of one block will initiate movement in a whole set of blocks.

The population of such movement sequences is then compared to a population of earthquakes. Although the motion of the system is entirely deterministic, analytic solutions are too complex to obtain in explicit form, and the behavior is irregular, with features of (pseudo)randomness. In particular a crude form of GR law generally results.

Many further mechanical models have been invented and studied, both in the laboratory and via numerical solution of the appropriate dynamical equations. Often the masses of the blocks or the spring parameters are selected at random from some plausible overall population. It is not this randomness which is the cause of the random-looking behavior, however, but rather the complexity of the movements (trajectory) of any particular system of blocks and springs or other similar elements.

One of the difficulties in matching such models to particular fault systems is to match the initial conditions, which are not easy to determine for real processes, even if they can be reproduced in the artificial system. It is precisely the relative simplicity of the stochastic model, which comes at the expense of describing many important physical details, which allows it to be fitted to catalogue data and then used to produce rough probabilistic forecasts.

In [70], one of the relatively few studies to fit point process models to the output from such mechanical systems, outputs were taken from the four different models described in [6] (see also [10]) and fitted by the simple stress release model. The four models illustrated different patterns of behavior, varying from highly random behavior with typical GR distributions, to regular, characteristic earthquake behavior. After suitable adjustment of the fre-

quency-magnitude law in the stress-release model, it was found that the stress release model could be fitted to all four versions, and used to describe the overall energy state of the mechanical system; as such it operated as a crude predictor of the next major event in the system.

Future Directions

There are many possible directions in which the applications of stochastic models in seismology could be extended and deepened. The fundamental limitations in the past have most commonly been limitations in data, and consequent limitations in the physical understanding of the processes. It has been the great improvement in data which has allowed the development of better and more insightful seismicity models in the last two decades; this tendency is continuing strongly and may well lead to unexpected new developments.

Two directions in particular appear to me to hold out scope for further development of stochastic models for earthquake occurrence. The first is related to the collection and integration of data on earth deformation. The extensive data now becoming available from GPS measurements have already led to new discoveries, for example 'slow earthquakes' which relieve strain without being registered on conventional seismometers. The underlying problem is how to link the data on strain to data on seismicity. This is not easy, and is likely to require new ideas on both the physical and statistical sides.

The second relates to the systematic collection and evaluation of data on potential precursor events. Many of the ideas initiated in the 1970's have been abandoned, with the result that for most potential phenomena there exists no substantial body of data by which their effectiveness can be adequately tested, or the underlying physical processes modeled. Many controversial and so far inadequately explained phenomena fall into this category. They have the potential to generate projects of interest and importance from the physical as well as the hazard estimation viewpoints. Unfortunately collecting and archiving such data is a long-term process with uncertain future outcomes, and therefore difficult to fund under current funding criteria. The balance of scientific, public and even government opinion may change, however, and further work in these fields may be anticipated.

Fundamental work on earthquake mechanism – still a largely unsolved problem – is also likely to attract attention during the next decades, and to require a combination of physical and statistical modeling.

If these are rather long-term developments, there are many smaller scale, more immediate problems that re-

quire further investigation. On the theoretical side, one important issue is the development of improved models and procedures for analyzing data with self-similar characteristics. Another area where further research is needed, particularly in subduction regions, is in the development of better physical and statistical models for deep earthquakes, including their possible links to different forms of activity (seismic, volcanic etc) closer to the surface.

The current interest in developing testing centers for probability forecasting is likely to promote more strenuous efforts to develop improved forecasting models, whether based just on catalogue data, or allied with deformation data, or with data on other precursory phenomena. Even the existing models are capable of producing time-varying forecasts which in principle could lead to significant reductions in earthquake risk. The problem here is how to realize these reductions in practice, for example through improved insurance or disaster mitigation activities. Many questions arise, of an operations research as much as a statistical character, which warrant further study and effort. The probability forecasts themselves, if they are to be useful in such contexts, need to incorporate the uncertainties in the underlying models and hence to take on a more explicitly Bayesian character, and to be presented in such a way that they can be related to the many additional factors that have to be borne in mind when making real-life decisions.

Acknowledgments

I am very grateful to friends and colleagues, especially David Harte, Mark Bebbington, David Rhoades and Yehuda Ben-Zion, for helpful discussions, correcting errors and plugging gaps.

Bibliography

1. Ambraseys NN, Melville CP (1982) *A History of Persian Earthquakes*. Cambridge University Press, Cambridge
2. Andersen PK, Borgan Ø, Gill RD, Keiding N (1993) *Statistical Models Based on Counting Processes*. Springer, New York
3. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94:15635–15637
4. Bebbington M, Harte DS (2003) The linked stress release model for spatio-temporal seismicity: formulations, procedures and applications. *Geophys J Int* 154:925–946
5. Bebbington M, Vere-Jones D, Zheng X (1990) Percolation theory: a model for earthquake faulting? *Geophys J Int* 100:215–220
6. Ben-Zion Y (1996) Stress, Slip and earthquakes in models of complex single-fault systems incorporating brittle and creep deformations. *J Geophys Res* 101:5677–5706
7. Ben-Zion Y, Dahmen K, Lyakhowsky V, Ertas D, Agnon A (1999) Self-driven mode-switching of earthquake activity on a fault system. *Earth Planet. Sci Lett* 172:11–21

8. Ben-Zion Y, Eneva M, Liu Y (2003) Large earthquake cycles and intermittent criticality on heterogeneous faults due to evolving stress and seismicity. *J Geophys Res* 108:2307V. doi:10.1029/2002JB002121
9. Ben-Zion Y, Lyakhovsky V (2002) Accelerated seismic release and related aspects of seismicity patterns on earthquake faults. *Pure Appl Geophys* 159:2385–2412
10. Ben-Zion Y, Rice J (1995) Slip patterns and earthquake populations along different classes of faults on elastic solids. *J Geophys Res* 100:12959–12983
11. Borovkov K, Vere-Jones D (2000) Explicit formulae for stationary distributions of stress release processes. *J Appl Prob* 37:315–321
12. Brémaud P, Massoulié L (2001) Hawkes branching processes without ancestors. *J App Prob* 38:122–135
13. Brillinger DR (1981) *Time Series: Data Analysis and Theory*, 2nd edn. Holden Day, San Francisco
14. Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57:341–371
15. Chelidze TL, Kolesnikov YM (1983) Modelling and forecasting the failure process in the framework of percolation theory. *Izvestiya Earth Phys* 19:347–354
16. Chong FS (1983) Time-space-magnitude interdependence of upper crustal earthquakes in the main seismic region of New Zealand. *J Geol Geophys* 26:7–24, New Zealand
17. Console R, Lombardi AM, Murru M, Rhoades DA (2003) Båth's Law and the self-similarity of earthquakes. *J Geophys Res* 108(B2):2128V. doi:10.1029/2001JB001651
18. Cox DR (1972) Regression models and life tables (with discussion). *Roy J Stat Soc Ser B* 34:187–220
19. Dahmen K, Ertas D, Ben-Zion Y (1998) Gutenberg-Richter and characteristic earthquake behavior in simple mean-field models of heterogeneous faults. *Phys Rev E* 58:1494–1501
20. Daley DJ, Vere-Jones D (2003) *An Introduction to the Theory of Point Processes*, 2nd edn, vol I. Springer, New York
21. Davison C (1938) *Studies on the Periodicity of Earthquakes*. Murthy, London
22. Diggle PJ (2003) *Statistical Analysis of Spatial Point Patterns*. 2nd edn. University Press, Oxford
23. Ebel JB, Chambers DW, Kafka AL, Baglivo JA (2007) Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California. *Seismol Res Lett* 78:57–65
24. Evison F, Rhoades D (2001) Model of long-term seismogenesis. *Annali Geofisica* 44:81–93
25. Felzer KR, Abercrombie RE, Ekström G (2004) A common origin for aftershocks, foreshocks and multiplets. *Bull Amer Seismol Soc* 94:88–98
26. Fisher RL, Dahmen K, Ramanathan S, Ben-Zion Y (1997) Statistics of earthquakes in simple models of heterogeneous faults. *Phys Rev Lett* 97:4885–4888
27. Griffiths AA (1924) Theory of rupture. In: *Proceedings 1st Int Congress in Applied Mech*, Delft, pp 55–63
28. Gutenberg B, Richter C (1949) *Seismicity of the Earth and Associated Phenomena*, 2nd edn. University Press, Princeton
29. Habermann RE (1987) Man-made changes of seismicity rates. *Bull Seismol Soc Am* 77(1):141–159
30. Hainzl S, Ogata Y (2005) Detecting fluid signals in seismicity data through statistical earthquake modelling. *J Geophys Res* 110. doi:10.1029/2004JB003247
31. Harte D (2001) *Multifractals: Theory and Applications*. Chapman and Hall/CRC, Boca Raton
32. Harte D, Li DF, Vreede M, Vere-Jones D (2003) Quantifying the M8 prediction algorithm: reduction to a single critical variable and stability results. *NZ J Geol Geophys* 46:141–152
33. Harte D, Li D-F, Vere-Jones D, Vreede M, Wang Q (2007) Quantifying the M8 prediction algorithm II: model, forecast and evaluation. *NZ J Geol Geophys* 50:117–130
34. Harte D, Vere-Jones D (2005) The entropy score and its uses in earthquake forecasting. *Pure Appl Geophys* 162:1229–1253
35. Hawkes AG (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika* 58:83–90
36. Hawkes AG, Oakes D (1974) A cluster representation of a self-exciting process. *J Appl Prob* 11:493–503
37. Helmstetter A, Sornette D (2002) Subcritical and supercritical regimes in epidemic models of earthquake aftershocks. *J Geophys Res* 107:2237. doi:10.1029/2001JB001580
38. Helmstetter A, Sornette D (2003) Båth's law derived from the Gutenberg-Richter law and from aftershock properties. *Geophys Res Lett* 103(20):2069. doi:10.1029/2003GL018186
39. Ishimoto M, Iida K (1939) *Bull Earthq Res Inst Univ Tokyo* 17:443–478
40. Iwata T, Young RP (2005) Tidal stress/strain and the b-values of acoustic emissions at the Underground Research Laboratory, Canada. *Pure Appl Geophys* 162:(6–7):1291–1308. doi:10.1007/s00024-005-2670-2 (P*1357)
41. Jackson DD, Kagan YY (1999) Testable earthquake forecasts for 1999. *Seismol Res Lett* 70:393–403
42. Jaeger JC, Cook NGW (1969) *Fundamentals of Rock Mechanics*. Methuen, London
43. Jaume SC, Bebbington MS (2004) Accelerating seismic moment release from a self-correcting stochastic model. *J Geophys Res* 109:B12301. doi:10.1029/2003JB002867
44. Jeffreys H (1938) Aftershocks and periodicity in earthquakes. *Beitr Geophys* 53:111–139
45. Jeffreys H (1939) *Theory of Probability*, 1st edn (1939), 3rd edn (1961). University Press, Cambridge
46. Jones LM, Molnar P (1979) Some characteristics of foreshocks and their possible relationship to earthquake prediction and premonitory slip on a fault. *J Geophys Res* 84:3596–3608
47. Kagan Y (1973) Statistical methods in the study of the seismic process. *Bull Int Stat Inst* 45(3):437–453
48. Kagan Y (1991) Seismic moment distribution. *Geophys J Int* 106:121–134
49. Kagan Y (1991) Fractal dimension of brittle fracture. *J Non-linear Sci* 1:1–16
50. Kagan Y (1994) Statistics of characteristic earthquakes. *Bull Seismol Soc Am* 83:7–24
51. Kagan Y, Jackson DD (1994) Probabilistic forecasting of earthquakes. *Geophys J Int* 143:438–453
52. Kagan Y, Knopoff L (1977) Earthquake risk prediction as a stochastic process. *Phys Earth Planet Inter* 14:97–108
53. Kagan Y, Knopoff L (1980) Spatial distribution of earthquakes: the two-point correlation function. *Geophys J Roy Astronom Soc* 62:303–320
54. Kagan Y, Knopoff L (1981) Stochastic synthesis of earthquake catalogues. *J Geophys Res* 86:2853–2862
55. Kagan Y, Knopoff L (1987) Statistical short-term earthquake prediction. *Sci* 236:1563–1567

56. Keilis-Borok VI, Kossobokov VG (1990) Premonitory activation of the earthquake flow: algorithm M8. *Phys Earth Planet Inter* 61:73–83
57. Kiremidjian AS, Anagnos T (1984) Stochastic slip predictable models for earthquake occurrences. *Bull Seismol Soc Am* 74:739–755
58. Knopoff L (1971) A stochastic model for the occurrence of main sequence earthquakes. *Rev Geophys Space Phys* 9:175–188
59. Kossobokov VG (1997) User manual for M8. In: *Algorithms for Earthquake Statistics and Prediction*. IASPEI Softw Ser 6:167–221
60. Kossobokov VG (2005) Earthquake prediction: principles, implementation, perspectives. Part I of *Computational Seismology* 36, “Earthquake Prediction and Geodynamic Processes.” (In Russian)
61. Kossobokov VG (2006) Testing earthquake prediction methods: The West Pacific short-term forecast of earthquakes with magnitude $M_w \geq 5.8$. *Tectonophysics* 413:25–31
62. Libicki E, Ben-Zion Y (2005) Stochastic branching models of fault surfaces and estimated fractal dimensions. *Pure Appl Geophys* 162:1077–1111
63. Lombardi A (2002) Probabilistic interpretation of Båth’s law. *Ann Geophys* 45:455–472
64. Lomnitz CA (1974) *Plate Tectonics and Earthquake Risk*. Elsevier, Amsterdam
65. Lomnitz-Adler J (1985) Asperity models and characteristic earthquakes. *Geophys. J Roy Astron Soc* 83:435–450
66. Lomnitz-Adler J (1985) Automaton models of seismic fracture: constraints imposed by the frequency-magnitude relation. *J Geophys Res* 95:491–501
67. Lomnitz-Adler J (1988) The theoretical seismicity of asperity models; an application to the coast of Oaxaca. *Geophys J* 95:491–501
68. Liu J, Chen Y, Shi Y, Vere-Jones D (1999) Coupled stress release model for time dependent earthquakes. *Pure Appl Geophys* 155:649–667
69. Loève M (1977) *Probability Theory I*, 4th edn. Springer, New York
70. Lu C, Vere-Jones D (2001) Statistical analysis of synthetic earthquake catalogs generated by models with various levels of fault zone disorder. *J Geophys Res* 106:11115–11125
71. Lu C, Harte D, Bebbington M (1999) A linked stress release model for Japanese historical earthquakes: coupling among major seismic regions. *Earth Planet. Science* 51:907–916
72. Macdonald II, Zucchini W (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London
73. Main IG, Burton PW (1984) Information theory and the earthquake frequency-magnitude distribution. *Bull Seismol Soc Am* 74:1409–1426
74. Mandelbrot BB (1977) *Fractals: Form, Chance and Dimension*. Freeman, San Francisco
75. Mandelbrot BB (1989) Multifractal measures, especially for the geophysicist. *Pure Appl Geophys* 131:5–42
76. Martínez VJ, Saar E (2002) *Statistics of the Galaxy Distribution*. Chapman & Hall/CRC, Boca Raton
77. Matsu’ura RS (1986) Precursory quiescence and recovery of aftershock activities before some large aftershocks. *Bull Earthq Res Inst Tokyo* 61:1–65
78. Matsu’ura RS, Karakama I (2005) A point process analysis of the Matsushiro earthquake swarm sequence: the effect of water on earthquake occurrence. *Pure Appl Geophys* 162 1319–1345. doi:10.1007/s00024-005-2762-0
79. Matthews MV, Ellsworth WL, Reasenberg PA (2002) A Brownian model for recurrent earthquakes. *Bull Seism Soc Amer* 92:2232–2250
80. Merrifield A, Savage MK, Vere-Jones D (2004) Geographical distributions of prospective foreshock probabilities in New Zealand. *J Geol Geophys* 47:327–339, New Zealand
81. Michael A (1997) Test prediction methods: earthquake clustering versus the Poisson model. *Geophys Res Lett* 24:1891–1894
82. Mogi K (1962) Study of elastic shocks caused by the fracture of heterogeneous materials and its relation to earthquake phenomena. *Bull Earthq Res Inst Tokyo Univ* 40:125–173
83. Mogi K (1985) *Earthquake Prediction*. Academic Press, Tokyo
84. Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Plan Int* 61:84–98
85. Molchan GM, Kagan YY (1992) Earthquake prediction and its optimization. *J Geophys Res* 106:4823–4838
86. Ogata Y (1988) Statistical models for earthquake occurrence and residual analysis for point processes. *J Amer Stat Soc* 83:9–27
87. Ogata Y (1998) Space-time point process models for earthquake occurrences. *Annals Inst Stat Math* 50:379–402
88. Ogata Y (1999) Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes. *J Geophys Res* 104:17995–18014
89. Ogata Y (2005) Detection of anomalous seismicity as a stress change sensor. *J Geophys Res* 110(B5):B05S06. doi:10.1029/2004JB003245
90. Ogata Y, Utsu T, Katsura K (1996) Statistical discrimination of foreshocks from other earthquake clusters. *Geophys J Int* 127:17–30
91. Ogata Y, Jones L, Toda S (2003) When and where the aftershock activity was depressed: Contrasting decay patterns of the proximate large earthquakes in southern California. *J Geophys Res* 108(B6):2318. doi:10.1029/2002JB002009
92. Omori F (1894) On aftershocks of earthquakes. *J Coll Sci Imp Acad Tokyo* 7:111–200
93. Otsuka M (1972) A chain reaction type source model as a tool to interpret the magnitude-frequency relation of earthquakes. *J Phys Earth* 20:35–45
94. Pietavolo A, Rotondi R (2000) Analyzing the interevent time distribution to identify seismicity patterns: a Bayesian non-parametric approach to the multiple change-point problem. *Appl Stat* 49:543–562
95. Pisarenko DV, Pisarenko VF (1995) Statistical estimation of the correlation dimension. *Phys Lett A* 197:31–39
96. Reasenberg PA (1999) Foreshock occurrence before large earthquakes. *J Geophys Res* 104:4755–4768
97. Reasenberg PA, Jones LM (1989) Earthquake hazard after a mainshock in California. *Sci* 243:1173–1176
98. Reid HF (1911) The elastic-rebound theory of earthquakes. *Bull Dept Geol Univ Calif* 6:413–444
99. Renyi A (1959) On the dimension and entropy of probability distributions. *Acta Math* 10:193–215
100. Rhoades DA (2007) Application of the EEPAS model to forecasting earthquakes of moderate magnitude in Southern California. *Seismol Res Lett* 78:110–115

101. Rhoades DA, Evison FF (2004) Long-range earthquake forecasting with every event a precursor according to scale. *Pure Appl Geophys* 161:147–171
102. Rhoades DA, Evison FF (2005) Test of the EEPAS forecasting model on the Japan earthquake catalogue. *Pure Appl Geophys* 162:1271–1290
103. Rhoades DA, Van Dissen RJ (2003) Estimation of the time-varying hazard of rupture of the Alpine Fault of New Zealand, allowing for uncertainties. *NZ J Geol Geophys* 40:479–488
104. Ripley BD (1988) *Statistical Inference for Spatial Processes*. University Press, Cambridge
105. Robinson R (2000) A test of the precursory accelerating moment release model on some recent New Zealand earthquakes. *Geophys J Int* 140:568–576. doi:10.1046/j.1365-246X.2000.00054.x
106. Robinson R, Benites R (1995) Synthetic seismicity models for the Wellington region of New Zealand: implications for the temporal distribution of large events. *J Geophys Res* 100:18229–18238. doi:10.1029/95JB01569
107. Rundle JB, Klein W, Tiampo K, Gross S (2000) Dynamics of seismicity patterns in systems of earthquake faults. In: *Geocomplexity and the Physics of Earthquakes*. Geophysical Monograph 120, American Geophysical Union
108. Saito M, Kikuchi M, Kudo M (1973) An analytical solution of: Go-game model of earthquakes. *Zishin* 26:19–25
109. Scholz CH (1968) The frequency-magnitude relation of micro-faulting in rock and its relation to earthquakes. *Bull Seism Soc Am* 58:399–415
110. Scholz CH (1990) *The Mechanics of Earthquakes and Faulting*. Cambridge University Press, New York
111. Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007) Earthquake likelihood model testing. *Seismol Res Lett* 78:17–29
112. Schuster A (1897) On lunar and solar periodicities of earthquakes. *Proc Roy Soc London* 61:455–465
113. Schwartz DP, Coppersmith K (1984) Fault behavior and characteristic earthquakes: examples from the Wasatch and San Andreas Faults. *J Geophys Res* 89:5681–5698
114. Shi YL, Liu J, Chen Y, Vere-Jones D (1999) Coupled stress release models for time-dependent seismicity. *J Pure Appl Geophys* 155:649–667
115. Shi Y, Liu J, Zhang G (2001) An evaluation of Chinese annual earthquake predictions, 1990–1998. *J Appl Prob* 38A:222–231
116. Shimazaki K, Nakata T (1980) Time-predictable recurrence model for large earthquakes. *Geophys Res Lett* 7:179–282
117. Smith WD (1986) Evidence for precursory changes in the frequency-magnitude *b*-value. *Geophys J Roy Astron Soc* 86:815–838
118. Smith WD (1998) Resolution and significance assessment of precursory changes in mean earthquake magnitude. *Geophys J Int* 135:515–522
119. Stoyan D, Stoyan H (1994) *Fractals, Random Shapes and Point Fields*. Wiley, Chichester
120. Tiampo KF, Rundle JB, Klein W, Ben-Zion Y, McGinnis SA (2004) Using eigenpattern analysis to constrain seasonal signals in Southern California. *Pure Appl Geophys* 16:19–10, 1991 V2003. doi:10.1007/s00024-004-2545-y
121. Turcotte DL (1992) *Fractals and Chaos in Geology and Geophysics*. Cambridge University Press, Cambridge
122. Utsu T (1961) A statistical study on the properties of aftershocks. *Geophys Mag* 30:521–605
123. Utsu T, Ogata Y (1997) *IASPEI Softw Libr* 6:13–94
124. Utsu T, Ogata Y, Matu'ura RS (1995) The centenary of the Omori formula for a decay law of aftershock activity. *J Phys Earth* 43:1–33
125. Vere-Jones D (1969) A note on the statistical interpretation of Båth's law. *Bull Seismol Soc Amer* 59:1535–1541
126. Vere-Jones D (1970) Stochastic models for earthquake occurrence. *J Roy Stat Soc B* 32:1–62
127. Vere-Jones D (1977) Statistical theories for crack propagation. *Pure Appl Geophys* 114:711–726
128. Vere-Jones D (1978) Space-time correlations of microearthquakes – a pilot study. *Adv App Prob* 10:73–87, supplement
129. Vere-Jones D (1978) Earthquake prediction: a statistician's view. *J Phys Earth* 26:129–146
130. Vere-Jones D (1995) Forecasting earthquakes and earthquake risk. *Int J Forecast* 11:503–538
131. Vere-Jones D (1999) On the fractal dimension of point patterns. *Adv Appl Prob* 31:643–663
132. Vere-Jones D (2003) A class of self-similar random measures. *Adv Appl Prob* 37:908–914
133. Vere-Jones D, Davies RB (1966) A statistical analysis of earthquakes in the main seismic region of New Zealand. *J Geol Geophys* 9:251–284
134. Vere-Jones D, Ozaki T (1982) Some examples of statistical inference applied to earthquake data. *Ann Inst Stat Math* 34:189–207
135. Vere-Jones D, Robinson R, Yang W (2001) Remarks on the accelerated moment release model for earthquake forecasting: problems of simulation and estimation. *Geophys J Int* 144:515–531
136. von Bortkiewicz L (1898) *Das Gesetz der kleinen Zahlen*. Teubner, Leipzig
137. Weibull W (1939) A statistical theory of the strength of materials. *Ingvetensk Akad Handl* no 151
138. Working Group on Californian Earthquake Probabilities (1990) Probabilities of earthquakes in the San Francisco Bay region of California. US Geological Survey Circular 153
139. Yin X, Yin C (1994) The precursor of instability for non-linear systems and its application to the case of earthquake prediction – the load-unload response ratio theory. In: Newman WI, Gabrielov AM (eds) *Nonlinear dynamics and Predictability of Natural Phenomena*. AGU Geophysical Monograph 85:55–66
140. Zheng X, Vere-Jones D (1994) Further applications of the stress release model to historical earthquake data. *Tectonophysics* 229:101–121
141. Zhuang J (2000) Statistical modelling of seismicity patterns before and after the 1990 Oct 5 Cape Palliser earthquake, New Zealand. *NZ J Geol Geophys* 43:447–460
142. Zhuang J, Yin X (2000) The random distribution of the loading and unloading response ratio under the assumptions of the Poisson model. *Earthq Res China* 14:38–48
143. Zhuang J, Ogata Y, Vere-Jones D (2004) Analyzing earthquake clustering features by using stochastic reconstruction. *J Geophys Res* 109(B5):B05301. doi:10.1029/2003JB002879
144. Zhuang J, Vere-Jones D, Guan H, Ogata Y, Ma L (2005) Preliminary analysis of precursory information in the observations on the ultra low frequency electric field in the Beijing region. *Pure Appl Geophys* 162:1367–1396. doi:10.1007/s00024-004-2674-3

Earthquake Scaling Laws

RAUL MADARIAGA

Ecole Normale Supérieure, Laboratoire de Géologie,
Paris, France

Article Outline

Glossary

Definition of the Subject

Introduction

Earthquakes and Seismic Radiation

Earthquake Fault Models:

The Scaling of Geometry and Stress

Earthquake Dynamics and the Scaling of Energy

Kinematics and Statistical Models for Fault Slip

Future Directions

Acknowledgments

Bibliography

Glossary

Seismic moment The most fundamental measure of the size of an earthquake. In the simplest situation it represents the moment of one of the couples of forces that make up a dipolar source. In more general cases it is a 3 by 3 symmetric tensor of elementary force couples.

Seismic radiation The seismic waves emitted by a seismic source. For point sources these are spherical P and S waves emitted by the point tensor source.

Seismic spectrum The absolute value of the Fourier transform of the displacement field radiated by an earthquake in the far field. For almost all earthquakes it has a common shape: flat at low frequencies and decays like the inverse squared power at high very high values of frequency.

Corner frequency The low and high frequency asymptotes of the earthquake spectrum intersect at a characteristic frequency, called the corner frequency. The corner frequency scales with the size of the earthquake measured by the seismic moment.

Radiated or seismic energy Total energy of the seismic waves radiated by a seismic source. It can be computed from the energy flow relatively far from the source of the earthquake.

Apparent stress Originally defined as the product of seismic efficiency times the average stress during earthquake slip. In practice, it is computed from the ratio of radiated energy to moment release of the earthquake multiplied by the shear modulus.

Energy release rate Amount of energy per unit surface used to make a rupture advance by a unit distance.

Static stress drop The static change in shear traction between the sides of the fault occurs during an earthquake. In principle, it could be determined by measuring stress before and after the earthquake. In practice stress drop is computed using very specific source models, like a circular crack.

Dynamic stress drop The stress change in shear traction as a function of time while the rupture is still growing. It can only be estimated from seismic records obtained in the near field by elaborate inversion schemes. The relation between static and dynamic stress drop can only be estimated once the friction law between the sides of the fault has been defined.

Definition of the Subject

Earthquake scaling laws provide some of the most basic knowledge about seismic sources. Since the end of the 70s, a very successful model for earthquakes was developed by seismologists. In this model earthquakes are due to rapid slip on pre-existing faults driven by steady load due to plate motion and resisted by friction between the fault walls. This model may be used to predict many of the general properties of seismic radiation that can be derived from a simple spectral shape of type omega-squared. In this article I derive general expressions for energy, moment and stress in terms of measured spectral parameters. The available data shows that earthquakes can be reduced to a single family in terms of three parameters: moment, corner frequency and radiated energy. Using specific models of rupture these three parameters can be reinterpreted in terms of moment, size and stress drop. Although details differ between the models proposed by seismologists, both seismic spectra and the wave-number spectra of slip distributions can be explained with a simple circular crack model. This does not mean that a circular crack is the best earthquake model; it means that the ensemble average of seismic sources has properties that are similar to those of simple circular shear cracks. A direct result of scaling laws is that total fracture energy must scale like radiated and strain release energy, so that fracture energy should scale with fault size as observed for many earthquakes and in certain laboratory experiments.

Introduction

It has been 40 years since Aki [4] published his seminal paper on the scaling law of earthquakes that established that to first order seismic moment scales like the third power of the fault size. This paper came just a year after Aki [3]

made the very first measurement of seismic moment, the torque of one of the couples that make up a basic source mechanism. Not much later, in 1970, Brune introduced a very simple source model and established a generalization of Aki's [4] spectral model of earthquakes that is now known as the omega-squared model of earthquakes. Almost simultaneously, Kostrov [43], Savage [66], Sato and Hirasawa [65] and Madariaga [46] developed models of a circular faults. Radiation from a circular crack explained well the omega-squared model. Digital data was not available at the time when these models were introduced. It is nowadays an almost standard observatory practice to study the scaling law, but as many recent studies have shown, proper estimation of spectral parameters like moment and corner frequency and high frequency decay for a large range of earthquake sizes is not trivial and it is often difficult to obtain from a single instrument. In the 1990s good quality digital data, both broad band seismograms and accelerograms, became available not just from surface instruments, but also from boreholes opening the way to a new appraisal of the scaling law.

We will briefly review some of these observations and we will try to establish general properties of earthquake radiation based on the recent work by McGarr [53], Abercrombie and Rice [2], Ide et al. [35], Prieto et al. [59], etc. Our purpose is not to review the very extensive literature on the determination of seismic source parameters; most of those papers assume specific scaling laws like circular cracks, Brune's relation between corner frequency and earthquake size, etc. Our purpose here is to derive the scaling law from same basic physical concepts and to test the validity of some common assumptions in seismology. Because different authors are interested in certain specific aspects of earthquake sources, often the data from different authors is hard to combine. Some authors use the Brune's [17] empirical model as a basis for the scaling law; others use the radiation from quasidynamic circular cracks as a model leading to substantial variations. Even more serious differences come from data processing, some authors using specific attenuation corrections in order to correct for Earth's Q, others use small events as empirical Green functions, etc. These corrections are very important, but they are beyond our goal which is to try to extract information about scaling and its inferences for earthquake physics. Two aspects will be particularly discussed: how to make model independent estimates of source parameters and how to establish model independent scaling laws. Some recent evidence indicates, for instance, that stress change during earthquakes as measured by apparent stress varies independently of Moment and size, so that earthquakes are probably quantified by three independent

parameters that need to be carefully chosen. The other will be an aspect that is often overlooked in the literature: Brune's [17] model, as well as Madariaga's [46] circular crack model make very specific statements about the partition of seismic energy at the source. In particular, the scaling law implies that fracture energy is not a constant but that it scales with earthquake size in a manner that was predicted on the basis of fracture dynamics [46]; friction experiments by Ohnaka and Shen [56] and Ohnaka [55]; arguments about scaling by McGarr and Fletcher [54], and dynamic seismic source inversions [34,57].

Earthquakes and Seismic Radiation

It is now well established that earthquakes are due to faulting and that the simplest way to measure them is to use the seismic moment M_0 , introduced by Aki [3], and given by

$$M_0 = \mu \bar{D} S \quad (1)$$

where μ is the shear or rigidity modulus of the material surrounding the fault, \bar{D} is the mean value of the final slip on the fault and S is the area of the fault rupture. M_0 is the moment of one of the couples that constitute a double couple, the simplest possible model of a point-like earthquake. Radiation from a point double couple source has been completely solved by seismologists and is thus the natural starting point for the development of earthquake scaling laws.

The far-field displacement \mathbf{u}_c radiated by a point double couple source can be written in the following form:

$$\mathbf{u}_c(\mathbf{r}, t) = \frac{1}{4\pi\rho c^3} \frac{1}{R} \mathbf{e}_c^T \cdot \dot{\mathbf{M}}_0 \cdot \mathbf{e}_R(t - R/c), \quad (2)$$

(see, e. g. [5]) where the subscript c stands for P or S waves, i. e. c is either α , the P wave speed, or β , the S wave speed; ρ is the density; R is the distance of the observation point from the source. \mathbf{e}_R is the radial unit vector in the direction of radiation. The unit vector \mathbf{e}_c is the polarization of the wave, that is $\mathbf{e}_c = \mathbf{e}_R$ for P waves, or \mathbf{e}_T the appropriate transverse unit vector for SH or SV waves. $\mathbf{M}_0(t)$ is the moment tensor of the source, a symmetric tensor of order three that describes the geometry and amplitude of the seismic source (see [5] for details).

Very often in seismology it is assumed that the geometry of the source can be separated from its time variation, so that the moment tensor can be written in the simpler form:

$$\mathbf{M}_0(t) = \mathbf{I}_0 M_0 s(t) \quad (3)$$

where \mathbf{I}_0 is a time-invariant tensor that describes the orientation of the source, M_0 is the scalar moment tensor of

the source, and $s(t)$ is the time variation of the moment, the source time function determined by seismologists. In the following we assume that $s(t)$ is causal, i. e. it is zero up to $t = 0$, and normalized so that

$$\int_0^\infty s(t)dt = 1.$$

Using (3) we can now write a simpler form of (2):

$$u_c(r, t) = \frac{1}{4\pi\rho c^3} \frac{\mathcal{R}_c}{R} \Omega(t - R/c). \quad (4)$$

For P waves, u_c is the radial component; for S waves, it is the appropriate transverse component for SH or SV waves. In (4) we have introduced the standard notation $\Omega(t) = M_0 ds(t)/dt$ for the source time function, the signal emitted by the source as seen in the far field. The term $\mathcal{R}_c(\theta, \phi)$ is the radiation pattern, a function of the takeoff direction of the ray from the source. In a spherical coordinate system (R, θ, ϕ) centered at the source, the radiation patterns are given by Aki and Richards [5]. We simply quote here the case of a so-called strike-slip earthquake where the fault is vertical, normal to the y axis, and slip is parallel to the x axis, so that only the $M_{xy} = M_{yx}$ components of the moment tensor are different from zero. In this case

$$\begin{aligned} \mathcal{R}_P &= \sin^2 \theta \sin 2\phi, \\ \mathcal{R}_{SV} &= \frac{1}{2} \sin 2\theta \sin 2\phi, \\ \mathcal{R}_{SH} &= \sin \theta \cos 2\phi. \end{aligned}$$

On the $z = 0$ plane ($\theta = \pi/2$), there are no SV waves. On the other hand, on this plane, the radiation patterns of P and SH waves have typical quadrupole distributions proportional to $\sin 2\phi$ and $\cos 2\phi$ respectively.

Spectral Domain Approach

At high frequencies, the signals radiated by earthquakes may become quite complex because of multipathing, scattering, etc., so that the actually observed seismogram $u(t)$ resembles the source time function $\Omega(t)$ only at long periods. It is usually verified that complexities in the wave propagation affect much more the phase of seismic waves than the spectral amplitudes in the Fourier transformed domain. The spectral domain approach was introduced by Aki [4], Wyss and Brune [73] and Brune [17]. Radiation from a simple point moment-tensor source can be obtained from (4) by Fourier transformation. Displacement pulses radiated from a point moment tensor in the Fourier transformed domain is then

$$u_c(r, \omega) = \frac{1}{4\pi\rho c^3} \frac{\mathcal{R}_c}{R} \tilde{\Omega}(\omega) e^{-i\omega R/c}. \quad (5)$$

Where $\tilde{\Omega}(\omega)$ is the Fourier transform of the source time function $\Omega(t)$. A well-known property of the Fourier transform is that

$$\lim_{\omega \rightarrow 0} \tilde{\Omega}(\omega) = M_0 * \int_0^\infty \dot{s}(t)dt = M_0 \quad (6)$$

so that the in the low-frequency limit of the source time function spectrum approaches the scalar moment.

From the observation of many earthquake spectra, and from the computation of magnitudes in different frequency bands, Aki [4] and Brune [17] concluded that the seismic spectra had a universal shape with a flat low frequency asymptote given by (6), a certain characteristic frequency called **corner frequency** by seismologists, and a decay at high frequencies that tends asymptotically to ω^{-2} . I will not repeat here the arguments that led to this model: Aki [4] proposed this spectral shape in order to explain the differences in magnitude determined from seismic waves of different frequencies. Brune's [17] argument was based on simple concepts about the high frequency radiation and the amount of energy radiated by a seismic event. The omega-squared model has now been confirmed by numerous observations and detailed studies of seismic radiation. As an example, Fig. 1 shows the displacement record of a M5 earthquake that occurred inside the Nazca plate, 99 km under Santiago de Chile on 7 January 2003. The spectrum shown at the bottom presents the typical low frequency asymptote proportional to the seismic moment, and high frequency decay proportional to ω^{-2} . The corner frequency is approximately 1.2 Hz for this event.

In its simpler form the ω^{-2} spectrum is [17]:

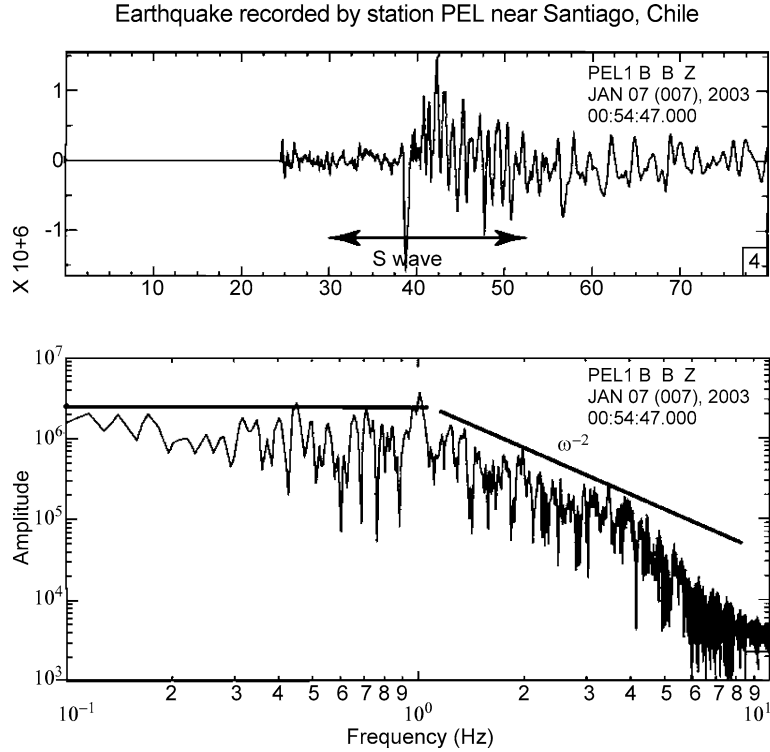
$$\tilde{\Omega}(\omega) = \frac{M_0}{1 + \omega^2/\omega_0^2} \quad (7)$$

where ω_0 is the corner frequency. Based on considerations about the spectrum of random signals, Boatwright [12] proposed the alternative model:

$$\tilde{\Omega}(\omega) = \frac{M_0}{[1 + \omega^4/\omega_0^4]^{1/2}}. \quad (8)$$

In these simple omega-squared models, seismic sources are characterized by only two independent scalar parameters: the seismic moment M_0 and the corner frequency ω_0 .

Brune's model (7) explains well the spectrum of S waves, it describes also that of P waves, but the corner frequencies are different for P and S. For a long time observations were not able to distinguish between the P and S waves spectra. The first clear observations of the corner frequency ratio were made by Hanks [26]. Recent work by Abercrombie and Rice [2], Prieto et al. [59], Ide et al. [35]



Earthquake Scaling Laws, Figure 1

Example of a seismic signal and its spectrum. Recording at the PEL broad-band station of the GEOSCOPE network of a $M = 5$ intermediate depth earthquake. The event occurred inside the subducted Nazca plate under Santiago, Chile at 99 km depth. The top panel shows the displacement, integrated from the broad band velocity record. The shear wave section used for computing the spectrum is shown with arrows. The bottom panel shows the amplitude Fourier spectrum. The low frequency and high frequency trends are indicated by the thick straight lines. The corner frequency is located at the intersection of the two asymptotes

has confirmed that P spectra have a higher corner frequency than S waves and that very roughly

$$\omega_0^P \cong 1.6\omega_0^S. \quad (9)$$

This is very close to $\sqrt{3}$, the ratio of P to S wave speed. This ratio is similar to that predicted for the quasi-dynamic circular crack model [46].

As mentioned earlier, not all earthquakes have displacement spectra as simple as (7), but the omega-squared model is a simple starting point for understanding seismic radiation.

From (7), it is possible to compute the spectra predicted for ground velocity:

$$\tilde{\dot{z}}(\omega) = \frac{i\omega M_0}{1 + \omega^2/\omega_0^2}. \quad (10)$$

Ground velocity spectra are characterized by a peak situated roughly at the corner frequency ω_0 . In actual earthquake ground velocity spectra, this peak is usually

broadened and contains oscillations and secondary peaks, but (10) is a good approximation to the spectra of ground velocity for frequencies lower than a certain cut-off frequency called f_{\max} by Hanks [27], and is close to 6–7 Hz in many areas. At frequencies higher than f_{\max} , attenuation, propagation and scattering modify the velocity spectrum.

Seismic Energy Radiated by Point Moment-Tensor Sources

In order to establish the most basic scaling relationship for seismic sources we have to compute the energy radiated by an omega-squared source like (8). This was actually the way Brune originally calibrated his source time function and Aki established the variation of corner frequency with moment. Assuming that the source is embedded in a homogeneous medium, and that the observation point is sufficiently far from the source, the energy flow per unit solid angle, e_r is proportional to the square of the particle velocity v_c (see [11,12,17,18,31]), so that the total flow per unit

solid angle is:

$$e_r^c = \rho c R^2 \int_0^\infty v_c^2(t) dt \quad (11)$$

where ρc is the P or S wave impedance, $v_c(t)$ is the ground velocity and R is again the distance of the observation point from the source. We can compute the radiated energy density replacing $v_c(t)$ by the time derivative of the far field displacement (4), and get:

$$e_r^c = \frac{1}{16\pi^2 \rho c^5} \mathcal{R}_c^2 \int_0^\infty \dot{\Omega}^2(t) dt. \quad (12)$$

As expected the energy flow per unit solid angle does not depend on the distance from the source R . We can now apply Parseval's theorem to express the energy in terms of the source spectral amplitude

$$\int_0^\infty \dot{\Omega}^2(t) dt = \frac{1}{\pi} \int_0^\infty \omega^2 |\tilde{\Omega}(\omega)|^2 d\omega$$

and compute the total radiated energy, E_r , for each type of wave. Integrating (12) over the angles θ and ϕ we get

$$E_r^c = \frac{1}{4\pi^2 \rho c^5} \langle \mathcal{R}_c^2 \rangle \int_0^\infty \omega^2 |\tilde{\Omega}(\omega)|^2 d\omega \quad (13)$$

where

$$\langle \mathcal{R}_c^2 \rangle = \frac{1}{4\pi} \iint_{\Omega} \mathcal{R}_c^2(\theta, \phi) \sin \theta d\theta d\phi$$

is the mean-squared radiation pattern. This expression for the total radiated energy is very interesting because it does not depend on any assumption about earthquake dynamics, just on the shape of the spectra.

For the ω -squared model (7) the integral over circular frequency in (13) can be evaluated exactly to $(\pi/4 M_0^2 \omega_0^3)$, so that the radiated energy is simply

$$E_r^c = \frac{1}{16\pi} \langle \mathcal{R}_c^2 \rangle \frac{M_0^2 \omega_0^3}{\rho c^5} \quad (14)$$

where we grouped in the last term all the dimensional variables. Let us remark that the numerical factor $1/16\pi$ depends on the particular model assumed for the spectrum near the corner frequency. Thus, for the Boatwright model (8), the coefficient is slightly larger ($\sqrt{2}/16\pi$).

Since radiated energy and moment have the same dimensional units it is customary to rewrite this expression in the non-dimensional form:

$$\frac{E_r^c}{M_0} = \frac{\langle \mathcal{R}_c^2 \rangle}{16\pi} \frac{M_0}{\rho} \frac{\omega_0^3}{c^5}. \quad (15)$$

Very often this expression is written in terms of frequency f_0 instead of the circular frequency ($\omega_0 = 2\pi f_0$), so that

$$\frac{E_r^S}{M_0} = \frac{\pi^2 \langle \mathcal{R}_c \rangle^2}{2} \frac{M_0}{\rho} \frac{f_0^3}{c^5}. \quad (16)$$

The average radiation patterns are well known, $\langle R_p \rangle^2 = 4/15$ and $\langle R_s \rangle^2 = 6/15$, see, e.g. Haskell [31], so that for S waves most authors use the following expression to quantify the ratio between the S wave radiated energy and the seismic moment

$$\frac{E_r^S}{M_0} = 1.9739 \frac{M_0}{\mu} \frac{f_0^3}{\beta^3}. \quad (17)$$

Where we used the definition of S wave speed ($\mu = \rho\beta^2$). This non-dimensional relation makes no assumptions about the rupture process at the source except that the spectrum decays like ω^{-2} at high frequencies. Note that the numerical coefficient is smaller by a factor of four from that computed by Singh and Ordaz [69]. The factor of four seems to be a misprint.

Apparent Stress

A very important parameter of seismic sources that can be computed independently of any particular source geometry is the apparent stress. Originally, apparent stress was defined as the product of the seismic efficiency η by the average stress $\bar{\sigma}$ that acts across the fault during the earthquake

$$\sigma_a = \eta \bar{\sigma},$$

efficiency η was in turn defined as the ratio between the radiated energy E_r and the total released energy W . Unfortunately, neither W nor the average stress can be directly inverted from seismic observations because seismic waves have no information about the average stress that acts on the fault. Wyss and Brune [73] proved that for uniform average stress, W could be written as

$$W = \frac{\bar{\sigma}}{\mu} M_0$$

so that apparent stress can be defined as

$$\sigma_a = \frac{\mu E_r^S}{M_0}.$$

This expression, originally derived for uniform average stress has become one of the most useful measures of stress

on seismic sources (see, e. g. [53]). Using the expression for radiated energy (17) we get:

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 1.9739 M_0 \frac{f_0^3}{\beta^3} \quad (18)$$

an expression that depends only on three measurable quantities: total S wave radiated energy, seismic moment and corner frequency. Because the energy flow can usually be computed only for those directions where stations are available, (18) can never be evaluated very accurately. This problem still persists at present; in spite of the deployment of increasingly denser instrumental networks, there will always be large areas of the focal sphere that remain outside the domain of seismic observations because the waves in those directions are refracted away from the station networks, energy is dissipated due to attenuation, etc.

Equation (18) shows that energy moment ratio is a non-dimensional number that depends only on observable quantities (E_r , M_0 and f_0) and wave speed. No assumption is made in Eq. (18) about particular models to convert corner frequencies into source dimensions. Let us finally remark that if σ_a is computed from the estimated radiated energy E_r , then the apparent stress may be considered as an independent parameter that may be used to test the relation (18).

Time Domain Approach

In the previous section we approached seismic radiation from the spectral point of view. An alternative way to understand radiation is to approach it from the time domain. The obvious question is what is the time domain signal associated with ω^{-2} spectrum (17)? Brune [17] proposed one of them:

$$\Omega(t) = M_0 \omega_0^2 t e^{-\omega_0 t} \quad \text{for } t > 0 \quad (19)$$

which is a causal function (i. e., it is zero for $t < 0$). It is also normalized so that

$$\int_0^\infty \Omega(t) dt = M_0.$$

The high frequency content for this function is controlled by the slope discontinuity at the origin. This is however only one of many functions sharing the same spectral amplitude described by (7) and (8), leaving sufficient freedom for variations in the source time function shape.

Kanamori and Rivera [40] defined a function of finite duration T , finite moment and minimum radiated energy. They found such signal by solving a variational problem

for fixed signal duration T . The time signal is the parabola defined by

$$\Omega(t) = \frac{6M_0}{T^3} t(T-t) \quad (20)$$

for $0 < t < T$ and 0 elsewhere. The radiated energy for this signal can be computed using the time domain expression (12). Using the integral

$$I_v = \int_0^\infty \dot{\Omega}^2(t) dt = \frac{12}{T^3} M_0^2$$

we can compute the total radiated energy as a function of T . In order to write the energy in the same form as the frequency domain expressions we determine the corner frequency from the Fourier transform of (20):

$$\tilde{\Omega}(\omega) = \frac{6M_0}{\omega^3 T^3} \left[(\omega T - 2i) + (\omega T + 2i)e^{i\omega T} \right]. \quad (21)$$

At low frequencies this expression tends to M_0 , as expected, while at high frequencies it behaves like

$$\lim_{\omega \rightarrow \infty} \tilde{\Omega}(\omega) = \frac{12M_0}{\omega^2 T^2} \frac{[1 + e^{i\omega T}]}{2}.$$

So that its envelope decreases asymptotically as $12M_0(\omega T)^{-2}$ that is, it has the same inverse omega-squared behavior as Brune's model (7). The corner frequency for this signal computed from the intersection of the asymptotes is $\omega_0 = \sqrt{12}/T$. Inserting into (17) we get the following energy moment ratio:

$$\frac{E_r^c}{M_0} = \frac{\langle \mathcal{R}_c \rangle^2}{4\sqrt{12}\pi} \frac{M_0}{\rho} \frac{\omega_0^3}{c^5}. \quad (22)$$

which has a numerical coefficient that is slightly larger than that of (17). This is apparent contradiction to the method used by Kanamori and Rivera, who derived (20) from the condition that E_r be a minimum for a given moment. The reason these two results are not contradictory is that Brune's signal has infinite duration and therefore the variational principle posed by Kanamori and Rivera does not apply to it. Thus the two expressions give very similar answers, but clearly the signal (20) is not the only one that minimizes radiated energy. It is interesting to observe that all these signals decay like ω^{-2} at high frequencies. The reason is that high frequency radiation in these models is controlled by the slope discontinuities in displacement. In Brune's signal (19) the slope discontinuity is at the origin, while in (20) there are two slope discontinuities at the origin and at $t = T$.

In conclusion, the radiation models proposed by seismologists share the following properties: (1) the amplitude is controlled by the seismic moment, (2) the displacement spectrum decreases like ω^{-2} at high frequencies, (3) the spectral shape has a corner frequency f_0 and (4) the radiated energy to moment ratio satisfies the relation

$$\frac{E_r^c}{M_0} = C_r \frac{M_0}{\rho} \frac{f_0^3}{c^5}, \quad (23)$$

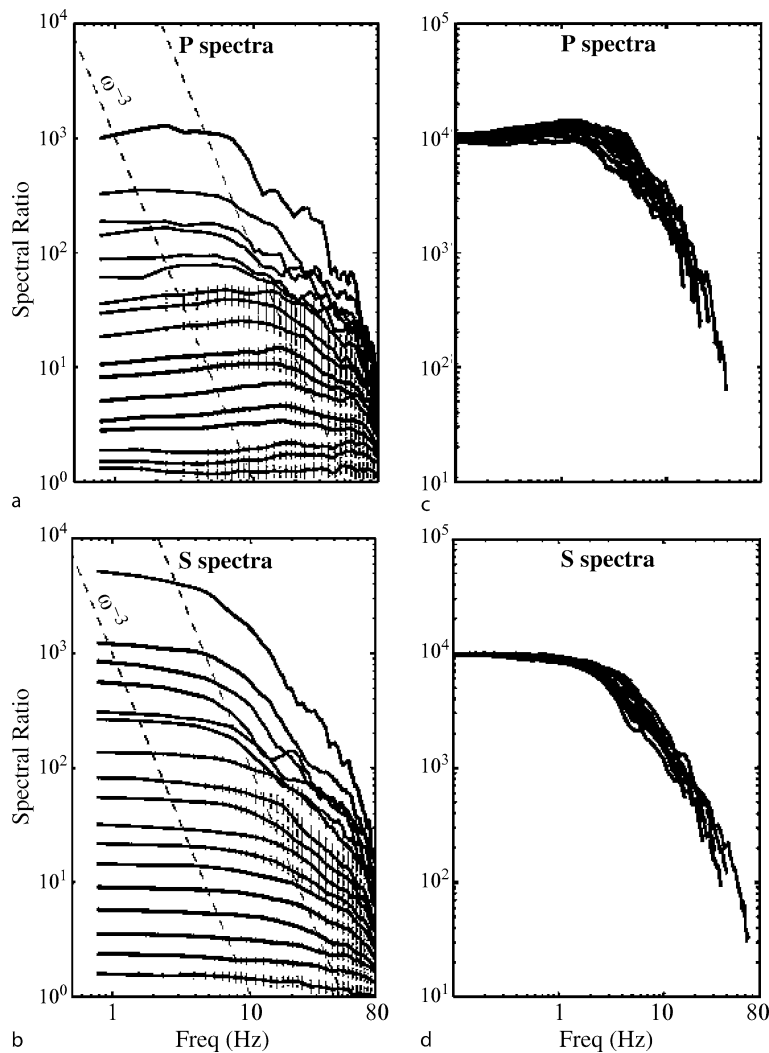
where C_r is a numerical constant on the order of two. Its consequences for energy balance are quite interesting as we shall promptly discuss.

Aki's Scaling Law

From observation of seismic data most authors (for recent data see, e. g., Abercrombie [1], Ide and Beroza [33] Ide et al. [35], McGarr [53]). Abercrombie and Rice [2], Prieto et al. [68] have concluded that apparent stress σ_a is almost independent of moment for most earthquakes. If that is correct, its immediate consequence is that moment scales like the inverse third power of the corner frequency:

$$M_0 \propto f_0^{-3}. \quad (24)$$

If apparent stress is constant, seismic moment is inversely proportional to the cube of the corner frequency. This result



Earthquake Scaling Laws, Figure 2

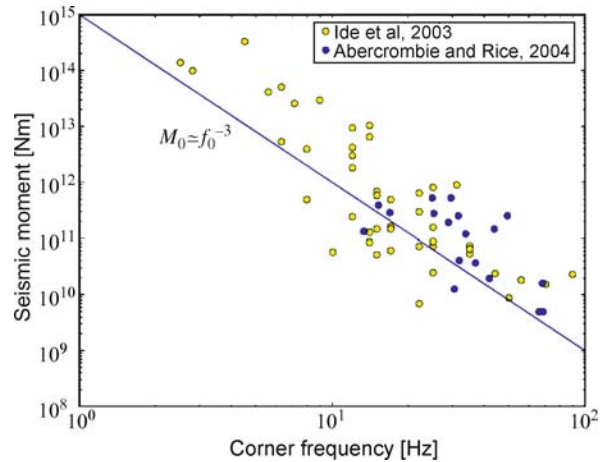
Scaling of the displacement spectrum in the far field as illustrated by several spectra for different size earthquakes reported by Prieto et al. [59]. Body wave spectra have the typical Brune [17] spectrum and can be collapsed into a single scaling figure by gliding the corner frequencies along an ω^{-3} line

is independent of the particular source model used and can be tested directly from seismic observations. Figure 2 from Prieto et al. [59] illustrates this scaling law. On the left-hand side a series of spectra for different size earthquakes are shown. The corner frequencies align along a line with slope ω^{-3} . Letting all the spectra glide along this line they computed the spectra stack shown at the right-hand side. The same properties are shared by P and S wave spectra. The P wave corner frequencies are higher by a factor of 1.6 than, those of shear waves (9).

Originally, the scaling law (24) was proposed by Aki [4] following a study of spectral data from several collocated earthquakes of different magnitude. Comparing their spectra, he concluded that corner frequencies scaled with seismic moment like (24). This is the so-called **scaling law of seismic spectrum**. This law has been tested by numerous authors with increasingly reliable digital data. Figure 3 shows an example derived from data published by Ide et al. [35] and Abercrombie and Rice [2]. In this Figure I plotted the corner frequency of S waves as measured by the authors as a function of seismic moment. The moment vs. corner frequency plot clearly follows the trend of Eq. (24). The fact that in Fig. 3 moment and corner frequency scale like (24) is often taken as a proof that earthquakes scale with a single parameter: the seismic moment. This is however not sufficient to prove scaling because seismic sources require at least three independent parameters for their quantification. It has been traditional to add an additional model dependent relation in order to derive length, time and stresses from (23). The most common assumption is that the corner frequency is related to the radius of an equivalent circular crack using the frequency radius relation proposed by Brune [17]. Some authors use other similar relations derived from quasidynamic models (e.g. [46,66]). With that assumption, corner frequency can be converted into fault size and stress drop can be derived from moment and source radius.

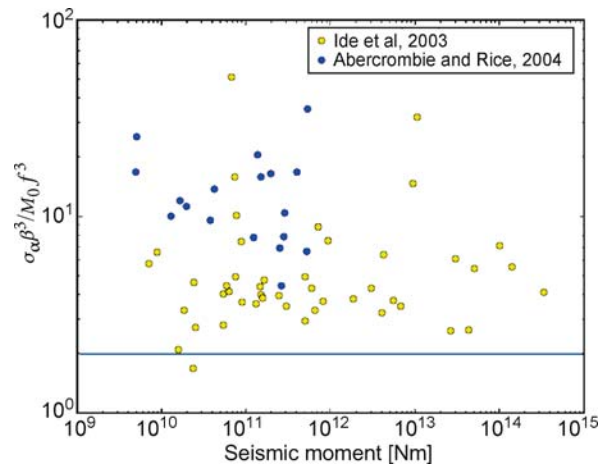
In order to test whether earthquakes scale with a single parameter, it is necessary to obtain an additional objective measure of seismic sources. The best current candidate is the radiated energy of S waves defined by (13). Originally proposed by Boatwright [11,12,13], estimates of seismic energy have become quite common but they are still difficult to obtain as discussed by many authors including Boatwright and Fletcher [15], Abercrombie [1], McGarr [53], Ide et al. [32] Singh and Ordaz [69], etc. In the following I will test the scaling law of earthquakes by computing the non dimensional ratio $C_r = \mu E_r \beta^3 / (M_0 f_0^3)$ from published data.

In Fig. 4 I test expression (18) for the data studied by Ide et al. [35] and Abercrombie and Rice [2]. The fig-



Earthquake Scaling Laws, Figure 3

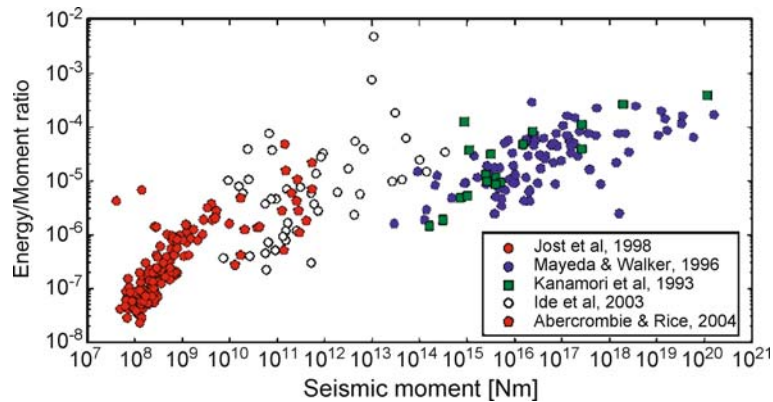
Self-similarity of earthquake spectra. The figure includes data from two different studies by Ide et al. [32] and by Abercrombie and Rice [2]. These authors measured all the quantities needed for testing the energy/moment scaling relation (23). The line labeled $M_0 \approx f_0^{-3}$ indicates the trend of variation of moment with corner frequency predicted by Aki's scaling law



Earthquake Scaling Laws, Figure 4

The nondimensional coefficient C_r defined by expression (23) plotted as a function of seismic moment. Data from Abercrombie [2] and Ide et al. [32]. The non dimensional coefficient predicted by Brune's spectral model should be equal to 1.9739 as indicated by the horizontal line. Actually it varies over almost two orders of magnitude in this data set. This may be due to errors in the measurement of radiated energy, moment and corner frequencies or it may reflect an important departure from single parameter scaling of seismic sources

ure plots the non dimensional ratio $C_r = \mu E_r \beta^3 / M_0 f_0^3$ as a function of seismic moment. According to (23), for strict scaling of seismic sources with a single parameter, this non-dimensional ratio should be a constant, indepen-



Earthquake Scaling Laws, Figure 5

Energy moment ratio as a function of moment for 14 orders of magnitude of moment. This figure is inspired on similar plots by McGarr [53] and Ide and Beroza [33]. We used data from Jost et al. [36], Mayeda and Walker [52], Kanamori et al. [41], Ide et al. [35] and Abercrombie and Rice [2]. The data of Jost et al. [36] was not corrected for attenuation as suggested by Ide and Beroza [33]. At this scale it is obvious that moment is the most broadly variable parameter, but there is also a large spread of apparent stress, over at least three orders of magnitude

dent of the seismic moment. Figure 4 shows that there is substantial scattering of the values of the non-dimensional ratio. There are many reasons for this variation; the most obvious one is experimental error due to uncorrected path or site effects, like attenuation, scattering and site amplification. I believe that even if those errors were corrected there would remain some variation due to source complexity that is not fully explained by the assumption that all earthquakes scale with a single parameter. In the literature authors generally assume that the scaling law applies and proceed to compute model-dependent quantities like static or dynamic stress drop using very specific models of rupture (see also p. 36 in [10]).

Clearly moment, energy/moment ratio and apparent stress are broadly distributed. This is illustrated in Fig. 5, inspired by previous figures of the same kind by McGarr [53] and Ide and Beroza [33]. The main difference between the results reported by these two authors is that Ide and Beroza [33] introduced a correction for attenuation at high frequencies. The figure shows the energy/moment ratio as a function of moment over 15 orders of magnitude of seismic moment. The sources of data [2,35,36,41,52] are not the same as those used by McGarr [53] and Ide and Beroza [33], because not all the data they used is published. I did not include small mine earthquakes from South Africa because it is not clear whether those events are due to frictional slip of pre-existing surfaces (see, e.g. [63]). For such a broad range of moments, the variation in energy-moment ratio is bounded, but it ranges over close to three orders of magnitude. It is clear that for the group of earthquakes reported in Fig. 5, apparent stress changes

were important and deserve further work. This is not completely unexpected: a multitude of observations point out that stresses in the seismic zones are highly variable as well as the geometry of faulting.

Earthquake Fault Models: The Scaling of Geometry and Stress

The previous discussion focused on the properties of seismic radiation from moment tensor sources and the time and frequency dependence of the moment rate function. Actually, the Brune spectrum and Aki's scaling law can be retrieved from seismic waves without any reference to a particular fault model. In order to understand how the moment is related to source dimensions and the origin of omega-squared radiation, we have to introduce a specific fault model. We will proceed in two steps: first we will study a simple source model that explains most of the observations and, in a second step, we will discuss how this model can be generalized.

A fault is defined as a rupture in the earth crust with a relative displacement of its two sides. The relative displacement between the two sides of the fault (or fault slip) will be denoted by $\mathbf{D}(\mathbf{x}, t)$, a vector function of position on the fault (\mathbf{x}) and time (t). Thus, in general, \mathbf{D} may vary in amplitude and direction over the fault plane and at each point is a function of time. The scalar seismic moment of an earthquake defined in (1) depends on source area and slip on the fault. It is essentially a static measurement of earthquake size. Corner frequency, on the other hand, is a measure of the duration of the earthquake signal which

is controlled by the time it takes for the rupture front to propagate across the fault. Thus, corner frequencies depend clearly on fault size, but one can expect this relation to be complex and very dependent on the details of the rupture process. This is often ignored in practical work, and simple source models are adopted in order to express the scaling law (23) in terms of source dimension. The argument in favor of this approach is that in order to build scaling laws that extend over several orders of magnitude of moment we may ignore the details of slip and geometry. Let us start by studying a simple circular crack, probably the simplest realistic fault model one can consider.

In a simplified model of fracture, the relative slip D of the two sides of a fault is produced by the relaxation of the shear stress transmitted across the fault. Shear stress changes with time due to the slow motion of plates, orogeny and a number of other processes that transfer stresses in the Earth's crust. When shear stress exceeds the strength of the material or the friction that maintains the fault locked, slip on the fault starts and, simultaneously, shear stress relaxes to a lower value until all motion on the fault ceases. This process is obviously very complex; detailed studies of even the simplest models of faulting show that stress relaxation at any point on the fault is complex function of not just local stress release on the fault but of other slipping points and of wave propagation on the fault. Solving such a complex problem is only worth the effort for very special, very well recorded events. For most other earthquakes we want to make global estimates of stress relaxation, and related them to the simple spectral model we studied in the previous section.

Let the shear stress acting on the fault plane before and after the occurrence of an earthquake be T_0 and T_f , respectively. We define stress drop $\Delta\sigma$, as the difference

$$\Delta\sigma = T_0 - T_f. \quad (25)$$

The stress drop represents the part of the acting stress which is used to produce the slip of the fault so that $\Delta\sigma$ is related to slip D . The relation between stress drop and slip will be in general very complex: it will depend on the geometry of the fault, but also on certain fundamental assumptions about the stress field in the earth. In general, stress will be much more heterogeneous than slip because, at least in the static case, stress is a generalized derivative of stress. We will discuss these properties briefly in Sect. "Earthquake Dynamics and the Scaling of Energy". We will assume here that the stress distribution has to be such that a finite amount of energy is released during faulting. This assumption is the basis of fracture mechanics, leading to the condition that stress may have at most inverse square root singularities on the fault surface

(see, e. g. [22,45,60]). In the early work on faults, many authors assumed that earthquakes were due to dislocations, slip distributions that present slip discontinuities at their borders. The best known example of such model is the Haskell [30] rectangular dislocation model. This model produces non-integrable stress concentrations around the edges of the fault that store an infinite amount of strain energy [48]. Thus, even if this model radiates a finite amount of energy it can not be used to estimate energy balance during seismic rupture. This paradox was well known in mechanics, where dislocations and cracks are treated as very different phenomena.

The Static Circular Crack

A simple model that may be used to explain many of the scaling laws observed in earthquake seismology is that of a static circular ("penny shaped") crack of radius a lying on the x, y plane. We assume that the fault is loaded by a uniform initial shear stress T_0 , and that T_f , the final stress, is also uniform inside the fault. The slip on the fault produced by a constant static stress drop $\Delta\sigma$ inside a circular crack was computed by Eshelby [20] and Keilis-Borok [42]

$$D(r) = \frac{24}{7\pi} \frac{\Delta\sigma}{\mu} \sqrt{a^2 - r^2}, \quad (26)$$

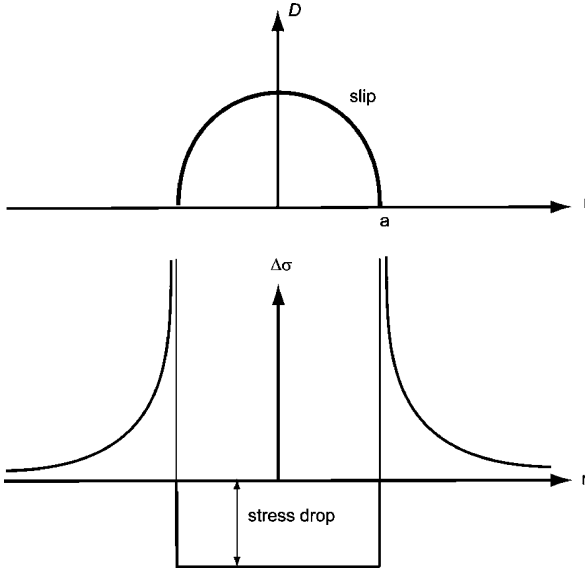
where r is the radial distance from the center of the crack on the (x, y) plane, a is the radius of the crack, and μ is the elastic rigidity of the medium surrounding the crack. Slip has the typical elliptical shape associated with cracks. The distribution of slip and the stress change for a circular crack are schematically shown in Fig. 6. Using the definition of the seismic moment (1) we can determine the scalar seismic moment for this circular fault:

$$M_0 = \frac{16}{7} \Delta\sigma a^3 \quad (27)$$

so that the moment is the product of the stress drop times the cube of the fault size. This simple relation will be used to explain the seismic scaling law in terms of fault radius and stress drop. Other fault geometries produce somewhat different scaling laws, including products of fault length and fault width. Unfortunately, as far as I know, no other geometry can be solved in such a simple closed form as the circular crack.

We can also compute the static strain energy change in the elastic medium surrounding the circular fault. This is defined as

$$\Delta W = \frac{1}{2} \int_S \Delta\sigma D dS. \quad (28)$$



Earthquake Scaling Laws, Figure 6

The simple static circular crack. The upper panel shows the slip distribution as a function of radius. The bottom panel shows the stress change produced by the slip at the top

From simple thermodynamic considerations, ΔW should be negative so that stress drop and slip should have opposite signs. For the circular crack this can be easily computed replacing the slip distribution (26) in this integral. Integrating, we find

$$\Delta W = \frac{8}{7} \frac{\Delta \sigma^2}{\mu} a^3 \quad (29)$$

where we have omitted the negative sign, so that ΔW should be interpreted as the reduction of strain energy from the elastic body caused by the earthquake. Dividing (27) into (29) we find that the strain energy to moment ratio for the circular crack is just.

$$\frac{\Delta W}{M_0} = \frac{1}{2} \frac{\Delta \sigma}{\mu} \quad (30)$$

In the very early studies of seismic rupture it was sometimes assumed that radiated energy was equal to strain energy change, i.e. $\Delta W = E_r$. In that case apparent stress $\sigma_a = 1/2 \Delta \sigma$ his assumption is sometimes referred to as the Orowan [58] model. This model is very unlikely to hold for real earthquakes: if all the available energy were radiated, there would be no energy left for producing rupture propagation and consequently rupture should propagate exactly at the P wave or the S wave speeds or should stop immediately in front of any obstacle (see [21,44]).

The circular crack model has been used to quantify numerous earthquakes for which the moment was estimated from the amplitude of seismic waves, and the source radius was estimated from corner frequencies, surface deformation, etc. The result is that for shallow earthquakes in the seismogenic zones like the San Andreas Fault, or the North Anatolian Fault in Turkey, average stress drops are of the order of 1–10 MPa. For deeper events in subduction zones, stress drops can reach several tens of MPa. Thus, average stresses do not vary much, compared to the variation of moment over more than 15 orders of magnitude.

Brune's [17] Model of Seismic Radiation

Brune developed a model of seismic radiation based on the observation that seismic spectra had the omega-squared spectral shape (7). Brune [17] proposed a model only for shear waves; although, as we already mentioned, P waves have a similar spectral shape with a corner frequency that is different from that of S waves. In his 1970 paper proposed also a relation between the corner frequency f_0 and the source radius a of a circular fault:

$$f_0 = 0.3724 \frac{\beta}{a} \quad (31)$$

That is, the corner frequency is inversely proportional to the size of the fault. The origin of the coefficient 0.3724 that appears in (31) is very important because it has major consequences for the partitioning of elastic energy. Although the steps followed by Brune were different, we can obtain his results in the following way.

Replacing the expression for moment of a circular crack (27) into the expression for energy moment ratio (17) we get

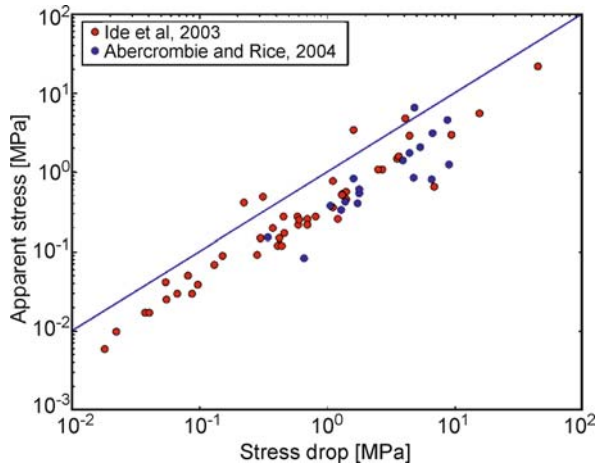
$$\frac{E_r^S}{M_0} = 4.5118 \frac{\Delta \sigma}{\mu} \frac{a^3 f_0^3}{\beta^3} \quad (32)$$

inserting Brune's expression for the corner frequency (31) into (32) we get the apparent stress

$$\sigma_a = \frac{\mu E_r^S}{M_0} = 0.2331 \Delta \sigma \quad (33)$$

So that apparent stress drop in Brune's model is proportional to the static stress drop $\Delta \sigma$. This result, derived by Singh and Ordaz [69], has a very interesting consequence for the energy balance in earthquakes. Indeed, inserting the relation (30) between moment and strain energy change during an earthquake into (34), we get

$$E_r^S = 0.466 \Delta W \quad (34)$$



Earthquake Scaling Laws, Figure 7

Apparent stress versus stress drop for the data reported by Ide et al. [32] and Abercrombie and Rice [2]. This figure suggests that apparent stresses computed directly from radiated energy and moment are not independent of stress drop computed for a static circular crack model

That is Brune's model makes the implicit assumption that the energy radiated in the form of S waves is 46.6% of the available strain energy. This was computed by Brune [17]; he found that S waves carried 44% of the available energy in his Equation (40). The number cited here, 46.6% is due to Brune's [18] correction of (31). I think that this is a very important consequence of the corner frequency source radius relationship (31), that is not often cited in the literature.

We can now test whether stress drops computed using the static circular crack model produces are or not independent of the apparent stresses computed from the energy moment ratio E_r/M_0 .

In Fig. 7 we plot the ratio of apparent stress to stress drop for the data of Ide et al. [35] and Abercrombie and Rice [2]. The relation is roughly linear but the ratio is not well defined, although it is clearly less than 1 for most of the events.

Earthquake Dynamics and the Scaling of Energy

Earthquakes are dynamic processes in which rupture propagates under the control of friction that acts between the two sides of the fault as they slip. The study of the friction law that actually operates on seismic faults is a major problem in seismology and fracture mechanics. Laboratory experiments, seismic observations and field studies are needed to solve this complex problem. In this section we will attempt to establish some general

properties of seismic sources without getting involved with fine details about friction and rupture propagation (see the contribution by Ampuero in the present volume for a fuller discussion). The main question in this context is: can we establish some general properties of seismic ruptures that are independent of the details of friction? Do the observations of seismic spectra and scaling laws constrain in any way the overall properties of seismic sources? This approach has been taken in recent years by many authors, some have tried to convert slip models inverted from near field seismic observations to determine energy balance [16,34]; others have tried to do the same by remarking that dynamic ruptures only propagate at reasonable rupture speeds for a very limited range of seismic parameters [57]; or, very recently, have tried to derive general properties of the friction law from the scaling of seismic spectra [2]. We will follow the latter approach because I believe that it is very promising.

The Dynamic Circular Crack Model

Perhaps, the simplest fault model that can be imagined is a circular crack that grows from a point at a constant or variable rupture speed and then stops at the rim of the fault, arrested by the presence of unbreakable barriers. This model is the natural dynamic equivalent to the static circular crack discussed in the previous section. The circular crack problem is posed in terms of stresses not of slip, but the rupture process is fixed in advance so that rupture does not develop spontaneously. This is the only unrealistic feature of this model, hence it is considered as quasidynamic, that is, rupture is kinematically defined, but slip is computed solving the elastodynamic equations. This model was carefully studied by a number of authors in the 1970s [43,46,47,62,65].

Let us consider a rupture that starts from a point and then spreads self-similarly at constant rupture speed v_r without ever stopping. Slip on this model is driven by stress drop inside the fault. The solution of this problem is somewhat difficult to obtain because it requires very advanced use of self-similar solutions to the wave equation and its complete solution for displacements and stresses must be computed using the Cagniard de Hoop method [62]. Fortunately, the solution for slip across the fault found by Kostrov [43] is surprisingly simple. Slip in the circular fault is parallel to the direction of stress drop on the fault and it has the typical elliptical shape:

$$D(r, t) = C(v_r) \frac{\Delta\sigma}{\mu} \sqrt{v_r^2 t^2 - r^2} \quad (35)$$

where r is the radius in a cylindrical coordinate system centered on the point of rupture initiation. $v_r t$ is the instantaneous radius of the rupture at time t . $\Delta\sigma$ is the dynamic stress drop assumed to be constant inside the rupture zone, μ is the elastic rigidity, and $C(v_r)$ is a slowly varying function of the rupture velocity v_r . For most practical purposes $C \sim 1$. This simple solution constitutes a key result containing one of the most important properties of circular cracks. Slip inside the fault scales with the ratio of stress drop over rigidity times the instantaneous radius of the fault. As rupture develops, slip increases with the size of the rupture zone.

Energy Release Rate for a Dynamic Circular Crack

We can determine the energy release rate for Kostrov's model (35) from the behavior of stresses near the edge of the crack. At time t , the fault radius is $r = v_r t$, the slip velocity field derived from (35) has the form

$$V(r, t) = C(v_r) \frac{\Delta\sigma}{\mu} \frac{v_r^2 t}{\sqrt{v_r^2 t^2 - r^2}} \quad (36)$$

so that, near the rupture front, the velocity field presents the well known inverse-squared root singularity predicted by dynamic crack theory [22,45]. We can then approximate the singularity in slip rate in the general form

$$V(r, t) = \frac{V_d}{\sqrt{2\pi}} \frac{1}{\sqrt{v_r t - r}} \quad (37)$$

where V_d is the velocity intensity factor, a measure of the amplitude of the square root singularity in slip velocity that moves with the rupture front. This velocity singularity is associated with a dynamic stress concentration ahead of the rupture front

$$\Delta\sigma(r, t) = \frac{K_d}{\sqrt{2\pi}} \frac{1}{\sqrt{r - v_r t}} \quad (38)$$

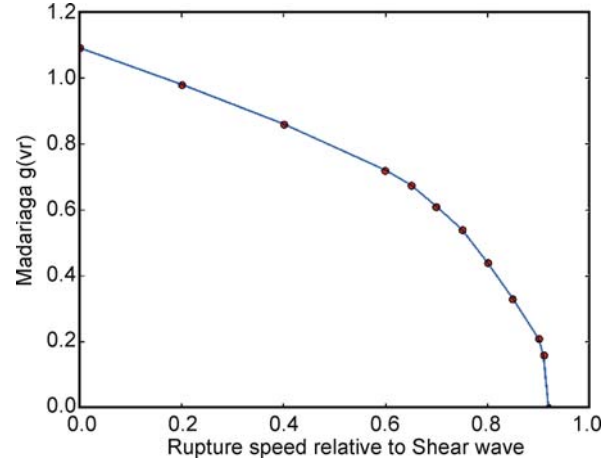
which is also of the inverse square root type. K_d is the dynamic stress intensity factor. The amplitudes V_d and K_d are linearly related to each other with a coefficient that is different for fracture modes II and III. Avoiding details that are discussed by Freund [22], we can write

$$V_d = a(v_r) \frac{K_d}{\mu} v_r \quad (39)$$

where $a(v_r)$ is a coefficient that depends on the instantaneous rupture velocity v_r .

We can compute the energy release rate near the border of the fault directly from these expressions, using several results from Kostrov [43] and Madariaga [47]:

$$G_c(v_r) = \frac{K_d V_d}{v_r} = a(v_r) \frac{K_d^2}{2\mu} \quad (40)$$



Earthquake Scaling Laws, Figure 8

The $g(v_r)$ function of rupture velocity. This function controls the fraction of strain energy that is used as fracture energy in a quasi-dynamic circular shear crack that ruptures at a constant rupture speed

that is, the energy release rate G_c near the crack front is proportional to the square of the dynamic stress intensity factor. Since the dynamic stress intensity factor K_d tends to zero at high rupture speeds, the energy release rate G_c also decreases at high speed rates. Thus, the faster the rupture, the less energy is spent in making the rupture advance.

For the circular crack, the energy release rate is not uniform around the perimeter of the circular fault because it is different in mode II (in-plane) and mode III (anti-plane). From results by Madariaga [46] we get for a circular crack

$$G_c(v_r, r) = \frac{g(v_r)}{3} \frac{\Delta\sigma^2}{\mu} r. \quad (41)$$

Where $g(v_r)$ is a monotonically decreasing function of rupture speed shown in Fig. 8. An exact expression for $g(v_r)$ was proposed by Madariaga [46], but his Eq. (30) contains a misprint that was corrected by Ide [32].

The Scaling of Energy Release Rate with Earthquake Size

We can now estimate the energy used for the propagation of the seismic rupture using the previous estimate for Brune's [17] model (34). We can establish the following global energy balance for an earthquake that is well modeled by Brune calibration of the shear wave spectra.

As rupture propagates, strain energy released by faulting ΔW is used in part to produce seismic waves, and in part to make fracture advance (measured by the energy release

lease rate G_c). Assuming that G_c is constant over the fault surface we find the following earthquake energy balance

$$\Delta W = E_r + G_c S \quad (42)$$

where S is the area of the fault. In order to estimate G_c , we need to estimate the total radiated energy E_r . Brune's model is only suitable to compute S wave radiated energy which as shown by Eq. (34) is 46.6% of the available strain energy. Energy transported by P waves can be computed using either theoretical arguments or observations of P to S energy ratios. This question was examined by Boatwright and Fletcher [15] who concluded that, theoretically at least, a crack-like source should produce about 15 times more energy carried by S waves than P wave energy. This number is confirmed by observations reported by Abercrombie and Rice [2], Prieto et al. [59]. We can thus estimate that radiated energy in the Brune model is roughly 50% of the strain energy released by the earthquake; the other 50% goes into rupture work, i. e.

$$G_c S \approx \frac{1}{2} \Delta W. \quad (43)$$

Using the expression (29) for ΔW we get

$$G_c \approx \frac{1}{2} \frac{\Delta W}{\pi a^2} = \frac{4}{7\pi} \frac{\Delta \sigma^2}{\mu} a. \quad (44)$$

That is, if we adopt Brune's model, *energy release rate scales with fault size*.

In the previous calculation we assumed that G_c was uniform inside the fault. A better assumption would be that, as the rupture propagates, G_c grows with the radius as in the simple quasi-dynamic model of Madariaga [46],

$$G_c(r) = \frac{\Delta \sigma^2}{3\mu} r g(v_r) \quad (45)$$

where $g(v_r)$ will be determined for Brune's model. The total energy release during rupture is then

$$\int_S G_c(r) dS = \frac{2\pi}{9} \frac{\Delta \sigma^2}{\mu} g(v_r) a^3.$$

Using (43) again we get

$$g(v_r) = \frac{18}{7\pi} = 0.818. \quad (46)$$

Thus, Brune's model is equivalent to a circular quasidynamic shear crack propagating such that the energy release rate grows with fault radius like

$$G_c = \frac{6}{7\pi} \frac{\Delta \sigma^2}{\mu} r.$$

Thus, whether we use a constant energy release rate on the fault (44), or a more realistic model where energy release grows with the radius of the fault, we find that energy release rate grows with fault radius in Brune's model. This result confirms that energy release rate scales with the fault radius and that it adjusts as the fault grows [55,56].

The relation between energy release rate and fault size was studied by Abercrombie and Rice [2] using their own data and data from a number of previous studies. They reached the conclusion that G_c grows with radius roughly like $r^{0.4}$, not like the radius as in (41) and (47). Abercrombie and Rice estimated G_c from the expression

$$G_c = \frac{1}{2} (\Delta \sigma - \sigma_a) D$$

where D is slip. This expression is entirely compatible with ours, so that the reason G_c scales with the power 0.4 of the radius is that stress drops scale with earthquake size, specially if $\Delta \sigma$ is not linearly related to σ_a as shown in Fig. 7. After checking Fig. 8 of Abercrombie and Rice [2], I conclude that they obtained scaling with a power of 0.4 using the full data set, including much larger earthquakes. For the Cajon pass earthquakes studied by Abercrombie and Rice the power of radius in scaling is much closer to one. Seismic data have now reached the quality necessary to test different hypothesis about the scaling of rupture energy in order to see whether larger earthquakes are really different from smaller ones.

Scaling of Energy, Magnitude and Moment

Kanamori [38] introduced the so-called moment magnitude, M_w , assuming that all available strain energy was converted into seismic waves, i. e. that $E_r \approx \Delta W$. This assumption means that no energy is used to propagate fracture so that $G_c = 0$. Using the definition of strain energy change, (28), we get Kanks and Kanamori [28]:

$$E_r = \frac{1}{2} \Delta \sigma \bar{D} S, \quad (47)$$

where \bar{D} is the average slip of the earthquake and S its source area.

This expression shows that from the radiated energy we only have information about the stress drop, not about the absolute stress level acting on the fault during faulting. Rewriting expression (27)

$$M_0 = \frac{16}{7\pi^{3/2}} \Delta \sigma S^{3/2} \quad (48)$$

and taking logarithms

$$\log M_0 = \frac{3}{2} \log S + \log \left(\frac{16\Delta\sigma}{7\pi^{3/2}} \right). \quad (49)$$

From this equation it follows that, for constant stress drop scaling, $\log S$ should be proportional to $2/3 \log M_0$. This hypothesis had been shown empirically to be valid for a large range of values of M_0 [39]. We noticed however that more recent data (Fig. 6) shows that stress drop varies over 3 orders of magnitudes, at least in the data reported by Ide and Beroza [33].

Assuming constant stress drop, Kanamori [38] defined the moment magnitude M_w based on the empirical relation of Gutenberg and Richter [23,24] between surface wave magnitude and seismic energy (in Joules): $\log E_r = 1.5M_S + 4.8$. Hanks and Kanamori [28] proposed the moment magnitude scale

$$M_w = \frac{2}{3} \log M_0 - 6.07 \quad (50)$$

where M_0 is measured in Nm. The moment magnitude has become the standard way to measure the size of earthquakes. Both M_w and seismic moment M_0 (Nm) can be related to other magnitude measurements by a number of empirical relations. (see, e. g. [39]).

Radiated seismic energy can not be computed directly from (47) because it needs to be corrected for the part of the strain energy that is used to propagate the fracture (see, (42)).

More General Scaling Relations

Derived from the Scaling Law of Earthquake Spectra

In the two previous sections several scaling relations have been established which relate the parameters involved in the fracture process of earthquakes. It has been shown that slips and slip velocities scale linearly with stress drop, which is the most fundamental scaling parameter (pp. 202–211 in [67]). If the average stress drop measured over the whole fault plane is roughly constant for all earthquakes, the slip on the fault should scale with the dimensions of the fault (L) which for small earthquakes represents the length (L) or radius of the fault and for large earthquakes the width (W). An unsolved issue, due mostly to lack of data for very long strike slip earthquakes, is that of a possible difference in the scaling of the seismic moment with fault length, between large and small earthquakes. Strike slip earthquakes with seismic moment less than 10^{21} Nm ($M_w < 8$) should scale with L^3 , while larger ones with L^2 .

Many other relations can be established starting from the basic scaling laws discussed earlier in this chapter. For instance, maximum and average slip for earthquakes scale like the cubic root of moment for most earthquakes [53,54], but these scaling relations can be derived from the basic relations discussed earlier.

More Realistic Radiation Model

In reality earthquakes occur in a complex medium that is usually heterogeneous and dissipative. Seismic waves become diffracted, reflected, and in general suffer from multipathing in those structures. Accurate seismic modeling would require perfect knowledge of Earth's structure. It is well known and understood that structural complexities dominate signals at certain frequency bands. For this reason the simple model presented here can be used to understand the main features of earthquakes at long wavelengths, while the more sophisticated approaches that attempt to model every detail of the wave form are reserved only for more advanced studies. Here, like in many other areas of geophysics, a balance between simplicity and concepts must be kept against numerical complexity that may not always be warranted by lack of knowledge of the details of Earth's structure. If the simple approach were not possible, then many standard methods to study earthquakes would be impossible to use. A good balance between simple, but robust concepts and the sophisticated reproduction of the complex details of real wave propagation is a permanent challenge for seismologists.

Why Does the Spectrum Decay Like ω Squared?

We have seen that seismic data is in very broad agreement with the general features of Brune's spectral model (7). We have explained the scaling of low frequencies in terms of simple static source models, the corner frequency and the high frequency decay are explained by the energy balance Eq. (32). Seismic energy must be finite and a well defined fraction of the available strain energy. These conditions require that the spectrum of seismic energy is integrable in expression (13). General properties of Fourier transform can be invoked to demonstrate that in the time domain displacement signals are continuous functions of time with discontinuous derivatives. That is, the velocity field emitted by a seismic source in the far field contains jumps in particle velocity as is the case with the seismic signals proposed by Brune (19) or by Kanamori and Rivera (20). For the dynamic circular crack studied in this section, the velocity jumps are emitted when rupture stops abruptly at the rim of the circle of radius a . The

nature of these stopping phases has been carefully studied by a number of authors, including Madariaga [46,47], Boatwright [12], Spudich and Frazer [71] and Bernard and Madariaga [8]. Their study is very complex and beyond the purpose of the present article, we will use a simpler approach based on a scaling argument.

Let us now consider how the stopping phases scale with earthquake size and stress drop. In the omega squared model, the high frequency decay of the far-field displacement produced by shear waves can be written in the very general form

$$u_s(r, \omega) = C(\theta, \varphi) \frac{M_0 \omega_0^2}{\mu \beta} \frac{1}{R} \omega^{-2},$$

where we lumped the numerical coefficients and the radiation pattern in the single non-dimensional coefficient $C(\theta, \varphi)$. R , θ and ϕ are spherical coordinates at a reference point on the fault. Using the expressions for M_0 and corner frequency in terms of fault radius and stress drop we obtain

$$u(r, \omega) = C(\theta, \varphi) \frac{\Delta \sigma a}{\mu} \frac{\beta}{R} \omega^{-2} \quad (51)$$

with a slightly different dimension-less coefficient. The far field waves scale at high frequencies like the product of stress drop times the radius of the fault. That is exactly what was predicted by Madariaga [47] for a circular crack. It is interesting to remark that the factor $\Delta \sigma a$ actually comes from the product of the stress intensity factor K and the square root of the radius a . Although the scaling of high frequencies was derived here for a very special circular crack model, it can be easily generalized to ruptures of any shape, splitting the factor $\Delta \sigma a$ into a stress intensity factor and the square root of the local radius of curvature of the wave front of the stopping phase.

The previous model for the radiation of omega squared high frequency waves can be extended to more complex source models, in particular to source models that contain a number of subfaults (see, e. g., [14]). The high frequency seismic waves emitted by such a model are due to stopping phases emitted all along their propagation process. Each such stopping phase contributes to enriching the high frequency contents of the seismic waves. The incoherent sum of those phases produces a total spectrum that scales with fault radius as in (51). In this sense, omega squared decay is the signature of the presence of cracks on the fault. Recent work has shown that omega squared waves are emitted every time the rupture front changes rupture velocity, or that the rupture is deviated from a plane by the presence of fault kinks or discontinuities.

Kinematics and Statistical Models for Fault Slip

So far we have discussed a dynamic crack approach to understanding earthquake scaling. Another method to describe seismic sources radiation was introduced by Haskell [30]. He assumed that earthquakes could be described by simple propagating dislocations leaving a constant slip in their trail. The most common such model is that of a flat rectangular fault with constant slip in it. Such model is mechanically impossible, because it needs infinite amount of energy to be created. Curiously, though, the Haskell model produces a finite amount of radiated energy and the radiated field can be computed exactly both in the far [31] and near field [48]. Because the strain energy change produced by Haskell's model is infinite, an energy balance equation like (42) can not be established. In Haskell model all of the energy released from the medium is absorbed by the dislocation motion and seismic radiation is just a secondary feature of the source processes.

Dislocation models have appeared in a different form, derived from statistical considerations about the distribution of slip on faults. In his seminal paper, Andrews [6] established some basic statistical properties of slip distributions that are actually based on Haskell's [31] original study of the power spectra of slip and the correlation functions of slip. His analysis is based on some general features of fractal surfaces and distributions; here we will look at these scaling relationships from the point of view of the circular crack model that we discussed in previous sections.

The slip function of a circular crack was defined in (26). This is a function of radius only so that its Fourier transform is very easy to compute using the Hankel transform:

$$\tilde{D}(k) = 2\pi \int_0^\infty D(r) J_0(kr) r dr \quad (52)$$

where k is the radial wave-number, J_0 is the Bessel function of degree zero. Inserting the expression (26) in (52) and integrating we get

$$\tilde{D}(k) = \frac{48}{7} \frac{\Delta \sigma}{\mu} a^3 \frac{\sin(ak) - ak \cos(ak)}{k^3 a^3}. \quad (53)$$

At low wave-number, when k tends to 0, the spectrum (53) tends to the value

$$\tilde{D}(0) = \frac{16}{7} \frac{\Delta \sigma}{\mu} a^3 = \frac{M_0}{\mu}$$

that is, the low wave number limit of the slip spectrum is the seismic moment, just as the low frequency limit of Brune's spectrum is the seismic moment. This is of course

not a coincidence but a consequence of the fact that the seismic moment is the source of low frequency waves. At high wave numbers, when $ka \gg 1$, the spectrum $D(k)$ behaves like

$$\lim_{k \rightarrow \infty} \tilde{D}(k) = -\frac{48}{7} \frac{\Delta\sigma}{\mu} a^3 \frac{\cos(ak)}{a^2 k^2}. \quad (54)$$

The slip spectrum decays like k^{-2} at high wave numbers, a property that seems to be as universal as the high frequency decay of seismic spectra with omega-squared.

The important issue is why is it k^{-2} ? The origin of the high frequency behavior of the slip spectrum can be determined with some simple properties of the Hankel transform (52). Take a circular fault of finite radius a . Then for different types of slip discontinuity we get the following asymptotic behavior:

	$\lim_{r \rightarrow a} D(r)$	$\lim_{k \rightarrow \infty} D(k)$
Constant	1	$k^{-3/2}$
Crack-like	$(a-r)^{1/2}$	k^{-2}
Conical	$(a-r)$	$k^{-5/2}$
Smooth	$(a-r)^{3/2}$	k^{-3}

Thus the high wave number behavior is a reflection of the discontinuity of slip at the border of the fault. Andrews [6] studied a slip distribution that behaves like $(a-r)^{3/2}$ near the edge of the fault. In his case the high wave number decay is k^{-3} . Thus the high wave number behavior of slip distributions is controlled by the discontinuities of slip, a crack like discontinuity producing a k -squared distribution. Note that the spectral behavior for two-dimensional distributions is quite different than for two dimensional slip distributions. Haskell [30] used the properties of 2D Fourier transforms to derive several conclusions about earthquake spectra that do not apply to circular cracks. For a two-dimensional plane or anti-plane crack the spectrum decays like $k^{-3/2}$, such a spectrum is inadmissible in 3D because it would imply non-integrable stress distributions as we will show now.

The stress field associated with the circular shear crack slip can be computed in a straightforward way using the expressions provided by Eshelby [20] or Sneddon [70]. Let the fault be located on the plane (x,y) and slip be parallel to the axis x , (i. e. $D(x) = \Delta u_x(x)$). In the spectral domain the associated stress drop can be computed from slip by

$$\Delta\sigma_{xz}(k) = -\frac{\mu}{2k} \left(\frac{2(\lambda + \mu)}{\lambda + 2\mu} k_x^2 + k_z^2 \right) D(k) \quad (55)$$

(see [64]). This is a relatively simple expression, but it is not easy to compute analytically because stress drop for the circular crack does not have cylindrical symmetry. Andrews [6] provided a simplification that we will use

here: assuming that the elastic constant $\lambda = 0$ i. e. that is the elastic medium is incompressible), we get in Fourier domain:

$$\Delta\sigma_{xz}(k) = -\mu k D(k). \quad (56)$$

Multiplying (53) by $-k$ and doing the inverse Hankel transform (see [70]) we get, approximately

$$\begin{aligned} \Delta\sigma_{xz}(r) &= -\Delta\sigma \quad \text{for } r < a, \\ \Delta\sigma_{xz}(r) &\cong K/\sqrt{r-a} \quad \text{for } r > a. \end{aligned} \quad (57)$$

That is, inside the crack, stress drop is constant while outside stress drop exhibits an inverse square root singularity typical of cracks as discussed in (38). This explains why the spectrum of a circular crack decays as k^{-2} in the wavenumber domain. The k^{-2} spectrum is the signature of the presence of a crack.

Mai and Beroza [51] computed the correlation lengths, fractal dimensions, Hurst exponents and wave number spectra of 42 earthquakes for which the slip distribution on the fault planes were available from inversion of seismic and geodetic data. They reached the conclusion that the high frequency decay had an average fractal dimension of 2.29 ± 0.23 that implies a high wave number asymptotic decay of $k^{1.71}$, that is slip distributions determined from slip inversions tend to be rougher than the spectra of circular cracks. The origin of this exponent needs to be carefully scrutinized in terms of fault segmentation.

Bernard and Herrero [7] and Bernard et al. [9] proposed a model linking seismic radiation to the spectral properties of the distribution of slip on the fault. In their model rupture propagates in a single space direction at constant speed. In order to obtain an omega squared far field spectrum they assumed that rise time is essentially a delta function or that it scales with wave number, which is equivalent. This result needs to be confronted with dynamic simulations propagating at finite speeds and finite energy release rate. In the circular crack model, the high frequency spectrum was controlled by stopping phases, which do not exist in the Bernard et al. model. This problem needs careful attention, specially in the presence of geometrical heterogeneity that may produce stopping phases.

Future Directions

Most properties of ensemble averaged seismic spectra and slip distributions can be explained by a simple circular crack model. Seismic waves as well as slip distributions determined from seismic and geodetic inversions carry

the signature of the crack models that are at the base of earthquake ruptures. Whether the earthquake can be modeled as a simple circular crack or as the complex sum of a distribution of such cracks the result is the same: slip is of k -squared type and seismic radiation is of omega-squared type. Departure from these models can be expected if stress drop scales with fault size. There is no clear-cut evidence for such behavior because of the difficulties in estimating radiated energy mentioned several times in this review. The variations of stress drop required to explain observations may be an intrinsic variation of stresses depending on fault maturity, the position of the fault in the seismic cycle, etc.

In recent years the quality and quantity of seismic data has improved very significantly with the deployment of digital instruments in many active areas of the earth. This is a unique opportunity to test the self-similarity of earthquakes. If tests like those of Fig. 4 are applied to new data, the problem of the variability of radiated energy/moment ratio (or equivalent of apparent stress and stress drop) will be addressed more carefully. I would not be surprised if we finally concluded that apparent stress varies significantly among different earthquakes as suggested by Fig. 4.

Acknowledgments

This research was partially funded by the SEISMULATORS contract with ANR under program Catastrophes Telluriques et Tsunamis, and by the Research training network SPICE of the 7th PCRD of the European Union. I am deeply indebted to Luis Rivera, Martin Mai and Art McGarr for their careful and patient review of an initial version of this paper.

Bibliography

1. Abercrombie RE (1995) Earthquake source scaling relationships from -1 to 5 ML using seismograms recorded at 2.5 km depth. *J Geophys Res* 100:24015–24036
2. Abercrombie RE, Rice JR (2005) Can observations of earthquake scaling constrain slip weakening? *Geophys J Int* 162:406–424
3. Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, (1964) part 2. Estimation of earthquake movement, released energy, and stress-strain drop from G wave spectrum. *Bull Earthq Res Inst Univ Tokyo* 44:23–88
4. Aki K (1967) Scaling law of seismic spectrums. *J Geophys Res* 72:1217–1231
5. Aki K, Richards PG (2002) *Quantitative Seismology*, 2nd edn. University Science Books, Sausalito
6. Andrews DJ (1980) A stochastic fault model: 1. Static case. *J Geophys Res* 85:3867–3877
7. Bernard P, Herrero A (1994) A kinematic self-similar rupture process for earthquakes. *Bull Seismol Soc Am* 84:1216–1228
8. Bernard P, Madariaga R (1984) A new asymptotic method for the modeling of near-field accelerograms. *Bull Seismol Soc Am* 74:539–557
9. Bernard P, Herrero A, Berge C (1996) Modeling directivity of heterogeneous earthquake ruptures. *Bull Seismol Soc Am* 86:1149–1160
10. Beroza G, Kanamori H (2007) Comprehensive overview in *Treatise on geophysics*. In: Schubert G (ed) *Earthquake Seismology*. Elsevier, Amsterdam
11. Boatwright J (1978) Detailed spectral analysis of two small New York State earthquakes. *Bull Seismol Soc Am* 68:1177–1131
12. Boatwright J (1980) A spectral theory for circular seismic sources: Simple estimates of source dimension, dynamic stress drop, and radiated energy. *Bull Seism Soc Am* 70:1–26
13. Boatwright J (1982) A dynamic model for far field acceleration. *Bull Seism Soc Am* 72:1049–1068
14. Boatwright J (1988) The seismic radiation from composite models of faulting. *Bull Seism Soc Am* 78:489–508
15. Boatwright J, Fletcher JB (1985) The partition of radiated energy between P and S waves. *Bull Seismol Soc Am* 75:361–376
16. Bouchon M (1997) The state of stress on some faults of the San Andreas Fault system as inferred from near-field strong-motion data. *Bull Seismol Soc Am* 58:367–398
17. Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
18. Brune JN (1971) Correction. *J Geophys Res* 76:5002
19. Brune JN, Archuleta RJ, Hartzell S (1979) Far-field S wave spectra, corner frequencies, and pulse shapes. *J Geophys Res* 84:2262–2272
20. Eshelby JD (1957) The elastic field of an ellipsoid inclusion and related Problems. *Proc R Soc Lond A* 241:376–396
21. Freund LB (1972) Energy flow into the tip of an extending crack in an elastic solid. *J Elast* 2:341–348
22. Freund LB (1989) *Fracture Dynamics*. Cambridge University Press, Cambridge
23. Gutenberg B, Richter CF (1942) Earthquake magnitude, intensity, energy, and acceleration. *Bull Seism Soc Am* 32:163–191
24. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy, and acceleration (second paper). *Bull Seism Soc Am* 46:105–145
25. Hanks TC (1979) b values and w^{-g} seismic source models: Implications for tectonic stress variations along active crustal fault zones and the estimation of high-frequency ground motion. *J Geophys Res* 84:2235–2242
26. Hanks TC (1981) The corner frequency shift, earthquake source models, and Q . *Bull Seismol Soc Am* 71:597–612
27. Hanks TC (1982) f_{max} . *Bull Seismol Soc Am* 72:1867–1879
28. Hanks TC, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
29. Hanks T, Thatcher W (1972) A graphical representation of seismic source parameters. *J Geophys Res* 77:4393–4405
30. Haskell NA (1964) Total energy spectral density of elastic wave radiation from propagating faults. *Bull Seism Soc Am* 54:1811–1841
31. Haskell NA (1966) Total energy spectral density of elastic wave radiation from propagating faults: Part II. A statistical source model. *Bull Seism Soc Am* 56:125–140
32. Ide S (2002) Estimation of Radiated energy of finite-source earthquake models. *Bull Seism Soc Am* 92:2294–3005

33. Ide S, Beroza GC (2001) Does apparent stress vary with earthquake size? *Geophys Res Letters* 28:3349–3352
34. Ide S, Takeo M (1997) Determination of constitutive relations of fault slip based on seismic wave analysis. *J Geophys Res* 102(27):379–391
35. Ide S, Beroza GC, Prejean SG, Ellsworth WL (2003) Apparent Break in Earthquake Scaling Due to Path and Site Effects on Deep Borehole Recordings. *J Geophys Res* 108(B5):2271. doi:10.1029/2001JB001617
36. Jost M, Busselberg LT, Jost O, Harjes HP (1998) Source parameters of injection-induced microearthquakes at 9 km depth at the KTB deep drilling site, Germany. *Bull Seismol Soc Am* 88:815–832
37. Joyner WB (1984) A scaling law for the spectra of large earthquakes. *Bull Seism Soc Am* 74:1167–1188
38. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2921–2987
39. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65:2981–2987
40. Kanamori H, Rivera L (2004) Static and Dynamic Scaling Relations for Earthquakes and their implications for Rupture Speed and Stress Drop. *Bull Seismol Soc Am* 94:314–319. http://www.gps.caltech.edu/faculty/kanamori/static_dynamic_scaling_relations.pdf
41. Kanamori H, Mori J, Hauksson E, Heaton TH, Hutton LK, Jones LM (1993) Determination of earthquake energy release and M_L using Terrascope. *Bull Seismol Soc Am* 83:330–346
42. Keilis-Borok VI (1957) Investigation of the mechanism of earthquakes. *Tr Inst Geofis Akad Nauk, SSSR* 40 (in Russian). (1960) *Sov Res Geophys Ser* 4 (Engl transl)
43. Kostrov BV (1964) Self-similar problems of propagation of shear cracks. *J Appl Math Mech* 28:1077–1087
44. Kostrov BV (1974) Seismic moment and energy of earthquakes and seismic flow of rock. *Izv Earth Phys* 1:23–40
45. Kostrov B, Das S (1988) *Principles of Earthquake Source Mechanics*. Cambridge University Press, Cambridge
46. Madariaga R (1976) Dynamics of an expanding circular fault. *Bull Seism Soc Am* 66:639–666
47. Madariaga R (1977) High frequency radiation from crack (stress drop) models of earthquake faulting. *Geophys J R Astr Soc* 51:625–651
48. Madariaga R (1978) The dynamic field of Kaskell's rectangular dislocation fault model. *Bull Seismol Soc Am* 68:869–887
49. Madariaga R (1979) On the relation between seismic moment and stress drop in the presence of stress and strength heterogeneity. *J Geophys Res* 84:2243–2250
50. Mai PM, Beroza GC (2000) Source-scaling properties from finite-fault rupture models. *Bull Seis Soc Am* 90(3): 604–615. <http://www.seismo.ethz.ch/staff/martin/papers/BSSA00scalingRP.pdf>
51. Mai PM, Beroza GC (2002) A spatial random-field model to characterize complexity in earthquake slip. *J Geophys Res* 107:2308. doi:10.1029/2001JB000588. <http://www.seismo.ethz.ch/staff/martin/papers/JGR02MaiSlipComplex.pdf>
52. Mayeda K, Walker WR (1996) Moment, energy, stress drop, and source spectra of western United States earthquakes from regional coda envelopes. *J Geophys Res* 101:11195–11208
53. McGarr A (1999) On relating apparent stress to the stress causing earthquake fault slip. *J Geophys Res* 104:3003–3011
54. McGarr A, Fletcher JB (2003) Maximum Slip in Earthquake Fault Zones, Apparent Stress, and Stick-Slip Friction. *Bull Seismol Soc Am* 93:2355–2362
55. Ohnaka M (2003) A constitutive scaling law and a unified comprehension for frictional slip failure, shear fracture of intact rock, and earthquake rupture. *J Geophys Res* 108:B2080. doi:10.1029/2000JB000123
56. Ohnaka M, Shen L-F (1999) Scaling of the shear rupture process from nucleation to dynamic propagation: Implications of geometric irregularity of the rupturing surfaces. *J Geophys Res* 104:817–844
57. Olsen KB, Madariaga R, Archuleta RJ (1997) Three-dimensional dynamic simulation of the 1992 Landers Earthquake. *Science* 278:834–838
58. Orowan E (1960) Mechanism of seismic faulting. *Geol Soc Am Memoir* 79:323–345
59. Prieto GA, Shearer PM, Vernon FL, Kilb D (2004) Earthquake source scaling and self-similarity estimation from stacking P and S spectra. *J Geophys Res* 109:B08310. doi:10.1029/2004JB003084. <http://pangea.stanford.edu/~gprieto/publication/scaling.pdf>
60. Rice JR (1980) The mechanics of earthquake rupture. In: Dziewonski AM, Boschi E (eds) *Physics of the Earth's Interior*. Proceedings of the International School of Physics "Enrico Fermi", Course 78, 1979, pp 555–569. North Holland, Amsterdam
61. Rivera L, Kanamori H (2005) Representations of the radiated energy in earthquakes. *Geophys J Int* 162:148–155
62. Richards PG (1973) The dynamic field of a growing plane elliptical shear crack. *Int J Solids Struct* 9:843–861
63. Richardson E, Jordan TH (2002) Seismicity in deep gold mines of South Africa: Implications for tectonic earthquakes. *Bull Seismol Soc Am* 92:1766–1782
64. Ripperger J, Mai PM (2004) Fast computation of static stress changes on 2D faults from final slip distributions. *Geophys Res Lett* 31:L18610. doi:10.1029/2004GL0594
65. Sato T, Hirasawa T (1973) Body wave spectra from propagating shear cracks. *J Phys Earth* 21:415–431
66. Savage JC (1974) Relation of corner frequency to fault dimensions. *J Geophys Res* 77:3788–3795
67. Scholz CH (2002) *The mechanics of earthquakes and faulting*. Cambridge University Press, Cambridge
68. Shearer PM, Prieto GA, Hauksson E (2006) Comprehensive analysis of earthquake source spectra in southern California. *J Geophys Res*, 111, B06303, doi:10.1029/2005JB003979. http://pangea.stanford.edu/~gprieto/publication/scsn_spectra.pdf
69. Singh SK, Ordaz M (1994) Seismic energy release in Mexican subduction zone earthquakes. *Bull Seismol Soc Am* 84:1533–1550
70. Sneddon IN (1951) *Fourier Transforms*. McGraw-Hill, New York
71. Spudich P, Frazer LN (1984) Use of ray theory to calculate high-frequency radiation from earthquake sources having spatially variable rupture velocity and stress drop. *Bull Seismol Soc Am* 74:2061–2082
72. Vassiliou MS, Kanamori H (1982) The energy released in earthquakes. *Bull Seism Soc Am* 72:371–387
73. Wyss M, Brune JN (1968) Seismic moment, stress, and source dimensions for earthquakes in the California-Nevada region. *J Geophys Res* 73:4681–4684

Earthquakes, Dynamic Triggering of

STEPHANIE G. PREJEAN¹, DAVID P. HILL²

¹ US Geological Survey, Alaska Science Center,
Anchorage, USA

² US Geological Survey, Volcano Hazards Program,
Menlo Park, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Review of Dynamic Triggering Observations

Characteristics of Dynamic Triggering

Physical Models of Dynamic Triggering

Future Directions

Bibliography

Glossary

Dynamic stress change a transient, often oscillatory change in the Earth's stress field.

Static stress change a permanent, step-like change in the Earth's stress field.

Definition of the Subject

Geoscientists have long sought understanding of how earthquakes interact. Can earthquakes trigger other earthquakes? The answer is clearly yes over short time and distance scales, as in the case of mainshock – aftershock sequences. Over increasing time and distance scales however, this question becomes more difficult to answer. The study of dynamically triggered earthquakes explores the most distant boundaries over which earthquakes trigger other earthquakes.

Dynamic triggering to temporary and oscillatory fluctuations in the stress/strain regime in a volume of the Earth's crust. Dynamic stress fluctuations are associated with ground shaking resulting from either anthropogenic activities or natural sources. Dynamic triggering occurs as seismic waves from an initial earthquake propagate through the Earth's crust, triggering secondary earthquakes. Once the seismic wave train has passed and ground shaking ends in a given locale, the crust returns to its previous stress state modified by the combined stress drops associated with any locally triggered earthquakes.

Dynamic triggering includes wide ranging phenomenon, both geographically and in its characteristics. It has been observed across the globe in a variety of geologic

and tectonic environments. It has been shown to occur at distances from the initial earthquake rupture varying from meters [21,27,52] to over 11,000 km [103]. In the most distant cases, earthquake triggering results from dynamic stress perturbations as low as 0.01 MPa. Earthquakes have also been shown to trigger other earthquakes at a variety of time scales. In many cases triggering of earthquakes occurs during or within minutes to hours following the responsible seismic waves (e.g. [26,39,74,103]). In other cases, earthquakes occurring weeks to months after the initial earthquake have been interpreted as a delayed response to dynamic triggering (e.g. [41,89,102]). Delayed triggered responses may reflect a more complex series of physical processes. For example, dynamic waves may trigger an aseismic process such as fault creep or changes in a volcanic system, which subsequently triggers earthquakes secondarily [2,4,37,50].

This field has been an area of extensive research in the past twenty five years. It offers a potentially important key to improving our understanding of earthquake nucleation in that, in principle, we can determine in-situ perturbations in the local stress field that lead to earthquake nucleation and rupture. In particular, the availability of broadband seismic data near sites of triggered seismicity allows us to calculate the time history of stress field fluctuations responsible for earthquake nucleation given adequate knowledge of the local seismic velocity structure [28,38,103].

The study of dynamically triggered earthquakes can also help better characterize the physical condition of the Earth's crust at seismogenic depths. Many researchers were surprised that earthquakes could be triggered by stress perturbations as small as 0.01 MPa. This observation indicates the Earth's crust is on the verge of failure in areas with triggered responses to distant earthquakes. This field of research may also provide clues to the hydrologic regime at depth. It has long been recognized that water tables change in response to earthquakes thousands of km distant [17]. The link between dynamically triggered earthquakes and dynamically triggered hydrological changes is an active area of research [8,19,78].

Within the context of complexity and system science, [82] suggest that remotely triggered seismicity may reflect large activation correlation lengths (ACL) in fault systems and stress fields that have reached a state of self-organized criticality. This statistical physics approach to earthquake occurrence focuses on the exploration of both analog and computational models that can mimic observed dynamical space-time patterns spanning a wide range of spatial-temporal scales. It is not concerned with inferred (or "non-observable") physical models for the lo-

cal processes linking dynamic stresses and brittle failure (triggered earthquakes) on faults (see [82]). In this review, however, we focus on these physical models together with a description of documented patterns of remote dynamic triggering.

Introduction

Introduction to Stress Triggering of Earthquakes

Earthquake triggering refers to a process by which any change in fault properties or the processes acting on a fault leads to rupture initiation. More specifically, stress triggering occurs when a change in the stress field acting on a fault leads to rupture. Stress triggering of earthquakes can result from stresses applied over a variety of time scales and with a variety of frequencies, which generally fall into three partially overlapping categories, 1) static stress triggering, 2) quasi-static stress triggering, and 3) dynamic stress triggering. In the case of static stress changes, the state of stress acting across a fault is permanently perturbed. This form of stress triggering is important in the near field of an earthquake where fault displacement significantly alters the stress field in the surrounding crust. Static stress triggering is commonly regarded as the dominant factor controlling aftershock generation (e.g. [54]). Because static stress changes decay rapidly with distance from the earthquake rupture (as d^{-3} , where d is distance from the earthquake epicenter), they are generally thought to be significant only within two to three fault lengths of the earthquake rupture. The role of static stress changes in triggering aftershocks and other earthquakes has been a vigorous and productive area of research in the past two decades (for reviews see [34,53,92,93]). The relative importance of static vs. dynamic triggering of aftershocks in the near field, however, has recently become an actively debated topic (e.g. [21,71]).

Because static stress changes in the near field develop essentially simultaneously with earthquake rupture, simple static stress triggering must appeal to other mechanisms to explain the time delay associated with many aftershocks and subsequently triggered earthquakes. In contrast, quasi-static stress triggering results from viscoelastic relaxation of the crust after an earthquake. Because quasi-static stress changes decay as d^{-2} and because viscoelastic relation is a time dependent process, these stress changes may explain triggered earthquakes more distant from an initial earthquake and triggered earthquakes with delay times from years to decades [72].

Dynamic stress changes decay more slowly with distance than either static stress changes or quasi-static stress changes (as $d^{-1.5}$ for surface waves). Thus dynamic

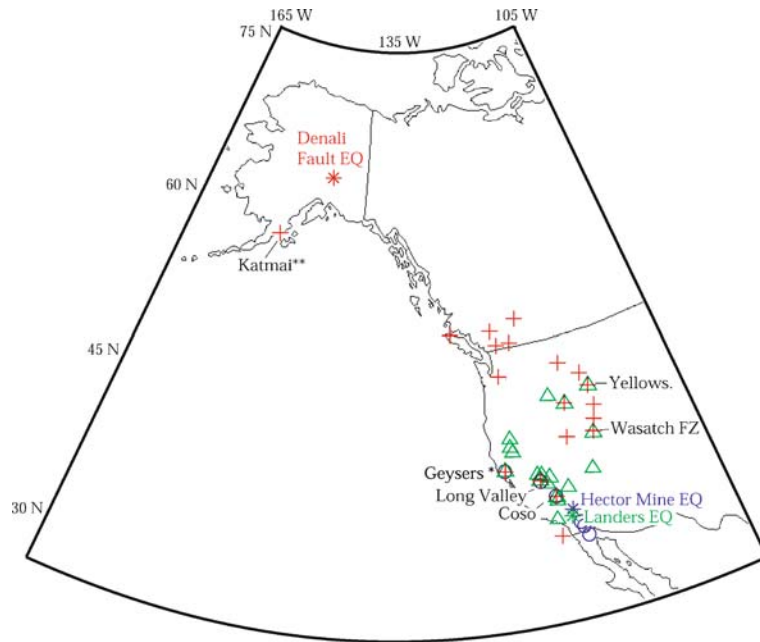
stress changes become increasingly dominant with increasing distance from the fault rupture. In this review we discuss dynamic triggering of earthquakes resulting from ground shaking due to the passing seismic wavetrain of other earthquakes. We focus on dynamic triggering due to remote earthquakes (greater than two fault lengths distance from the earthquake rupture), though we will briefly discuss the active research topic of dynamic triggering in an earthquake's aftershock zone as well. We also limit discussion to frequencies of ground shaking above ~ 0.01 Hz (periods less than ~ 100 s), though some work has been done on earthquakes triggered by longer wavelength fluctuations. For example, [16,95] recently found evidence that solid Earth tides can modulate background seismicity rates. Other reviews of dynamic stress triggering can be found in [23,37,92] and [38].

Brief History of Dynamic Stress Triggering Research

The ability of earthquakes to trigger other earthquakes at great distances has been discussed in scientific literature throughout the latter half of the 20th century (see [38] for review). However, making a credible case for a causal link between two earthquakes remains a major challenge in this field. Beyond the realm of aftershock zones, it was difficult to justify statistically that one earthquake triggered another until the 1980s and 1990s. By then continuously recording telemetered seismic networks and automated processing of data became commonplace, providing reliable spatial and temporal records of earthquake occurrence at $M \geq 1 - 2$ and the statistical leverage associated with large numbers of small earthquakes.

Dynamic triggering of earthquakes was widely accepted in the scientific community following the 1992 M 7.3 Landers earthquake in southern California. In the minutes to days following the Landers earthquake, earthquake rates increased dramatically across the western United States at distances well beyond the aftershock zone [39]. Earthquakes were triggered throughout California, Nevada, Utah, Wyoming, and Idaho at distances of up to 1250 km (Fig. 1). Although time delays of triggered events ranged from seconds to 33 hours after the arrival of the Landers earthquake wavetrain, the sudden increase in seismicity across the Western United States could not be ignored. This earthquake spawned a plethora of studies into the nature of earthquake triggering and remains one of the best studied triggering episodes to date.

The geophysical community had a unique research opportunity when the 1999 M 7.1 Hector Mine earthquake occurred. Because it was an earthquake with similar magnitude to the Landers earthquake occurring in a similar



Earthquakes, Dynamic Triggering of, Figure 1

Map showing sites of triggered seismicity in western North America from the Landers (green triangles), Hector Mine (blue circles), and Denali Fault (red crosses) earthquakes. Modified from [38], Treatise on Geophysics

location, it provided leverage to tune ideas about dynamically triggered seismicity and the underlying physical processes. Although the Hector Mine earthquake triggered seismicity at some of the same locations as the Landers earthquake, the Hector Mine earthquake had a much more limited triggered response (Fig. 1) [30]. The difference is likely due in part to differences in seismic radiation patterns between the two earthquakes [30,43]. The Landers earthquake ruptured unilaterally to the north, whereas the Hector Mine earthquake ruptured bilaterally, primarily to the south.

Following the Landers and Hector Mine earthquakes, the search for dynamic triggering began in earnest. Researchers across the globe began scanning earthquake catalogs and waveform data searching for dynamic triggering in a wide variety of environments (see “Sect. [Review of Dynamic Triggering Observations](#)”). At the Geysers, CA alone [91] identified 7 episodes of dynamic triggering between 1988 and 1994, making this among the most frequently triggered locations known.

Like the Landers earthquake, the 2002 M 7.9 Denali Fault earthquake triggered a widespread response across western Canada and the United States (Fig. 1) [27,33,45,64,68,74]. The increase in the number of broadband and high-dynamic range seismometers by 2002 allowed scientists to visually scan on-scale seismic data during the earth-

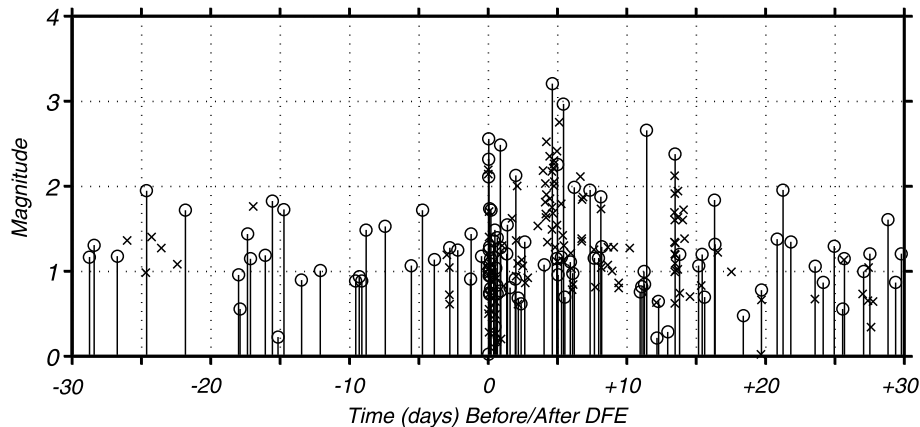
quake’s wavetrain. Thus, many triggered events were detected which were absent from earthquake catalogs. As an example of how instrumentation improvements increase our ability to detect dynamic triggering, [45] point out that the triggered response of the Yellowstone caldera to the Denali Fault earthquake could not have been detected at the time of the Hector Mine earthquake only three years earlier.

Review of Dynamic Triggering Observations

Detection of Dynamic Triggering

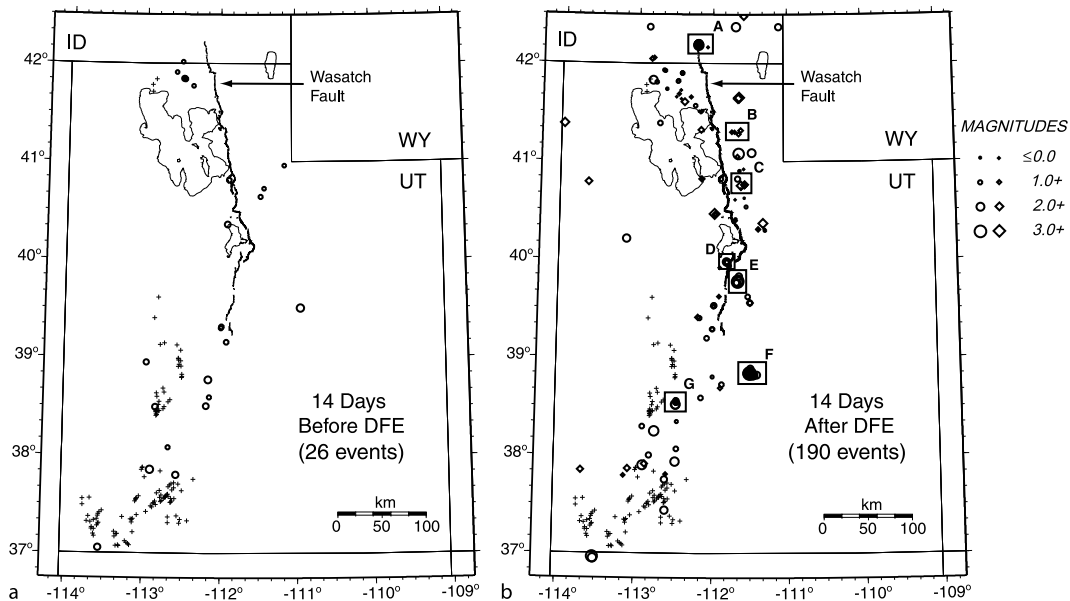
Dynamic triggering has been observed in a variety of locations around the globe. Some suggest that dynamic triggering of earthquakes is a ubiquitous process in the Earth’s crust (e.g. [26,41]). Others suggest that some areas are more likely to experience dynamic triggering of earthquakes than others (e.g. [39,64]). Observations of dynamic triggering are limited geographically due to uneven seismic network coverage and the effort applied to examining seismic data.

Following the 1992 M 7.3 Landers earthquake, dynamic triggering was recognized by the sudden increase in the number of earthquakes located through standard network processing across the western US in the days to weeks after the large earthquake. Searching earthquake



Earthquakes, Dynamic Triggering of, Figure 2

Plot of earthquake magnitude versus time in the area of the Wasatch Front, Utah, 30 days before and after the Denali Fault earthquake. Circles represent independent events. Crosses indicate secondary events determined by declustering the earthquake catalog. Figure reprinted from [68], BSSA

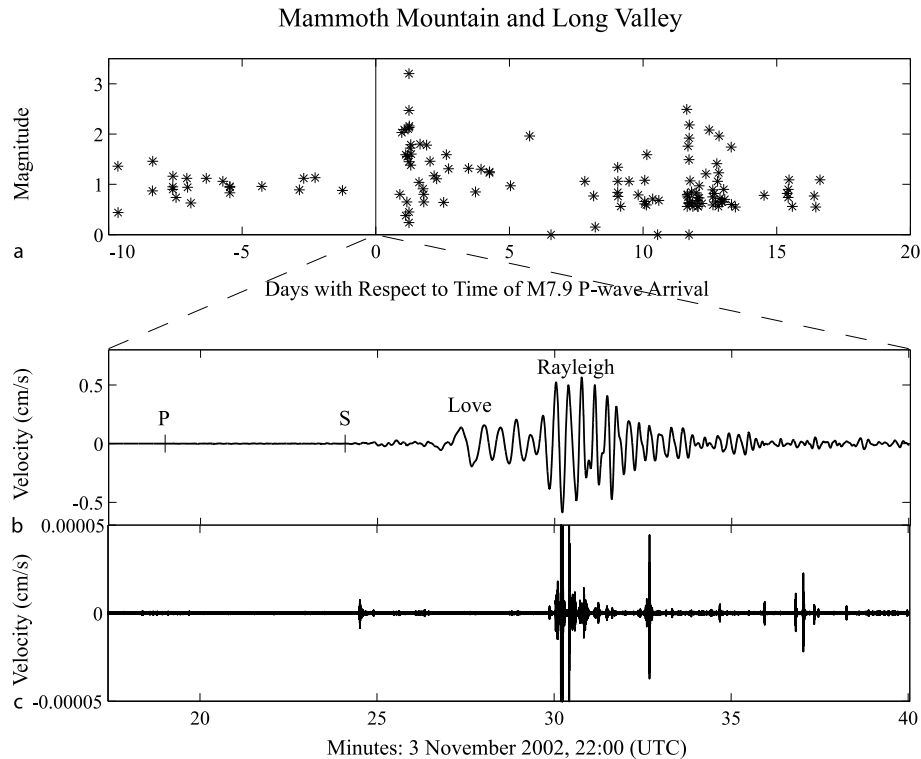


Earthquakes, Dynamic Triggering of, Figure 3

Seismicity in Utah 14 days before and after the Denali Fault earthquake. Diamonds in b show earthquakes occurring in the first 24 hours after the arrival of the wave train from the Denali Fault earthquake. Cross are locations of quaternary volcanic vents [3]. Figure reprinted from [68], BSSA

catalogs for sudden increases in seismicity after a large earthquake is one common method of identifying dynamic triggering (Figs. 2 and 3) (e.g. [6,30,39,41,68]). Identifying triggered seismicity using earthquake catalogs simplifies interpretation of triggered response with respect to background seismicity rates, as earthquake catalogs often provide stable long-term records of earthquake occurrence at a consistent threshold.

A second commonly used method of detecting dynamic triggering involves visually scanning continuous seismic data shortly before and after a large earthquake to identify a sudden increase in earthquakes too small to be detected and located through standard network processing (Fig. 4). This latter method is effective at identifying triggered seismicity in sparsely instrumented areas, identifying very small triggered earthquakes, and identifying



Earthquakes, Dynamic Triggering of, Figure 4

Seismicity triggered at Mammoth Mt. b-c and within the Long Valley caldera, California, a following the Denali Fault earthquake. a Catalog from NCEDC showing two swarms following the Denali fault earthquake in the caldera's south moat. The two lower panels show data from this very small earthquake swarm recorded on the broadband UNR/USGS station OMM from rotated to transverse direction, showing Denali earthquake wavetrain at Long Valley. Major arrivals are labeled. c Record from b high pass-filtered, showing small local earthquakes occurring during Denali wavetrain. Modified from [74], BSSA

earthquakes that occurred during the wavetrain from the initial earthquake (e.g. [26,47,64,74,103]). This method of detecting triggering has become more common with increasing availability of continuously recorded high-dynamic-range seismic data.

Possible instances of dynamic triggering have also been proposed based on historical accounts [40,42,44,59]. Because these studies rely on felt reports, they are generally limited to moderate to large sized triggered earthquakes that are separated in time by days to months.

With any method of detecting dynamic triggering, it must be shown that one earthquake is likely causally linked to the dynamic waves radiating from a previous earthquake, rather than by coincidence. Earthquakes near each other in time are more likely to be related than earthquakes separated in time. Additionally, earthquakes unlikely to occur randomly (e.g. large events in seismically quiet areas) are more likely to be related than earthquakes occurring commonly as background seismicity (e.g. small earthquakes in a seismically active area). To calculate the

probability that two earthquakes are related, one must first calculate the probability of each occurring randomly. This is usually done using patterns of earthquake occurrence based on local earthquake catalogs. The most commonly used statistical test to identify whether an increase in number of earthquakes is statistically significant is the Beta statistic [58]. Pankow et al. [68] also employ a binomial distribution analysis to this end. These techniques have potential pitfalls however, as they rely on assumptions about earthquake distributions and compare snapshots of seismicity in time in regions where seismicity rates fluctuate regularly [58]. Objectively determining whether one earthquake is genetically related to another remains a challenge.

Because spatial-temporal clusters of earthquakes are less common than isolated events, clusters of earthquakes temporally coincident with dynamic stresses are more easily identified as being triggered than isolated earthquakes. In the case of earthquake clusters, however, it may be difficult to discriminate between earthquakes directly trig-

gered by dynamic stresses from a remote earthquake and secondary aftershocks to directly triggered events [9]. To address this, earthquake catalogs are frequently ‘declustered’ (e.g. [45,68]). This process involves decomposing a catalog into primary and secondary earthquakes based on statistical patterns of aftershock sequences (e.g. [76]). [68] and [9] show that in some cases triggered seismicity is modeled well as an aftershock sequence. In other cases, however, triggered seismicity swarms cannot be dismissed as secondary aftershock sequences (e.g. [74,103]). In such cases it is likely that most events in a swarm were triggered directly by the dynamic waves radiated from a distant earthquake or perhaps as a secondary response to some aseismic process (e.g. fluid flow, or local deformation associated with fault creep) triggered by the dynamic waves.

Dynamic Triggering in Volcanic and Geothermal Regimes

Although dynamic triggering has been observed in a variety of environments, many of these observations are from areas with active volcanic and hydrothermal systems (Table 1) [6,73,74]. These areas typically have high background seismicity rates indicating that the crust habitually hovers near failure and thus is particularly susceptible to dynamic triggering. Furthermore, because these areas tend to be well instrumented, dynamically triggered earthquakes may be unusually easy to detect. Here we summarize observations of triggered seismicity in volcanic and hydrothermal areas.

The Geysers geothermal field in northern California is among the most frequently triggered sites known with 9 cases of dynamic triggering documented in the past 20 years [28,74,91]. Earthquakes that have caused triggering at the Geysers range in magnitude from 6.9 to 7.9 and in distance from 212 km to 3120 km. The Coso geothermal field in southern California has also experienced repeated episodes of dynamic triggering following the M 7.3 Landers, the M 7.1 Hector Mine, and the M 7.9 Denali Fault earthquakes (Fig. 1) [39,74]. Following the Hector Mine earthquake, dynamically triggered earthquakes and ground deformation were observed near a third geothermal field – Cerro Prieto, Baja California [24,30].

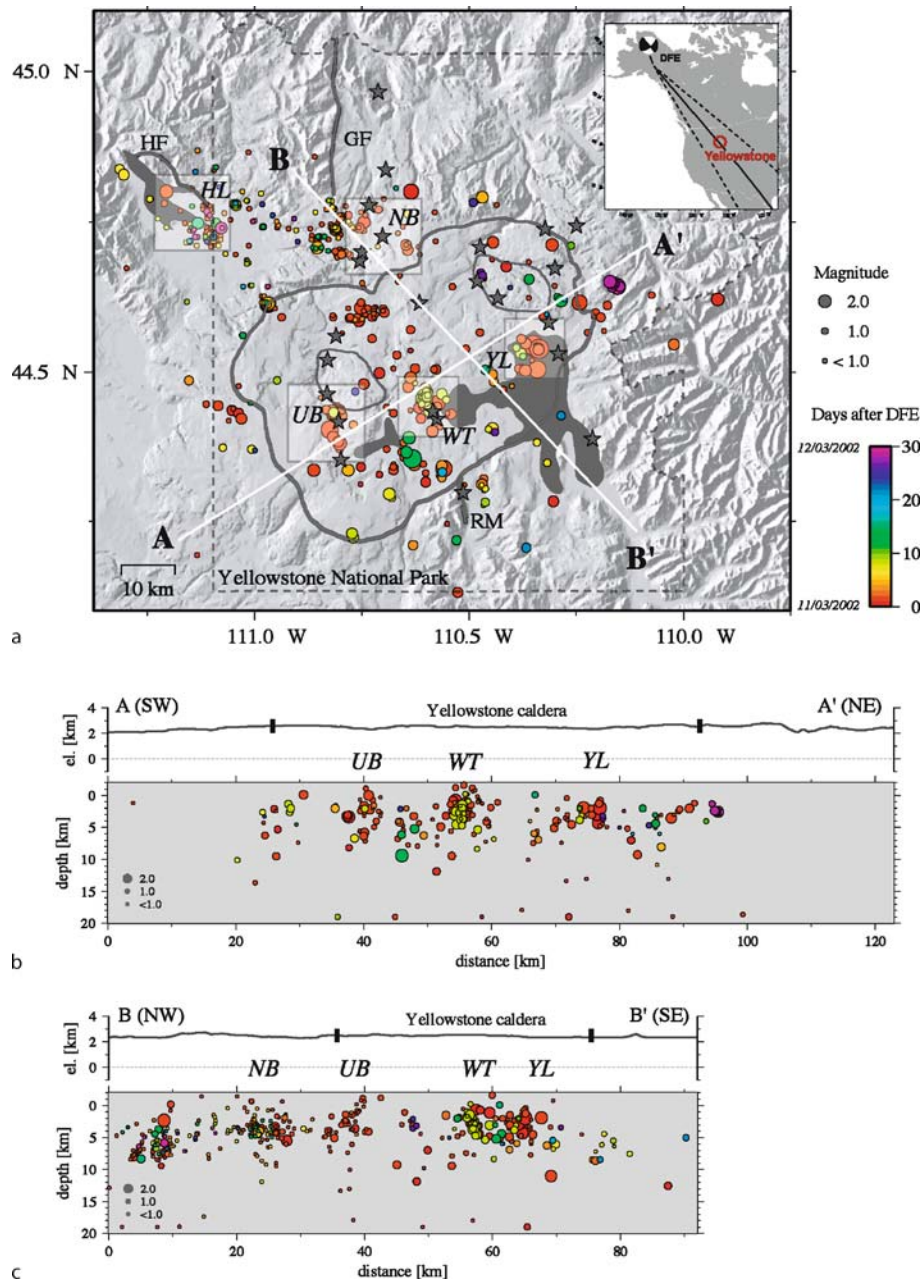
Yellowstone, Wyoming is a larger and more complicated system than the geothermal fields mentioned above, as it is a caldera system characterized by ongoing magmatic and tectonic activity, in addition to hydrothermal activity. Yellowstone had a triggered response to the M 7.3 Landers earthquake [39] and the M 7.9 Denali Fault earthquake [45,47]. The triggered response to the Denali Fault

earthquake was particularly dramatic (Figs. 5 and 6). Seismicity increased immediately following the arrival of surface waves from the Denali Fault earthquake and remained unusually high for 30 days with magnitudes ranging from < 0.0 to M 3.2 [45]. The time scale of the triggered response was spatially variable (Figs. 5 and 6). Because the Denali Fault earthquake led to immediate triggering in the area of geysers and affected periodicity of geysers at Yellowstone, it is likely that changes in the hydrothermal regime induced a triggered response in some areas ([45,46]. In other areas however, the development of triggered earthquake sequences was delayed and similar to commonly observed tectonic activity [45].

Like the Yellowstone caldera, the Long Valley caldera experiences dynamic triggering with complex characteristics. The area responded to the Landers earthquake [39], the Hector Mine earthquake [50], and the Denali Fault earthquake [50,74] both seismically and geodetically, although each response varied in location and intensity (Fig. 7). The Landers earthquake produced the largest triggered response with 340 earthquakes in seven days up to M 3.4 throughout the south moat of the caldera [39]. The seismic response to the Hector Mine earthquake was comparatively short lived and limited to the region of Mammoth Mountain. After the Denali fault earthquake, the caldera area experienced two phases of triggered seismicity. A burst of $\sim 60 M \leq 0.8$ earthquakes occurred beneath Mammoth Mountain during and shortly after the arrival of the surface waves from the Denali Fault earthquake [74]. Twenty-four hours later a larger swarm of earthquakes of $M \leq 3.4$ occurred in the Long Valley caldera’s south moat [74]. All three episodes of dynamic triggering in the Long Valley caldera were accompanied by deformation transients with geodetic moments an order of magnitude larger than the cumulative seismic moment of the triggered seismicity [36,50], though the time history and magnitude of each deformation response varied.

Iwo Jima, Japan, a volcanic island hosting a Holocene eruption, geothermal activity, and historic phreatic (steam) eruptions is a third complex caldera system which has experienced dynamic triggering of local earthquakes. [98] examined continuous waveform data of $21 M > 7$ earthquakes < 3000 km distance from Iwo Jima, and identified 4 cases of resulting increased local seismicity. In all cases earthquakes were triggered locally during surface wave arrivals and persisted for 6–15 minutes.

Dynamic triggering of earthquakes has been observed at shallow depths in volcanic edifices at a variety of locales (Table 1). In the Pacific Northwest, Mt. Rainier experienced 6–8 $M < 0$ earthquakes during the wavetrain of the Denali Fault earthquake and 8 $M \leq 0.9$ earthquakes in



Earthquakes, Dynamic Triggering of, Figure 5

Seismicity within one month of the Denali Fault earthquake at Yellowstone caldera, *color* coded with time: **a** earthquake locations, **b** cross sections along AA', **c** cross section along BB'. Specific areas labeled: HL, Hebgen Lake area; NB, Norris geyser basin; UB, Upper geyser basin; WT, West Thumb geyser basin; YL, northern end of Yellowstone Lake. Large normal faults are represented with *thick black lines*: RM, Red Mountain fault zone; GF, Gallatin fault; HF, Hebgen and Red Canyon faults. *Inset* shows location of Denali Fault earthquake and Yellowstone. *Solid and dashed lines* in *inset* show the great circle path ± 10 degrees along the strike of the Denali Fault earthquake. Figure reprinted from [45], BSSA

Earthquakes, Dynamic Triggering of, Table 1

Published occurrences of discrete remotely triggered earthquakes

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	M _{max}	Reg	Env
Aso, Japan	[61]	1400	Chi-Chi, 1999	7.7	During P waves	–	E	V
British Columbia	[26]	1800–2200	Denali Fault, 2002	7.9	During surface and coda waves	–	–	N
Burney, CA	[39]	900	Landers, 1992	7.3	23 hour	2.8	E	N
Central and Southern, CA	[41]	variable	15 Central and Southern CA earthquakes, 1988–2004	5.3–7.1	Within 1 month	–	–	–
Cerro Prieto, Mexico	[24]	260	Hector Mine, 1999	7.1	–	4.1	E	V, G
Coso, CA	[39]	165–205	Landers, 1992	7.3	~ 3 hour	4.4	E	G
Coso, CA	[74]		Hector Mine, 1999	7.1	–	–	E	G
Coso, CA	[74]	3,660	Denali Fault, 2002	7.9	15 min	2.3	E	G
Geysers, CA	[28]	2,500	Gulf of Alaska, 1988	7.6	–	0.2–2.5	E	G
Geysers, CA	[28]	212	Loma Prieta, 1989	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	443	Off Oregon Coast, 1991	6.9	–	0.2–2.5	E	G
Geysers, CA	[28]	390	Gorda Plate, CA, 1991	7.1	–	0.2–2.5	E	G
Geysers, CA	[28]	202	Petrolia, CA, 1992	7.0	–	0.2–2.5	E	G
Geysers, CA	[39]	740	Landers, 1992	7.3	3 min	1.6	E	G
Geysers, CA	[91]	635	Northridge, 1994	6.6	–	–	E	G
Geysers, CA	[28]	308	Cape Mendoceno, CA, 1994	6.9	–	0.2–2.5	E	G
Geysers, CA	[74]	3,120	Denali Fault, 2002	7.9	12 min	2.5	E	G
Greece	[6]	400–1000	Izmit, 1999	7.4	After surface waves	3.5	E	G
Iceland	[2]	64–78	South Iceland Seismic Zone, 2000	6.5	< 5 min	5	E	V, G
Idaho, Cascade	[39]	1100	Landers, 1992	7.3	33 hour	1.7	E	G
Idaho, Cascade	[26,47]	2300	Denali Fault, 2002	7.9	During Rayleigh waves	4.6	E	G
Iwo Jima, Japan	[98]	≤ 2009 km	4 earthquakes, 1983–1993	7.1–8.0	During surface waves	< 2	–	V, G
Katmai volcanoes	[73]	115	1999	7.0	< 3 min	2.3	–	V, G
Katmai volcanoes	[64]	122	2000	6.8	–	0.9	–	V, G
Katmai volcanoes	[64]	161	2001	7.0	< 2 min	1.5	–	V, G
Katmai volcanoes	[64]	161	2001	6.8	–	–	–	V, G
Katmai volcanoes	[64]	740	Denali Fault, 2002	7.9	3.9 min	2.0	–	V, G
Katmai volcanoes	This paper	3620	Kurile, 2007	8.2	During surface waves	–	–	V, G
Lassen, CA	[39]	840	Landers, 1992	7.3	12 min	2.8	E	V, G
Little Skull Mt., NV	[39]	240	Landers, 1992	7.3	1.5 hour	5.6	E	N
Long Valley, CA	[39]	415	Landers, 1992	7.3	9 min.	3.4	E	V, G
Long Valley, CA	[74]	3,454	Denali Fault, 2002	7.9	23.5 hour	3.0	E	V, G
Mammoth, CA	[50]	450	Hector Mine, 1999	7.1	20 min.	–	E	V, G
Mammoth, CA	[74]	3,454	Denali Fault, 2002	7.9	17 min.	0.8	E	V, G
Mono Basin, CA	[39]	450	Landers, 1992	7.3	19 hour	3.1	E	N
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	12 min.	0.0	E	V
Mt. Rainier, WA	[74]	3,108	Denali Fault, 2002	7.9	2.5 hour	0.9	E	V
Nanki Trough, Japan	[60]	900–1400	Tokachi-oki, 2003	8.1	After surface waves	–	S	–

the following days. In response to the Landers earthquake, Mt. Lassen in northern California hosted 14 earthquakes of $M \leq 2.8$. Volcanoes in the Katmai Volcanic Cluster,

Alaska, have experienced triggered seismicity on at least seven occasions since 1999 ([64,65,73], this chapter). The largest of these triggered responses included 17 earth-

Earthquakes, Dynamic Triggering of, Table 1
(continued)

Locations	Citation	Distance (km)	Triggering Earthquake	M	Onset	M _{max}	Reg	Env
The Netherlands Roer Valley	[12]	40	Roermond, 1992	5.4	–	3.7	E	N
New Madrid, MO	[41]	1000	1811–1812 New Madrid	~ 7.8	–	–	C	N
Offshore Southern CA	[74]	4,003	Denali Fault, 2002	7.9	Mainshock coda	2.5	E	N
Salton Sea, CA	[43]	120–150	Hector Mine, 1999	7.1	–	4.7	E	V,G
Syria – Lebanon border	[62]	500	Gulf of Aqaba, 1995	7.3	2 hr 47 min	3.7	C	N
Taiwan	[102]	variable	9 earthquakes, 1978–1994	6.5–7.1	≤ 15 days	≥ 4.0	–	V,G
Tonga Trench	[96]	290–313	Tonga Region, 2002	7.6	2 min.	7.7	S	–
Utah, Cedar City	[39]	490	Landers, 1993	7.3	39 min.	4.1	E	G
Utah, Wasatch Front	[68]	3000–3500	Denali Fault, 2002	7.9	During surface waves	3.2	E	G
Valley of Mexico	[89]	303–588	7 earthquakes	7.6–8.0	–	~ 4.0	E	V,G
Western Nevada	[1,39]	450–650	Landers, 1993	7.3	9 min.	4.0	E	G
White Mts., CA	[39]	380–420	Landers, 1993	7.3	11.6 hour	3.7	E	N
Mt. Wrangell, AK	[45]	11,000	Denali Fault, 2002	7.9	During Rayleigh waves	1.0	–	V, G
Yellowstone	[39]	1250	Landers, 1993	7.3	1.8 hour	2.1	E	V, G
Yellowstone	[45]	120	Hector Mine, 1999	7.1	During surface waves	–	E	V,G
Yellowstone	[45]	3150	Denali Fault, 2002	7.9	During Love waves	3.2	E	V, G

Location is location of triggered seismicity. Distance is distance from location of triggering to triggering earthquake (mainshock) epicenter. M is magnitude of mainshock. Onset is onset time of triggered activity with respect to arrival of waves from mainshock. M_{max} is magnitude of the largest triggered earthquake. Reg describes stress regime: E-extensional or transtensional, C-compressional or transpressional, S-subduction zone. Env describes if the area is volcanic (V), geothermally active (G), or neither (N). – indicates data not available or inconclusive.

quakes of $M \leq 2.3$. During the wavetrain of the 2006 M 8.7 Sumatra–Andaman Islands earthquake, Mt. Wrangell, Alaska had triggered 14 earthquakes [103]. With the exception of the Mt. Lassen response following the Landers earthquake and the delayed events at Mt. Rainier following the Denali Fault earthquake, these earthquakes triggered in volcanic edifices were too small to be detected and located by automatic processing systems.

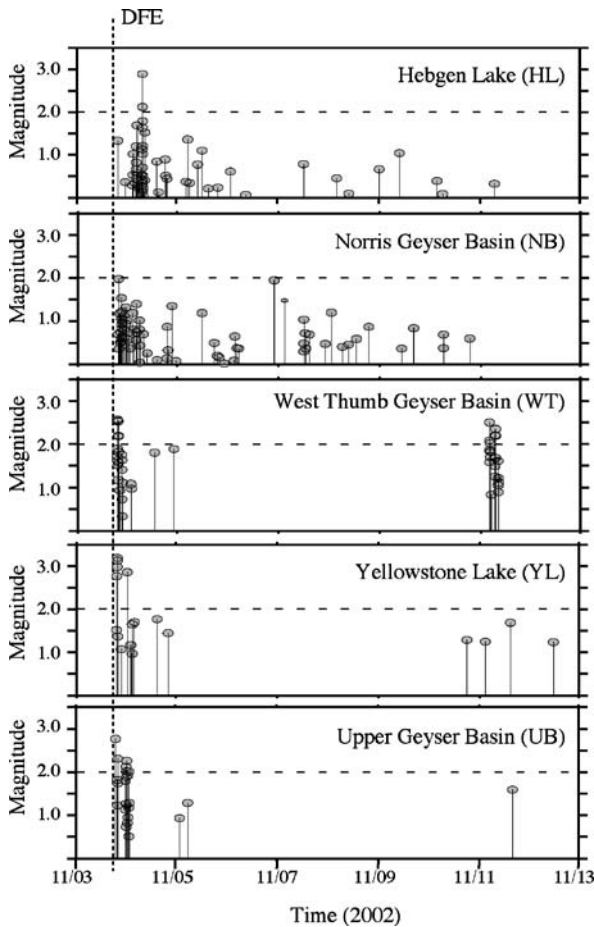
The Valley of Mexico is a large volcanically and geothermally active area located in the Trans Mexican Volcanic Belt. [89] searched for dynamically triggered earthquakes in the Valley of Mexico following 18 $M \geq 7.0$ Mexican earthquakes between 1920 and 1998. In seven cases, they found evidence for dynamic triggering of earthquakes within 2 days of a large earthquake. In four additional cases, seismicity increased after a large earthquake, but was delayed by up to one month. Because this study used only one station however, the potentially triggered events can only be located to within some ill-defined region surrounding the station.

The South Iceland Seismic Zone is a transform zone in a volcanically and geothermally active area. In 2000, a $M_w = 6.5$ earthquake in the South Iceland Seismic Zone

triggered widespread seismicity, including three $M_w \sim 5.0$ earthquakes within 5 minutes of its occurrence. Coulomb failure stress calculations indicate that the two $M > 5$ earthquakes located ~ 100 km to the west on the Reykjanes Peninsula are beyond the range where static stress changes are significant [2], and thus appear to have been dynamically triggered. Furthermore, one of these $M > 5$ earthquakes had a geodetic moment significantly larger than its seismic moment suggesting that deformation associated with aseismic fault creep may have indirectly triggered many of the smaller earthquakes in the area [2].

Dynamic Triggering in Regimes with Limited Volcanic and Geothermal Activity

Extensional and Transtensional Environments The majority of occurrences of triggered seismicity documented to date have been in extensional or transtensional tectonic regimes (Table 1). In the western United States dynamic triggering following the M 7.3 Landers earthquake occurred exclusively in transtensional tectonic regimes, many of which were also volcanically or geother-



Earthquakes, Dynamic Triggering of, Figure 6

Plot of earthquake magnitude versus time of seismicity for selected areas in Yellowstone caldera. See Fig. 5 for locations of these areas. Dashed line DFE as the origin time of the Denali Fault Earthquake. Figure reprinted from [45], BSSA

mally active, to distances of up to 1250 km [1,39]. These locations included Little Skull Mountain, Nevada; western Nevada, White Mountains, California, Mono Basin California, Cedar City Utah, the Wastach Front in central Utah, Burney, California, and Cascade, Idaho. The onset of triggering ranged from during the passage of the Landers wavetrain to 33 hours after the Landers earthquake. The largest of these earthquakes was a $M = 5.6$ earthquake triggered beneath Little Skull Mountain, Nevada. Otherwise, triggered earthquakes had $M \leq 3.0$. The most vigorous responses containing tens to hundreds of triggered earthquakes occurred near Cedar City Utah, Western Nevada, and Cascade Idaho.

The $M_w = 7.1$ Hector Mine earthquake also led to an impressive display of triggered earthquakes in exclusively extensional, transtensional, and geothermal envi-

ronments in the western United States. Triggered earthquakes began during the passage of the wavetrain near the Salton Trough in Indio and at the southern end of the Salton Sea [30,43]. In general, the triggered response to Hector Mine was less extensive and energetic than that of the Landers earthquake [30].

Following the Denali Fault earthquake seismicity was triggered in several extensional and transtensional areas in the western United States. [47] detected a $M 4.6$ earthquake triggered during the Denali Fault earthquake wavetrain in Cascade, Idaho. Seismicity remained elevated for 25 days along a 500 km stretch of the Intermountain seismic belt in Utah, on the border of the Basin and Range province (Figs. 2 and 3) [68].

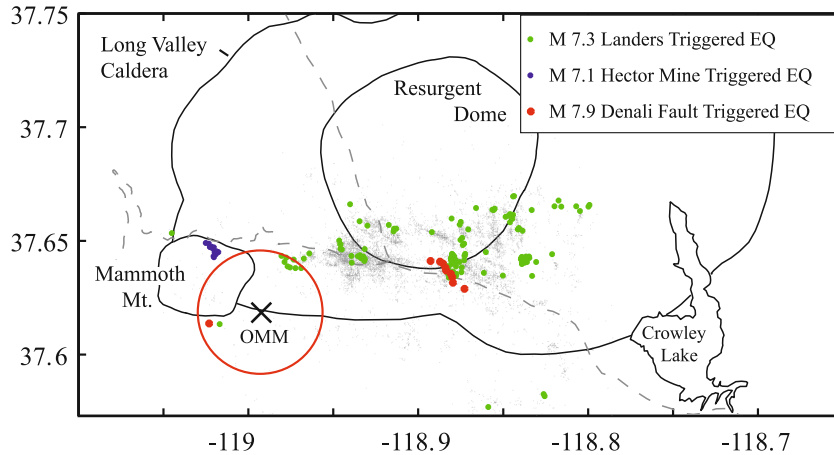
Through examining historical documents [59] identify several earthquakes in extensional/transtensional environments that may have been dynamically triggered by the $M_w = 7.8$ 1906 San Francisco earthquake, including a $M 3.5$ and $M 4.5$ earthquake in western Nevada and a $M 6.1$ earthquake in the Brawley Seismic Zone near the Salton Sea in Southern California. These events are within 400–700 km from the fault rupture, thus beyond the after-shock zone of the San Francisco earthquake.

In the day following the arrival of surface waves from the $M_w = 7.4$ Izmit, Turkey earthquake, catalog seismicity rates throughout continental Greece, 400–1000 km from the epicenter, increased significantly [6]. Greece is an area of active extension and hosts significant hydrothermal activity. Although [6] did not address a possible correlation to hydrothermal activity systematically, at least some clusters of dynamically triggered seismicity occurred in areas with active hot springs.

A second report of dynamically triggered seismicity in Europe comes the Roer Valley, the Netherlands. This area is an actively extending northern branch of the Rhine Graben System. Following a $M_w = 5.4$ earthquake in 1992, [12] determine that a large cluster of aftershocks occurred at distance of 40 km from the mainshock. They conclude that these events are dynamically triggered because they are located beyond the zone where static stress changes are significant.

Transpressional and Compressional Environments

Although dynamic triggering is not commonly observed in compressional environments, several studies suggest it does occur. Less than three hours after a $M_s = 7.3$ earthquake in the Gulf of Aqaba 1995, an earthquake swarm began 500 km distant from the mainshock epicenter in a restraining bend of the Dead Sea transform fault on the Syria–Lebanon border [62]. The swarm consisted of 21 earthquakes of $M_d \leq 3.7$.



Earthquakes, Dynamic Triggering of, Figure 7

Map of triggered seismicity beneath Long Valley caldera and Mammoth Mountain, California, for the Landers (green), Hector Mine (blue), and Denali Fault (red) earthquakes. Gray dots show background seismicity from 1997–1998. The red circle centered on station OMM indicates area within which the earthquakes triggered by the Denali Fault earthquake must be located based on S-P phase arrival times. The single red dot was large enough to be located [74]. Modified from [38], *Treatise on Geophysics*

The central United States is a transpressional environment with low strain rates. Dynamic triggering in the central US has not yet been observed instrumentally. However, [40,44,66] suggest that dynamic triggering occurred during the 1886 Charleston, South Carolina earthquake and 1811–1812 New Madrid earthquakes based on examination of historical felt reports. Similarly, [42] describe historical evidence for dynamic triggering of a $M \sim 7$ earthquake following the 1905 Kangra earthquake in India.

The stress state in Taiwan is variable, but generally transpressional [104]. [102] searched for dynamically triggered seismicity in the Taiwan region following 12 regional $M \geq 6.5$ earthquakes occurring between 1973 and 1994. They identify 9–10 cases of increased seismicity following a large event, although the increase is small in all cases, with 1–7 $M \geq 4.5$ earthquakes more in the 15 days following the large earthquake than in the 15 days before.

Dynamic Triggering in Subcrustal Environments

The occurrence of dynamically triggered earthquakes in subcrustal environments has been investigated in subduction zones in South America and Japan. [96] found that a $M \geq 7.6$ earthquake at 598 km depth in the Tonga trench in 2002 was followed by $M \geq 5.9$ and $M \geq 7.7$ earthquakes at 647 and 664 km depth within 2 and 7 minutes of the initial earthquake, respectively. By investigating the rupture history and Coulomb stress change resulting from the initial event, they conclude that the secondary events were triggered dynamically. They highlight 4 additional earth-

quakes of > 450 km depth which have similar large aftershocks that may be dynamically triggered. During the surface waves of the $M \geq 8.1$ Tokachi-oki earthquake, [60] identified deep low frequency earthquakes triggered in the Nankai subduction zone through analyzing Hi-Net borehole seismic data and use of the Beta statistic. The triggering occurred during a slow slip event in a region of the subduction zone which was active with deep low frequency tremor.

Triggered Tremor

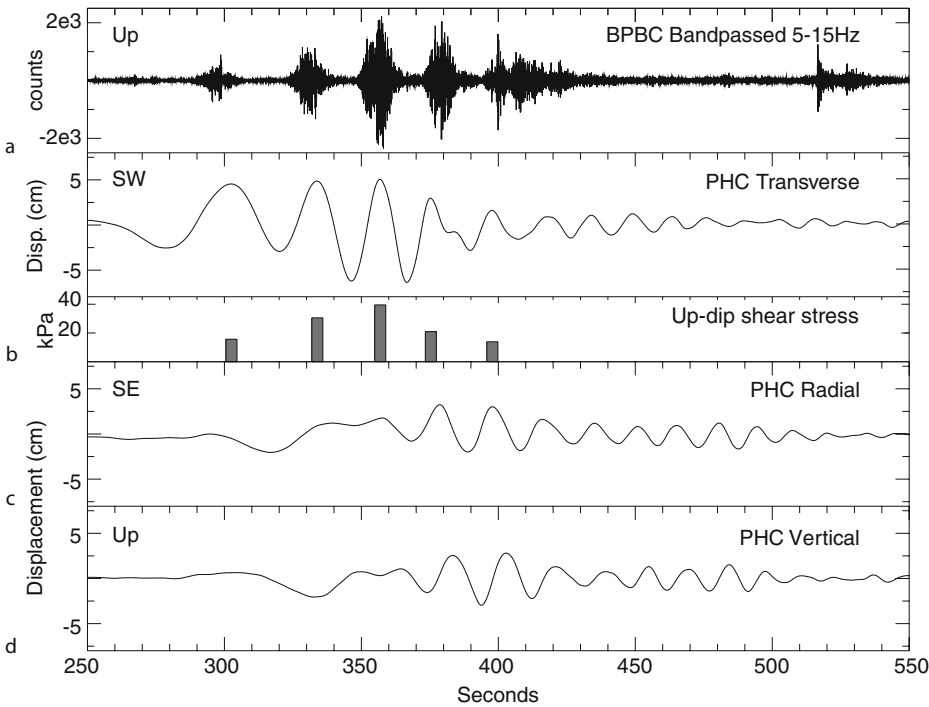
With the exception of triggered subduction zone seismicity described above, the majority of dynamically triggered earthquakes described in this review are typical brittle failure earthquakes. For example, [43] show that the earthquakes triggered by the Hector Mine earthquake near the Salton Sea had typical spectra and stress drops, consistent with standard brittle failure source mechanism. In the last few years however researchers have demonstrated that volcanic tremor and deep non-volcanic tremor respond to dynamic waves from regional and teleseismic earthquakes as well as typical crustal earthquakes (Table 2) ([32,60,61,81]). These findings emphasize that dynamic triggering can occur in a wide variety of environments and affect multiple seismic processes in addition to brittle failure of crustal rock. They provide an intriguing new perspective on the triggering processes.

At Aso volcano, Japan, [61] identify dynamically triggered earthquakes and volcanic tremor following the 1999

Earthquakes, Dynamic Triggering of, Table 2
Published occurrences of dynamically triggered tremor

Site	Citation	Triggering Earthquake	M**	Type of tremor	Responsible phase
Aso volcano, Japan	[61]	Chi-Chi, 1999	7.7	Shallow volcanic	P waves
Cascadia subduction zone, Canada	[81]	Denali Fault, 2002	7.9	Non-volcanic subduction zone	Love waves
7 sites throughout California	[32]	Denali Fault, 2002	7.9	Non-volcanic	Surface waves

** M is the magnitude of the triggering earthquake.



Earthquakes, Dynamic Triggering of, Figure 8

Time series showing tremor triggered by Love waves from the Denali Fault Earthquake in the Cascadia subduction zone: **a** Tremor at station BPBC, time adjust to correct for travel time from source to seismometer, **b–d** Displacement seismograms for transverse, radial, and vertical components at station PCH, the closest 3 component broadband station to the tremor, time adjust to correct for travel time from source to seismometer. Tremor occurs when the Love wave displacement is to the SW. Figure reprinted from [81], Nature

M 7.7 Chi-Chi earthquake. To test the uniqueness of these observations, they searched for triggered tremor at Aso following 20 other $M_w \geq 7$ earthquakes occurring within 3000 km distance between 1995 and 2002. Five of these earthquakes triggered tremor following P wave arrivals at Aso. All occurred between 1998 and 1999, a time with usually high heat supply to the volcano’s crater. As yet, this is the only documented episode of dynamically triggered volcanic tremor.

On the other side of the Pacific Ocean and a different tectonic environment, [81] identified episodes of deep non-volcanic tremor in the Cascadia subduction zone,

Canada, which were triggered by the Love waves of the M 7.9 Denali Fault earthquake. In this case tremor amplitude modulates perfectly with strain amplitude from the incident Love waves (Fig. 8).

More recently, [32] identified triggered non-volcanic tremor in seven locations in California following the Denali Fault earthquake. In all cases tremor amplitude modulates with strain amplitude from incident surface waves. Five of these are strike-slip faulting regimes. These observations are the first reported cases of non-volcanic tremor beyond subduction zones (e. g. [80]) and the San Andreas fault [67].

Lack of Triggering Observations

Interestingly, some areas of high ambient seismicity show a notable lack of dynamically triggered seismicity. For example, the San Andreas fault near Parkfield, California showed no triggered response to the Landers earthquake [90]. Japan boasts high rates of shallow background seismicity, frequent large earthquakes from the subduction zone, high seismic network density, and a variety of crustal stress environments and volcanic and geothermal regions. However, through examining both earthquake catalogs and waveform data from individual seismic stations before and after nine large remote events, [33], show that dynamic triggering in Japan is not common, as it is in extensional regimes of the Western United States.

Similarly, Alaska abounds with crustal and subduction zone seismicity and active volcanic and geothermal systems, although network density is far lower than in Japan. Though the Katmai Volcanic cluster appears to be particularly susceptible to triggering [64,73] and dynamic triggering has been observed at Mt. Wrangell [103], dynamic triggering is rare compared to the western United States. [64] suggest that this results from unknown differences in the magmatic and hydrothermal systems of the volcanoes. [83] document a decrease in seismicity at Mt. Wrangell and Veniaminof volcanoes following the M 7.9 Denali Fault earthquake. To date these are the only documented examples of seismicity repression from a large distant earthquake.

Characteristics of Dynamic Triggering

Environmental Controls on Dynamic Triggering

Extensional and transtensional tectonic regimes with high levels of background seismicity are highly susceptible to dynamic triggering [38]. This may reflect the ease with which fluids can migrate upwards in these stress environments [32,38]. Because such fluids are often hot with high concentrations of dissolved solids, rapid precipitation may form high pressure compartments over rapid time scales, further enhancing a tendency toward failure. Faults in extensional stress regimes are also inherently weak compared to those in compressional environments [43,88]. [26,41] suggest that dynamic triggering is a ubiquitous process in the crust which is detected more commonly in certain areas due to high instrumentation and scrutiny levels. Only one study to date has carefully addressed this question. By comparing seismicity rates on the San Andreas fault in California and the Western United States Basin and Range Province, [90] show that the San Andreas fault is less likely to experience dynamic trigger-

ing than similarly instrumented areas with similar levels of background seismicity in the Western United States Basin and Range province. More studies like [90] are necessary to resolve whether triggering is truly ubiquitous or favored in specific tectonic environments.

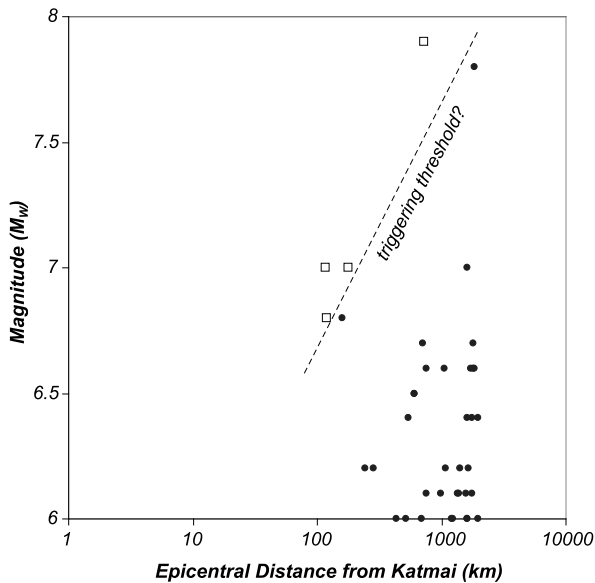
Triggering Thresholds and Recharge Times

In most reports of remote dynamic triggering, seismicity is triggered by earthquakes of M 6.5 or greater (Table 1). Dynamic triggering responses are strongest in areas that experience strong directivity [26,30,39]. These first order observations suggest that strength of triggered response is a function of ground shaking amplitude. Although amplitude-based triggering thresholds have been suggested for some areas [28,29,31,64], a consistent triggering threshold that applies throughout the crust has not been established [38]. Large earthquakes regularly occur without dynamically triggering seismicity beyond their aftershock zones.

Lack of triggering reports below M 6.5 may reflect subtle triggered responses. [41] uses the Beta statistic to give evidence of small seismicity increases at distances of 70–110 km in the month following 14 moderate (M 5.5–7) earthquakes in California. Because this distance corresponds with where a large SMS phase should arrive, [41] suggests that the SMS phase is responsible for the triggered response in these cases.

If a simple amplitude-of-shaking threshold is required to dynamically trigger earthquakes, we would expect that even moderate earthquakes trigger seismicity near their epicenters. [21,27,52,71] give strong evidence that dynamic triggering occurs in the near field. Because it is difficult to distinguish the influence of static and dynamic stress changes in the near field, many studies of dynamic triggering have limited their investigation to the realm beyond the aftershock zone.

Because many aftershocks in the near field are likely dynamically triggered, [31] include aftershocks in a search for an amplitude-based triggering threshold. They find that peak dynamic stress distributions correlate well with aftershock and remotely triggered seismicity distributions, except in the Long Valley caldera, CA. The result of [31] is consistent with failure thresholds found in laboratory studies [49] and independent of frequency of shaking. [64] also find evidence for a ground shaking amplitude-based triggering threshold at the Katmai Volcanic Cluster, Alaska by comparing magnitude and distance of mainshock with triggered response (Fig. 9). Their magnitude-distance relationship is similar to that proposed by [28] for the Geysers, CA. However, the triggering



Earthquakes, Dynamic Triggering of, Figure 9

Plot of magnitude vs. distance from Mageik volcano in the Katmai Volcanic Cluster for all $M_w > 6$ earthquakes between 1996 and 2003 located within 2000 km of Katmai. Hollow squares triggered seismicity in the KVC. Solid circles did not. Dashed line represents possible triggering threshold. Figure reprinted from [64], BSSA

threshold at Katmai appears to be higher than that suggested for the Geysers.

In other cases, a simple amplitude-of-shaking threshold is not consistent with data, and large amplitude ground shaking is neither a necessary nor sufficient condition to cause dynamically triggered earthquakes. [31] show that the Long Valley caldera appears to be more susceptible to triggering than other areas they studied. Because their study was based on catalog seismicity, it did not include triggered earthquakes that were too small to appear in earthquake catalogs, such as those at the Coso geothermal field in response to the Denali Fault earthquake. These events were triggered by dynamic stresses of < 0.01 MPa [74] and, like Long Valley, would not fit the thresholds proposed by [29] and [31].

By comparing spectra of all earthquakes with high amplitude ground shaking in the Long Valley caldera [7] find that in this area, high-amplitude low-frequency shaking is more likely to trigger seismicity than high-amplitude high-frequency shaking. [1] come to the same conclusion after examining strong ground shaking spectra of earthquakes which did and did not trigger seismicity in the Western Great Basin. Longer wave lengths associated with low frequency ground shaking favor triggering by larger earthquakes in at least some locales. Whether remote dy-

namic triggering in both the near and far field results from the same physical process or processes remains an open question.

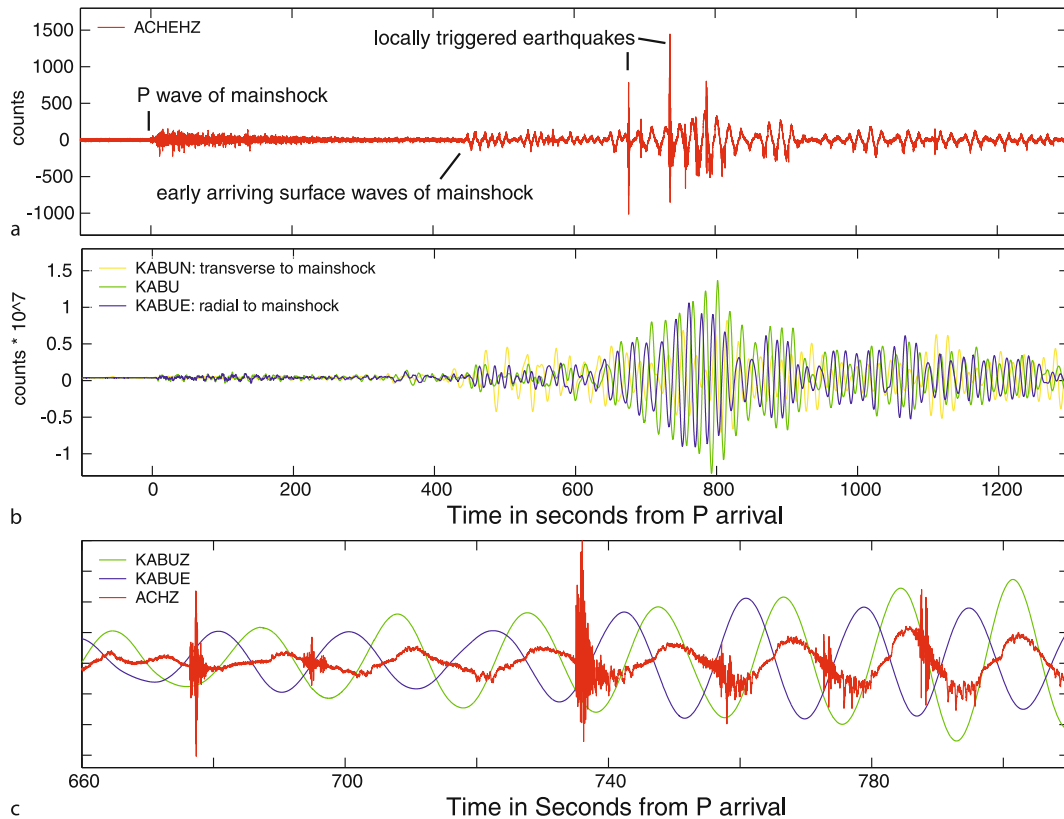
One parameter that may complicate the search for a triggering threshold in amplitude and/or frequency is recharge time. Because the occurrence of earthquakes releases stored strain energy, it may take time for an area to re-accumulate strain energy sufficiently to be primed for failure again following local earthquake activity or previous episodes of remotely triggered seismicity [38]. However some areas, such as the Geysers geothermal field require little to no time to recharge, as triggered seismicity episodes have been separated by time intervals of months or less [28]. Recharge times are dependent on many parameters including earthquake history, regional tectonic strain rates, and mass and heat advection rates in areas of hydrothermal and volcanic activity.

Time Scales of Dynamic Triggering and Responsible Phases

Remote dynamic triggering of earthquakes occurs over a variety of time scales following the onset of dynamic stressing. At Aso Volcano, Japan [61] triggered tremor begins with the P-wave arrival from distant large earthquake. In the case of discrete earthquakes however, it is most common for triggering to begin during surface wave arrivals (Table 1), leading many to suggest that the specific low-frequency large-amplitude ground motions associated with surface waves initiate the failure process [1,7,38,103]. Although the onset of dynamic triggering at remote locations is most commonly observed during Rayleigh wave arrivals [38], clear cases of remote triggering of tremor on the Love wave exist as well [81].

In some cases, dynamic triggering begins hours to days after an initial stress perturbation (e.g. [39,89,102]), hinting that the physical process responsible for initiating earthquake failure evolve with time. For example the largest triggered event following the M 7.3 Landers earthquake, a M 5.6 at the Little Skull Mountain, Nevada, occurred 33 hours after the mainshock [39]. In the case of Long Valley caldera's south moat and Mt. Rainier after the Denali Fault earthquake, delayed earthquake swarms began 24 hours and 2 hours respectively after the passage of the dynamic waves from the mainshock (Fig. 4) [74]. Both of these areas also had much smaller triggered swarms during the mainshock's wavetrain.

Determining the duration and decay time of triggered swarms is more difficult than detecting their onsets, particularly in areas of high ambient seismicity. Many triggered earthquakes may be triggered secondarily as aftershocks



Earthquakes, Dynamic Triggering of, Figure 10

Phase modulated dynamically triggered earthquakes in the Katmai Volcanic Cluster following the 2007 M 8.2 Kurile earthquake: **a** short period record from station ACH showing both wavetrains for the Kurile earthquake and the larger amplitude, locally triggered earthquakes, **b** broadband record from station KABU showing wave motion of the Kurile earthquake, **c** time series from ACH and KABU zoomed in to show how local earthquakes seen clearly in red are occurring on a specific phase of the wavetrain from the Kurile earthquake

to earlier triggered earthquakes [9]. The Yellowstone response to the Denali Fault earthquake and the Long Valley caldera response to Landers are fit well with an Omori-type law decay [45]. In some cases however, triggered swarms end abruptly after the dynamic stress perturbation stops (e. g. [103]). Although our understanding of decay rates of triggered swarms is incomplete emphasizing that the subject deserves further investigation, decay rates give strong constraints on physical processes responsible for triggering.

Phase Modulated Triggering

Recent findings show that earthquakes can be triggered during specific phases of the wavetrain. At Mt. Wrangell, Alaska, triggered earthquakes occurred preferentially during phases of the largest positive vertical ground displacement from the 2004 M 9.0 Sumatra earthquake [103]. Sim-

ilarly at Katmai Volcanic Cluster, Alaska, triggered earthquakes occurred only during specific phases of Rayleigh waves from the 2007 M 8.2 Kurile earthquake (Fig. 10). Such observations will allow us to resolve the precise dynamic stress field perturbations at the moment of earthquake nucleation on specific failure planes (e. g. [38]).

Physical Models of Dynamic Triggering

The wide variations in the characteristics of dynamic triggering and the limited data for individual response instances admit a spectrum of competing models for the physical processes linking dynamic stresses from a large, distant earthquake to the locally triggered response. Broadly considered, published models fall into three partially overlapping categories: 1) those involving some form of stress-driven brittle failure across local fractures, 2) those involving the activation of hydrous or mag-

matic fluids, and 3) those involving some form of localized aseismic relaxation (deformation). The brittle failure models are generally consistent with the onset of locally triggered seismicity during dynamic stressing (rapid-onset triggering), including the possibility that seismicity may persist as aftershocks for some time after the dynamic stressing has stopped [21]. Under the latter two categories, the onset of local seismicity represents a second-order phenomenon driven by a first order response to dynamic stressing in the form of fluid activation or transient deformation. In principle, models under these two categories admit a significant delay in the onset of the triggered seismicity with respect to the dynamic stresses generated by a distant earthquake. Because the dynamic stress amplitudes that trigger a response at remote distances are typically an order of magnitude or more below background tectonic stress levels, all models carry the implicit assumption that a crustal volume susceptible to dynamic stress triggering must be in a near-critical stress state prior to a triggered response.

Brittle Failure

Brittle failure models are based on the premise that the dynamic stresses propagating with the seismic waves from a distant earthquake are sufficient to nudge the local stress acting on a pre-existing dislocation beyond the threshold for the particular failure mode. This threshold may be the Griffith criteria for the tensile strength of a partially healed crack or the Coulomb criteria for frictional strength of a fault [86,87]. Crustal fluids play an important passive role in all brittle failure models by counteracting the rock matrix stress acting on a fracture through pore pressure, p , according to

$$\sigma' = \sigma - Ip$$

where σ' and σ are the effective and rock matrix stress tensors, respectively, and I is the identity tensor. Thus, pore pressure reduces the effective normal stress, σ'_n , acting on a fracture by opposing the rock matrix normal stress as $\sigma'_n = \sigma_n - p$. Alternatively, for pressure-sensitive friction models the role of pore pressure can be expressed in terms of an effective coefficient of friction as $\mu' = \mu(1 - \lambda_p)$, where $\lambda_p = p/\sigma_n$. Elevated pore pressures lower the effectively strength by moving the background stress state closer to extensional or shear failure thresholds thereby increasing vulnerability for failure by imposition of small dynamic stress perturbations.

In the simplest frictional failure model, a triggered earthquake occurs when the stress acting on a fault exceeds the Coulomb threshold for static friction, or $CFF(t) = 0$,

and friction abruptly drops from static to dynamic values with $\mu_s > \mu_d$, respectively. Here, $CFF(t)$ is the Coulomb Failure Function defined as

$$CFF(t) = |\tau(t)| - \mu_s \sigma'_n(t) - C, \quad \text{or its equivalent} \\ = |\tau(t)| - \mu'_s \sigma_n(t) - C$$

where σ_n , σ'_n , μ_s , μ'_s are defined in the preceding paragraph, τ is the shear, and C is the cohesive strength ([34], and references therein). This simple case implies rapid-onset triggering with the triggered seismicity beginning promptly when $CFF(t)$ first becomes positive for a fault optimally oriented for failure in the background stress field. The combination of dynamic stress components $\Delta\tau$ and $\Delta\sigma_n$ for which $CFF > 0$ will depend on the wave type (e.g. Love or Rayleigh wave) and its incidence angle on the optimally oriented fault [35]. Although details vary, Love waves will generally have a greater triggering potential than Rayleigh waves when incident on vertical, strike-slip faults while the opposite is the case for incidence on inclined, dip-slip faults.

The Coulomb failure criterion applies to more elaborate non-linear friction models as well (see [18,25,31,70,100]). Because the behavior of non-linear models depend on factors such as slip history and slip rate, however, the failure threshold for static friction may vary with time, and the triggered earthquake may be delayed with respect to the time the failure criterion was first exceeded (e.g. [69]). Susceptibility to dynamic triggering may result when a dynamic stress is imposed on quasi-static loading under a conditionally stable regime (e.g. [84]). Based on their analysis of the dynamic triggering observed at Long Valley caldera, [7] conclude that this mechanism requires near-lithostatic pore pressures to be effective.

Models based on the non-linear response of granular media to dynamic stresses may apply to dynamic triggering of mature faults with a well-developed core of fault gouge. [49] document an abrupt decrease in the modulus of fault gouge under low effective normal stress (σ'_n 0.1 MPa) when excited by dynamic strains $> 10^{-6}$ in the laboratory. Thus, this model also requires near-lithostatic pore pressures to be effective.

Sub-critical crack growth, or stress corrosion, is another non-linear form of brittle failure that has a potential role in dynamic triggering. Under this model, a sudden increase in differential stress or an oscillatory stress applied to a pre-existing crack can lead to crack growth due to weakening of the crack tip by chemical corrosion. This can shorten the time to earthquake rupture. This process will be enhanced in an environment with fluids at elevated temperatures. It turns out that the equations governing sub-critical crack growth have the same mathemat-

ical form as rate-state friction equations above [51]. Thus near-lithostatic pore pressure appears to be a requirement for each of these non-linear brittle-failure models, at least as they apply to dynamic triggering at remote distances.

Fluid Activation Models

In addition to their passive role in reducing the effective strength of a rock volume through ambient pore pressure, fluids may play an active role in the dynamic triggering process. Fluid activation models generally appeal to either 1) pore-pressure re-distribution associated with changes in permeability and fluid transport, or 2) state changes induced in multi-phase fluids.

Dynamic stressing may be capable of physically disrupting permeability barriers separating volumes of differing pore pressure. For example, dynamic stress may shake accumulated detritus from clogged fractures or opening partially healed fractures by extensional failure. In either case, fluid diffusion down the pressure gradient will result in a re-distribution of pore pressure with the potential for triggering seismicity in previously under-pressured volumes in a near-critically stressed state. The evolution of triggered seismicity in this case will be governed by the diffusion length for a given permeability and the proximity of the pre-existing stress state to brittle failure. [8] proposed the clogged fracture model as an explanation for the hydrologic response of water wells in southern Oregon to surface waves from $M > 7$ earthquakes at distances of 300 km and 3850 km.

Geothermal areas may be particularly susceptible to dynamic triggering through pore pressure re-distribution. In these areas fractures are rapidly sealed by precipitation from circulating, solute-rich geothermal fluids and plastic deformation of quartz-rich rocks under elevated temperatures tend to isolate pockets of elevated pore pressure. Most active geothermal systems are located in areas of extensional tectonism. In these areas normal stresses induced by Rayleigh waves on vertical planes may open vertical fractures, allowing high-pore-pressure fluids access to shallower crustal volumes with lower pore pressure [35]. The hydraulic surge model described by [22] for volcanic and geothermal systems is a version of this process in which the brittle-plastic transition at the base of the seismogenic crust serves as a low-permeability barrier separating near-lithostatic pore pressures in the plastic regime from a hydrostatic regime in the overlying seismogenic crust. Rupturing the permeability seal by dynamic stresses would release near-lithostatic pore pressures into the brittle, seismogenic crust thereby inducing a surge in triggered seismicity.

Models for bubble excitation by dynamic stresses in a two-phase fluid (multi-phase in a partially crystallized magma) offers interesting possibilities for remotely triggered responses in geothermal and volcanic systems. This is a particularly intriguing concept for remote triggering in volcanic systems because of the importance of bubbles in eruption dynamics [57] and the source mechanisms of long-period volcanic earthquakes [13], ► [Volcanoes, Non-linear Processes in](#). Advective overpressure and rectified diffusion were the first bubble models proposed as explanations for remotely triggered seismicity [10,55,94]; although subsequent work has shown that both hold less promise as viable explanations than initially thought [56].

Under the advective overpressure model, the pressure in a gas-saturated, incompressible fluid confined in a rigid container increases as $\rho g \Delta h$ as a pre-existing bubble adhering to the wall of the container is shaken loose by passing seismic waves. The bubble ascends buoyantly a distance Δh through a fluid of density ρ where g is the acceleration of gravity [55]. The resulting pressure increase in the container (magma body) deforms the surrounding rock inducing small earthquakes. This model is criticized on the basis that assumptions of a ridged container and an incompressible fluid seriously violate realistic conditions in the earth [75].

Under rectified diffusion, pressure oscillations imposed on a gas-saturated fluid with pre-existing bubbles pump gas into the bubbles over multiple cycles. Gas exolves from the fluid into the bubble during the dilatational phase, when bubble surface area is maximal, and out of the bubble back into solution during compressional phase, when the bubble surface area is minimal [94]. The implied pressure gain integrated over multiple cycles is then transmitted to the surrounding rock inducing small earthquakes. [48] point out, however, that the effectiveness of this model is limited by reasonable gas diffusion rates in hydrous fluids or magma with respect to the frequencies of seismic waves driving the pressure oscillations.

More promising bubble models appeal to the strong sensitivity of bubble nucleation rate to the supersaturation pressure [56] and the results of numerical models by ► [Volcanoes, Non-linear Processes in](#) and [14,85], indicating that a small pressure drop imposed on a densely packed matrix of tiny bubbles can lead to rapid, diffusion-driven bubble growth. The implications of these models, however, have yet to be more fully explored in the context of dynamic triggering.

Two more speculative models involve magma instabilities triggered by dynamic stresses. In one, a loosely held crystal mush accumulated on the walls of a crystal-

lizing magma body may be dislodged by dynamic shaking. The sinking crystal mush would induce a convective plume as it displaced hotter, less dense magma. In the case of volatile-rich magma, buoyant convection would be enhanced by bubble nucleation and growth as confining pressure drops with decreasing depth [37]. Under suitable conditions, the resulting pressure increase within the magma body could evolve over days [56]. If the magma chamber was already in a near critical state, the culmination could be magma intrusion into the overlying crust or the onset of an eruption. Whether this process culminates in a simple pressure increase, an intrusion, or an eruption, the sensible onset of locally triggered seismicity and deformation might be delayed by hours to perhaps days with respect to the passing seismic waves from the distant earthquake. A second, even more speculative model appeals to dynamic stresses disrupting the solid matrix of a partially crystallized magma body thereby releasing any differential tectonic stress sustained by the solid matrix [36,37]. As the magma body relaxes with a time constant governed by the effective viscosity of the disrupted crystal mush, stress would be transferred to the surround crust inducing deformation and local seismicity. In essence, this model corresponds to the relaxation of an Eshelby inclusion in an elastic medium [20].

Aseismic Deformation

The relaxing magma body of the previous paragraph is one example of aseismic deformation with the potential of triggering local deformation and the onset of secondary seismicity. A less speculative example involves aseismic creep on faults triggered by dynamic stressing. Deformation associated with fault creep transfers stress to the adjacent crust, which in turn triggers local seismicity, as in the example involving seismicity triggered on the Reykjanes Peninsula following the $M = 6.5$ earthquake in the South Iceland Seismic Zone in 2000 [2]. [4] document aseismic fault slip (creep) on faults in the Salton Trough of southern California triggered by the three $M > 6$ earthquakes in the Landers, California sequence of 1992 (the $M = 6.1$ Joshua Tree, $M = 7.3$ Landers, and the $M = 6.2$ Big Bear earthquakes). In this case, all instances of triggered slip were on faults within 150 km of the $M > 6$ earthquakes. In these examples and observations from triggering in Long Valley caldera and Cienega Prieta geothermal field in Baja California [38,50] the geodetic moment for triggered aseismic deformation exceeds the cumulative seismic moment for the triggered earthquakes by a factor of two or more. This emphasizes the importance of high-resolution deformation monitoring in ar-

ear susceptible to dynamic triggering for resolving the role of aseismic deformation in the dynamic triggering process.

Future Directions

In the last 25 years, in the wake of the Landers earthquake, the study of dynamically triggered seismicity has given us new insight into earthquake initiation and the failure regime in the Earth's crust. Some argue that the state of stress in the crust is highly spatially variable [77]. Given this, the likelihood of triggering seismicity would also be spatially variable. [97] and [105], however, argue that the Earth's crust is critically stressed and on the verge of failure nearly everywhere. If this were the case, one might expect triggering due to small dynamic stress perturbations to be a ubiquitous phenomenon. In either case, the study of remotely triggered seismicity provides clues to spatial distribution of critically stressed crustal volumes.

Unfortunately, although we can measure stress field perturbations from dynamic waves from earthquakes, we rarely have a detailed understanding of the background stress field these perturbations are modulating. In addition, dynamically triggered earthquakes are often too small or occur in too sparsely instrumented areas to resolve reliable focal mechanisms. Because of these limitations, our understanding of how dynamic stresses from remote earthquake wavetrains induce a given crustal volume to respond with triggered seismicity remains incomplete. Advances will require more cases of dynamically triggered seismicity captured by both spatially dense seismic networks and continuous, high-resolution deformation monitoring networks.

Recent observations of phase modulated dynamic triggering offer powerful datasets of the precise time history of dynamic stress triggering. Because similarities exist between phase modulated dynamic triggering of seismicity in the shallow crust of a volcano's edifice [103] and deep in a subduction zone [81], the emerging study of non-volcanic tremor may provide new leverage on understanding how dynamic stresses influence seismic slip.

As new observations of dynamically triggered seismicity are reported, one conclusion is becoming increasingly evident: multiple causative processes exist. The wide variety in time scales over which triggering occurs and the spatial and temporal characteristics of triggered seismicity sequences and associated deformation responses cannot be fit with any one model yet proposed. Rather, different models are consistent with different episodes of triggering. For example, fluid activation and stress corrosion models

are most applicable in volcanically and geothermally active environments. In some cases, such as the complex triggered response of Yellowstone, Mt. Rainier, and the Long Valley caldera areas to the Denali Fault earthquake, multiple processes may be occurring simultaneously in the same locale, yet on different time scales. Of the physical models described above, seismicity triggered instantaneously or within seconds of the dynamic stress perturbation is consistent with models based on simple brittle failure, brittle failure with nonlinear friction effects, stress corrosion, unclogging of fractures, or rectified diffusion, whereas triggered seismicity delayed by hours to days is more consistent with models involving aseismic deformation, advective overpressure, sinking crystal plumes, or a relaxing magma body.

In the few cases where hydrologic and high-sample rate strain data are available, dynamically triggered seismicity is accompanied by changes in water levels in wells [79] and significant deformation signals [2,25,36,50]. A complete understanding of dynamic triggering will require research approaches that integrate seismic, deformation, and hydrologic datastreams. To this end, we challenge Earth scientists to broaden their thinking and tap these observations to better understand the initiation of earthquake failure.

Bibliography

Primary Literature

- Anderson JG, Brune JN, Louie JN, Zeng Y, Savage M, Yu G, Chen Q, de Polo D (1994) Seismicity in the western Great Basin apparently triggered by the Landers, California, earthquake, 28 June 1992. *Bul Seismo Soc Am* 84:863–891
- Arnadottir T, Geirsson H, Einarsson P (2004) Coseismic stress changes and crustal deformation on the Reykjanes Peninsula due to triggered earthquakes on 17 June 2000. *J Geophys Res*. doi:10.1029/2004JB003130
- Blackett RE, Wakefield S (2002) Geothermal resources of Utah, Utah geological survey open file report 397, ISBN 1-55791-677-2
- Bodin P, Bilham R, Behr J, Gomberg J, Hudnut K (1994) Slip triggered on southern California faults by the 1992 Joshua Tree, Landers, and Big Bear earthquakes. *Bul Seismo Soc Am* 84:806–816
- Bodin P, Gomberg J (1995) Triggered seismicity and deformation between the Landers, California and Little Skull Mountain Nevada earthquakes. *Bul Seismo Soc Am* 84:835–843
- Brodsky EE, Karakostas V, Kanamori H (2000) A new observation of dynamically triggered regional seismicity: earthquakes in Greece following the August, 1999, Izmit, Turkey earthquake. *Geophys Res Lett* 27:2741–2744
- Brodsky EE, Prejean SG (2005) New constraints on mechanisms of remotely triggered seismicity at Long Valley Caldera. *J Geophys Res*. doi:10.1029/2004JB003211
- Brodsky EE, Roeloffs E, Woodcock D, Gall I, Manga M (2003) A mechanism for sustained groundwater pressure changes induced by distant earthquakes. *J Geophys Res*. doi:10.1029/2002JB002321
- Brodsky EE (2006) http://www.pmc.ucsc.edu/~brodsky/reprints/Sus5_merged.pdf. Long-range triggered earthquakes that continue after the wavetrain passes. *Geophys Res Lett* 33:L15313
- Brodsky EE, Sturtevant B, Kanamori H (1998) Earthquakes, volcanoes, and rectified diffusion. *J Geophys Res* 103:23827–23838
- Brune J (1970) Tectonic stress and the spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Camelbeeck T, van Eck T, Pelzing R, Ahorner L, Loohuis J, Haak HW, Hoang-Trong P, Hollnack D (1994) The 1992 Roermond earthquake, the Netherlands, and its aftershocks. *Geologie en Mijnbouw* 73:181–197
- Chouet B (1992) A seismic model for the source of long-period events and harmonic tremor. In: Gasparini P, Scarpa R, Aki K (eds) *Volcanic seismology*. IAVCEI Proceedings in Volcanology. Springer, Berlin, pp 133–156
- Chouet B, Dawson P, Nakano M (2006) Dynamics of diffusive bubble growth and pressure recovery in a bubbly rhyolitic melt embedded in an elastic solid. *J Geophys Res*. doi:10.1029/2005JB004174
- Cocco M, Rice JR (2002) Pore pressure and poroelasticity effects in Coulomb stress analysis of earthquake interactions. *J Geophys Res* doi:10.1029/2002JB002319
- Cochran ES, Vidale JE, Tanaka S (2004) Earth tides can trigger shallow thrust fault earthquakes. *Science* 306:1164–1166
- Cooper HH, Bredehoeft JD, Papadopoulos S, Bennett RR (1965) The response of well-aquifer systems to seismic waves. *J Geophys Res* 70:3915–3926
- Dieterich JH (1979) Modeling of rock friction 1, Experimental results and constitutive equations. *J Geophys Res* 84:2161–2168
- Elkhoury JE, Brodsky EE, Agnew DC (2006) Seismic waves increase permeability. *Nature* 441:1135–1138
- Eshelby JD (1957) The determination of the elastic field of an ellipsoidal inclusion, and related problems. *Proceedings of the Royal Society of London A* 241:376–396
- Felzer KR, Brodsky EE (2006) <http://www.nature.com/nature/journal/v441/n7094/full/nature04799.html>. Decay of after-shock density with distance indicates triggering by dynamic stress. *Nature* 441:735–738
- Fournier RO (1999) Hydrothermal processes related to movement of fluid from plastic to brittle rock in the magmatic-epithermal environment. *Economic Geology* 94:1193–1211
- Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annual Rev Earth and Planet Sci* 33:1255–1256
- Glowacka E, Nava AF, Cossio DD, Wong V, Farfan F (2002) Fault slip, seismicity, and deformation in the Mexicali Valley, Baja California, Mexico, after the M 7.1 Hector Mine earthquake. *Bul Seismo Soc Am* 92:1290–1299
- Gomberg J, Blanpied ML, Beeler NM (1997) Transient triggering of near and distant earthquakes. *Bul Seismo Soc Am* 87:294–309
- Gomberg J, Bodin P, Larson K, Dragert H (2004) Earthquakes nucleated by transient deformations caused by the M = 7.9 Denali, Alaska, earthquake. *Nature* 427:621–624

27. Gombert J, Bodin P, Reasenber PA (2003) Observing earthquakes triggered in the near field by dynamic deformations. *Bul Seismo Soc Am* 93:118–138
28. Gombert J, Davis S (1996) Stress/strain changes and triggered seismicity at The Geysers, California. *J Geophys Res* 101:733–749
29. Gombert J, Johnson P (2005) Dynamic triggering of earthquakes. *Nature* 437:830
30. Gombert J, Reasenber PA, Bodin P, Harris R (2001) Earthquakes triggering by seismic waves following the Landers and Hector Mine earthquakes. *Nature* 411:462–465
31. Gombert J, Reasenber PA, Cocco M, Belardinelli ME (2005) A frictional population model of seismicity rate change. *J Geophys Res.* doi:10.1029/2004JB003404
32. Gombert J, Rubinstein JL, Peng Z, Creager KC, Vidale JE, Bodin P (in press) Widespread triggering of non-volcanic tremor in California. *Science* 319:117
33. Harrington RM, Brodsky EE (2006) The absence of remotely triggered seismicity in Japan. *Bul Seismo Soc Am* 96:871–878
34. Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. *J Geophys Res* 103:24347–24358
35. Hill DP (2008) Dynamic stresses, Coulomb failure, and remote triggering. *Bul Seismo Soc Am* 98:66–92
36. Hill DP, Johnston MJS, Langbein JO (1995) Response of Long Valley caldera to the Mw = 7.3 Landers, California, earthquake. *J Geophys Res.* doi:10.1029/1005GL024753
37. Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. *Physics Today* 55:41–47
38. Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) *Geophysical treatise, earthquake seismology*. Elsevier, Amsterdam
39. Hill DP, Reasenber PA, Michael A, Arabaz WJ, Beroza G, Brumbaugh D, Brune JN, Castro R, Davis S, dePolo D, Ellsworth WL, Gombert J, Harmsen S, House L, Jackson SM, Johnston MJS, Jones L, Keller R, Malone S, Munguia L, Nava S, Pechmann JC, Sanford A, Simpson RW, Smith RB, Stark M, Stickney M, Vidal A, Walter A, Wong A, Zollweg J (1993) Seismicity remotely triggered by the magnitude 7.3 Landers, California, earthquake. *Science* 260:1617–1622
40. Hough SE (2001) Triggered earthquakes and the 1811–1812 New Madrid, central United States, earthquake sequence. *Bul Seismo Soc Am* 91:1547–1581
41. Hough SE (2005) Remotely triggered earthquakes following moderate mainshocks (or why California is not falling into the ocean). *Seismological Research Letters* 76:58–66
42. Hough SE, Billham R, Ambraseys N, Field N (2005) Revisiting the 1897 Shillong and 1905 Kangra earthquakes in northern India: site response, Moho reflections and a triggered earthquake. *Current Science* 88:1632–1638
43. Hough SE, Kanamori H (2002) Source properties of earthquakes near the Salton Sea triggered by the 16 October 1999 Mw 7.1 Hector Mine, California, earthquake. *Bul Seismo Soc Am* 92:1281–1289
44. Hough SE, Seeber L, Armbruster JG (2003) Intraplate triggered earthquakes: observations and interpretation. *Bul Seismo Soc Am* 93:2212–2221
45. Husen S, Taylor R, Smith RB, Healsen H (2004) Changes in geyser eruption behavior and remotely triggered seismicity in Yellowstone National Park produced by the 2002 Mw 7.9 Denali fault earthquake, Alaska. *Geology* 32:537–540
46. Husen S, Wiemer S, Smith RB (2004) Remotely triggered seismicity in the Yellowstone National Park region by the 2002 Mw 7.9 Denali Fault earthquake, Alaska. *Bul Seismo Soc Am* 94:S317–S331
47. Husker AL, Brodsky EE (2004) Seismicity in Idaho and Montana triggered by the Denali Fault earthquake: a window into the geologic context for seismic triggering. *Bul Seismo Soc Am* 94:S310–S316
48. Ichihara M, Brodsky EE (2006) <http://www.pmc.ucsc.edu/~brodsky/reprints/2005GL024753.pdf>. A limit on the effect of rectified diffusion in volcanic systems. *Geophys Res Let.* doi: 10.1029/2005GL024753
49. Johnson P, Jia X (2005) Nonlinear dynamic, granular media and dynamic earthquake triggering. *Nature* 437:871–874
50. Johnston MJS, Prejean SG, Hill DP (2004) Triggered deformation and seismic activity under Mammoth Mountain in Long Valley caldera by the 3 November 2002 Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:S360–S369
51. Kanamori H, Brodsky EE (2004) http://www.pmc.ucsc.edu/~brodsky/reprints/rpp4_8_R03.pdf. The physics of earthquakes, Reports on Progress in Physics 67:1429–1496
52. Kilb D, Gombert J, Bodin P (2000) Triggering of earthquake aftershocks by dynamic stresses. *Nature* 408:570–574
53. King GCP, Cocco M (2001) Fault interactions by elastic stress changes: new clues from earthquake sequences. *Advances in Geophysics* 44:1–38
54. King GCP, Stein RS, Lin J (1994) Static stress changes and the triggering of earthquakes. *Bul Seismol Soc Am* 84:935–953
55. Linde AT, Sacks IS, Johnston MJS, Hill DP, Billham RG (1994) Increased pressure from rising bubbles as a mechanism for remotely triggered seismicity. *Nature* 371:408–410
56. Manga M, Brodsky EE (2006) Seismic triggering of eruptions in the far field: volcanoes and geysers. *Annual Rev Earth and Planet Sci* 34:263–291
57. Mangan M, Sisson T (2000) Delayed, disequilibrium degassing in rhyolite magma: decompression experiments and implications for explosive volcanism. *Earth and Planetary Science Letters* 183:441–455
58. Matthews MV, Reasenber PA (1988) Statistical methods for investigating quiescence and other temporal seismicity patterns. *Pure Appl Geophys* 126:357–372
59. Meltzner AJ, Wald DJ (2003) Aftershocks and triggered events of the great 1906 California earthquake. *Bul Seismo Soc Am* 93:2160–2186
60. Miyazawa M, Mori J (2005) Detection of triggered deep low-frequency events from the 2003 Takachi-oki earthquake. *Geophys Res Let* 32:L10307
61. Miyazawa M, Nakanishi I, Sudo Y, Ohkura T (2005) Dynamic response of frequent tremors at Aso volcano to teleseismic waves from the 1999 Chi-Chi, Taiwan earthquake. *J Vol Geotherm Res* 147:173–186
62. Mohamad RA, Darkal N, Seber D, Sandoval E, Gomez F, Barazangi M (2000) Remote earthquake triggering along the Dead Sea Fault in Syria following the 1995 Gulf of Aqaba earthquake ($M_s = 7.3$). *Seismol Res Let* 71:47–52
63. Moran SC (2003) Multiple seismogenic processes for high-frequency earthquakes at Katmai National Park, Alaska: evidence from stress tensor inversions of fault plane solutions. *Bul Seismo Soc Am* 93:94–108
64. Moran SC, Power JA, Stihler SD, Sanchez JJ, Caplin-Auerbach J (2004) Earthquake triggering at Alaskan volcanoes follow-

- ing the 3 November 2002 Denali Fault earthquake. *Bul Seismo Soc Am* 94:S300–S309
65. Moran SC, Zimbelman DR, Malone SD (2003) A model for the magmatic-hydrothermal system at Mount Rainier, Washington, from seismic and geochemical observations. *Bull Volcanol* 61:425–436
 66. Mueller K, Hough SE, Bilham R (2004) Analysing the 1811–1812 New Madrid earthquakes with recent instrumentally recorded aftershocks. *Nature* 429:284–288
 67. Nadeau R, Dolenc D (2005) Nonvolcanic tremors deep beneath the San Andreas Fault. *Science* 300:1942–1943
 68. Pankow KL, Arabasz WJ, Pechmann JC, Nava SJ (2004) Triggered seismicity in Utah from the 3 November 2002 Denali Fault earthquake. *Bul Seismo Soc Am* 94:S332–S347
 69. Parsons T (2005) A hypothesis for delayed dynamic earthquake triggering. *Geophys Res Lett* 32:L04302
 70. Perfettini HJ, Schmittbuhl J, Cochard A (2003) Shear and normal load perturbations on a two-dimensional continuous fault: 2. dynamic triggering. *J Geophys Res.* doi:10.1029/2002JB001805
 71. Pollitz FF, Johnston MJS (2006) Direct test of static-stress versus dynamic-stress triggering of aftershocks. *Geophys Res Lett* 33:L15318
 72. Pollitz FF, Sacks IS (2002) Stress triggering of the 1999 Hector Mine earthquake by transient deformation following the 1992 Landers earthquake. *Bul Seismo Soc Am* 92:1487–1496
 73. Power JA, Moran SC, McNutt SR, Stihler SD, Sanchez JJ (2001) Seismic response of the Katmai volcanoes to the 6 December 1999 magnitude 7.0 Karluk Lake earthquake, Alaska. *Bul Seismo Soc Am* 91:57–63
 74. Prejean SG, Hill DP, Brodsky EE, Hough SE, Johnston MJS, Malone SD, Oppenheimer DH, Pitt AM, Richards-Dinger KB (2004) Remotely triggered seismicity on the United States west coast following the Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:S348–S359
 75. Pyle DM, Pyle DL (1995) Bubble migration and the initiation of volcanic eruptions. *J Vol Geotherm Res* 67:227–232
 76. Reasenber P (1985) Second-order moment of central California seismicity, 1969–1982. *J Geophys Res* 90:5479–5495
 77. Rivera L, Kanamori H (2002) Spatial heterogeneity of tectonic stress and friction in the crust. *Geophys Res Lett* 29:10.1029/2001GL013803
 78. Roeloffs E (1998) Poroelastic techniques in the study of earthquake-related hydrologic phenomena. *Adv Geophys* 37:135–195
 79. Roeloffs E, Sneed M, Galloway DL, Sorey ML, Farrar CD, Howle JF, Hughes J (2003) Water level changes induced by local and distant earthquakes at Long Valley Caldera, California. *J Vol Geotherm Res* 127:269–303
 80. Rogers G, Dragert H (2003) Episodic tremor and slip on the Cascadia subduction zone: The chatter of silent slip. *Science* 296:1679–1681
 81. Rubinstein JL, Vidale JE, Gombert J, Bodin P, Creager KC, Malone S (2007) <http://www.nature.com/nature/journal/v448/n7153/full/nature06017.html>. Non-Volcanic Tremor Driven by Large Transient Shear Stresses. *Nature* 448:579–582
 82. Rundle JB, Turcotte D, Shcherbakov R, Klein W, Sammis C (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41:4/1019
 83. Sanchez JJ, McNutt SR (2004) Intermediate-term declines in seismicity at Mt. Wrangell and Mt. Veniaminof volcanoes, Alaska, following the 3 November 2002 Mw 7.9 Denali Fault earthquake. *Bul Seismo Soc Am* 94:S370–S383
 84. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391:37–42
 85. Shimomura Y, Nishimura T, Sato H (2006) Bubble growth processes in magma surrounded by an elastic medium. *J Vol Geotherm Res* 155:307–322
 86. Sibson R (2000a) A brittle failure mode plot defining conditions for high-flux flow. *Economic Geology* 95:41–48
 87. Sibson R (2000b) Fluid involvement in normal faulting. *Geodynamics* 29:469–499
 88. Sibson RH (1982) Fault zone models, heat flow, and the depth distribution of earthquakes in the continental crust of the United States. *Bul Seismo Soc Am* 72:151–163
 89. Singh SK, Anderson JG, Rodriguez M (1998) Triggered seismicity in the Valley of Mexico from major Mexican earthquakes. *Geofiscia International* 37:3–15
 90. Spudich P, Steck LK, Hellweg M, Fletcher JB, Baker LM (1995) Transient stresses at Parkfield, California, produced by the M 7.4 Landers earthquake of June 28, 1992: observations from the UPSAR dense seismograph array. *J Geophys Res* 100:675–690
 91. Stark MA, Davis SD (1996) Remotely triggered microearthquakes at The Geysers geothermal field, California. *Geophys Res Lett* 23:945–948
 92. Steacy S, Gombert J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. *J Geophys Res.* doi:10.1029/2005JB003692
 93. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
 94. Sturtevant B, Kanamori H, Brodsky E (1996) Seismic triggering by rectified diffusion in geothermal systems. *J Geophys Res* 101:25269–25282
 95. Tanaka S, Ohtake M, Sato H (2003) Tidal triggering of earthquakes in Japan related to the regional tectonic stress. *Earth Planets and Space* 56:511–515
 96. Tibi R, Wiens DA, Inoue H (2003) Remote triggering of deep earthquakes in the 2002 Tonga sequence. *Nature* 424:921–925
 97. Townend J, Zoback MD (2000) How faulting keeps the crust strong. *Geology* 28:399–402
 98. Ukawa M, Fujita E, Kumagai T (2002) Remote triggering of microearthquakes at the Iwo-Jima volcano. *J Geography* 111:277–286
 99. Unruh JR, Hauksson E, Monastero FC, Twiss RJ, Lewis JC (2002) Seismotectonics of the Coso Range – Indian Wells Valley region, California: Transtensional deformation along the southeastern margin of the Sierran microplate. *Geol Soc Am Mem* 195:277–294
 100. Voisin C (2002) Dynamic triggering of earthquakes: the nonlinear slip-dependent friction case. *J Geophys Res* 107(B12):10.1–10.11
 101. Weaver CS, Hill DP (1978/79) Earthquake swarms and local crustal spreading along major strike-slip faults in California. *Pageoph* 117:51–64
 102. Wen KL, Beresnev IA, Cheng S (1996) Moderate-magnitude seismicity remotely triggered in the Taiwan Region by large

earthquakes around the Philippine Sea Plate. *Bul Seismo Soc Am* 86:843–847

103. West M, Sanchez JJ, McNutt SR (2005) Periodically triggered seismicity at Mount Wrangell, Alaska, after the Sumatra earthquake. *Science* 308:1144–1146
104. Yabe Y, Song S, Wang C (2005) Stress state around Chelungpu Fault, Taiwan, Estimated from boring core samples. *EOS Trans* 86:T51A–1316
105. Zoback MD, Zoback ML (2002) State of stress in the Earth's lithosphere In: Lee WH, Kanamori H, Jennings PC, Kisslinger C (eds) *International handbook of earthquake and engineering seismology, Part A*. Academic Press, Amsterdam, pp 559–568

Books and Reviews

- Freed AM (2005) Earthquake triggering by static, dynamic, and postseismic stress transfer. *Annual Rev Earth and Planet Sci* 33:1255–1256
- Harris RA (1998) Introduction to a special section: Stress triggers, stress shadows, and implications for seismic hazards. *J Geophys Res* 103:24347–24358
- Hill DP, Pollitz F, Newhall C (2002) Earthquake-volcano interactions. *Physics Today* 55:41–47
- Hill DP, Prejean SG (2007) Dynamic triggering. In: Kanamori H (ed) *Geophysical treatise, earthquake seismology*. Elsevier, Amsterdam
- Manga M, Brodsky EE (2005) Seismic triggering of eruptions in the far field: volcanoes and geysers. *Annual Rev Earth and Planet Sci* 34:263–291
- Steady S, Gombert J, Cocco M (2005) Introduction to special section: Stress transfer, earthquake triggering, and time-dependent seismic hazard. *J Geophys Res*. doi:10.1029/2005JB003692

Earthquakes, Electromagnetic Signals of

SEIYA UYEDA¹, MASASHI KAMOGAWA²,
TOSHIYASU NAGAO¹

¹ Earthquake Prediction Research Center,
Tokai University, Shizuoka, Japan

² Department of Physics, Tokyo Gakuji University,
Koganei-shi, Japan

Article Outline

Glossary

Definition of the Subject

Introduction

Telluric Current Anomalies and Natural Time

Ultra Low Frequency (ULF) Anomalies

Higher Frequency Electromagnetic Emission
and Earthquake Light

Lithosphere-Atmosphere-Ionosphere (LAI) Coupling
Mechanism of Pre-Seismic EM Phenomena

Future Directions

Bibliography

Glossary

Earthquake prediction Place of epicenter, time of occurrence, and magnitude are the three main items of earthquake prediction. Occurrence time is the most difficult to predict. Depending on the concerned time scales, prediction is usually classified as long term (\sim tens of years), intermediate term (\sim a few years), and short term (months to days) predictions. Electromagnetic signals of earthquakes are mainly concerned with the short term prediction.

Piezo-electric effect Piezo-electricity is the electric polarization produced in certain crystals and ceramics by the application of mechanical stress. Among rock-forming minerals, quartz is most strongly piezo-electric, but its effect is much reduced because quartz crystals are usually randomly oriented. Moreover, stress-induced piezo-electric polarization in rocks is kept canceled by compensating charges. At rapid stress drop, bulk polarization appears as the compensating charge cannot disappear instantly and decays with a time constant $\tau = \epsilon/\sigma$, where ϵ is dielectric constant and σ electric conductivity.

Electro-kinetic effect Electro-kinetic effect, also called streaming potential, is caused by the presence of the solid-liquid interface. The double layer consists of ions (anions in most cases of rock-water system) that are firmly anchored to the solid phase and ions of the opposite sign (cations) in the liquid phase attracted to them near the boundary. The liquid phase is in surplus of cations so that when the liquid flows due to a pressure gradient, an electric potential gradient is formed. It is expressed as $\text{grad } V = -(\epsilon\zeta/\eta\sigma) \text{ grad } P$, where ϵ , σ and η are the dielectric constant, electric conductivity, and viscosity of the fluid, whereas ζ is a constant called zeta potential. Thus, the streaming potential is small for high conductive and viscous liquid.

Telluric current Electric current flowing in the surface layer of the earth's crust is called telluric current. Mainly it consists of the current induced by extra-terrestrial geomagnetic field variations (called magneto-telluric or MT current) and the current as a part of the global circuit between ionosphere and ground. MT current carries information on the electrical structure of the earth's interior: higher (lower) frequency for shallower (deeper) structure. Telluric current can also be of man-made origin leaking from such electric sources as factories and trains. Telluric current is mea-

sured by dipoles of electrodes inserted into the ground at separate points. It has been postulated that transient anomalous telluric currents are observed before earthquakes.

Frequency bands of electromagnetic waves

Electromagnetic waves are classified by frequency bands as follows:

ULF (< a few Hz), ELF (a few Hz \sim 3 kHz), VLF (3–30 kHz), LF (30–300 kHz), MF (300–3000 kHz), HF (3–30 MHz), VHF (30–300 MHz), UHF (300–3000 MHz), SHF (3–30 GHz). Not only ULF to VHF bands, but also infrared ($\sim 10^{13}$ Hz) and visible ($\sim 10^{14}$ Hz) bands are considered to be involved in earthquake-related electromagnetic waves.

Skin effect The intensity of electromagnetic wave decreases exponentially with distance in a conductive medium. In a simple case, the distance where the intensity becomes $1/e$, called the skin depth δ , is expressed as $\delta = \sqrt{2/\mu\sigma\omega}$, where μ and σ are magnetic permeability, and electric conductivity of the medium and ω is the angular frequency of the wave.

Ionosphere The upper atmosphere, where electrons are stripped off from oxygen and nitrogen atoms by solar radiation, is called the ionosphere. It consists of a D-layer (60–90 km), E-layer (90–130 km), F_1 -layer (130–210 km), and F_2 -layer (210–1000 km). Electron density is highest in the F_2 -layer. The electron density of the ionospheric lower layer can be measured by ground-based ionosonde, whereas total electron content (TEC) of the whole ionosphere is estimated by global position system (GPS). Electric currents in the ionosphere produce transient variations of geomagnetic field. The suggestion has been made that the ionosphere is affected before earthquakes.

Definition of the Subject

Throughout most of human history, electromagnetic phenomena associated with earthquakes have been repeatedly told. A typical one is earthquake light. Until rather recently, however, most records were in the realm of folklore [31,71]. Since earthquakes are understood as a catastrophic event to occur when slowly increasing tectonic stress in the earth's crust reaches a critical level, it may well be expected that the same stress may give rise to some electric, magnetic, or electromagnetic phenomena (EM phenomena hereafter) and some persistent research on them was initiated more or less simultaneously in varied parts of the world in the 1980s in two main streams. One was monitoring of possible emissions from focal regions in a wide range of frequency from DC to VHF, whereas the other

was to monitor the anomalous transmission of man-made EM waves of varied frequencies over focal regions. Theoretical and experimental studies on the mechanism of EM phenomena have also been made. This relatively new branch of science is now called Seismo-Electromagnetics.

These EM phenomena attract high attention for their possible usefulness in earthquake prediction, which is of immense societal importance and considered as one of the last frontiers in earth sciences. Because many of the EM phenomena are observed prior to earthquakes, they may serve as their precursors, which have been difficult to find by usual seismological and geodetic methods. The extremely interdisciplinary nature of the subject matter is the distinct feature of Seismo-Electromagnetics and the backgrounds of many research fore-runners are neither seismology nor geodesy, but other fields, e.g., general geophysics, solid state, statistical, and ionospheric physics, radio, space, and even biological sciences. This situation in turn tends to make their accomplishments difficult to be understood and accepted by the conventional earthquake community. Of course, EM phenomena do not cause earthquakes. Both EM phenomena and earthquakes are considered to be caused by regional or local tectonic stresses, but some EM phenomena seem to appear shortly before the occurrence of earthquakes. That EM phenomena do not cause earthquakes may be another reason why few seismologists are interested in them.

Introduction

Earthquake (EQ) related EM signals may be classified into two major groups, each covering wide frequency ranges. One is EM signals supposedly emitted from the focal zone and the other is anomalous transmission of EM waves over the epicentral region.

The emission type signals are reported for geo-electrical potential (telluric current) and geomagnetic field and for EM waves. The best known example of the former is the Seismic Electric Signals (SES) in the VAN-method which has been developed in Greece since the early 1980s and applied also in Japan since the 1990s [76,78,82]. SES are transient DC geo-electrical potential variations observed before EQs by dipoles of buried electrodes. Experimentally and theoretically, the VAN method is by far the best equipped method in this category and has survived debates [22,41]. Along with the current views that earthquakes are catastrophic events at a critical state of complex systems, a new time domain called Natural Time has been introduced to integrate SES with seismicity for short-term EQ prediction (e.g., [82]). Late in the 1980s, pre-seismic magnetic signals began to be reported. They were

ultra low frequency (ULF) anomalous changes observed before *M*7.1 Loma–Prieta EQ, 1988 [16] and *M*7.1 Spitak EQ, 1989 [39] followed by *M*8.0 Guam EQ, 1993 [29]. Among them the case of Loma–Prieta is considered as most convincing. For the higher frequency range, there have been reports of VLF signals received by aerial antennas [3,23,95]. There have been reports from Greece that even VHF signals have also been received [9]. For high frequency signals, considering the electric conductivity of the crust, their emergence to the surface from the source regions at depth presents problems to be solved.

Pre-seismic appearance of EM signals makes them useful for EQ prediction. It has often been questioned, however, why they appear only pre-seismically and not co-seismically. This point makes the scientific community dubious about EQ-related EM signals in general. Actually, co-seismic signals are routinely observed but not reported each time for the obvious reason that they are not useful for EQ prediction. However, all the co-seismic signals observed so far were found to occur at the time of the arrival of seismic waves. They are, therefore, co-seismic wave signals and not “true” co-seismic. The fact that no true co-seismic signals are observed may be an important clue in exploring the physical mechanism of signal generation.

The generation mechanism of EM signal emission may be different for different frequencies. They may involve the electro-kinetic effects and pressure-stimulated polarization effects [79] for DC to low frequency signals, and piezo-electric effects and exo-electron emission [10] for higher frequency ones. For the optical range like EQ light, a very different mechanism such as de-excitation of air molecules may be involved. Apart from these, a mechanism involving so-called positive holes (p-holes) in rock forming minerals under stress has also been proposed recently [18].

The second class of signals, i. e., the anomalous transmission of EM waves, began to be actively discussed in the late 1980s [24]. One of the best documented early cases may be at the 1995 Kobe earthquake in which received VLF radio waves for navigation purposes showed anomalies in both phase and amplitude a few days before the main shock [52]. Moreover, at the same Kobe EQ, it was first found that FM radio waves from stations beyond the line of sight can be received before main shocks [40].

The anomalous transmission of EM waves means that there are some anomalies in the path, i. e., ionosphere or atmosphere, over the epicentral region, which may be verified independently. Investigation to detect such changes has been vigorously conducted both by ionosonde [45] and by topside observation from satellites [63]. On the basis of these observations, the concept of Lithosphere–At-

mosphere–Ionosphere (LAI) coupling through which pre-seismic changes in the earth’s crust may be transferred to the upper atmosphere became one of the central issues of Seismo-EM studies.

This article does not deal with the aspects of EM studies of the earth devoted for elucidating the subterranean electrical structures. Although their EQ related structural time changes, if observed, would be of great interest, there has practically been no significant reported progress.

Telluric Current Anomalies and Natural Time

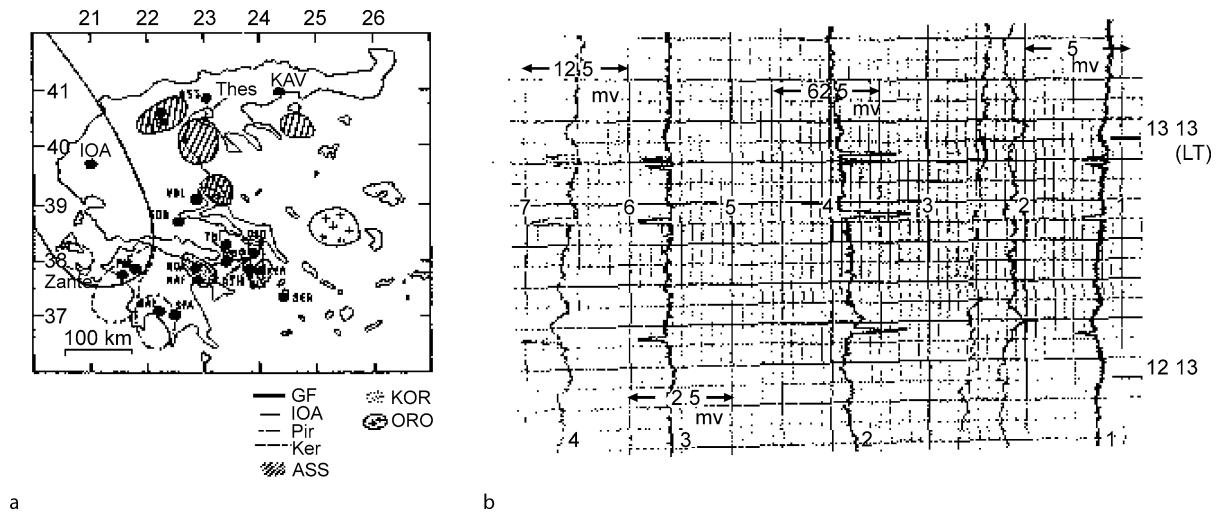
The VAN Method

The best example of modern research on DC electric signals is that of the VAN method [18,80,82], named after the initials of the founding Greek scientists, P. Varotsos, K. Alexopoulos, and K. Nomikos. The VAN group has been making actual short-term predictions of $M \geq 5$ Greek EQs during well over a couple of decades. The criteria for successful prediction imposed by themselves are: $< a$ few weeks in time, < 0.7 units in magnitude (M , hereafter), and < 100 km in epicentral distance. The length of time window depends on the type of signals.

The changes in the geo-electrical potential differences between buried electrodes, called Seismic Electric Signals (SES), are continuously monitored at many stations (Fig. 1a). At each station, several short (50–200 m) dipoles in both EW and NS directions and a few long dipoles (2–20 km) in appropriate directions are installed. Compared with all earlier works using only one or two dipoles, adoption of the multiple dipole system was a distinct progress in noise rejection.

Amplitude of SES is of the order of 1 mV/100 m. There are four types of signals, i. e., single SES, SES Activity, Gradual Variation of Electric Field (GVEF), and short duration pulse. Single SES, having duration 1/2 min ~ several hours, precedes single EQ, whereas SES Activity, which consists of a number of SES in a short-time, is followed by a series of EQs before the main shock (Fig. 1b). As will be explained later, SES Activity has been playing a major role in the recent VAN work related to Natural Time analysis. GVEF has amplitude an order of magnitude larger than usual SES, but is only rarely observed for large EQ. The last type, i. e., short duration pulses appear shortly (some minutes) before EQs. These pulses, with amplitude sometimes amounting orders stronger than SES, have received rather little attention mainly because their lead time of minutes has been considered too short for useful EQ prediction.

In the VAN type of observation, noise discrimination is critically important. To eliminate noise, they have developed a set of rules as follows:



Earthquakes, Electromagnetic Signals of, Figure 1

a Distribution of VAN stations and "Selectivity map" of several stations as of 1996 (after [74]). For Zante, see text. **b** An example of SES Activity recorded on three short dipoles (labeled 1, 3, and 4) and long dipole (2) at Ioannina (IOA in **a**) station on August 31, 1988. Note that intensity scales in mV are different for different dipoles. The Killini–Varthelomio EQs were predicted based on these data [82]

1. Changes with magneto-telluric origin can be eliminated because they appear at all the stations simultaneously.
2. SES must appear simultaneously on all of short and long dipoles, but only at the concerned station.
3. SES must satisfy the $\Delta V/L = \text{constant}$ relation for short parallel dipoles, where ΔV is the amplitude of SES and L the dipole length.
4. The polarity and amplitude of SES of short and long dipoles must be compatible with the assumption that the source is distant compared with the dipole lengths.

The VAN group made two major discoveries. One is the so-called "Selectivity" and the other is the so-called "VAN-relation". The Selectivity has two aspects. (1) There are only selected sites which are sensitive to SES (sensitive sites). They were found only through testing at many sites: Almost 90% of sites were insensitive. This fact gives another strong reason why earlier efforts to catch precursory electric signals failed. (2) A sensitive site is sensitive only to SES from some specific focal area(s), which are not always in close proximity. A map identifying those focal area(s), SES from which are sensed by a site, is called the "Selectivity map" of that site (Fig. 1a), which provides information on the epicentral location of the impending EQ when a SES is observed at the site. The Selectivity is considered to originate from the inhomogeneity of the subterranean electrical structures, i.e., SES goes only through conductive channels. The VAN group has presented many model

studies of channels [82]. So far, however, the real existence of such subterranean channels has not been verified by usual MT or other electric exploration techniques, possibly because the scales of the proposed channel structures are too small for the presently available resolving power.

The other discovery from the VAN research, i.e., the "VAN relation", is the following relationship among the focal distance, r , EQ magnitude, M and the observed intensity of SES, $\Delta V/L$.

$$\log(\Delta V/L \times r) = aM + b, \quad (1)$$

where a is a constant 0.34–0.37 and b is a site-dependent constant. Once the epicentral location is estimated from the Selectivity map mentioned above, M of the impending EQ can be assessed since both $\Delta V/L$ and r are known.

The VAN method has been a contentious subject (e.g., [22,41]). It is difficult, in principle, to prove the causal relationship between SES and EQ occurring at separate times and the only reasonable way to make the relationship credible would be to accumulate as many case studies as possible on one hand and to build plausible physical models on the other. Both endeavors have been published in papers too many to quote here and they are summarized in [82]. According to an independent evaluation (Uyeda et al. [75] and later additional check), out of 16 $m_b \geq 5.5$ EQs which occurred in the Greek region during Jan. 1, 1984–Jan. 1, 2004, 13 were successfully predicted. After the predictions of three large EQs in 1995,

there were no large EQs until another three $m_b \geq 5.5$ occurred in late 1997. It is remarkable that during the 2.5-year of quiescence no prediction was issued for the area and two out of the three 1997 events were predicted remarkably well. The score of the VAN method has been assessed by many authors. Most of the evaluations were in favor, but some were not. The low scores often resulted either when the assessors did not follow what the VAN group designated on the items such as allowable lead times for different type of SES, or the magnitude scales to use. Mulargia and Gasperini [53] claimed that “the apparent success of VAN predictions can be confidently ascribed to chance” and ignited heated debates (e.g., [22,41]). In the present authors’ view, however, VAN has well survived them. In Greece, VAN-type SES has been observed by other groups also.

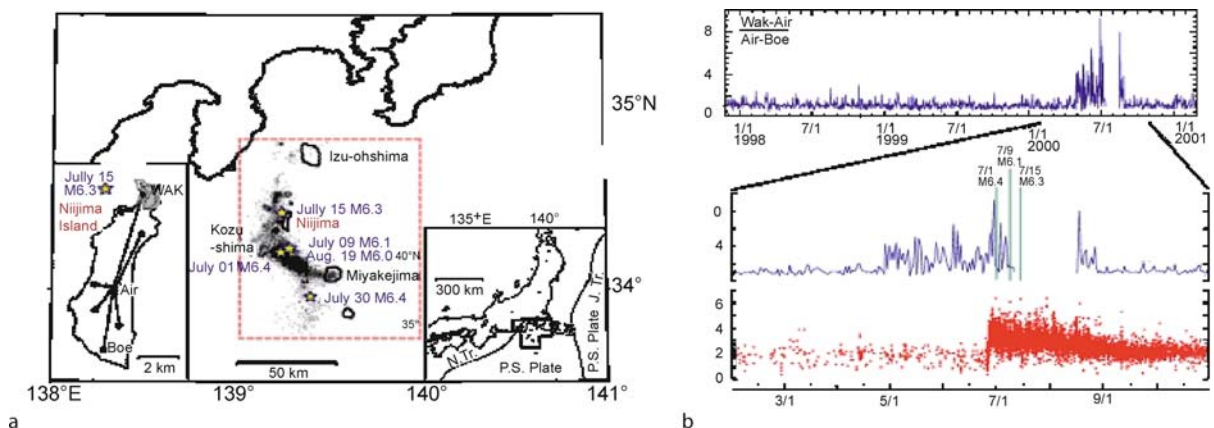
In Japan, VAN-type monitoring was initiated on a trial basis in the late 1980s and was expanded in 1996 [76]. Despite the serious problems caused by the high level of artificial noise, in particular from DC-driven electric trains, the existence of the VAN-type SES has been confirmed for $M > 5$ EQs occurring within ~ 20 km or so of a station. Moreover, phenomena attesting to “Selectivity” were discovered. In the year 2000, a two-month long seismic swarm, with $\sim 7,000$ $M \geq 3$ shocks and five $M \geq 6$ shocks, occurred in Izu Island region: See Fig. 2a. For this swarm activity, significant pre-seismic electric disturbances were observed [77]. From about 2 months before the swarm onset on June 26, innumerable clear, unusual geo-electrical potential changes started on Niiijima Island (Fig. 2). These anomalous changes appeared only in the northern part of the island, possibly reflecting the ex-

tremely heterogeneous underground structure of the volcanic region.

Co-seismic signals have been observed for many EQs. However, they always started with the arrival of seismic waves and not at the origin time of EQs. The changes are probably local effects of passing seismic waves. There may be many reasons why no true co-seismic signals are observed. One is that, as laboratory experiments show, signals generated at ruptures are in much higher frequency range, so that they cannot be registered by usual high-cut measurement (0.1–1 Hz sampling) and the second is that, even when a higher sampling rate is employed, the high frequency signals attenuate before reaching the receiver. In fact, the pre-seismic stress accumulating process giving rise to SES and the instantaneous stress releasing event are physically very different processes and there seems to be no compelling reason why they generate similar signals.

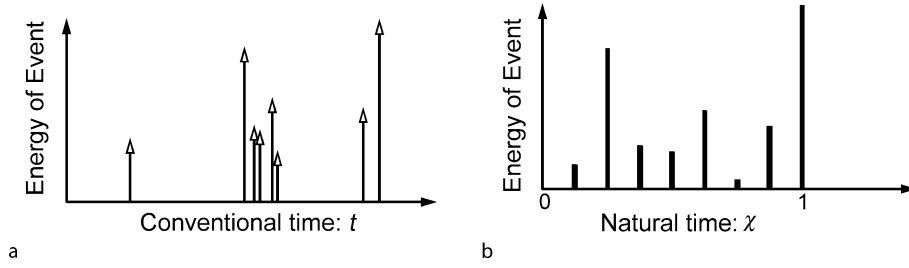
Natural Time

Seismicity as a critical phenomenon has been actively discussed by many authors (e.g., [4,38,65,68,73]). It has been shown that SES and EQs reveal dynamic evolution characteristics to the critical stage when their time series is analyzed in the framework of natural time χ , which was introduced by the Varotsos’ group (e.g. [82,83]). The symbol χ stands for the ancient Greek word $\chi\rho\nu\nu\sigma$, which means “time”. The possible usefulness of natural time analysis in predicting catastrophic events has been demonstrated not only for the subjects of our immediate concern, but also for other critical phenomena, including sudden cardiac death [84,85].



Earthquakes, Electromagnetic Signals of, Figure 2

a Seismic swarm activity in 2000 in Izu island region. *Inset in the left* shows the dipole configuration in Niiijima Island. Each end of the long dipole had short dipoles. Only Wak-Air dipole showed the pre-swarm signals shown in **b**. The *bottom panel* in **b** shows seismicity (modified from [77])



Earthquakes, Electromagnetic Signals of, Figure 3

Time series of events, **a** in conventional time t , and **b** in the natural time χ

In a time series comprised of N events, the natural time $\chi_k = k/N$ serves as the index for the occurrence of the k th event. In natural time analysis, the evolution of the pair of two quantities (χ_k, Q_k) is investigated where Q_k denotes the quantity proportional to the energy of the k th event. The time series of events as shown in Fig. 3a is expressed in natural time as in Fig. 3b. For the purpose of analysis, the following function $\Phi(\omega)$ was introduced.

$$\Phi(\omega) = \sum_{k=1}^N p_k e^{i\omega \frac{k}{N}}, \quad (2)$$

where $p_k = Q_k / \sum_{n=1}^N Q_n$ and $\omega = 2\pi\phi$, ϕ standing for the frequency in natural time (natural frequency). This $\Phi(\omega)$ should not be confused with the discrete Fourier Transform because ω is a continuous variable. If we regard p_k as the probability density function of χ , in analogy with probability theory, its Fourier transform $\Phi(\omega)$ may be regarded as the characteristic function of ω . The power spectrum of $\Phi(\omega)$, $\Pi(\omega) = |\Phi(\omega)|^2$, for the dynamical system approaching critical state with infinitely long-range temporal correlation was calculated to be as follows (see p.259–260 in [82]):

$$\Pi(\omega) = |\Phi(\omega)|^2 = \frac{18}{5\omega^2} - \frac{6 \cos \omega}{2\omega^2} - \frac{12 \sin \omega}{5\omega^3}. \quad (3)$$

Taylor expansion of Eq. (3) gives, for small values of ω ($\omega \rightarrow 0$),

$$\begin{aligned} \Pi(\omega) &= 1 - \kappa_1 \omega^2 + \kappa_2 \omega^4 + \kappa_3 \omega^6 + \kappa_4 \omega^8 + \dots \\ &= 1 - 0.07 \omega^2 + \dots \end{aligned} \quad (4)$$

Thus, a time series should show $\kappa_1 \approx 0.07$ when approaching the critical stage. The reason why the natural time domain is useful when information on intervals between events is lost while retaining only information on the order and relative importance of events is an intriguing question. As to this point, Abe et al. [1] have shown

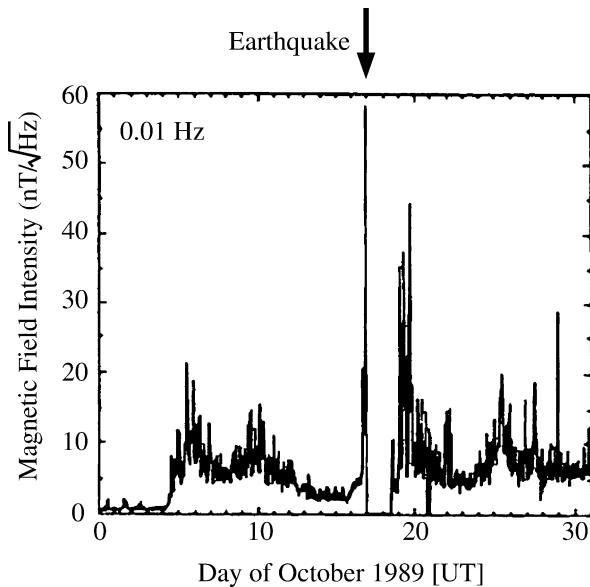
that this time domain in fact is optimal for enhancing the signals in time-frequency space.

In Greece, $\kappa_1 \approx 0.07$ was experimentally ascertained first for SES activities preceding four large EQs; 1995 $M_{6.6}$ Kozani–Grevena EQ, 1995 $M_{6.5}$ Eratini–Egio EQ, 1997 $M_{6.4}$ Strofades EQ, 2001 $M_{6.6}$ Aegean Sea EQ and later also for other major EQs, supporting this view [82,86]. Infinitely long-range temporal correlations of Greek SES were independently confirmed by Weron et al. [89].

In the case of seismicity, to investigate its time evolution, the power spectrum $\Pi(\omega)$ of the seismicity in natural time subsequent to associated SES activity was calculated as each consecutive EQ occurred. It was, then, shown that, for major Greek EQs, $\Pi(\omega)$ approached that of the critical state ($\kappa_1 = 0.07$) a few days before the main shocks [82]. This indicated that the seismicity approached the critical state at that time. This unexpected discovery may shed new light on the EQ generation mechanism itself. At the same time, this suggests the possibility of narrowing the time window of predicting EQs to a few days, when SES data are available. It may be added, albeit different from the Natural Time defined here, that an attempt was made to identify seismic quiescence with the viewpoint that the seismic process proceeds with its internal clock called “events time scale” [66].

Ultra Low Frequency (ULF) Anomalies

ULF generally means lower than several Hz. Research in this frequency range was started late in the 1980s. ULF signals are advantageous over those in higher frequencies because of their large skin depth. The best-known example is the case of the $M_{7.1}$ Loma-Prieta (California) EQ in 1989 [16]. Observation was made at a site which happened to be at 7 km from the epicenter. The amplitude of the horizontal component started anomalous enhancement at about 2 weeks prior to and a sharp increase a few hours before the EQ. Figure 4 shows the records at 0.01 Hz band. The disturbance lasted for about 3 months after the



Earthquakes, Electromagnetic Signals of, Figure 4

The amplitude of the geomagnetic horizontal component at 0.01 Hz band [16]

EQ. These anomalous changes were not of solar terrestrial origins because they were not observed at other distant stations. Moreover, these have never been observed at any other time during the whole period of observation of more than 15 years. It was, thus, concluded that the anomalies were related to the EQ. Reports of observing pre-seismic ULF geomagnetic anomalies have been made also for *M*6.9 Spitak (Armenia) EQ in 1988 [39] and *M*8.0 Guam (Marianas) EQ in 1993 [29]. Further efforts in Japan and elsewhere have been summarized by Hattori [28]. It seems, however, that a more rigorous approach is needed to make the ULF studies sufficiently credible to the scientific community.

Higher Frequency Electromagnetic Emission and Earthquake Light

Pre-seismic electromagnetic wave emission in the VLF–LF-range has been reported since the 1980s. Gokhberg et al. [23] reported pioneering observations as shown in Fig. 5. Emissions at 81 kHz increased one or two hours before *M*6.1 and *M*5.3 earthquakes took place and decreased after the second shock.

Asada and his group started investigation of EQ-related VLF emissions in the early 1990s [3]. They monitored the wave forms of two horizontal magnetic components of VLF waves, through which the apparent incoming direction of VLF pulses was determined. They found that, before *M*5 class land EQs within 100 km of

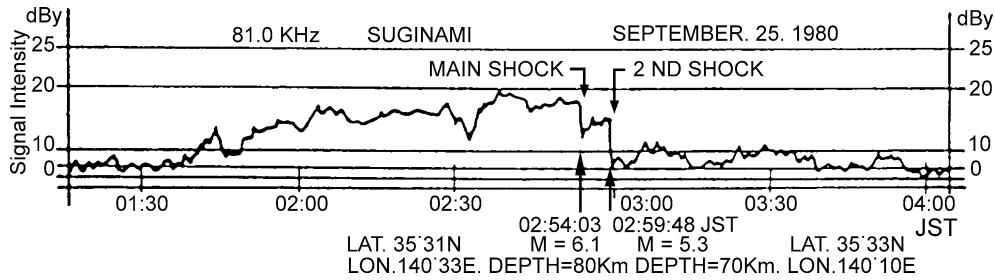
their stations, some pulses with a fixed incoming direction appeared and the EQs actually occurred in that direction, whereas the sources of overwhelmingly numerous and stronger noises were moving along with lightning sources (Fig. 6). Moreover, there is a well-documented report of undeniable noise in commercial MF radio bands, experienced by an automobile driver approaching Kobe, some minutes before the Kobe EQ of 1995 (see [55]).

Enomoto et al. [12] recorded anomalous pulses of geo-electrical current (HF-band) at Erimo station, Hokkaido, Japan, from February 2000 to March 2001 and from August to September 2003. The former anomalies occurred before and during the volcanic activity of Mt. Usu (200 km away), while the latter started one month before the 2003 September 26 *M*8.0 Tokachi-Oki EQ (80 km away). These were the only anomalies during their 10-year observation period.

For the Kobe EQ, while measuring sporadic Jovian decametric emissions with a radio interferometer at an observatory at about 80 km from the epicenter, unusual pulsed emissions at 22.2 MHz were detected tens of minutes both before and after the main shock [47]. Such unusual pulses have never been observed at other times and the possible source direction was estimated to be that of the main surface exposure of the EQ fault. There was no clear co-seismic radiation. Warwick et al. [88] reported a similar observation related to the 1960 Great *M*9.5 Chilean EQ.

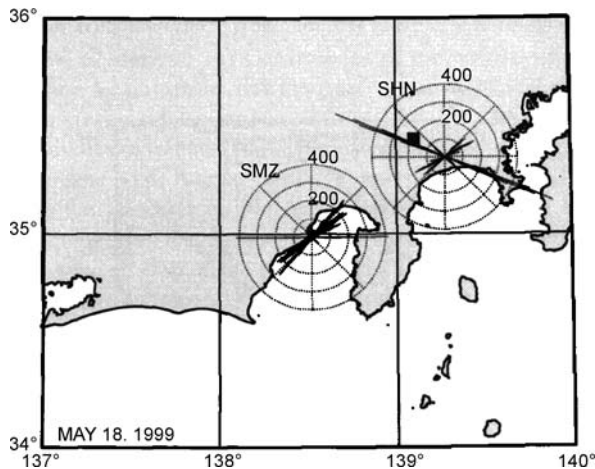
Also in the high frequency range, Eftaxias et al. [9] have reported results obtained in Greece. Since 1994, they have been running a station on Zante Island in the Ionian Sea (see Fig. 1a), where installation was performed of (i) six loop antennas in EW, NS and vertical magnetic field at both 3 kHz and 10 kHz; (ii) $\lambda/2$ electric dipoles for 41, 54, and 135 MHz and (iii) two Short Thin Wire Antennas for ULF (< 1 Hz) anomalies. Sampling rate is 1 Hz. They report that MHz–kHz EM anomalies have been detected during a few days to a few hours prior to near-surface land EQ with *M* > 6, i.e., the 1995 Kozani-Grevena EQ and 1999 Athens EQ. The MHz radiation appeared earlier than the kHz. They interpret the phenomena as due to small-scale cracking and assume that the more grown-up cracks generate the lower frequency anomalies. They also look to the EQ as a critical phenomenon and suggest that the shift from MHz to kHz activity corresponds to an anti-persistence to persistence shift. In their observations, there was no co-seismic anomaly.

As mentioned earlier, however, transmission of EM waves in the conducting earth beyond the skin depth distance is an important unresolved problem common to all the topics reported in this section.



Earthquakes, Electromagnetic Signals of, Figure 5

Change of 81 kHz electromagnetic wave observed in Tokyo [23]



Earthquakes, Electromagnetic Signals of, Figure 6

Rose diagram of incoming VLF signals observed on May 18, 1999 at two sites. M4.1 EQ occurred at black square point on May 22 [3]

EQ Light

Earthquake light (mostly co-seismic) has been reported all over the world from ancient Greek, Roman, and Chinese times. There is no doubt that the phenomena exist. Light emanates from the whole sky, or locally from the ground. The shape reported is like aurora, pole, flash, ball lightning, and so on, while the color widely ranged from blue and blue-white to red-yellow and orange. We, however, note that all reports may not be on natural phenomena but on some artificial effects such as sparks at power lines.

Galli [21] collected 148 eyewitness reports in late 19th Century Europe (see [71]). In Japan, Musha [54] collected about 2,000 eyewitness reports for 65 EQs, while Terada [69] discussed the theoretical aspects and suggested that the electro-kinetic effect may be a possible cause. From 1965 to 1967, there was a large EQ swarm at Matsushiro area in central Japan and numerous luminous phe-



Earthquakes, Electromagnetic Signals of, Figure 7

Photograph of EQ light at Matsushiro seismic swarm taken by Kuribayashi (after [91])

nomena were seen and photographs were taken as shown in Fig. 7 [91].

For the 1995, M7.3 Kobe EQ, Enomoto and Zheng [11] examined the trace of gas emission in the Awaji fault where the rupture started. They suggested that the gas plasma emission might have emitted the light. Kamogawa et al. [37] reported some independent witnesses that a luminous object moved a long distance a few seconds before the main shock in the direction of the rupture. Ikeya and Takaki [32] numerically showed that the screening charges neutralizing the polarized piezoelectrical rock may generate a strong co-seismic electric field, and the de-excitation of nitrogen molecules excited by collision of electrons accelerated by the electric field produce the blue EQ-light.

Lithosphere-Atmosphere-Ionosphere (LAI) Coupling

Pre-seismic atmospheric-ionospheric anomalies before EQs have been reported since the 1970s [2,20,24,27] and the concept of pre-seismic lithosphere-atmosphere-ionosphere coupling arose. Historical reviews and important works that are not introduced in this article may be tracked from references of Pulinets and Boyarchuk [61] and Kamogawa [33].

Liu et al. [43] found in Taiwan that the ionosonde measured critical plasma frequency, foF2, corresponding to the electron density of the ionospheric F2 layer, significantly decreased during afternoons within a few days before $M \geq 6$ EQs. For example, such ionospheric anomalies appeared 3 and 4 days before the 1999 $M7.7$ Chi-Chi EQ. Similar EQ-related electron density depression occurring above Taiwan Island was observed in the GPS total electron density (TEC) [44]. From such observations, Liu et al. [45] demonstrated that the appearance of the anomalies within 5 days was statistically significant at 5% level for the $M \geq 5.4$ EQs occurring within 150 km.

Sub-ionospheric anomalies before large EQs were reported by Gokhberg et al. [24] and Gufeld et al. [27]. They used VLF ship-navigation waves (10–20 kHz) and observed pre-seismic anomalies between the transmitter and the receiver during mid-night. Marenko [49] statistically supported the results of Gokhberg et al. [24], while Michael [50] obtained a less optimistic conclusion. Meanwhile, the studies have been further developed mainly in Russia, Japan, and Italy. For example, pre-seismic variations of terminator-times, i.e. the sunrise and sunset for VLF waves, were demonstrated [52]. Clilverd et al. [6], on the other hand, did not obtain similar positive results when they applied the terminator-time method to their 5-year data of reception at Faraday, Antarctica (receiver) of VLF waves transmitted from the northern United States. Maekawa et al. [48], measuring LF waves, statistically investigated the correlation between sub-ionospheric anomalies and $M \geq 6$ EQs in Japan and found that the amplitude and dispersion of received signals significantly decreased 2–6 days before the EQs. Thus, this issue is still controversial [34,64].

In the VHF range, Kushida and Kushida [40], while monitoring meteorites plunging into the high atmosphere by reflection of FM radio waves, detected anomalous reception, just a few days before the Kobe EQ, of the FM radio waves from distant (beyond the line-of-sight) stations. This was a new discovery and these authors consequently began extensive measurements on other EQs. With regard to the anomalous reception of VHF waves from transmitters beyond the line-of-sight, Fujiwara et al. [20] statistically showed significant enhancement of atmospheric anomalies lasting for a few minutes–several hours within 5 days before $M \geq 4.8$ EQs.

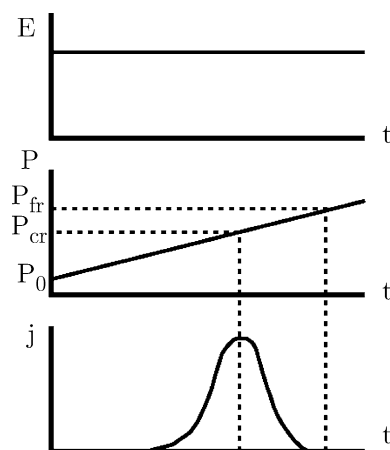
Mechanism of Pre-Seismic EM Phenomena

Generation Mechanism of EM Signals

Electro-Kinetic Effect The electro-kinetic effect can be a plausible source for SES (DC) and ULF emission. Mizu-

tani et al. [51] first proposed a model in which, during the dilatancy stage, pore pressure in the dilatant region decreases and water flows into this region from the surrounding area, generating electric and magnetic precursors of EQs. Since then, many models have been proposed (e.g., [15,93]). Fedorov et al. [13], however, suggested that the expected magnitude of seismo-EM signals in the ULF-VAN range from an electro-kinetic source may reach the detection level only for a favorable set of crustal parameters.

Models Related to Defects in Solids A SES-generation model by pressure-stimulated currents (PSC) was proposed by Varotsos and Alexopoulos [79]. Their model is based on the physics of the point defects in solids. The impurities and vacancies have excessive and opposite-sign effective charges and form local electric dipoles. The directions of the local electric dipoles usually distribute randomly. Under an electric field, dipoles will align in its direction. The alignment is an activation process in which the time constant is an Arrhenius-type function of stress as well as temperature. Therefore, an avalanche of alignment takes place when stress approaches a critical level (Fig. 8). It has been later suggested that, instead of electric field, inhomogeneous deformation mentioned in the next paragraph may work to align the dipoles in the direction of the stress gradient (see [14]). This model is unique among other models in that SES is generated spontaneously during gradual increase of stress without requiring any sudden change of stress such as micro-fracturing. For the SES to work as a precursor, it is assumed that the critical level



Earthquakes, Electromagnetic Signals of, Figure 8

Pressure-stimulated current j occurs at a critical pressure P_{cr} under the external electric field E . P_{fr} is the fracture pressure (after [79])

of stress for SES generation is lower than that of mechanical failure causing EQ. A thorough verification of the PSC model by laboratory pressure experiments is fatally lacking up to this stage.

In relation to SES generation, deformation-induced charged flow is an interesting possibility [56]. This flow was observed to take place as a result of inhomogeneous plastic deformation of ionic crystals, such as NaCl, in the direction of the stress gradient without applying electric field. It was interpreted that charge carriers are charged dislocations. Some experiments were conducted on rocks with similar results (see [82]). Independent of these, Freund and his colleagues have recently been proposing a unique mechanism for ULF electric signals ([18] and Ref. therein). They have discovered in the laboratory that when a block of igneous rock is put under stress locally, the rock turns into a battery without any external electric field (Fig. 9).

This striking phenomenon is interpreted as follows: A fraction of the oxygen anions in the rock-forming silicate minerals is not in their usual 2-valence state (O^{2-}) but in the 1-valence state (O^{1-}), which represent defect electrons, i. e., positive holes (p -holes). They are unstable and form more stable positive hole pairs (PHP), chemically equivalent to peroxy links, $O_3X/\cdot\cdot\cdot XO_3$, which are electrically inactive. These dormant PHPs, however, are awoken by deviatoric stress, and make the insulating host material a p -type semi-conductor. The p -holes flow out of the stressed volume because of mutual electrostatic repulsion. If scaling to earthquake size is allowed, the current thus produced may attain 10^3 – 10^5 A/km³.

Other Models For generation of high frequency signals, models related to micro-cracking have been proposed from laboratory experiments. They are (1) Discharge of screening charge of piezo-electric polarization [31,94], (2) electrification of fresh crack surfaces [90], (3) exo-elec-

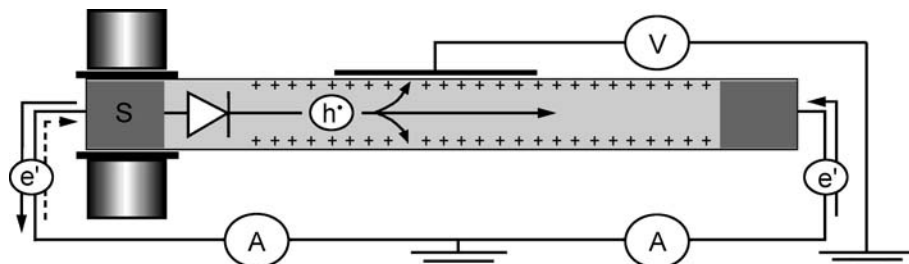
tron [10]. As to the occurrence of pre-main shock micro-cracking, there have been only a few reliable field reports. Furthermore, it might be pointed out that in these models much stronger co-seismic signals would be expected. Some ad hoc mechanism, therefore, would be needed to explain that no co-seismic signals have been observed so far in the field.

Transmission Mechanism of EM Signals

Even if EM signals are generated around a seismic focal region, signals except in the ULF range cannot be transmitted long distance in the conductive crust due to the decay caused by the skin effect, as long as the displacement current component is negligible. Even for DC signals, geometric decay would prohibit their long distance reception in a homogeneous or simple layered earth [5]. To overcome this difficulty, Varotsos et al. [81] proposed a conductive channel model, in which electric signals are transmitted through the conductive channel to a surface point close to the upper end of the channel. Freund [17] reported that, in the laboratory experiment, mobile positively charged holes (p -holes) appeared on the rock surface when a stress gradient was given to the rock sample. His results implied the possibility of appearance of a positively charged area at long distance on the surface before EQs. Kamogawa and Ohtsuki [35] proposed a model to explain how the higher frequency EM waves can be observed before EQs, i. e., longitudinal plasma waves excited by exo-electrons [10] may be transformed into EM waves by the surface roughness.

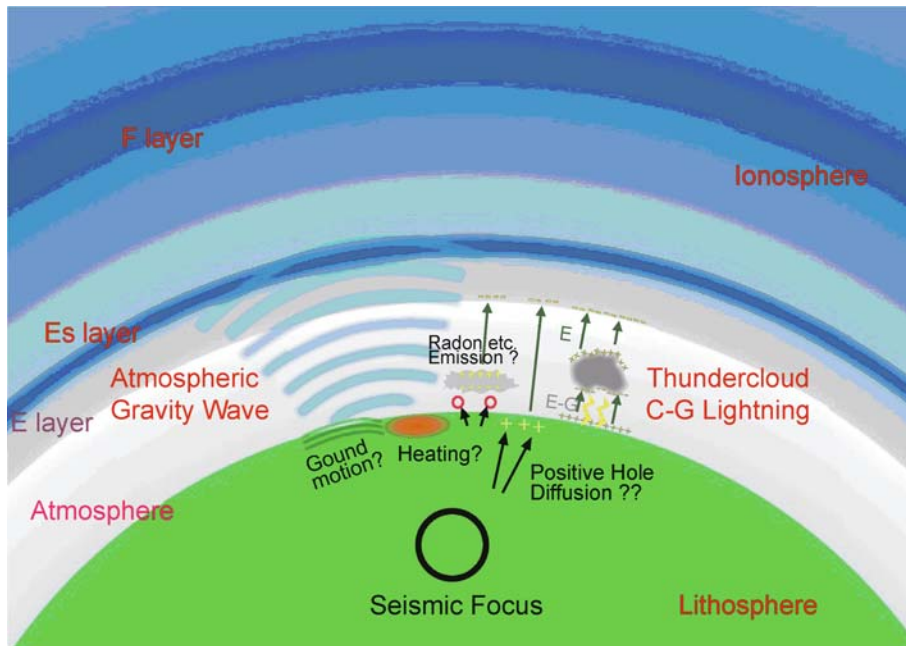
LAI Coupling Mechanism

If the pre-seismic atmospheric-ionospheric anomalies are real, some causative factors may be detected on the ground surface. Possible mechanisms of pre-seismic lithosphere-atmosphere-ionosphere coupling have been proposed by



Earthquakes, Electromagnetic Signals of, Figure 9

Conceptual diagram of the battery current carried by electrons which flow out of the stressed portion S (left) through the outer circuit and by p -holes which close the circuit by flowing through the unstressed portion and meeting the electrons at the far end (right), flowing out from the stressed portion S , (after [18])



Earthquakes, Electromagnetic Signals of, Figure 10
Concept of LAI coupling (modified from [33])

many researchers. They may be categorized in two groups as shown in Fig. 10.

First, some atmospheric electric field \vec{E} is generated on/near the ground surface during the pre-seismic period and it will cause the ionospheric anomalies [25,26,62]. Pulinets et al. [62] proposed that such an atmospheric electric field is caused by radon emission (see [30,87,92]). Alternatively, it is proposed that positively charged holes diffused from the seismic focal area to the ground surface generate the electric field [17]. However, such an electric field on the ground has not yet been observed even when pre-seismic ionospheric anomalies were detected [36].

Second, some researchers proposed that atmospheric gravity waves (AGW) propagate into the ionosphere, and disturb it before EQs [46,52,59]. The proposed source of AGW is the long-period ground oscillation or appearance of thermal anomalies on the ground. The former was inferred from some observations that co-seismic ground vibration actually excited AGW which propagated into the ionosphere (e.g. [8]). However, there is no report that long-period ground oscillations were detected at the pre-seismic stage even by high-sensitive superconducting gravimeter observations so far. The latter proposed source of AGW is the pre-seismic temperature rise, “thermal anomalies”, reaching 2–4°C or higher in a wide area around impending EQs, based on

satellite observation of enhanced infrared (IR) emission from the ground surface [7,70,72]. Many models have been put forward to explain the origin of the “thermal anomalies”, including latent heat release at condensing water vapor due to enhanced radon emission. Pulinets [60] develops a scenario where the “thermal anomalies” give rise to ionospheric anomalies. Freund et al. [19] cast another interpretation on the enhanced IR emission based on their p -hole model mentioned above. When p -holes appear on the surface of the unstressed area, they form a positive charge layer and recombine to form the more stable $O^+ - O^+$ bond, emitting IR as de-excitation energy.

Future Directions

It seems that, despite much circumstantial evidence, earthquake related electromagnetic signals, in particular those at the pre-seismic stage, have not yet been completely accepted as real physical quantities. Putting the common indifference and prejudice of the conventional scientific community against new science aside, it seems appropriate at this stage to recognize that there are legitimate reasons for the critical views. In fact, most of the problems of fundamental importance in seismo-electromagnetics are still unresolved.

To name a few, propagation of high-frequency EM signals in conductive earth has been proven unequivocally enough neither empirically nor theoretically. Techniques of direction finding of EM signals at various frequency ranges and atmospheric-ionospheric anomalies have not shown sufficiently credible results yet. Solving these problems will be important issues in near future investigation.

The mechanisms of signal generation are still far from established. The majority of proposed mechanisms attribute pre-seismic signals to effects such as piezo-electric, electro-kinetic, charged dislocations, p -holes, and exoelectrons, all induced by stress release at micro-fracturing in the last stage of EQ preparation. In such cases, critics demand that by far the largest signals should be observed at the instance of the main shocks when the largest stress drop takes place. However, as described above, all the “co-seismically” observed electric or ULF signals are associated with the arrivals of seismic waves (to be called co-seismic wave) and are not co-seismic in the true sense. For higher frequency signals, even co-seismic wave signals have not been confirmed. This fact is a popular basis for negating the EQ-related signals in general. However, this very fact, i. e., the non-observation of true co-seismic signals in any frequency range may present some important clues with regard to the mechanism of both the signal generation and earthquakes as follows.

Numerous lab-experiments show strong co-fracturing signals in the form of high frequency EM waves. They are very different from the low frequency signals observed during pre-seismic stages. Thus, one explanation for non-observation of co-seismic signals in the ULF range is (1) field observation uses a low-pass recording system to avoid high frequency noise, and (2) high frequency signals are attenuated in short distance in the earth. However, this explanation seems to suffer from a weak point as follows: Since it takes some seconds for the fault motion of a large EQ to terminate, the overall signals should contain a low frequency component powerful enough to be caught by the low-pass recording system. Moreover, even higher frequency wave monitoring systems have not captured any co-seismic signals. All these seem to speak for non-generation of co-seismic signals of any frequency in the field and researchers had to devise some ad hoc scenarios as to how to reconcile with lab results, often invoking overgrowth of micro-faults by the time of main shocks to produce signals. For pre-seismic signal generation, these suggested mechanisms assume pre-seismic micro-fractures, which in fact, are micro-EQs that may be observed by high sensitivity seismic networks. Although depending on the required size, there has been no such observation,

which constitutes another objection to pre-seismic EM signals.

The pressure-stimulated currents cited above regards the SES generation as a critical phenomenon. SES is supposed to be spontaneously generated when the gradually increasing tectonic stress level reaches a critical value. This seems to be the only proposed mechanism which needs no stress release by micro-fracturing or any special events, although it makes the causal relationship between SES and EQ less apparent. For the same reason, this mechanism does not need to generate any strong signals at EQ itself. Some of the other mechanisms, such as electro-kinetic, deformation-induced charged flow or p -holes flow mentioned earlier might be modified to fit the observation by incorporating the concept of critical state since they only need a development of stress gradient for signal generation.

Experimental verification of these mechanisms is urgently needed as it has been decisively inadequate. In any case, it should be kept in mind that the EQ preparation and EQ itself are different physical processes, the former being a gradual stress increasing process, whereas the latter is an instantaneous stress drop.

No true co-seismic signal in contrast to lab fracture experiments presents a question if the EQ is a fracture or not. It is now well-known that the EQ is a sudden sliding of pre-existing faults. However, according to Yoshida et al. [94], even stick-slip experiments on dry granite revealed strong signals at the time of slip which were understood as due to the piezo-electric effect. Non-generation of co-seismic EM signals still remains an important problem requiring further investigation.

Important unsolved questions are by no means confined to pre-seismic signal emissions. On the contrary, even more fundamental unsolved problems lie in LAI coupling. Here, the very origin on the ground to cause any of the suggested elementary agents, such as anomalous atmospheric electric field, atmospheric gravity wave, and thermal anomaly, is unknown. So far, observations on these features have been carried out by various researchers independently, so that the integration of fragmentary results for constructing a unified physical scenario of the whole process has been difficult. Since very active multi-national as well as multi-disciplinary cooperative research has been underway recently, involving GPS-TEC estimation and even topside measurement of the ionosphere by satellites of several nations, substantial progress in the upper end of LAI coupling is expected in the near future. Lately, active pursuit of EQ-related ionospheric anomalies has been made through topside and in-situ observation by satellites such as the French micro-satellite DEMETER [58]. How-

ever, the lower initiating side of the LAI coupling appears much more difficult to elucidate. It would require long sustained pre-seismic ground-based network observations on such phenomena as long-period ground motion and radon emission in as many earthquake prone areas as possible. But these tedious efforts should be enhanced on a global scale at all cost. Finally, it may be added that Kamogawa [33] pointed out that reported atmospheric-ionospheric anomalies might be caused by some EM phenomena which also trigger seismicity. For instance, suggestions have been made that geomagnetic storms [67] and cloud-to-ground lightning [42,57] may trigger EQs. It may be worthwhile to keep such a possibility of difference in cause and effect in mind in the future studies.

Bibliography

1. Abe S, Sarlis NV, Skordas ES, Tanaka HK, Varotsos PA (2005) Origin of the usefulness of the natural-time representation of complex time series. *Phys Rev Lett* 94:170601
2. Antselevich MG (1971) The influence of Tashkent earthquake on the Earth's magnetic field and the ionosphere. In: Tashkent earthquake 26 April 1966. FAN, Tashkent, pp 187–188 (in Russian)
3. Asada T, Baba H, Kawazoe K, Sugiura M (2001) An attempt to delineate very low frequency electromagnetic signals associated with earthquakes. *Earth Planets Space* 53:55–62
4. Bak P, Tang C (1989) Earthquakes as a self-organized critical phenomenon. *J Geophys Res* 94(15):15637–15639
5. Bernard P (1992) Plausibility of long distance electrotelluric precursors to earthquakes. *J Geophys Res* 97:17531–17536
6. Clilverd MA, Rodger CJ, Thomson NR (1999) Investigating seismo-ionospheric effects on a long subionospheric path. *J Geophys Res* 104(A12):28171–28179
7. Dey S, Singh RP (2003) Surface latent heat flux as an earthquake precursor. *Nat Haz Earth Syst Sci* 3:749–755
8. Ducic V, Artru J, Lognonne P (2003) Ionospheric remote sensing of the Denali Earthquake Rayleigh surface waves. *Geophys Res Lett* 30(18):1951. doi:10.1029/2003GL017812
9. Eftaxias K, Kapiris P, Polygiannakis J, Peratzakis A, Kopanas J, Antonopoulos G, Rigas D (2002) Experience of short term earthquake precursors with VLF-VHF electromagnetic emissions. *Nat Hazards Earth Syst Sci* 20:1–12
10. Enomoto Y, Hashimoto H (1990) Emission of charged particles from indentation fracture of rocks. *Nature* 346:641–643
11. Enomoto Y, Zheng Z (1998) Possible evidences of earthquake lightning accompanying the 1995 Kobe earthquake inferred from the Nojima fault gouge. *Geophys Res Lett* 25:2721–2724
12. Enomoto Y, Hashimoto H, Shirai N, Murakami Y, Mogi T, Takada M, Kasahara M (2006) Anomalous geoelectric signals possibly related to the 2000 Mt. Usu eruption and 2003 Tokachi-oki earthquake. *Phys Chem Earth* 31:319–324
13. Fedorov E, Pilipenko V, Uyeda S (2001) Electric and magnetic fields generated by electrokinetic processes in a conductive crust. *Phys Chem Earth C* 26:793–799
14. Fischbach DB, Nowick AS (1958) Some transient electrical effects of plastic deformation in NaCl crystals. *J Phys Chem Solids* 5:302–315
15. Fitterman DV (1978) Electrokinetic and magnetic anomalies associated with dilatant regions in a layered earth. *J Geophys Res* 83:5923–5928
16. Fraser-Smith AC, Bernardi A, McGill PR, Ladd ME, Helliwell RA, Villard OG Jr (1990) Low-frequency magnetic field measurements near the epicenter of the Ms 7.1 Loma Prieta earthquake. *Geophys Res Lett* 17:1465–1468
17. Freund F (2000) Time-resolved study of charge generation and propagation in igneous rocks. *J Geophys Res* 105:11001–11019
18. Freund F, Takeuchi A, Lau BES (2006) Electric currents streaming out of stressed igneous rocks – a step towards understanding pre-earthquake low frequency EM emissions. *Phys Chem Earth* 31:389–396
19. Freund FT, Takeuchi A, Lau BWS, Al-Manaseer A, Fu CC, Bryant NA, Ouzounov D (2007) Stimulated infrared emission from rocks: Assessing a stress indicator. *eEarth* 2:7–16
20. Fujiwara H, Kamogawa M, Ikeda M, Liu JY, Sakata H, Chen YI, Ofuruton H, Muramatsu S, Chuo YJ, Ohtsuki YH (2004) Atmospheric anomalies observed during earthquake occurrences. *Geophys Res Lett* 31:L17110. doi:10.1029/2004GL019865
21. Galli I (1910) Raccolta e classificazione de fenomeni luminosi osservati nei terremoti. *Bull Soc Sis Ital* 14:221–447 (in Italian)
22. Geller R (ed) (1996) Debate on “VAN”. *Geophys Res Lett* 23:1291–1452
23. Gokhberg MB, Morgounov VA, Yoshino T, Tomizawa I (1982) Experimental measurement of electromagnetic emissions possibly related to earthquakes in Japan. *J Geophys Res* 87(B9):7824–7828
24. Gokhberg MB, Gufeld IL, Rozhnoy AA, Marenko VF, Yampolsky VS, Ponomarev EA (1989) Study of seismic influence on the ionosphere by super long-wave probing of the Earth ionosphere wave-guide. *Phys Earth Planet Inter* 57:64–67
25. Gokhberg MB, Morgounov VA, Pokhotelov OA (1995) Earthquake prediction, seismo-electromagnetic phenomena. Gordon and Breach, Reading, p 289
26. Grimalsky VV, Hayakawa M, Ivchenko VN, Rapoport YG, Zadorozhnyi VI (2003) Penetration of an electrostatic field from the lithosphere into the ionosphere and its effect on the D-region before earthquakes. *J Atmos Solar-Terr Phys* 65:391–407
27. Gufeld IL, Rozhnoi AA, Tyumensev SN, Sherstuk SV, Yampolsky VS (1992) Radiowave disturbances in period to Rudber and Rachinsk earthquakes. *Phys Solid Earth* 28:267–270
28. Hattori K (2004) ULF geomagnetic changes associated with large earthquakes. *Terr Atmos Ocean Sci* 15:329–360
29. Hayakawa M, Kawate R, Molchanov OA, Yumoto K (1996) Results of ultra-low frequency magnetic field measurements during Guam earthquake of 8 August 1993. *Geophys Res Lett* 23:241–244
30. Igarashi G, Saeki S, Takahata N, Sumikawa K, Tasaka S, Sasaki Y, Takahashi M, Sano Y (1995) Ground-water radon anomaly before the Kobe earthquake in Japan. *Science* 269:60–61
31. Ikeya M (2004) Earthquakes and Animals. World Scientific, Singapore, 294 pp
32. Ikeya M, Takaki S (1996) Electromagnetic fault for earthquake lightning. *Jpn Jour Appl Phys Part 2* 35(3A):355–357
33. Kamagawa M (2006) Preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 87:417, 424
34. Kamogawa M (2007) Reply to comment on preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 88:248

35. Kamogawa M, Ohtsuki YH (1999) Plasmon model for origin of earthquake related electromagnetic wave noises. *Proc Japan Acad* 75(Ser. B):186–189
36. Kamogawa M, Liu JY, Fujiwara H, Chuo YJ, Tsai YB, Hattori K, Nagao T, Uyeda S, Ohtsuki YH (2004) Atmospheric field variations before the March 31 2002 M6.8 Earthquake in Taiwan. *Terr Atmos Ocean Sci* 15:445–461
37. Kamogawa M, Ofuruton H, Ohtsuki YH (2005) Earthquake light: 1995 Kobe earthquake in Japan. *Atmos Res* 76:438–444
38. Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Heidelberg, 335 pp
39. Kopytenko YA, Matishvili TG, Voronov PM, Kopytenko EA, Molchanov OA (1993) Detection of ultra-low-frequency emissions connected with the Spitak earthquake and its aftershock activity, based on geomagnetic pulsation data at Dusheti and Vardzia observatories. *Phys Earth Planet Inter* 77:85–95
40. Kushida Y, Kushida R (2002) Possibility of earthquake forecast by radio observations in the VHF band. *J Atmos Electr* 22:239–255
41. Lighthill J Sir (ed) (1996) *A critical review of VAN*. World Scientific, Singapore, 376 pp
42. Liu J, Chen Y, Ho Y (2004) A study of lightning activities and $M \geq 5.0$ Earthquakes in Taiwan during 1993–2002. *Eos Trans AGU* 85(47):T51B-0456 (Fall Meet. Suppl., Abstract)
43. Liu JY, Chen YI, Pulinet SA, Tsai YB, Chuo YJ (2000) Seismo-ionospheric signatures prior to $M \geq 6.0$ Taiwan earthquakes. *Geophys Res Lett* 27:3113–3116
44. Liu JY, Chen YI, Chuo YJ, Tsai HF (2001) Variations of ionospheric total electron content during the Chi-Chi earthquake. *Geophys Res Lett* 28:1383–1386
45. Liu JY, Chen YI, Chuo YJ (2006) A statistical investigation of pre-earthquake ionospheric anomaly. *J Geophys Res* 111:A05304. doi:10.1029/2005JA011333
46. Lizunov G, Hayakawa M (2004) Atmospheric gravity waves and their role in the lithosphere-troposphere-ionosphere interaction 1109. *IEEJ Trans Fundam Mater* 124-A:1109–1120
47. Maeda K, Tomisaka T (1996) Decametric radiation at the time of the Hyogo-ken Nanbu earthquake near Kobe in 1995. *Geophys Res Lett* 23:2433–2436
48. Maekawa S, Horie T, Yamauchi T, Sawaya T, Ishikawa M, Hayakawa M, Sasaki H (2006) A statistical study on the effect of earthquakes on the ionosphere, based on the subionospheric LF propagation data in Japan. *Ann Geophys* 24:2219–2225
49. Marenko VF (1989) Investigation of the relationship between seismic processes and disturbances to the lower ionosphere by means of VLF radio transmissions. Ph.D. Dissertation, USSR Academy of Sciences, Siberian Department, Irkutsk, 160 pp
50. Michael AJ (1997) Testing prediction methods: Earthquake clustering versus the Poisson model. *Geophys Res Lett* 24:1891–1894
51. Mizutani H, Ishido T, Yokokura T, Ohnishi S (1976) Electrokinetic phenomena associated with earthquakes. *Geophys Res Lett* 3:365–368
52. Molchanov OA, Hayakawa M (1998) Subionospheric VLF signal perturbations possibly related to earthquakes. *J Geophys Res* 100:1691–1712
53. Mulargia F, Gasperini P (1992) Evaluating the statistical validity beyond chance of VAN earthquake precursors. *Geophys J Int* 111:32–44
54. Musha K (1932) Investigations into the luminous phenomena accompanying earthquakes. *Bull Earthquake Res Inst Tokyo Univ* 10:666–673
55. Nagao T, Enomoto Y, Fujinawa Y, Hata M, Hayakawa M, Huang Q, Izutsu J, Kushida Y, Maeda K, Oike K, Uyeda S, Yoshino T (2002) Electromagnetic anomalies associated with 1995 Kobe earthquake. *J Geodynamics* 33:349–359
56. Norwick AS (1996) The golden age of crystal defects. *Ann Rev Mater Sci* 26:1–19
57. Ouzounov DP, Williams RG, Wohlman R (2000) A joint analysis of earthquake and lightning activity in the Southern California (1995–1999). *Eos Trans AGU* 81(19):S41B-08 (Spring Meet. Suppl. Abstract)
58. Parrot M (ed) (2007) First results of the DEMETER micro-satellite. *Planet Space Sci* 54(5):411–558
59. Pilipenko V, Shamimov S, Uyeda S, Tanaka H (2001) Possible mechanism of the over-horizon reception of FM radio waves during earthquake preparation period. *Proc Japan Acad* 77(Ser. B):125–130
60. Pulinet S (2007) Natural radioactivity, earthquakes, and the ionosphere. *Eos* 88:217–218
61. Pulinet S, Boyarchuk K (2005) *Ionospheric precursors of earthquakes*. Springer, p 315
62. Pulinet SA, Boyarchuk KA, Hegai VV, Kim VP, Lomonosov AM (2000) Quasielectrostatic model of atmosphere-thermosphere-ionosphere coupling. *Adv Space Res* 26:1209–1218
63. Pulinet SA, Legen'ka AD, Gaivoronskaya TV, Depuev VK (2003) Main phenomenological features of ionospheric precursors of strong earthquakes. *J Atmos Solar Terr Phys* 65:1337–1347
64. Rodger CJ, Clilverd MA (2007) Comment on preseismic lithosphere-atmosphere-ionosphere coupling. *Eos* 88:248
65. Rundle JB, Turcotte DL, Sammis C, Klein W, Shcherbakov R (2003) Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems. *Rev Geophys* 41(4). doi:10.1029/2003RG000135
66. Schreider SY (1990) Formal definition of premonitory seismic quiescence. *Phys Earth Planet Inter* 61:113–127
67. Sobolev GA, Zakrzhevskaya NA, Kharin EP (2001) On a relation between seismicity and magnetic storms. *Phys Earth* 11:66–72
68. Sornette D (2000) *Critical phenomena in natural sciences*. Springer, Berlin, 434 pp
69. Terada T (1931) On luminous phenomena accompanying earthquakes. *Bull Earthq Res Inst Tokyo Univ* 9:225–255
70. Tramutoli V, Di Bello G, Pergola N, Piscitelli S (2001) Robust satellite techniques for remote sensing of seismically active areas. *Annali di Geofisica* 44:295–312
71. Tributsch H (1982) *When the snakes awake*. MIT Press, Cambridge, 248 pp
72. Tronin AA (1996) Satellite thermal survey – a new tool for the study of seismoactive regions. *Int J Remote Sens* 41:1439–1455
73. Turcotte DL (1997) *Fractals and chaos in geology and geophysics*. Cambridge University Press, Cambridge, 398 pp
74. Uyeda S (1996) Introduction to the VAN method of earthquake prediction, a critical review of VAN. *World Scientific, London, Singapore*, pp 3–28
75. Uyeda S, Al-Damegh K, Dologlou E, Nagao T (1999) Some relationship between VAN seismic electric signals (SES) and earthquake parameters. *Tectonophysics* 304:41–55
76. Uyeda S, Nagao T, Orihara Y, Yamaguchi Y, Takahashi T (2000) Geoelectric potential changes: Possible precursors to earthquakes in Japan. *Proc Nat Acad Sci USA (PNAS)* 97:4561–4566

77. Uyeda S, Hayakawa M, Nagao T, Molchanov O, Hattori K, Orihara Y, Gotoh K, Akinaga Y, Tanaka H (2002) Electric and magnetic phenomena observed before the volcano-seismic activity 2000 in the Izu Island Region, Japan. *Proc Nat Acad Sci USA* (PNAS) 99(11):7352–7355
78. Varotsos P, Alexopoulos K (1984) Physical properties of the variations of the electric field of the earth preceding earthquakes. *Tectonophysics* 110:73–125
79. Varotsos P, Alexopoulos K (1986) Stimulated current emission in the earth and related geophysical aspects. In: Amelinckx S, Gevers R, Nihoul J (eds) *Thermodynamics of point defects and their relation with bulk properties*. North Holland, Amsterdam
80. Varotsos P, Kulhanek O (eds) (1993) *Measurements and theoretical models of the Earth's electric field variations related to earthquakes*. *Tectonophysics* 224:1–288
81. Varotsos P, Sarlis N, Lazaridou M, Kapiris P (1998) Transmission of stress induced electric signals in dielectric media. *J Appl Phys* 83:60–70
82. Varotsos PA (2005) *The physics of seismic electric signals*. TerraPub, Tokyo, 338 pp
83. Varotsos PA, Sarlis N, Skordas E (2002) Long-range correlations in the electric signals that precede rupture. *Phys Rev E* 66:011902
84. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2004) Entropy in the natural time domain. *Phys Rev E* 70:011106
85. Varotsos PA, Sarlis NV, Skordas ES, Lazaridou MS (2005) Natural entropy fluctuations discriminate similar-looking electric signals emitted from systems of different dynamics. *Phys Rev E* 71:011110
86. Varotsos PA, Sarlis NV, Skordas ES, Tanaka HK, Lazaridou MS (2006) Entropy of seismic electric signals: Analysis in natural time under time reversal. *Phys Rev E* 73:031114
87. Wakita H, Nakamura Y, Notsu K, Noguchi M, Asada T (1980) Radon anomaly: A possible precursor of the 1978 Izu-Oshima-Kinkai Earthquake. *Science* 207:882–883
88. Warwick JW, Stoker C, Meyer TR (1982) Radio emission associated with rock fracture: Possible application to the great Chilean earthquake of May 22 1960. *J Geophys Res* 87:2851–2859
89. Weron A, Burnecki K, Mercik S, Weron K (2005) Complete description of all self-similar models driven by Lévy stable noise. *Phys Rev E* 71:016113
90. Yamada I, Masuda K, Mizutani H (1989) Electromagnetic and acoustic emission associated with rock fracture. *Phys Earth Planet Int* 57:157–168
91. Yasui Y (1968) A study on the luminous phenomena accompanied with earthquakes (part 1). *Mem Kakioka Mag Obs* 13:25–61
92. Yasuoka Y, Igarashi G, Ishikawa T, Tokonami S, Shinogi M (2006) Evidence of precursor phenomena in the Kobe earthquake obtained from atmospheric radon concentration. *Appl Geochem* 21:1064–1072
93. Yoshida S (2001) Convection current generated prior to rupture in saturated rocks. *J Geophys Res* 106(B2):2103–2120
94. Yoshida S, Uyeshima M, Nakatani M (1997) Electric potential changes associated with slip failure of granite: Preseismic and coseismic signals. *J Geophys Res* 102:14883–14897
95. Yoshino T, Tomozawa I, Sugimoto T (1993) Results of statistical analysis of low-frequency seismogenic EM emissions as precursors to earthquakes and volcanic eruptions. *Phys Earth Planet Interi* 77:21–31

Earthquake Source: Asymmetry and Rotation Effects

ROMAN TEISSEYRE

Institute of Geophysics, Polish Academy of Sciences, Warsaw, Poland

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Asymmetric Continuum and Rotation Effects](#)

[Earthquake Source: Fracture Processes](#)

[Final Remarks](#)

[Acknowledgments](#)

[Bibliography](#)

Glossary

We use the tensor notation and the summation convention for the repeating indices:

$$T_{ss} = \sum_s T_{ss}, \quad A_k B_k = \sum_k A_k B_k.$$

In some places we underline the symmetric and antisymmetric properties of the tensors using the (.), [...] brackets for indices:

$$S_{(ik)} = S_{(ki)}, \quad S_{[ik]} = -S_{[ki]}.$$

The deviatoric part of a symmetric tensor, $T_{(ik)}^D$, having zero value of trace, is defined as

$$T_{(ik)}^D = T_{(ik)} - \frac{1}{3} \delta_{ik} T_{(ss)}, \quad T_{(ik)}^A = \frac{1}{3} \delta_{ik} T_{(ss)},$$

$$T_{(ss)}^D = \sum_s T_{(ss)}^D = 0,$$

where $T_{(ik)}$ is any symmetric tensor, while $T_{(ik)}^A$ is the axial tensor.

In some places we use for the partial differentiations the following notations equivalently:

$$\frac{\partial u_n}{\partial x_k} \leftrightarrow u_{n,k}, \quad \frac{\partial U}{\partial x_\alpha} \leftrightarrow U_{,\alpha},$$

where the indices with Roman characters run from 1 to 3, while those with Greek characters run from 1 to 4.

We use the fully antisymmetric tensor

$$\varepsilon_{lps} = \begin{Bmatrix} 1 \\ 0 \\ -1 \end{Bmatrix} \text{ for } \begin{Bmatrix} \text{even permutation of} \\ \text{repeating} \\ \text{odd permutation of} \end{Bmatrix} \text{ indices};$$

this fits to the tensor notation and helps to express some operations, e. g. curl:

$$\text{curl } \psi \Leftrightarrow \varepsilon_{lps} \frac{\partial}{\partial x_p} \psi_s .$$

Definition of the Subject

The problem of rotation waves becomes actual again due to the recent observations based on very precise instruments able to measure very small rotation time rates, and due to development and new approaches to the continuum theories.

Our aim is to present a consistent theory describing a continuum subjected to complex internal processes. First, we consider all possible kinds of point-related motions and deformations (strictly speaking, a single couple is not a point-source, as a displacement derivative only tends to a point, but a double couple enters into a family of point-deformations as it can be given by the string-string deformation, that is as a shear point-nucleus).

We define the complex rotation field which includes the spin and twist; the latter describes the angular oscillations of the shear axes and related amplitude variations. Twist point deformation can be represented by the string-string and string-membrane motions. A twist vector is defined as a vector perpendicular to a string-string plane; it becomes an important counterpart to spin and a key to presented theory, in which we shall also include the axial point deformation (e. g., the thermal one).

We believe that all point motions, displacement and rotation, and point deformations, axial and twist, shall be governed by some fundamental laws, and we intend to find the invariant forms of such relations in a frame of a modified continuum theory. Such a continuum theory may give us a new insight into the complexity of processes which can be included in a continual material description; we will demonstrate that interaction of these motions and deformations can lead us to a rich variety of internal processes.

We may mention also that the theory of continuum containing all these deformations with rotational motions (spin and twist), with the inner central motion and with defects may be projected on the intrinsic properties of non-Euclidean space. We confine ourselves only to a remark that application of differential geometry is extremely enlightening for the fundamental understanding of the nonlinear processes.

The independent rotation field can be related to the additional constitutive law joining the antisymmetric stresses and rotations, as proposed by Shimbo [1,2] in his considerations on the friction and fracturing processes; the intro-

duced motion equations are equivalent to the stress moment – angular velocity relation.

For a reference to our consideration, we recall also the Kröner theory and its modifications as introduced by Teisseyre and Boratyński [3,4].

The possible dual description of any motions by means of displacement and rotation fields are being discussed and their formal equivalence concerning the propagation processes is demonstrated, but not valid for the source phenomena. Thus, we point out the importance of physically independent rotation motions in the inner granulation and fracture processes, and different types of rotational deformations are analyzed.

Our considerations bring several results for the following subjects.

For seismology:

- New description of the source processes including the role of rotational processes, and explanation of co-action of the slip and rotation motions
- Theory of the seismic rotation waves
- Thermodynamical conditions for a seismic energy release

For continuum and fracture mechanics:

- Theory of asymmetric continuum with the balance equations for the symmetric and antisymmetric stresses
- The relations between the asymmetric stresses and dislocation field
- A new approach to fracture processes with the hypothesis of the twist-shear release following the extreme angular deformation related to the internal particles, or grains constituting a continuum
- A synchronization role of the specific wave fields in the fracture processes
- Formation of a mylonite zone adjacent to fracturing and its constitutive description

For fluid mechanics:

- Theory of the asymmetric fluid continuum with the non-vanishing rates of the asymmetric stresses
- Explanation of the extreme wave phenomena (solitons)

Asymmetric Continuum and Rotation Effects

Introduction

Earthquake rotation effects were observed and discussed at the time when the 19th century seismological science was formed. Some eminent scientists, e. g., Charles Lyell (1797–1875), Charles Darwin (1809–1882), Robert Mallet

(1810–1881), and Alexander von Humboldt (1769–1859), raised the problem of the vortical movements, or vortex motion, induced by earthquakes. After the Lisbon earthquakes (1755) and those of Calabria (1783), many scientists focused their attention on the effects induced by such “vortical” waves. Robert Mallet was the first who precisely explained the observed rotation effects of some surface objects, pointing out the roles of the center of adherence of these objects and their inertia moment in relation to forces twisting the objects (see: Kozak [5], and Ferrari [6]). Many scholars tried to design the instruments to record the “vortical motions”, but the first instrument prepared especially to record such motions was that constructed by Filippo Cecchi, the director of the Ximeniano Observatory of Florence, in 1875. Cecchi’s electrical seismograph used sliding smoked paper. However, at that time it was too early to construct an instrument sensitive enough to obtain any traces of such wave motions.

The problem of seismic rotation waves was apparently closed after the Gutenberg [7] statement (1926) that such waves cannot propagate as they will be immediately attenuated, even when generated at the source. Of course the rotation effects remained, with the related explanation by Mallet, as objects of studies especially in the domain of the macroseismic observations.

From the contemporary point of view, two groups of achievements shall be mentioned; first, related to continuum theories, and second, related to development of the modern very precise instruments, able to record extremely small rotation time rates.

The continuum elastic theory bears from its very origin the serious limitation that the angular motions and related moments are not included. In such a situation, there was no place for a constitutive law describing the reaction between the stress moments and rotation processes. The lack of such a law automatically denies the existence of the rotation waves. We will return to these problems further on.

Experimental Evidence

The modern instrumentation techniques and the obtained results presenting the rotation wave seismograms need more attention. First, we can mention that maybe the first rotation seismogram (see Teisseyre [8]) was achieved in an indirect way: the azimuth array of horizontal seismographs, installed in one of the coal mines in Upper Silesia, Poland, to record the nearby tremors permitted one to deduce the rotational component of motions. However, the first, fully documented, rotation seismograms were obtained at the two geodetic fundamental stations in

Germany (Wetzell) and in Australia (Cochard et al. [9]; Schreiber et al. [10]), equipped with ring-laser interferometers based on the Sagnac principle. These stations were established, primarily, to record very small deviations and disturbances of the Earth’s rotational motion. However, the instruments having sensitivity up to 10^{-9} rad/s were able to record the rotation motions related to many distant earthquakes.

Latter, sensors of another type – the fiber-optic interferometers – were used by Takeo [11] especially for seismic observations; one version of his sensors included the tri-axial system. Jaroszewicz et al. [12] followed this system of rotation seismographs for the study of Silesian seismic events.

In a more traditional way, Moriya (see Moriya and Teisseyre [13]) has constructed the first rotation seismograph system consisting of a pair of anti-parallel seismographs; such a system, with the common suspension of the anti-parallel pendulums was repeated in latter constructions (Wiszniewski [14]).

Data collected by the recording systems mentioned above brought at least two important results:

- Records of different events in the very near field indicate that some events, e. g., shallow volcanic and those of explosion type, differ from the common characteristics by the extremely small rotation components (Teisseyre et al. [15]).
- Correlations between the rotation seismograms obtained from the ring-laser system and the rotation motions, curl u , derived from the array of seismometers (located at the same site) show almost perfect fit (Cochard et al. [9]).

Following the Cosserat theory and the micropolar and micromorphic theories (see Subsect. “[Asymmetric Continuum Theory](#)”), the independent rotation field, e. g., rotation related to grains or points of a continuum, were considered by Shimbo [1,2] in relation to the friction and fracture processes; the related constitutive law, we will call it the Shimbo law, joins the antisymmetric stresses with rotations and leads us towards the asymmetric continuum theory. The Shimbo law was latter generalized for the spin and twist rotation motions (Teisseyre et al. [16]; twist motion is introduced as the equivalent to oscillations of the shear axes).

We shall remind the reader that Gutenberg [7], in the frame of the classical elasticity theory, has proved that the independent rotation waves must be immediately attenuated. Now we know that this statement is due only to the fact that in the classical theory the rotations are not related to the stress or stress moment response; we cannot intro-

duce the constitutive law joining the symmetric stresses with the antisymmetric rotations.

However, in the asymmetric continuum theory such a constitutive law is required and appears as a natural element of the theory.

Displacements and Rotations

We cannot deny that independent rotation waves do not exist in a continuum built by the point-particles; however, this question is reduced to the magnitude of the independent rotations (that is, independent of $\text{rot } \mathbf{u}$) generated in the seismic sources; the rotation wave motion is assured by the Shimbo constitutive law; this constitutive law joining the antisymmetric stresses and rotation relates directly to the friction as kind of material resistance:

$$S_{[ik]} = 2\mu^* \omega_{[ik]}, \quad S_{(ik)}^D = 2\mu E_{(ik)}^D, \quad (1)$$

where we have added the constitutive law for pure shear (further on we define the pure shear oscillations as twist), the tensors $S_{(ik)}^D$ and $E_{(ik)}^D$ are the deviatoric stress and strain tensors, and where μ is the rigidity modulus and the constant μ^* is defined as rotation rigidity, the constant entering in the antisymmetric stress – angular velocity relation (this constant is not equal to the rotation modulus in the stress moment – angular velocity relation); further on, we assume that both constants are equal, $\mu^* = \mu$, as it follows from the seismic wave observations.

We shall notice that both motions, displacements and rotations, are interrelated, which follows also from the fact that pure rotation, $\omega_{[s]}$, can be presented by means of the potentials represented by some displacement field, $\mathbf{u}^{\text{micro}}$; conversely, the displacements U (excluding those related to the scalar potential, e. g., those of thermal origin or that related to explosion process, that is, we put $\frac{\partial}{\partial x_i} U_i = 0$) can be described by the vector potentials represented by some rotation field $\Omega_{[i]}$:

$$\omega_{[i]} = \text{curl } U, \quad U = l^2 \text{curl } \Omega_{[i]} \quad (2)$$

and we arrive at the possible dual approach to the continuum mechanics.

When applying such equivalent approaches twice, e. g., from rotation field to displacement and again to rotations we obtain

$$\begin{cases} \omega_{[s]} = \varepsilon_{smn} \frac{\partial U_n}{\partial x_m}, & u_i = l^2 \varepsilon_{iks} \frac{\partial \omega_s}{\partial x_k} \\ \rightarrow u_i = -l^2 \Delta U_i & \text{at } \frac{\partial}{\partial x_i} U_i = 0. \end{cases} \quad (3a)$$

Or otherwise:

$$\begin{cases} \omega_{[i]} = \varepsilon_{iks} \frac{\partial U_s}{\partial x_k}, & U_s = l^2 \varepsilon_{smn} \frac{\partial \Omega_n}{\partial x_m} \\ \rightarrow \omega_{[i]} = -l^2 \Delta \Omega_i & \text{at } \frac{\partial}{\partial x_i} \Omega_i = 0, \end{cases} \quad (3b)$$

where l represents the basic intrinsic length measure.

The intrinsic length (Cosserat characteristic length) plays an important role in material properties; there is extensive literature related to this subject, however we limit ourselves to the remark that the displacement and rotation motions could be completely independent only for the case with $l = 0$, but such a case is excluded by quantum mechanics with the minimal Planck's length of the order of 10^{-34} m.

Excluding the axial motions, our consideration leads us to an apparent equivalence of the two descriptions, those by means of displacements and rotations. However, there remains the problem of the scales of these motions generated at the different fracture modes in the seismic source and also the scales of these motions observed at the Earth's surface.

The observed rotation fields and effects are usually much smaller than those related to the displacement field. There are, however, some exceptions leading to situations in which rotations play an important role, e. g.: the tilt motions, the rocking and tilting components related to building structures hit by strong ground motions and, as it will be discussed further on, the rotations generated by fracture which occurred under the prevailing compression load.

These equivalent descriptions of the motions in a continuum can be combined in the asymmetric theory, but we shall note that it could be constructed as well, the continuum theory neglecting completely the displacement fields and using only the rotation motions – such a case can be called a degenerated continuum.

Therefore, keeping in mind the above remarks that the rotations contribute to displacement derivatives and that displacements may contribute to rotations, we can state that these motions are interrelated; this statement is empirically supported by the above mentioned almost perfect fit between the derived rotations, $\text{curl } \mathbf{u}$, and the rotations obtained from the ring-laser system data. Therefore, we think that the problem related to existence of rotation waves appears as an irrelevant question. However, we shall stress that the displacements and rotation motions differ in general, especially when considering their physical origins and effects. Instead of that problem, we propose to consider the classification of rotation motions from the point of view of their origins, scales and effects produced.

We propose the following classification:

- The micro-rotations or rotations, ω , as related to the wave motions based on the internal friction processes (rotation rigidity), as well as to slip motions with friction/fracture processes
- The meso-rotations related to material granulation and formation of the mylonite zones under the shear load fracturing processes
- The total rotations, ω^T , (the nomenclature introduced by Kröner [17]) related to the displacement field, \mathbf{u} ;
- The macro-rotations as related to fragmentation of material at the fracturing under compression load
- The mega-rotation effects related to the ground tilts and tilting of high objects on the ground

The important counterpart of rotational processes in the mechanics of fracturing and the related energy release shall be underlined (Teisseyre et al. [18]). Both under confining pressure and under external shears, the role of micro-fracturing in the bond breaking process is similar; however, we observe here the essential differences for rotations in larger scales.

The confining load condition leads to formation of the induced opposite arrays of dislocations, resulting in fragmentation processes and related macro-rotations. On the other hand, shear load leads to more concentrated fracturing along some planes. In the thermodynamical fracture band theory, see Subsect. “[Earthquake Thermodynamics](#)”, we consider the additional super-lattice formed by dislocations and the properly defined vacant dislocations; with this advanced approach we express an effective role of dislocation band structure in the shear fracture thermodynamics. Similarly, the fragmentation and macro-rotation processes become more effective for the fracturing under confining pressure. Thus, we try to find a continuum description for a number of processes leading us from an elastic solid to that undergoing successive deterioration by crushing, granulation and fragmentation.

These considerations give us ground for the classification of the basic motions.

Basic Deformations and Simple Motions in Asymmetric Continuum

Basic and simple motions could be defined as those which may be reduced to the 3D point motion in the Cartesian or Riemann spaces, or those deformations conceived as the respective curvatures.

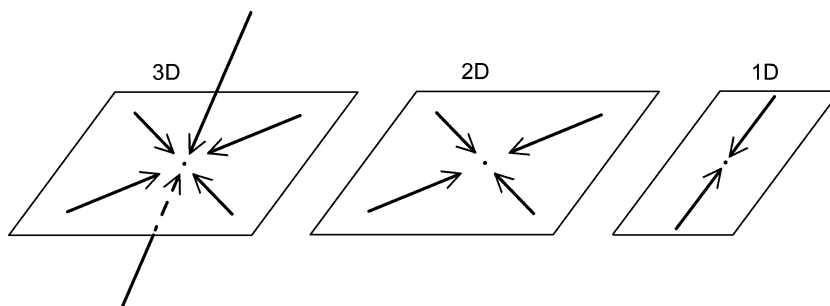
Considering basic motions, we can distinguish the simple motions. First the translation described by vector \mathbf{u} and the independent rotation, called spin; in a non-homogeneous continuum, the grains having different material parameters, can rotate due to an interaction of a displacement field with the related moment of inertia of the grains. Then, we shall pass to the tensorial motions/deformations:

Any antisymmetric tensor can be related to the vectorial field, e. g., to spin motion; thus, we come again to the equivalent vector field. However, this simple spin motion, $\omega_{[.]}$, shall be treated basically as independent of the displacement rotation, however, both contribute to the total spin field, $\omega_{[.]} + \text{curl } \mathbf{u}$, observed, e. g. in seismology. We have already mentioned that any symmetric tensor can be split into the axial and deviatoric tensors. The axial deformation tensor relates to the point deformations representing compression/dilatation nuclei, e. g., related to thermal anomaly.

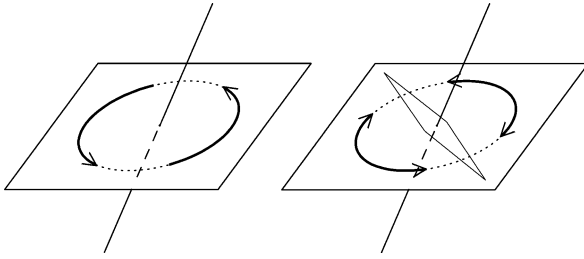
These total axial oscillation motions include the equal translation motions along all three axes and relate to that part of the displacement field which can be derived from a scalar potential (see Fig. 1).

For the point-like continuum with the axial deformations, we would obtain either the Riemannian curvature or, for more complicated cases, the Riemannian torsion tensor.

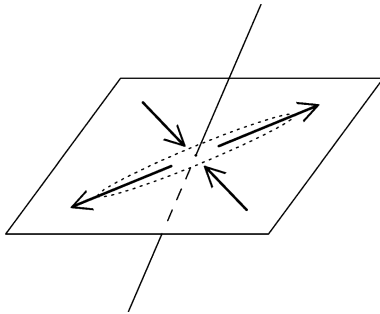
There remains the deviatoric field to consider. This field relates to pure shear deformations; it is possible to



Earthquake Source: Asymmetry and Rotation Effects, Figure 1
Axial basic deformations (3D, 2D and 1D)



Earthquake Source: Asymmetry and Rotation Effects, Figure 2
Rotational motions: spin and twist



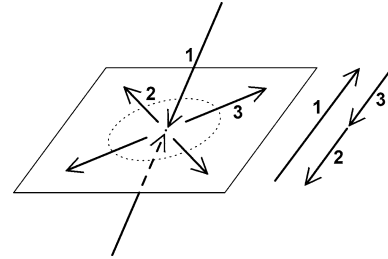
Earthquake Source: Asymmetry and Rotation Effects, Figure 3
String-string nucleus

show that the deviatoric field may be used to define the new antisymmetric tensor related to the simple deformations representing another kind of rotation motion – the twist deformation. These deviatoric deformations, for continua formed by particles/grains relate to pure shear oscillations (see Fig. 2; the left side shows a spin motion and the right side a twist one).

However, considering the point motions, it is better to relate these shear deformations to the equivalent twist tensor, $\omega_{(ks)}$; such motions contribute to the observed, e.g. in seismology, total shear/twist field. This field is directly related to the S-waves ($\mathbf{u}^S = l\omega$, where l is the effective radius of grains/particles forming the continuum and ω is the related rotation).

In the limit related to the point-like deformations, we would arrive at the string-string type motions (see Fig. 3) leading to the another representation of twist as vector, $\omega_{(k)}$, perpendicular to the string-string plane and having an appropriate magnitude of a string-string deformation (invariant representation of the string-string vector is discussed in Subsect. “Spin and Twist Motions”, while determination of its amplitude Subsect. “Recording Spin and Twist Angle Variation”).

A combination of the axial oscillation motions and those related to twist deformations leads to different forms



Earthquake Source: Asymmetry and Rotation Effects, Figure 4
String-membrane

of the point deformations: Fig. 4 relates to the string-membrane oscillations.

We shall repeat that all motions, except axial deformation, may be described by the displacement field, but this is not true for the point related spin and string-string deformation originated as the independent source motions.

Of course, there remains a number of the first, and higher order moments of these basic motions and deformations; such source models tend in a limit to a point source, but cannot be treated as exact point nuclei.

Asymmetric Continuum Theory

We shall add to the above considerations some arguments forming the fundamentals of the asymmetric continuum theory which is based on the asymmetric stresses:

- When studying the elastic field of an edge dislocation, we find some asymmetry in relation to its components in the plane perpendicular to its line (wedge direction); in confrontation with the symmetry of shears, this fact results in the asymmetry of stresses for a continuous distribution of dislocations (for screw dislocations such a contradiction does not exist). Therefore, a direct differential relation between any density of dislocations and the related stresses cannot be adequately found in a symmetric continuum.
- Fracture usually reveals its asymmetric pattern with a main slip plane; we shall believe that the premonitory processes, as described by deformations in a continuum with defects, develop also in an asymmetric way.
- In the classical continuum, the balance of angular momentum holds only if the stresses are symmetric; here the angular motions can be introduced only artificially with the help of a length element and a reference rotation point. This classical theory has also many other limitations and therefore many trials have been undertaken to generalize it. The asymmetric theory of elasticity with asymmetric stresses and couple-stresses

was founded by Nowacki [19]. However, a first generalization to include the moments in a continuum is due to Voigt in 1887 and a complete theory, including the asymmetry of stress and strain, is that known as the Cosserat theory of elasticity with the displacement vector and rotation vector [20]. Micropolar and micromorphic elastic theories were developed by Eringen and his co-workers and Mindlin (see: Eringen and Suhubi [21] and Mindlin [23]). Teisseyre [22] proposed a simpler version of the asymmetric theory which includes asymmetric stresses, strains and rotations, but in which the equations for the antisymmetric stresses differ from those of the couple moments in the Nowacki theory; their roles are interchanged, but both systems remain almost equivalent.

- Usually, when searching the fault slip solutions, we rely on classical elasticity with the friction constitutive laws introduced additionally in accordance with the experimental data. The obtained results well explain the observational data. Instead, we consider the consistent elastic continuum with asymmetric stresses and defects; such an approach enables one to study the defect interactions and elastodynamic solutions describing a slip propagation along a fault, including friction effects and related seismic radiation.
- The asymmetry of fields follows also from the notion of antisymmetric stresses considered by Shimbo [1,2] in relation to the friction processes and rotation of grains. Fracture processes develop usually along the main fault plane; hence there appears the initial asymmetry of the fracture pattern [24]; because of friction, the rotation of grains adjacent to the main slip plane causes an appearance of the antisymmetric part of the stresses and twist deformations. Following Shimbo [1,2], we introduced the constitutive law joining the antisymmetric stresses with the rotation nuclei (self-rotation field); without such a constitutive law any theory reduces both the rotation motions (except the rotation of displacements) and the related rotation waves to zero.
- In the asymmetric continuum, defined as that including both the symmetric stresses and the antisymmetric stresses, there appear also the rotational motions/deformations which split into pure spin and twist motions, the latter relate to the shear deformations of the grains; when considering the point-like nuclei, the twist deformation passes into 3D space torsion (Riemannian space).
- Experimental evidence for an appearance of spin and twist motions in a seismic field is based on the records of seismic rotation fields. For spin motion we shall be aware that the recorded rotation contains two el-

ements: a rotation of displacements and an independent spin motion. Both these elements co-act in motion propagation and represent its dual description, but differ in their origin, depending on the source processes and material properties. At fracturing under a confining load, we deal rather with a high rotation release process and therefore the spin motion for the very near seismic events usually distinctly overpasses the displacement rotation. For some events of an explosive nature, or for some near-surface volcanic events, both the pure spin motion and rotation of displacement almost disappear; some observed effects might be related to a nearby $P \Leftrightarrow S$ conversion. For the strong motions which include a tilting component, the rotation of displacements exceeds a spin motion. In engineering seismology we observe that the rotation of displacements may exceed the pure spin motion; such an effect is due to magnification of a horizontal rotation of displacements and to the appearance of a rocking/tilting component of displacement rotation caused by the geometry of construction, especially for high buildings.

Self-Field Nuclei: Deviations from Classical Elasticity

Any continuum could be described using the Kröner approach [17] based on a concept of internal fields excited by a density of defects and internal nuclei; stresses and strains are related by the unique constitutive law for the ideal elasticity. This approach is equivalent to another approach in which we change the constitutive law in a way appropriate to describe the plastic, viscous and relaxation effects. In the Kröner continuum with a density of the internal point-like nuclei, the elastic strains, rotations and stresses can be expressed as differences between total and self-fields.

Following the Kröner approach, we can keep the ideal elastic relation for the stresses and strains, supplemented by the constitutive law joining the antisymmetric stresses with rotations, and we introduce the self/inner stresses, strains and rotations as related to the internal nuclei or defects: $\mathbf{S}^S, \mathbf{E}^S, \omega^S$. We distinguish between the total stresses, strains and rotations related to the displacement field: $\mathbf{S}^T, \mathbf{E}^T, \omega^T$, and the asymmetric elastic stresses, strains and rotations $\mathbf{S}, \mathbf{E}, \omega$:

$$\begin{aligned} \mathbf{E} &= \mathbf{E}^T - \mathbf{E}^S, & \omega &= \omega^T - \omega^S, & \beta &= \beta^T - \beta^S, \\ \mathbf{S} &= \mathbf{S}^T - \mathbf{S}^S, \end{aligned} \quad (4a)$$

$$\mathbf{E}_{ki}^T = \mathbf{u}_{(i,k)}, \quad \omega_{i,k}^T = \mathbf{u}_{[ik]}. \quad (4b)$$

The elastic and self-deformations, strains and rotations, and stresses can be, in general, asymmetric ones (see: Teis-

seyre and Boratyński [4]); under the conditions that the antisymmetric parts of the stresses and strains, as well as the symmetric parts for elastic and self-rotations, be mutually compensated for:

$$\mathbf{E}_{[ik]} + \mathbf{E}_{[ik]}^S = 0, \quad \mathbf{S}_{[ik]} + \mathbf{S}_{[ik]}^S = 0, \quad \omega_{(ik)} + \omega_{(ik)}^S = 0. \quad (5)$$

However, referring to our earlier papers (see: Teisseyre and Boratyński [4]) we shall then assume that the respective self-parts of the asymmetric strain and rotation are equal to each other:

$$E_{[ik]}^S = \omega_{[ik]}^S, \quad \omega_{(ik)}^S = E_{(ik)}^S, \quad (6)$$

where symmetric rotation is related to the shear axes oscillations (comp.: twist definition and Figs. 2 and 3). The elastic fields \mathbf{S} , \mathbf{E} , ω represent the physical fields, while the total fields \mathbf{S}^T , \mathbf{E}^T , ω^T relate, according to the compatibility condition, to the displacement motions u_i , and the self-fields relate to the internal nuclei, defect densities and continuum structure.

Any deviations from the symmetry properties of fields, and any deviations from the ideal elasticity, can be described by suitable forms of the self-field, represented by the internal nuclei for both the defects and interaction fields.

The defects, dislocation and disclination densities can be defined, following Kossecka and De Witt [25], by considering the total disclosure and twist along a closed circuit (the Burgers vector and the Frank vector) and the appropriate form of the twist-bend tensor:

$$B_l = -\oint [E_{(kl)} - \varepsilon_{lqr} \chi_{kq}^S x_r] dl_k, \quad (7)$$

$$\Omega_q = -\oint \chi_{kq}^S dl_k = \theta_{pq} ds_p$$

and the definitions of the dislocation and disclination densities, α and θ , become based on the self-fields $E_{(kl)}^S$ and χ_{kq}^S :

$$\alpha_{pl} = -\varepsilon_{pmk} \left(\frac{\partial E_{(kl)}^S}{\partial x_m} + \varepsilon_{klq} \chi_{mq}^S \right), \quad (8)$$

$$\theta_{pq} = -\varepsilon_{pmk} \frac{\partial \chi_{kq}^S}{\partial x_m}.$$

After Teisseyre [26], the total twist-bend tensor can be defined as follows:

$$\chi_{mq}^T = \varepsilon_{ksq} \frac{\partial \omega_{mk}^T}{\partial x_s}, \quad \chi_{mq}^T = \chi_{mq} + \chi_{mq}^S \quad (9)$$

where, for the continuum with the asymmetric part of the stresses, we are not restricted to the compatibility condition for the twist-bend tensor (Kleman [27]).

The compatibility conditions for the asymmetric stresses and strains lead us to the physical equations for the dislocation and disclination densities in relation to the elastic fields of strain $E_{(kl)}$ and twist-bend χ_{kq} :

$$\alpha_{pl} = \varepsilon_{pmk} \left(\frac{\partial E_{(kl)}}{\partial x_m} + \varepsilon_{klq} \chi_{mq} \right), \quad \theta_{pq} = \varepsilon_{pmk} \frac{\partial \chi_{kq}}{\partial x_m}. \quad (10)$$

Further on we will not rely on the Kröner approach, instead we will confine ourselves to a simpler approach given by the standard asymmetric continuum theory.

Asymmetric Continuum: Standard Theory

In opposition to the Kröner approach presented above, we may construct the asymmetric standard theory entirely related to the displacement field. Such a theory shall be based both on the symmetric and asymmetric stresses and on the related constitutive laws and motion equations. The asymmetric deformations contain the symmetric strain and antisymmetric rotation. Thus, our theory is based on two groups of relations; for the symmetric and antisymmetric fields:

$$S_{kl} = S_{(kl)} + S_{[kl]}, \quad E_{kl} = E_{(kl)}, \quad \omega_{kl} = \omega_{[kl]},$$

$$D_{ks} = E_{ks} + \omega_{ks}, \quad (11a)$$

where D_{ks} means the asymmetric deformation tensor.

However, when introducing the new material parameters (material structure indices): e^0 , χ^0 , we may join these deformation fields in an independent way, with some reference displacement motion:

$$E_{kl} = e^0 \frac{1}{2} \left(\frac{\partial u_l}{\partial x_k} + \frac{\partial u_k}{\partial x_l} \right), \quad \omega_{kl} = \chi^0 \frac{1}{2} \left(\frac{\partial u_l}{\partial x_k} - \frac{\partial u_k}{\partial x_l} \right). \quad (11b)$$

For an internal energy stored in such a medium we obtain:

$$E = S_{(ks)} E_{ks} + S_{[ks]} \omega_{ks}.$$

The indices e^0 , χ^0 are not new constitutive constants, but they define families of solutions describing the complexity of deformation processes in continua; their ratio determines the phase shift between strain and rotation tensors. In this sense, the strain and rotation can be shifted in phase as follows from the particular deformations considered.

For the particular cases of these index values, e^0 , χ^0 , we have:

- The classic elasticity, obtained for $\chi^0 = 0$;
- For $e^0 = 0$ we obtain a granular/crushed medium filled with rigid spheres interacting by friction; when applying a torque load on its surface boundary, e.g., a cylindrical one, we would obtain only some angular deformation, and torque energy stored given as $E = S_{[ks]}\omega_{ks}$;
- The cases with $e^0 = \chi^0$ relate to the elastic continua with friction and different kinds of internal defects – different kinds of dislocation densities and the granulated materials;
- A continuum densely filled with the edge dislocations is described by the case $e^0 = -1$, $\chi^0 = -1$; while that of a partial content of that density $\alpha^E = \{0, 1\}$ by $e^0 = 1 - 2\alpha^E$, $\chi^0 = \alpha^E$;
- A continuum filled densely with the screw dislocations would be given by $e^0 = 2$, $\chi^0 = 2$; while that of a partial content of that density $\alpha^S = \{0, 1\}$ by $e^0 = 1 + \alpha^S$, $\chi^0 = 2\alpha^S$.

Further on, we will consider a more general continuum with the constitutive laws, including also the time rates processes; for such cases we might discuss in a similar way the different particular cases of the material structure indices including dynamic objects.

For the symmetric part of stresses we can assume the classical constitutive relation:

$$S_{(kl)} = \lambda \delta_{kl} E_{ss} + 2\mu E_{kl} . \quad (12)$$

But there is no problem to include in it the appropriate linear deviations related to visco-plastic effects.

To construct the asymmetric theory, we assume, after Shimbo [1,2], the appropriate constitutive law for the antisymmetric part of stresses. It joins the friction/fracture rotations with the antisymmetric stresses:

$$S_{[kl]} = 2\mu \omega_{kl} , \quad (13)$$

where rigidity constant μ plays the role of rotation rigidity entering in the antisymmetric stress-angular velocity relation (while the rotation modulus enters in the stress moment-angular velocity relation and may be considered as product of the rigidity μ and the Cosserat characteristic length l).

The motion equation for antisymmetric stresses $S_{[ni]}$ shall replace the balance law for the stress moments. To this end, we take the divergence of the rotation force moment acting on a body element due to the antisymmetric stresses (rotational moment of forces per infinitesimal arm length corresponding to stress moments), and, on the other hand, the balancing term – the acceleration related

to angular momentum [3]:

$$\varepsilon_{lki} \frac{\partial^2}{\partial x_k \partial x_n} S_{[ni]} = \rho \frac{1}{2} \varepsilon_{lki} \frac{\partial^2}{\partial t^2} \left(\frac{\partial u_i}{\partial x_k} - \frac{\partial u_k}{\partial x_i} \right) + \varepsilon_{lki} \rho K_{[ki]} , \quad (14)$$

where we have put $e^0 = 1$ and we have introduced the body couple $K_{[ki]}$ equivalent to body moment $K_{[l]} = \varepsilon_{lki} \rho K_{[ki]}$.

With the compatibility condition introduced in a similar way as for the symmetric strains:

$$I_{[ij]} = \varepsilon_{imk} \varepsilon_{jns} \frac{\partial^2}{\partial x_m \partial x_n} \omega_{ks} = 0$$

we obtain from Eqs. (11b), (13) and (14):

$$\begin{aligned} \frac{\partial^2 S_{[ki]}}{\partial x_s \partial x_s} &= 2\rho \frac{\partial^2 \omega_{ki}}{\partial t^2} + 2\rho K_{[ki]} , \quad \text{or} \\ \mu \frac{\partial^2 \omega_{ki}}{\partial x_s \partial x_s} - \rho \frac{\partial^2 \omega_{ki}}{\partial t^2} &= \rho K_{[ki]} , \end{aligned} \quad (15a)$$

where we have introduced also the body couple $K_{[ki]}$ or body moment $K_{[l]} = \varepsilon_{lki} \rho K_{[ki]}$.

Otherwise, we can write:

$$\mu \frac{\partial^2}{\partial x_k \partial x_k} \omega_{[l]} - \rho \frac{\partial^2}{\partial t^2} \omega_{[l]} = \rho K_{[l]} , \quad (15b)$$

where the left-hand side of this form presents the basic expression for the resulting stress moment divergence.

These relations are equivalent to the following ones:

$$\begin{aligned} \frac{1}{l^2} \frac{\partial}{\partial x_k} M_{lk} &= \varepsilon_{lki} \frac{\partial^2}{\partial x_k \partial x_n} S_{[ni]} = \varepsilon_{lki} \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_n} S_{[ki]} , \\ \frac{1}{l^2} M_{lk} &= \varepsilon_{lki} \frac{\partial}{\partial x_n} S_{[ni]} \end{aligned}$$

or defining the angular moment Ξ_i , we obtain:

$$\frac{\partial}{\partial x_k} M_{ik} = 2\mu \Xi_i , \quad \Xi_i = l^2 \varepsilon_{iks} \frac{\partial}{\partial x_n} \frac{\partial}{\partial x_s} \omega_{[kn]} .$$

From the motion equation for the symmetric part of stresses

$$\frac{\partial}{\partial x_k} S_{(kl)} = \rho \frac{\partial^2}{\partial t^2} u_l + F_l$$

and using the scalar and vector potentials

$$\begin{aligned} u_l &= l^2 \frac{\partial}{\partial x_l} \varphi + l^2 \varepsilon_{lps} \frac{\partial}{\partial x_p} \psi_s , \\ F_l &= l^2 \frac{\partial}{\partial x_l} \Phi + l^2 \varepsilon_{lps} \frac{\partial}{\partial x_p} \Psi_s \end{aligned}$$

we obtain:

$$\begin{aligned} (\lambda + 2\mu) \frac{\partial^2}{\partial x_k \partial x_k} \varphi &= \rho \ddot{\varphi} + \Phi, \\ \mu \frac{\partial^2}{\partial x_k \partial x_k} \psi_s &= \rho \ddot{\psi}_s + \Psi_s, \end{aligned} \quad (16)$$

where according to Eq. (11b) we have introduced the index e^0 and we assume $\frac{\partial}{\partial x_s} \psi_s = 0$, $\frac{\partial}{\partial x_s} \Psi_s = 0$ and where we have introduced the intrinsic length unit l .

Here the potential ψ_s may be interpreted as rotation vector motions in another scale than that defined by relation $\text{curl } v = \omega$:

$$\varepsilon_{mql} \frac{\partial}{\partial x_q} u_l = -l^2 \frac{\partial^2}{\partial x_k \partial x_k} \psi_m,$$

where ω means the micro-rotations, and ψ – the meso-rotations related to the granulated material (mylonite) and shear processes.

The strain tensor and its trace can be presented with the help of the introduced potentials as follows:

$$\begin{aligned} E_{lq} &= l^2 \frac{\partial^2}{\partial x_l \partial x_q} \varphi + \frac{1}{2} l^2 \varepsilon_{lps} \frac{\partial^2}{\partial x_p \partial x_q} \psi_s \\ &+ \frac{1}{2} l^2 \varepsilon_{qps} \frac{\partial^2}{\partial x_p \partial x_l} \psi_s. \end{aligned}$$

We can divide this expression into the axial and deviatoric parts:

$$\begin{aligned} E_{kk} &= l^2 \frac{\partial^2 \varphi}{\partial x_s \partial x_s}, \\ E_{lq}^D &= l^2 \left(\frac{\partial^2 \varphi}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \frac{\partial^2 \varphi}{\partial x_s \partial x_s} \right. \\ &\quad \left. + \frac{1}{2} \frac{\partial}{\partial x_p} \left(\varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \psi_s \right). \end{aligned} \quad (17)$$

Returning to our wave Eqs. (16) we arrive at the wave equations for the axial and deviatoric strain parts:

$$(\lambda + 2\mu) \Delta E_{kk} - \rho \frac{\partial^2 E_{kk}}{\partial t^2} = l^2 \Delta \Phi, \quad (18a)$$

$$\begin{aligned} (\lambda + \mu) \left(\frac{\partial^2 E_{ss}}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \frac{\partial^2 E_{ss}}{\partial x_k \partial x_k} \right) \\ + \mu \frac{\partial^2 E_{lq}^D}{\partial x_k \partial x_k} - \rho \frac{\partial^2 E_{lq}^D}{\partial t^2} \\ = l^2 \left(\frac{\partial^2 \Phi}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \Delta \Phi \right. \\ \left. + \frac{\varepsilon_{lps}}{2} \frac{\partial^2 \Psi}{\partial x_p \partial x_q} + \frac{\varepsilon_{qps}}{2} \frac{\partial^2 \Psi}{\partial x_p \partial x_l} \right). \end{aligned} \quad (18b)$$

In terms of the potentials we obtain

$$\begin{aligned} \left(\mu \Delta - \rho \frac{\partial^2}{\partial t^2} \right) \left(\left(\frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \Delta \right) \varphi \right. \\ \left. + \frac{\partial}{2 \partial x_p} \left(\varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \psi_s \right) \\ = \left(\left(\frac{\partial^2}{\partial x_l \partial x_q} - \frac{\delta_{lq}}{3} \Delta \right) \Phi \right. \\ \left. + \frac{\partial}{2 \partial x_p} \left(\varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \Psi_s \right) \end{aligned}$$

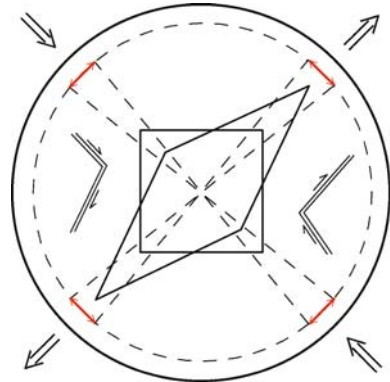
Spin and Twist Motions

The spin motion is governed by Eq. (15a), or equivalently by its vector form Eq. (15b). We may consider the system related to the main shear axes or that related only to the off-diagonal components; in the latter case the motion equation for the deviatoric strains Eq. (18b) can be presented in the form of the rotation vector motion – the twist, $\omega_{(s)}$:

$$\{\omega_{(s)}\} = \{E_{23}^D, E_{31}^D, E_{12}^D\}. \quad (19a)$$

The defined twist motion, $\omega_{(s)}$, means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor (corresponding to oscillation of the main shear axes), E_{lq}^D , accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by the internal fracturing processes (see Fig. 5).

Once having defined the twist vector field we can maintain its form due to the invariant properties of the



Earthquake Source: Asymmetry and Rotation Effects, Figure 5

Twist motion: rotational oscillations of the off-diagonal shear axes and internal fractures as the sources of perturbations; in the center we present an external shear deformation, while arrows along the circle give possible oscillations of the shear axes as influenced by some intrinsic processes, e. g. the fractures marked inside

Dirac tensors applied to the symmetric off-diagonal tensor $\omega_{(ik)}$ in its 4D form:

$$\begin{aligned}\omega_{(\lambda\kappa)} &= \omega_{(1)}\gamma^1 + \omega_{(2)}\gamma^2 + \omega_{(3)}\gamma^4\gamma^2\gamma^3 \\ &= \begin{bmatrix} 0 & -\omega_{(3)} & -\omega_{(2)} & -\omega_{(1)} \\ -\omega_{(3)} & 0 & \omega_{(1)} & -\omega_{(2)} \\ -\omega_{(2)} & \omega_{(1)} & 0 & -\omega_{(3)} \\ -\omega_{(1)} & -\omega_{(2)} & -\omega_{(3)} & 0 \end{bmatrix},\end{aligned}\quad (19b)$$

where

$$\begin{aligned}\gamma^1 &= \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \\ \gamma^2 &= \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix}, \\ \gamma^3 &= i \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}, \\ \gamma^4 &= i \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix},\end{aligned}\quad (19c)$$

and

$$\gamma^4\gamma^2\gamma^3 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

In a similar way, we may define the external off-diagonal part of the right-side expression of Eq. (18b):

$$\begin{aligned}\mathbf{Y}_{(lq)} &= \\ l^2 &\left(\frac{\partial^2}{\partial x_l \partial x_q} \Phi + \frac{\partial}{2\partial x_p} \left(\varepsilon_{lps} \frac{\partial}{\partial x_q} + \varepsilon_{qps} \frac{\partial}{\partial x_l} \right) \Psi_s \right).\end{aligned}$$

For its 4D form we can write:

$$\begin{aligned}\mathbf{Y}_{(\lambda\kappa)} &= \mathbf{Y}_{(12)}\gamma^1 + \mathbf{Y}_{(13)}\gamma^2 + \mathbf{Y}_{(23)}\gamma^4\gamma^2\gamma^3 \\ &= \begin{bmatrix} 0 & -\mathbf{Y}_{(12)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(23)} \\ -\mathbf{Y}_{(12)} & 0 & \mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} \\ -\mathbf{Y}_{(13)} & \mathbf{Y}_{(23)} & 0 & -\mathbf{Y}_{(12)} \\ -\mathbf{Y}_{(23)} & -\mathbf{Y}_{(13)} & -\mathbf{Y}_{(12)} & 0 \end{bmatrix}.\end{aligned}$$

Using these definitions for the off-diagonal form Eq. (19b) we obtain

$$\mu \frac{\partial^2 \omega_{(\lambda\kappa)}}{\partial x_k \partial x_k} - \rho \frac{\partial^2 \omega_{(\lambda\kappa)}}{\partial t^2} = \mathbf{Y}_{(\lambda\kappa)}. \quad (20)$$

The defined 4D twist motion, $\omega_{(\lambda\kappa)}$, means the rotational oscillation of the off-diagonal shear axes of the deviatoric tensor, E_{lq}^D , accompanied by the changes of the shear magnitude; such perturbation of the shear load may be caused by internal fracturing processes (Fig. 5).

The spin and twist motions form the complex rotation field defined as:

$$\omega_s = \omega_{[s]} + i\omega_{(s)} \quad (21)$$

From the balance relation (see: Subsect. “Recording Spin and Twist Angle Variation”) we obtain the relations joining the spin and twist motions.

Defects: Dislocation and Disclination Densities

The classical approach to the dislocation and disclination densities is based on the Kröner description of continuum with the self-fields (compare Subsect. “Self-field Nuclei: Deviations from Classical Elasticity”). In the asymmetric homogeneous continuum, the defect density can be introduced using the modified definition of disclosure, B_l , and the following definition of the twist-bend vector (compare Eqs. (7–9)) we define:

$$B_l = \oint [E_{kl} - \omega_{kl}] dl_k, \quad \Omega_q = \oint \chi_{kq}^T dl_k = \iint \theta_{kq} ds_k, \quad (22a)$$

where for

$$\chi_{mq}^T = \varepsilon_{ksq} \frac{\partial \omega_{mk}}{\partial x_s} \quad (22b)$$

the disclination density vanishes due to the compatibility conditions:

$$\theta_{pq} = \varepsilon_{pmk} \frac{\partial \chi_{kq}^T}{\partial x_m} = \chi^0 \varepsilon_{pmk} \varepsilon_{qns} \frac{\partial^2 \omega_{ks}}{\partial x_m \partial x_n} = 0.$$

For the dislocation field we obtain (compare Eq. (8)):

$$\begin{aligned}\alpha_{pl} &= \varepsilon_{pmk} \left(\frac{\partial E_{kl}}{\partial x_m} - \frac{\partial \omega_{kl}}{\partial x_m} \right) \\ &= \varepsilon_{pmk} \frac{\partial}{\partial x_m} \left(\frac{e^0}{2} \left(\frac{\partial u_l}{\partial x_k} + \frac{\partial u_k}{\partial x_l} \right) - \frac{\chi^0}{2} \left(\frac{\partial u_l}{\partial x_k} - \frac{\partial u_k}{\partial x_l} \right) \right).\end{aligned}\quad (23)$$

With the help of the constitutive relations (12) and (13) we arrive at the relation between the dislocation density and asymmetric stresses:

$$\alpha_{pl} = \frac{\varepsilon_{pmk}}{2\mu} \frac{\partial}{\partial x_m} \left((S_{(kl)} - \frac{\nu}{1+\nu} \delta_{kl} S_{ii}) - S_{[kl]} \right). \quad (24)$$

We shall note that the material constants, e^0 and χ^0 , define the types of defects and types of rotation nuclei; the complex constants will mean the constant phase shift between the fields.

Note that in the classic theory with defects, we distinguish also the different definitions for a dislocation field, e. g., the Burgers and Nye dislocations (comp: [28]).

For some particular case, $e^0 = 1$, $\chi^0 = -1$: we obtain a vanishing of defects, like:

$$B_l = \oint [E_{kl} - \omega_{kl}] dl_k = \oint \frac{\partial u_l}{\partial x_k} dl_k = 0, \quad \alpha_{pl} = 0. \quad (25)$$

This case may represent an extreme shear deformation.

We give also relations for another simple case, $e^0 = \chi^0$, which leads to dislocation density:

$$B_l = \oint [E_{kl} - \omega_{kl}] dl_k = \chi^0 \oint \frac{\partial u_k}{\partial x_l} dl_k, \quad (26)$$

$$\alpha_{pl} = \chi^0 \varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}.$$

We can consider two particular cases, the first giving a relation between the asymmetric stresses and the edge type dislocations ($e^0 = -1$, $\chi^0 = -1$):

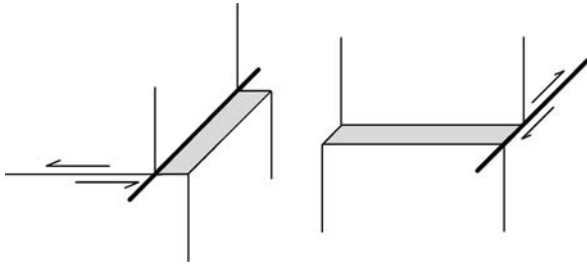
$$\alpha_{pl} = -\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l} \quad (27)$$

and the other which may describe the relation between the screw-type dislocations and asymmetric stresses ($e^0 = 2$, $\chi^0 = 2$):

$$\alpha_{pl} = 2\varepsilon_{pmk} \frac{\partial^2 u_k}{\partial x_m \partial x_l}, \quad (28)$$

where $p = l = s$, no summation over indices p and l .

Both of the considered cases could relate to the formation of the respective slip-discontinuities – Fig. 6.



Earthquake Source: Asymmetry and Rotation Effects, Figure 6
The edge and screw dislocation types

We find that for a suitable choice of the disclosure definition, we may arrive at different definitions of the dislocation and disclination densities; in particular, we note that for the edge and screw dislocations we shall consider different rotation nuclei.

The case (Eq. (25)) presents the extreme deformation while the further cases present the standard source models and related rotations.

Balance Laws for the Rotation Field and the EM Analogy

The complex rotation field (21), $\omega_s = \omega_{[s]} + i\omega_{(s)}$, may be presented in the tensor form:

$$\varepsilon_{kis} \omega_s = \omega_{ki} = \begin{pmatrix} 0 & \omega_{[3]} + i\omega_{(3)} & -\omega_{[2]} - i\omega_{(2)} \\ -\omega_{[3]} - i\omega_{(3)} & 0 & \omega_{[1]} + i\omega_{(1)} \\ \omega_{[2]} + i\omega_{(2)} & -\omega_{[1]} - i\omega_{(1)} & 0 \end{pmatrix}. \quad (29)$$

We can write the balance condition as

$$\iint \varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s ds_k = \iint \left(\frac{\partial}{\partial t} \omega_k + \frac{4\pi}{V} J_k \right) ds_k, \quad (30)$$

where we introduce the current field J_k and velocity V .

Hence, we obtain the field equations for the complex rotation motions:

$$\varepsilon_{kps} \frac{\partial}{\partial x_p} \omega_s - \frac{\partial}{\partial t} \omega_k = \frac{4\pi}{V} J_k \quad (31a)$$

or for spin and twist motions explicitly:

$$\varepsilon_{kps} \frac{\partial \omega_{[s]}}{\partial x_p} - \frac{1}{V} \dot{\omega}_{(k)} = \frac{4\pi}{V} J_k, \quad \varepsilon_{kps} \frac{\partial \omega_{(s)}}{\partial x_p} + \frac{1}{V} \dot{\omega}_{[k]} = 0. \quad (31b)$$

These equations lead us to the related wave forms:

$$\Delta \omega_{[n]} - \frac{1}{V^2} \ddot{\omega}_{[n]} = \frac{4\pi}{V} \varepsilon_{npk} \frac{\partial}{\partial x_p} J_k, \quad (32)$$

$$\Delta \omega_{(n)} - \frac{1}{V^2} \ddot{\omega}_{(n)} = 4\pi \frac{\partial}{\partial x_n} \varepsilon - \frac{4\pi}{V^2} \dot{J}_n,$$

where $\omega_{[s],s} = 0$ and $\omega_{(s),s} = 4\pi \varepsilon$, and under the condition that the velocity, V , is simultaneously transformed according to relativistic rules for a sum of velocities.

The obtained wave equations coincide with those derived previously (compare Eqs. (15, and 18b) with the definition for twist – Eq. (19)).

We shall note that in the asymmetric elastic continuum, the bonds related to rotational deformations are considered as comparable to those related to elastic rigidity moduli. More complicated situations with the material constant appears in the micropolar and micromorphic theories with the infinitesimally small nuclei (Eringen [21]). In the asymmetric continuum theory, presented in our treatise, the displacements and rotations appear as equally and similarly treated independent fields. Here enter also in a natural way, the axial deformation fields, with a structure similar to that of a thermal field (comp. Eq. (18a)).

Finally, we shall note that these wave fields correspond with $v \rightarrow c$ to the EM fields, $\omega_{[n]} \rightarrow B_n$, $\omega_{(n)} \rightarrow E_n$. The form of the rotation complex tensor (Eq. (29)) is fully analogous to the definition of the complex electromagnetic field $F_s = B_s + iE_s$.

Recording Spin and Twist Angle Variation

We shall find the suitable links between the defined fields and experimental data.

The spin motion can be precisely recorded by means of the Sagnac type interferometers (up to 10^{-9} rad/s); there are different types of such systems, e. g., ring laser and fiber optic, to record spin motion.

The angular twist oscillations and shear-twist motions we can record using a system of rotation seismometers. Such a system is based on the rotation seismographs that can record simultaneously the spin and twist angular motions [14,15,29].

In order to obtain the rotation motions, e. g., spin and twist, around the vertical axis we need the data from two parallel horizontal pendulums of opposite orientation. The observations collected clearly indicate that both the mean values of the spin and those of twist angular motions show the seismic oscillations with the same order of magnitudes [13,15,29].

We stress that the twist field, measured in this way gives only the angular variations of the off-diagonal axes of shears (19a); however, we may note that both the spin motion and the twist variation are mutually joint (see Eq. (31)) and therefore, we might theoretically derive knowledge of the shear state from the spin observations.

We shall add that when measuring the shear deformations with the help of a system of strainmeters, we can achieve more reliable and independent data on the shear-twist variations. Moreover, the strainmeter system can measure also the axial deformations.

Finally, we shall reply to the question of how we could compare the invariant twist field (19a and 19b) with the

observed shear variations. An exact procedure requires the following: the 6 components of the shear strain determined in an observation site system shall be transformed, at each time moment, into the off-diagonal system:

$$\{E_{11}, E_{22}, E_{33}, E_{23}, E_{31}, E_{12}\} \rightarrow \{E_{23}^D, E_{31}^D, E_{12}^D\} \\ = \{\omega_{(s)}\}.$$

Conclusions

In the standard asymmetric continuum theory the defects defined are not the material defects, but only those related to the structural deformations. This standard asymmetric theory permits one to find the differential relation between the dislocation density and the asymmetric stress field. Moreover, in this theory we may consider also other deviations related to other defects or interaction fields; to this end we could apply the Kröner approach with the elastic, self- and total fields [4].

The other important conclusion is that the influence of rotational processes in earthquake sources spreads outward, because these waves are not attenuated strongly, as it was believed according to classical ideal elasticity.

Earthquake Source: Fracture Processes

Introduction

We start our considerations with the thermodynamical conditions related to seismic energy release, and then we consider the rotation counterpart in the fracturing.

We shall be aware that the rotation processes of different nature and scale take part in such extremely complicated fracture phenomena, in which the dynamic processes proceed together with the simultaneous changes of material properties (see Teisseyre [30]). We shall recall the special role of rotations in the energy release effectiveness under different load conditions, and further on we shall include the rotation impact on the granulation processes accompanying the material crushing.

The constitutive laws must undergo simultaneously considerable changes, from the rigid elastic to plastic, and further, to mylonite-type material (in tectonics the mylonite means the crushed, granulated and even partly melted material in zone adjacent to fracture plane). In the narrow zones adjacent to fracturing, the shear stresses break the molecular bonds and in the crashed rock material the stresses immediately drop to a much lower level, while together with the advancing material granulation, we shall include a rapid increase of the stress and strain rates. Finally, in that narrow zone adjacent to fracturing, the stresses and strains may be gradually neglected

and progressively replaced by their time-rates. To describe these processes we shall simultaneously introduce the changes into the related constitutive relations. In result, the rock properties in this zone may even approach those characteristic for fluid. Such conditions may permit one to include in the fracture description the transport Navier–Stokes relations. The fracturing transport process, the bond breaking and granulation processes force us to include in the fracturing description, the hypothesis that the twist-shear deformations leading to the bond breaking precede the rebound rotation motion by $\pi/2$ in phase; this means that the difference between the shear motion and spin motion shall reach minimum when the latter is shifted by $\pi/2$ in phase.

We shall underline that the considered conditions in the mylonite zone can serve as the basis to formulate the asymmetric fluid theory with the extreme motion phenomena and dynamic defect objects.

A counterpart to the rotations and rotation energy release at fracture processes (e. g., in an earthquake source) explains fragmentation and spall processes and makes it possible to estimate the efficiency of different fracturing modes. Again we shall underline that in any theoretical approach, the elastic rotation energy can be considered only when assuming the constitutive law joining rotations with the antisymmetric stresses or stress moments.

Teisseyre et al. [18] have reexamined Dietrich's compression experiments [31], coming to the conclusion that under the compression load, there arise at some centers in the source region the induced precursory shear stresses; at a fracturing event we would arrive at the coseismic rebound compensation leading to a release of the induced stresses by the rebound process. Similarly, the precursory rotations associated with the newly formed dislocations or cracks shall have an opposite orientation to that related to the coseismic process. At the precursory stage these repeated processes lead to micro-fracturings, while during the seismic event there will occur under compression load, the fracturing with the rock fragmentation and the rebound macro-rotations at the inner centers where the precursory induced stresses accumulate.

Earthquake Thermodynamics

Basic thermodynamic relations for line defects (dislocations and vacant dislocations) are derived under the assumption of a dense network of defects forming a kind of super-lattice [32,33,34,35]. The thermodynamic functions of line defects can be associated with defects in the super-lattice. Let us confine our considerations to the irreversible (plastic) deformations of solids.

To distinguish the thermodynamic functions used here from those used under pressure conditions, we will use the symbols with hat and we will consider only a pure shear work under shear load $S_{(\cdot)}$ and under induced friction stress moment $S_{[\cdot]}$ (see: Eqs. (13) and (14)) – we consider deformations at a constant volume; the work $d\hat{W}$ done on a body (per unit volume), the internal energy change $d\hat{U}$ and the heat received in an exchange with the surrounding dQ are related:

$$\begin{aligned} d\hat{W} &= SdE = S_{(\cdot)}dE_{(\cdot)} + S_{[\cdot]}d\omega_{[\cdot]} \geq 0, \\ d\hat{U} &= dQ + SdE, \end{aligned} \quad (33)$$

where $dE_{(\cdot)}$ and $d\omega_{[\cdot]}$ are increments of strain and spin.

For the Helmholtz free energy \hat{F} and Gibbs free energy \hat{G} we have:

$$\begin{aligned} \hat{F} &= \hat{U} - T\hat{S}, \quad \hat{G} = U - SE - T; \\ Td\hat{S} &\geq dQ; \quad d\hat{S} \geq 0, \end{aligned} \quad (34)$$

where T is the absolute temperature and \hat{S} is the entropy; $d\hat{S}$ would be the entropy production due to the irreversible processes occurring inside the system.

The local formulation of the second law of thermodynamics requires that the entropy production be positive wherever an irreversible process occurs [34]. It is postulated that even outside equilibrium, the entropy depends only on the same variables as at equilibrium. In order to derive the expression for the entropy production, Prigogine [36] introduced some additional assumptions. Namely, he assumed that the entropy production can be determined for conditions near equilibrium.

Formation of a dislocation gives negative contribution to the Gibbs energy [37] and therefore it is not possible to find a minimum of the Gibbs function with respect to the number of dislocations. Thus, the dislocation distribution cannot exist as a thermodynamically stable system, since the Gibbs free energy has no minimum of any equilibrium concentration of dislocations.

However, for a dense dislocation distribution there enter the repulsive interactions between dislocations, and a kind of dislocation super-lattice can be considered [32,34,35]. The “ideal super-lattice” can be treated as a reference state and the real super-lattice, in the case of dense distribution of dislocations, can be in the equilibrium state. We define the vacant dislocations in the following way: to the randomly formed network of dislocations we shall add a number of line vacancies – the vacant dislocations – in such a way that as a result we obtain the super-lattice filled by dislocations and vacant dislocations. In this situation, a real distribution of dislocations can be

described as a departure from the state of ideal super-lattice, given by the amount of vacant dislocations.

These processes are accompanied by an internal friction related to displacement formed by dislocations and hence a spin motion appears as inherently present there.

The Gibbs energy minimum can now exist as the equilibrium number of the vacant dislocations. We can consider the structure of a cross-zone consisting of bands of layerlets; such a structure favors the appearance of some macroscopic dislocations under conditions of shearing deformation. The particular values of the Burgers vector become related to particular layer thicknesses. In this sense we suppose that a fine band structure could play the role of a quantization kind factor; this problem is related to the earthquake shear band model.

Consider a continuum that contains a regular (cubic) super-lattice of dislocation lines with a certain super-lattice parameter Λ ($\Lambda \gg \lambda$; λ relates to a basic rock lattice). The notion of the super-lattice is directly related to the shear band model of fracturing [34], see Fig. 7.

We associate the thermodynamic functions of line defects with the defects in a super-lattice; the Gibbs free energy may have a minimum corresponding to the equilibrium concentration of the vacant dislocations in the super-lattice. Many results can be now transferred from the thermodynamics of point defects (Varotsos and Alexopoulos [38]). The regular super-lattice, which includes the dislocations and vacant dislocations, may be described in a very rough approximation by a characteristic distance Λ (super-lattice constant). For the ideal super-lattice (no vacant dislocations), the mean value of distances following from distribution of dislocations defines the reference dislocation density $\alpha^0 = \lambda/\Lambda^2$, while for a real body with n dislocations we may add to it other \hat{n} vacant dislo-

cations in such a way that the whole set $n + \hat{n} = N$ (dislocations and vacant dislocations) fits to a regular super-lattice with the smallest error. For the density of dislocations α , and vacant dislocations $\hat{\alpha}$, we can write [34]:

$$\alpha = \left(1 - \frac{\hat{n}}{N}\right) \frac{\lambda}{\Lambda^2}, \quad \hat{\alpha} = \frac{\hat{n}\lambda}{N\Lambda^2} \frac{\lambda}{\Lambda^2} \exp\left(-\frac{\hat{g}^f}{kT}\right), \quad (35)$$

where the number \hat{n} can be identified with an equilibrium value in relation to the formation energy of vacant dislocation \hat{g}^f per length of the crystal lattice λ .

The stress field and the resistance stress (e. g., the drag resistance in a dislocation motion and the friction stress in a crack motion) are defined as [37]:

$$S = \frac{\partial \hat{W}}{\partial E}, \quad S_F \equiv \frac{\partial \hat{F}}{\partial E},$$

while the Gibbs function for a crystal containing the vacant dislocations can be written as

$$\hat{G} = \hat{G}^0 + \hat{n}\hat{g}^f - T\hat{S}_c,$$

where \hat{S}_c is the configuration entropy.

Near the equilibrium state under a constant local shear S and temperature T the Gibbs energy is close to its minimum and the equilibrium values could be found as follows:

$$\begin{aligned} \left. \frac{\partial \hat{G}}{\partial \hat{n}} \right|_{S,T} &= 0, \quad \hat{n}^{\text{eq}} = N \exp\left(-\frac{\hat{g}^f}{kT}\right), \\ \hat{\alpha} &= \frac{\lambda}{\Lambda^2} \exp\left(-\frac{\hat{g}^f}{kT}\right), \quad \hat{S}_c = \hat{n} \left(k + \frac{\hat{g}^f}{T}\right), \end{aligned} \quad (36a)$$

while the Gibbs energy function becomes

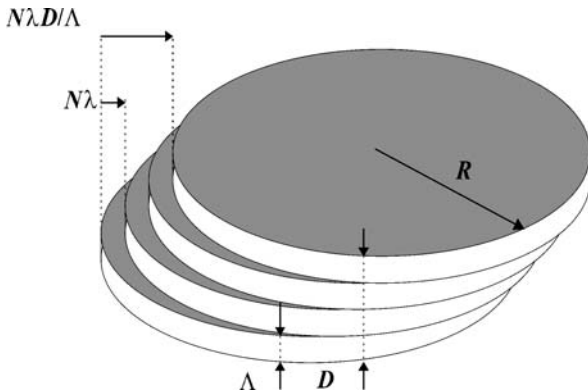
$$\hat{G} = \hat{G}^0 - \hat{n}kT. \quad (36b)$$

The equilibrium free energy is less than that for an ideal super-lattice \hat{G}^0 ; the difference is kT per line vacancy, per length of crystal lattice.

For the point defect thermodynamics, Varotsos and Alexopoulos [38] have introduced the so-called $CB\Omega$ theory approximating the contribution to the Gibbs energy from the formation of point defects.

For the line vacancies, a change of the Gibbs energy depends on the stress level and resistance stress. Therefore, we postulate for the approximative value of such change per unit element (formation energy of vacant dislocation) the following expression defining the $C\mu b\lambda^2$ model:

$$\hat{g}^f = C\mu b\lambda^2, \quad \hat{n}^{\text{eq}} = N \exp\left(-\frac{C\mu b\lambda^2}{kT}\right), \quad (37)$$



Earthquake Source: Asymmetry and Rotation Effects, Figure 7
Shear band model

where C is constant; \hat{g}^f becomes here independent on stress load and resistance, μ is the rigidity, b is the Burgers vector of dislocation.

Concluding, a body containing some number of dislocations cannot be in a state of equilibrium; there is no minimum of the Gibbs function, because when reducing the number of dislocations we always get a smaller value of the free energy. For a dense distribution of dislocations we can assume, due to their interaction, that there exists a certain super-lattice composed of dislocations.

The equilibrium density of the vacant dislocations may be written now with help of Eq. (37)

$$\hat{\alpha} = \frac{\lambda}{\Lambda^2} \exp\left(-\frac{C\mu b\lambda^2}{kT}\right) \quad (38)$$

and becomes useful, when looking for the most probable density value of defects after the energy release in a fracturing process. The density $\alpha^0 = \lambda/\Lambda^2$ may be identified here with the reference density.

We can assume that before an earthquake a super-lattice is almost completely filled in by dislocations ($n \approx N$ and $\hat{n} \approx 0$). The maximum number of dislocations in arrays could reach the value $(\Lambda/\lambda)^2$ per area Λ^2 . The total moment for an area $\Delta s = N\Lambda^2$ affected by the arrays of dislocation along the slip planes becomes:

$$\tilde{M} = \mu\lambda\Delta s = \mu\lambda N\Lambda^2 \left(\frac{\Lambda}{\lambda}\right) = \mu N\Lambda^3.$$

After an earthquake, the number of vacant dislocations \hat{n} shall increase, probably to the equilibrium value (37) and hence we can express the seismic moment by the number of coalescence processes related to surface element Λ^2 as equal to $\Delta\hat{n} = \frac{\Lambda}{\lambda}\hat{n}^{eq}$; the factor Λ/λ expresses a maximum concentration of dislocations in the arrays.

We obtain for the seismic moment

$$\begin{aligned} \tilde{M}_0 &= \tilde{M}\Delta\hat{n} = \mu N\Lambda^3 \left(\frac{\Lambda}{\lambda}\right) \Delta\hat{n} \\ &= \mu N\Lambda^3 \left(\frac{\Lambda}{\lambda}\right) \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right), \end{aligned}$$

where C is constant for given structure.

Using the expression for a change of the free energy values we may include the formation of dislocation arrays along the glide planes and we put

$$G = G^0 + \Delta\hat{n} \left(\frac{\Lambda}{\lambda}\right) kT$$

According to these results, the total energy release ΔE and seismic moment are:

$$\begin{aligned} \Delta E &= G - G^0 = \Delta\hat{n} \left(\frac{\Lambda}{\lambda}\right) kT \\ &= \left(\frac{\Lambda}{\lambda}\right) NkT \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right) \text{ and} \\ \tilde{M}_0 &= \mu\Lambda^3 \frac{\Delta E}{kT}. \end{aligned} \quad (39)$$

This formula is an important relation between the energy release density and seismic moment density; for instance, for a given ΔE the elementary seismic moment \tilde{M}_0 decreases with temperature. Free energy related to defect formation, \hat{g}^f , is proportional to $\mu\lambda\Lambda^2$ being constant for a given structure; with growing value of Λ the seismic moment becomes greater.

Neglecting the term related to the formation entropy, we can write for entropy density change:

$$\Delta\tilde{S} = kN \left(\frac{\Lambda}{\lambda}\right) \left(1 + \frac{C\mu b\lambda^2}{kT}\right) \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right).$$

All of these relations concern the quantities referred to the multiple of the cubic volume $N\Lambda^3$ thus, we can correct these quantities to that related to a given source volume by introducing the factor $\pi R^2 D/N\Lambda^3$:

$$\begin{aligned} M_0 &= \mu\pi R^2 D \left(\frac{\Lambda}{\lambda}\right) \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right), \\ \Delta E &= \pi R^2 D \frac{kT}{\Lambda^2\lambda} \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right) \text{ and} \\ \frac{M_0}{E^{\text{rad}}} &= \frac{\mu\Lambda^3}{\eta kT}, \\ \Delta\tilde{S} &= \pi R^2 D \frac{k}{\Lambda^2\lambda} \left(1 + \frac{C\mu\lambda\Lambda^2}{kT}\right) \exp\left(-\frac{C\mu\lambda\Lambda^2}{kT}\right), \end{aligned} \quad (40)$$

where $\eta\Delta E = E^{\text{rad}}$, η is the seismic efficiency; E^{rad} is radiated energy.

In the above consideration we took into account both energies related to slip and friction processes, and thus, the total released energy includes that related to stress drop and that related to heat caused by friction processes.

Further on, we will consider the fracturing processes in an earthquake source.

Synchronization and Fracturing

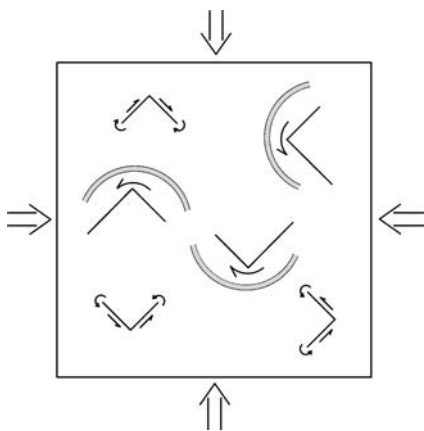
The inner stress accumulation relates to the formation of defect densities; due to the interaction between dislocations, we arrive at stress concentration at the first blocking dislocations of the formed dislocation arrays.

In the compression case with no initial shear field and due to the lower value of shear resistance, we have to assume that inside a body there appear regions with induced shear stresses of opposite signs, and induced antisymmetric stresses. The earthquake process and its energy release relate to a coalescence of dislocation arrays of opposite signs and related rotation release motion. Of course, we shall consider a fracture process as a chain of events; let us consider the micro-fracture centers formed on two perpendicular plane fragments; the induced shear stresses will be opposite on those plane fragments, but will have the common orientation of a spin motion – see Fig. 8; hence, the shears will be almost compensated for, while the spin field will remain unchanged.

The spin field, $\omega_{[s]}$, propagates and influences the processes in the adjacent regions; we believe that this propagation synchronizes the spin motions in the adjacent centers, in such a way that the sense of spin motion becomes the same over the whole fracture region. That means, the spin propagation assures a synchronization of fracture processes, especially under compression load where the energy release relates to the fragmentation revealed by rotation and granulation processes.

Reversely, under the shear load, while common shear deformation, E_{kn}^D (or expressed as twist $\omega_{(s)}$), progresses, the spins on the main fracture differ from those on the adjacent perpendicular fragments and attenuate fracture progress on those fragments.

Accordingly, we can believe that at the compression load, the total shear stress drop will be relatively small, while the rebound rotations will release an important amount of rotation energy. At the shear load the release of shear stresses will prevail.



Earthquake Source: Asymmetry and Rotation Effects, Figure 8
Compression load: induced shear centers and formation of fragments with related rotations

Concluding, the rotation processes in fragmentation and fracturing under compression load play an essential role. Under prevailing shear load the rebound process releases shear load with the regional stress drop, while the rotation processes play a minor role.

Further on, we will discuss the importance of the granulation processes related to rotations in meso-scale, which we can place between the bond breaking processes in the micro-scale and material fragmentation in the macro-scale.

Granulation and Formation of Mylonite Zones

The fracturing process, especially under the action of shearing load, is accompanied by material granulation adjacent to shear fracture planes; thus it becomes spectacular at the formation of narrow, long mylonite zones. In this process we shall take into account a special role of rotations – the meso-rotations of different scales; these rotations are related to bond breaking and friction processes.

Co-action of the spin and twist-shear motions in bond breaking, granulation and formation of mylonite material can effectively help us to explain the fracture process; the simultaneous formation of the adjacent mylonite zone appears due to such a co-action of spin and shears and of the fracture transport phenomena.

Based on the standard asymmetric continuum theory, as presented in the first part, we would like to consider the material undergoing a progressive crushing process. We may arrive even at the conditions more similar to fluid material, and thus finally shall enter into our considerations the Navier–Stokes transport equations.

Starting with the description of the rock continuum following from the standard asymmetric theory of continuum ($S_{ik} = S_{(ik)} + S_{[ik]}$, $E_{ik} = E_{(ik)}$, $\omega_{ik} = \omega_{[ik]}$), we approach the final stage of the crushing/granulation process in zones adjacent to fracture planes. In these zones, simultaneously with dynamic processes, there occur changes of material properties from hard rocks to mylonite granulated material.

Approaching the final stage, the stresses, strains and rotations presented in the description of the standard asymmetric continuum become gradually neglected and progressively replaced by the constitutive relations for time-rates of stresses and strains.

The constitutive laws for rock asymmetric continuum – the relations (1) written for the deviatoric fields and for the antisymmetric fields

$$S_{(kl)}^D = 2\mu E_{kl}^D, \quad S_{[kl]} = 2\mu\omega_{kl} \quad (41)$$

– will gradually change during fracturing to those including the time dependent processes:

$$\begin{aligned}\sigma S_{(ik)}^D + \tau \dot{S}_{(ik)}^D &= 2\mu E_{ik}^D + 2\eta \dot{E}_{ik}^D, \\ \sigma S_{[ik]} + \tau \dot{S}_{[ik]} &= 2\mu E_{ik} + 2\eta \dot{\omega}_{ik}.\end{aligned}\quad (42)$$

The introduced material constants are related to magnitudes of the slip, u , and slip rate, v .

When in a narrow zone the huge shear stresses break the molecular bonds, the stresses crushing rock material immediately drop down to the low values and in the crushed mylonite material we observe the immediate increase of the stress and strain rates to such degree that the stresses and strains may be neglected in the respective constitutive relations for that narrow zone. Finally, these changes will lead to the constitutive laws for the melt and granulated parts of mylonite material in which, practically, will remain only the field time rates:

$$\dot{S}_{(ik)} = 2\eta \dot{E}_{ik}, \quad \dot{S}_{[ik]} = 2\eta \dot{\omega}_{ik}. \quad (43)$$

The direct observation of the gauge zone of the Kobe (Japan, 1995) earthquake at the Avaji island suggests that the size of an inner completely melted part of the mylonite zone ranges around couple of centimeters (private communication W. Debski).

Further on, we will assume for the sake of simplicity, that during the fracturing the mylonite material remains incompressible.

In such a way, the nucleation progresses and fracture propagates simultaneously with the granulation process in the intact material (or in the compact zone previously crushed) – see Fig. 9.

In this new description, the shear rates create the dynamic angular deformations then lead to the bond

breaking processes, and finally to the fracturing transport process.

We pass to the final stage; for the crushed incompressible mylonite or sand, similarly to incompressible fluids, where tensor $\dot{\omega}_{ik}^T$ is related to spin motion (not to rotation of displacement), the mylonite viscosity is η , and the mylonite relaxation time is denoted by τ .

The relations (42) define the ideal quasi-viscous mylonite for an incompressible crushed material.

Further, we assume the coincidence/identity of rotation of velocity field $\dot{\mathbf{u}}$ with the point rotation field ω . So, we assume that rotations of particles (micro-rotation ω) coincide with macro-rotations ($\text{rot } \mathbf{u}$). For mylonite, such a coincidence between the micro-rotation and macro-rotation seems reasonable, and thus, our assumption that viscosity η coincides with rotation viscosity η^* may be correct.

For our narrow mylonite zone, existing already near the pre-slip planes or just simultaneously formed, we may, further on, apply the Navier–Stokes transport equation. Referring to our former considerations on the asymmetric continuum theory (see Subsect. “Spin and Twist Motions”) we may add to the spin rotational motions the oscillations of the strain shear rates (called twist motion). Such motions, especially in an earthquake source zone, are due to the friction processes.

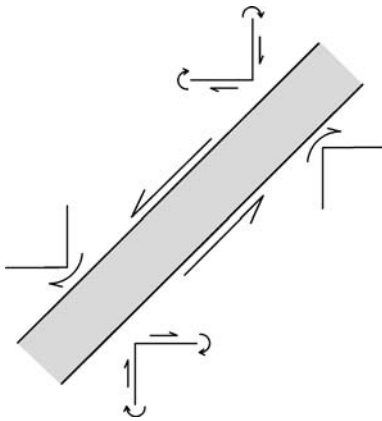
We may note that when including these complex rotational motions in the theory we may replace the friction constitutive laws, as based on the experimental data, by the constitutive law joining the asymmetric stresses with spin and shears field oscillations or otherwise with spin and twist.

Slip Propagation and Spin Release Hypothesis

While searching the fault slip solutions, we use the classical elasticity tools with an additional friction constitutive law based on experimental data. When instead of it we consider the asymmetric elastic continuum, we are able to include the defect interaction and we can derive the elastodynamic fault solution describing slip propagation with fracturing process and related seismic radiation.

The angular deformations preceding the bond breaking process lead to the efficient rise of the angular moments around material grains. In the narrow mylonite zone, we arrive at the equivalence between this expression and the laws introduced in the considerations on the friction resistance and slip.

The co-action of the rotation ($\text{rot } \mathbf{u}$, or spin $\omega_{[.]}$) and shear ($E^D(\mathbf{u})$, or twist – $\omega_{(.)}$) motions can lead further to the slip fracturing motion. We assume that the bond



Earthquake Source: Asymmetry and Rotation Effects, Figure 9
Mylonite zone and neighboring deformations

breaking process and granulation of material precede the slip movement: just after the bond breaking micro-process there, we would have the released rebound spin motion retarded in phase.

This hypothesis is supported by the following solution of the homogeneous wave equations for the twist and spin in a mylonite zone (see Eq. (31)):

$$\begin{aligned}\omega_{(s)} &= i\omega_{[s]} ; \quad \omega_{[s]} = \omega_{[s]}^0 \exp[i(k_i x_i - \omega t)] , \\ \omega_{(s)} &= \omega_{(s)}^0 \exp[i(k_i x_i - \omega t)] ,\end{aligned}\quad (44)$$

where the six constants in $\omega_{[s]}^0 = \text{abs}(\omega_{[s]}^0) \exp(i\psi_s)$, $\omega_{(s)}^0 = \text{abs}(\omega_{(s)}^0) \exp(i\varphi_s)$ shall fulfill the six conditions.

We may consider the following 2D solution of Eq. (44) in the systems $\{r, \varphi, z\}$:

$$\omega_{(\varphi)}(r) = i\omega_{[\varphi]}(r) , \quad \square\omega_{[\varphi]}(r) = 0 .$$

The related solution corresponds to a turbulence structure. Thus, from a dislocation-slip structure formed in the earthquake premonitory domain (see: Subject “[Earthquake Thermodynamics](#)”), gradually destroyed during a fracture process by a spin release motion, we can arrive at a turbulence structure appearing in a melted or fully granulated material.

With the introduced waves, $\omega_{(s)} = i\omega_{[s]}$ we arrive at the possibility to study the dynamic defect objects and to explain the synchronization of the micro-fracturing processes due to an influence of the propagating waves. For the fracture processes under compression such a synchronization will assure the common sense of the induced twist and spin motions, while under shear load – the formation of a long shearing fracturing. In the last case, the spin waves related to a given slip on the main fracture plane attenuate those with the opposite spins generated at the perpendicular fragments, and due to the conjugate solution (see Eq. (44)) reduce the slip motions on those fragments.

The presented conjugate solution Eq. (44) suggests that the spin rebound motion is delayed in phase by $\pi/2$ (as we have $\exp[i(k_i x_i - \omega t)] = \exp[ik_i x_i - i(\omega t - \pi/2)]$); when slip starts due to breaking of bonds, the micro-spin motions are released.

Following this assumption we expect that such a correlation between the recorded twist motions and spin motions shifted by $\pi/2$ in phase can exist in some wavelets.

Now we can propose the following description of the fracture process.

- First, according to external load conditions the stresses rise while the disclosure and dislocation field can be neglected – the case given by relations (25).

- Next, approaching the fracture process we may observe the “accumulation” phase with the co-action of the twist and spin – the case given by relations (26).
- Finally, fracturing processes start and when entering into the time rate domain we can describe the “release” phase of the process as follows (compare: Eqs. (21–26 and 43):

$$\dot{B}_l = \oint [E_{(kl)} + i\omega_{[kl]}] dl_k , \quad (45)$$

$$\begin{aligned}\dot{\alpha}_{pl} &= \varepsilon_{pmk} \left(\frac{\partial \dot{E}_{kl}}{\partial x_m} + i \frac{\partial \omega_{kl}}{\partial x_m} \right) \\ \dot{\alpha}_{pl} &= \frac{\varepsilon_{pmk}}{2\mu} \frac{\partial}{\partial x_m} \left(\left(\dot{S}_{(kl)} - \frac{v}{1+v} \delta_{kl} \dot{S}_{ii} \right) + i \dot{S}_{[kl]} \right) ,\end{aligned}\quad (46)$$

where we have the dynamic disclosure and v -dislocation density under the conditions formed by solution (see Eq. (44)), supplemented with the relations between the asymmetric stress rates and the dynamic dislocation objects (v -dislocations).

- This case presents a formation of dynamic discontinuities and the related dynamic processes in which the slip and bond breaking leads to the rebound spin motions delayed in phase by $\pi/2$.

The co-action of the spin and twist motions leads to the “accumulation” phase, while the conjugate solution (see Eq. (44)) presents a fracture process – “release” phase.

We might suppose that the fracture process could proceed with the consecutive accumulation and release micro-processes; in such a situation the related twist and spin motions will appear consecutively as pairs of wavelets in phase (or anti-phase) and those differing in phase by $\pi/2$. Figure 10 presents an example of the coincidence of the spin and twist motions after the Hilbert transformation shifting the angular twist record ahead in phase by $\pi/2$.

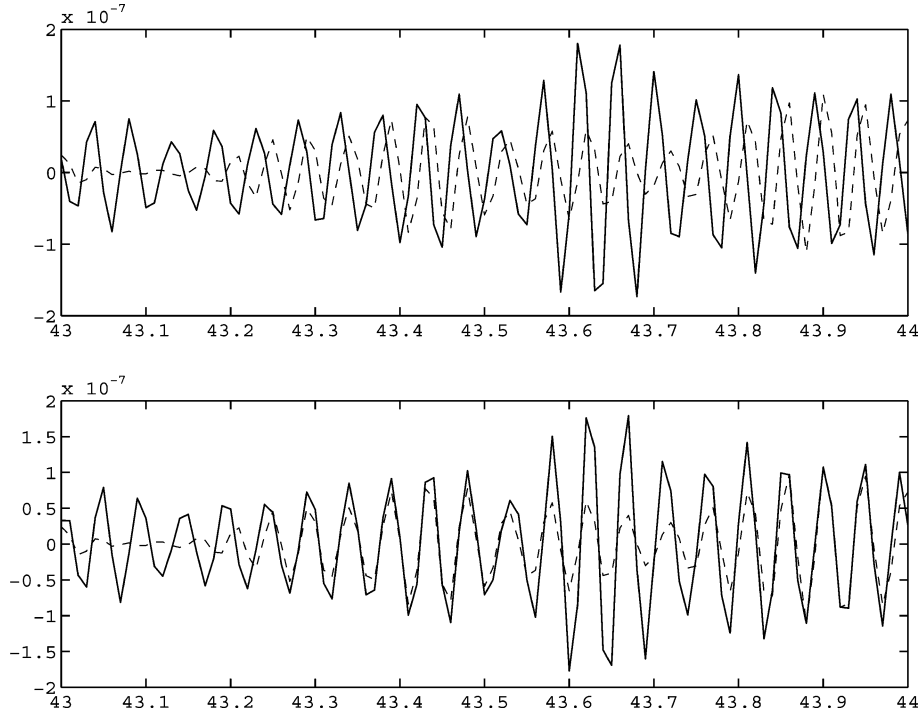
The presented theory, due to its simplicity, could be very useful for some problems, among others those in which macro-rotation takes an important role in the asymmetric fluid dynamics.

We can call the solution (44) as the fracture synchronization waves.

Finally, we shall notice that similar solutions may exist for the electric and magnetic induction vectors:

$$D_s = iB_s \leftrightarrow D_s^0 = iB_{[s]}^0 , \quad (47)$$

where this relation shall be assured by the appropriate material constants.



Earthquake Source: Asymmetry and Rotation Effects, Figure 10

Example of coincidence of the spin and twist angular motions (rad/s versus s) after the Hilbert transformation shifting the twist record ahead in phase by $\pi/2$; *upper part* – the original records, *lower part* – the twist record transformed (from the original seismic record obtained by the system of the rotation seismometers; L'Aquila Observatory, 17.02.2006; the *continuous line* – twist, the *broken line* – spin)

Towards Asymmetric Fluid Theory and Extreme Phenomena

Approaching the conclusions, we shall specify how we could formulate the asymmetric theory of the fluid continuum in which the stress, strain and rotation fields vanish, but their rates exist as related to the velocity field v :

$$\begin{aligned}\dot{E}_{ik} &= e^0 \frac{1}{2} \left(\frac{\partial v_k}{\partial x_i} + \frac{\partial v_i}{\partial x_k} \right), \\ \dot{\omega}_{ik} &= \chi^0 \frac{1}{2} \left(\frac{\partial v_k}{\partial x_i} - \frac{\partial v_i}{\partial x_k} \right).\end{aligned}\quad (48)$$

The velocity field v shall obey the Navier–Stokes **transport equation**.

For the sake of simplicity, let us consider the incompressible fluid ($\dot{E}_{ss} = 0$, $\dot{E}_{ik} = \dot{E}_{ik}^D$), the basic constitutive laws are assumed similarly to those for the asymmetric continuum:

$$\dot{S}_{(ik)} = \eta \dot{E}_{ik}, \quad \dot{S}_{[ik]} = \eta \dot{\omega}_{ik}, \quad (49)$$

where η is viscosity.

In a similar manner, the structural dynamic objects can then be defined as defects in the standard asymmetric continuum.

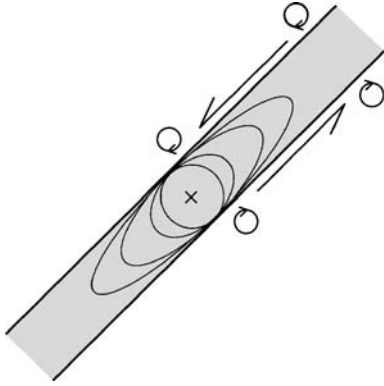
The extreme motion phenomena, related to the shear rate and spin, $\dot{E}_{(kl)} + \dot{\omega}_{(kl)}$, could be expected for the following case (compare: Eqs. (22–26 and 43):

$$\begin{aligned}\dot{B}_l &= \oint [\dot{E}_{kl} + \dot{\omega}_{kl}] dl_k = \oint \frac{\partial v_l}{\partial x_k} dl_k = 0, \\ \dot{\alpha}_{pl} &= 0, \quad \dot{\Omega}_q = 0,\end{aligned}\quad (50)$$

while formation of the dynamic defect objects can be described as:

$$\begin{aligned}\dot{B}_l &= \oint [\dot{E}_{kl} - \dot{\omega}_{kl}] dl_k =, \\ \dot{\alpha}_{pl} &= \frac{1}{2\mu} \varepsilon_{pmk} \frac{\partial}{\partial x_m} [\dot{S}_{(kl)} - \dot{S}_{[kl]}].\end{aligned}\quad (51)$$

The former case (see Eq. (50)) presents an extreme shear rate deformation, like soliton waves, while this case (see Eq. (51)) would relate to a formation of the v -slip-discontinuity.



Earthquake Source: Asymmetry and Rotation Effects, Figure 11

Extreme motions: soliton wave will be related to a given deformation of the circle as follows for the parameter $C = e^0$; Amplitude plot (Mathematica 5.0): with *AspectRatio* \rightarrow *Automatic*:
 $A = \text{Plot}[y = -Cx/2 \pm 0.5\sqrt{[(Cx) \wedge 2 - 4x \wedge 2 + 4]}],$
 $\{x, -\sqrt{[4/(4 - C \wedge 2)]}, \sqrt{[4/(4 - C \wedge 2)]}\}$

In 2D we can show an effect of the co-action of the macro spin and twist motions in the following way; let us put $C = e^0$ and $\chi^0 = 1$

$$\begin{aligned}\dot{E}_{12} + \dot{\omega}_{12} &= Cv_{(2,1)} + v_{(2,1)} \\ &= C \frac{1}{2} \left(\frac{\partial v_k}{\partial x_i} + \frac{\partial v_i}{\partial x_k} \right) + \frac{1}{2} \left(\frac{\partial v_k}{\partial x_i} - \frac{\partial v_i}{\partial x_k} \right).\end{aligned}$$

For $C = 0$ we would have only the spin motion, while a full coincidence will occur at $C = 1$.

The effect of such a superposition of the spin and twist motions is presented on Fig. 11.

We believe that this approach might explain some extreme fluid phenomena related to atmosphere and oceans.

The balance equations for field rates (48) at $e^0 = \chi^0$ may lead us to the wave equations for the related spin and twist rate fields, $\dot{\omega}_{[s]}$ and $\dot{\omega}_{(k)}$, as defined similarly to the relations derived in Subsect. “Balance Laws for the Rotation Field and the EM Analogy” (see: Eqs. (29 and 31):

$$\begin{aligned}\varepsilon_{kps} \frac{\partial}{\partial x_p} \dot{\omega}_{[s]} - \frac{1}{V} \frac{\partial}{\partial t} \dot{\omega}_{(k)} &= \frac{4\pi}{V} j_k, \\ \varepsilon_{kps} \frac{\partial}{\partial x_p} \dot{\omega}_{(s)} + \frac{1}{V} \frac{\partial}{\partial t} \dot{\omega}_{[k]} &= 0.\end{aligned}\quad (52)$$

The appearance of such coupled waves transversal to the transport motion brings physical background for diffraction in fluids as explained usually by the Huygens principle.

Conclusions

Under both the confining pressure and external shear, the role of micro-fracturing in the bond breaking process is

similar; however, we observe the essential differences for rotations in larger scales.

The confining condition leads to formation of induced opposite arrays of dislocations, resulting in fragmentation processes and chaotically oriented macro-rotations, leading therefore to a rotation release process.

The shear condition leads to more concentrated fracturing along some planes, high shear strain release and correlated rotations.

Both cases include formation of narrow mylonite zones adjacent to the fracturing planes or their fragments, but these processes prevail rather under shear conditions.

We draw attention to the importance of the rotations in meso-scales – between the micro-scale bond breaking process and that related to macro-rotation at material fragmentation. The meso-scale rotations are related to material granulation and become observed in any fracturing process; such motions may be revealed in the spectacular formation of the narrow, long mylonite zones under shear load conditions. Coincidence and co-action of the spin and twist-shear motions in bond breaking and formation of mylonite material help one to understand the fracture motion; the simultaneous formation of the mylonite zones appears due to common action of these motions and to fracture transport phenomena.

Rotations at source zones help to understand geometry of fracturing and releases of stress and rotation counterparts as a result of precursory and rebound processes.

We have presented also a new idea how to construct the asymmetric fluid theory with the asymmetric stress rate field.

Final Remarks

With the additional constitutive law joining the rotations with antisymmetric part of stresses, we have proved that the rotation waves exist, even in a homogeneous elastic continuum. We have defined the twist motion as the rotational oscillation of the main shear axes including the shear magnitude variations. The derived wave equations for the twist and spin motions have been considered in relation to the processes in seismic source.

Only in the presented approach with the standard asymmetric theory may we study the co-action of the independent motions and deformations; this is due to the new relations joining the spin and slip motions. A possible phase shift between these motions leads to different families of deformation and related solutions.

We have shown how the rotations at source zones help us to understand physics and geometry of fracturing

and release of stresses in the precursory and rebound processes.

The derived wave equations for spin and twist motions are similar to the EM wave equations.

Our considerations show the importance of the simultaneous recording of the translational and rotational earthquake motions, and also the strains (at least the deviatoric strains).

Finally, we have shown how to construct the asymmetric fluid theory in which, with a help of the asymmetric stress rates, the various extreme phenomena, including soliton waves, can be theoretically explained.

Acknowledgments

Work done under the support of the INTAS Project 05-1000008-7889.

Bibliography

Primary Literature

- Shimbo M (1975) A geometrical formulation of asymmetric features in plasticity. *Bull Fac Eng Hokkaido Univ* 77:155–159
- Shimbo M (1995) Non-Riemannian geometrical approach to deformation and friction. In: Teisseyre R (ed) *Theory of earthquake premonitory and fracture processes*. PWN, Polish Scientific Publishers, Warszawa, pp 520–528
- Teisseyre R, Boratyński W (2003) Continua with self-rotation nuclei: evolution of asymmetric fields. *Mech Res Commun* 30:235–240
- Teisseyre R, Boratyński W (2006) Deviations from symmetry and elasticity: Asymmetric continuum mechanics. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 31–42
- Kozak JT (2006) Development of earthquake rotational effect study. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 3–10
- Ferrari G (2006) Note on the historical rotation seismographs. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 367–376
- Gutenberg B (1926) *Grundlagender Erdbebenkunde*. Univ. Frankfurt a/M, Frankfurt
- Teisseyre R (1973) Earthquake processes in a micromorphic continuum. *Pure Appl Geophys* 102:15–28
- Cochard A, Igel H, Schuberth B, Suryanto W, Velikoseltsev A, Schreiber U, Wassermann J, Scherbaum F, Vollmer D (2006) Rotational motions in seismology: Theory, observation, simulation. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 391–411
- Schreiber KU, Stedman GE, Igel H, Flaws A (2006) Ring laser gyroscopes as rotation sensors for seismic wave studies. In: Teisseyre et al (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 377–389
- Takeo M (2006) Rotational motions excited by earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 131–156
- Leszek R, Jaroszewicz LR, Krajewski Z, Solarz L (2006) Absolute rotation measurement based on the Sagnac effect. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 413–438
- Moriya T, Teisseyre R (2006) Design of rotation seismometer and non-linear behaviour of rotation components of earthquakes. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 439–450
- Wiszniowski J (2006) Rotation and twist motion recording – couple pendulum and rigid seismometers system. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 451–470
- Teisseyre R, Suchcicki J, Teisseyre KP, Wiszniowski J, Palangio P (2003) Seismic rotation waves: basic elements of the theory and recordings. *Ann Geophys* 46:671–685
- Teisseyre R, Białecki M, Górski M (2006) Degenerated asymmetric continuum theory. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 43–56
- Kröner E (1981) Continuum theory of defects. In: Balian R, Kleman M, Poirer JP (eds) *Physique des défauts/physics of defects* (Les Houches, Session XXXV, (1980)). North Holland Publ Com, Dordrecht
- Teisseyre R, Górski M, Teisseyre KP (2006) Fracture-band geometry and rotation energy release. In: Teisseyre R, Takeo M, Majewski E (eds) *Earthquake source asymmetry, structural media and rotation effects*. Springer, Berlin, pp 169–184
- Nowacki W (1986) *Theory of asymmetric elasticity*. PWN, Warszawa and Pergamon Press, Oxford, New York, Toronto, Sydney, Paris, Frankfurt, p 383
- Cosserat E, Cosserat F (1909) *Theorie des Corps Déformables*. A. Hermann, Paris
- Eringen AC, Suhubi ES (1964) Non-linear theory of simple micro-elastic solids. I *Int J Eng Sci* 2:189–203
- Mindlin RD (1965) On the equations of elastic materials with microstructure. *Int J Solids Struct* 1:73
- Teisseyre R (2005) Asymmetric continuum mechanics: Deviations from elasticity and symmetry. *Acta Geophys Polon* 53:115–126
- Teisseyre R, Kozak JT (2003) Considerations on the seismic rotation effects. *Acta Geophys Polon* 51:243–256
- Kossecka E, DeWitt R (1977) Disclination kinematic. *Arch Mech* 29:633–651
- Teisseyre R (2001) Evolution, propagation and diffusion of dislocation fields. In: Teisseyre R, Majewski E (eds) *Earthquake thermodynamics and phase transformations in the earth's interior*. Academic Press (Vol. 76 of International Geophysical Series), San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, pp 167–198
- Kleman M (1980) The general theory of disclinations. In: Nabarro FRN (ed) *Dislocations of solids, vol 5. Other effects of dislocations: Disclinations*. North-Holland Publ. Comp., Amsterdam, pp 243–297

28. Nabarro FRN (1967) Theory of crystal dislocations. Clarendon Press, Oxford
29. Moriya T, Marumo T (1998) Design for rotation seismometers and their calibration. *Geophys Bull Hokkaido Univ* 61:99–106
30. Teisseyre R (1996) Shear band thermodynamical earthquake model. *Acta Geophys Polon* 44:219–236
31. Dietrich JHJ (1978) Preseismic fault slip and earthquakes prediction. *J Geophys Res* 83(B8):3940–3954
32. Teisseyre R, Majewski E (1990) Thermodynamics of line defects and earthquake processes. *Acta Geophys Polon* 38:355–373
33. Teisseyre R, Majewski E (1995) Earthquake thermodynamics. In: Teisseyre R (ed) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, pp 586–590
34. Teisseyre R, Majewski E (2001) Thermodynamics of line defects and earthquake thermodynamics. In: Teisseyre R, Majewski E (eds) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, pp 261–278
35. Majewski E, Teisseyre R (1997) Earthquake thermodynamics. *Tectonophysics* 227:219–233
36. Prigogine I (1979) Irreversibility and randomness. *Astron Phys Space Sci* 65:371–381
37. Kocks UF, Argon AS, Ashby MF (1975) Thermodynamics and kinetics of slip. Pergamon Press, Oxford, New York, p 288
38. Varotsos PA, Alexopoulos KD (1986) Thermodynamics of point defects and their relation with bulk properties. North-Holland, Amsterdam, New York, p 474
- Teisseyre R, Majewski E (2002) Physics of earthquakes. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) International handbook of earthquake and engineering seismology, Part A. Academic Press, Amsterdam, pp 229–235
- Teisseyre R, Takeo M, Majewski E (eds) (2006) Earthquake source asymmetry, structural media and rotation effects. Springer, Berlin, p 582
- Thoft-Christensen P (ed) (1974) Continuum mechanics aspects of geodynamics and rock fracture mechanics, NATO Advance Study Institutes Series C, vol 12. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, p 273

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning

BARRY HIRSHORN, STUART WEINSTEIN
NOAA/NWS/Pacific Tsunami Warning Center,
Ewa Beach, USA

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Tsunami Warning Center Operations](#)
[Seismic Methods](#)
[Earthquake Source Parameters](#)
[Future Directions](#)
[Acknowledgments](#)
[Bibliography](#)

Glossary

CMT centroid moment tensor The CMT represents the displacement of the Earth's crust that best reproduces the observed wave-field generated by an earthquake and gives the average location in time and space of the earthquake energy release. The seismic moment can be determined from the CMT.

Convolution Convolution is a type of integral transform combining two signals to form a third signal or output. It is the single most important technique in Digital Signal Processing. In the case of Seismology, the two signals can be e.g., the ground motion as a function of time and the response of the seismometer, and the output is the seismogram.

Deconvolution Does the reverse of convolution. In the case of Seismology, one uses deconvolution to remove the instrument response from the seismogram to recover the actual ground motion.

Books and Reviews

- Bridgman W (1950) The thermodynamics of plastic deformation and generalized entropy. *Rev Mod Phys* 22:56–63
- Eringen AC (1999) Microcontinuum field theories I: Foundations and solids. Springer, Berlin, p 325
- Eringen AC (2001) Microcontinuum field theories II: Fluent media. Springer, Berlin, p 340
- Drazin PG (1983) Solitons. Cambridge University Press, Cambridge
- Infeld E, Rowlands G (2000) Nonlinear waves, solitons and chaos. Cambridge University Press, Cambridge
- Muskhelishvili NT (1953) Some basic problems of the elasticity. Noordhoff, Groningen
- Newell A (1985) Solitons in mathematics and physics. Society for Industrial and Applied Mathematics, Philadelphia
- Prigogine I (1978) Thermodynamics of irreversible processes, 3rd edn. Wiley, New York
- Teisseyre R (1974) Symmetric micromorphic continuum: wave propagation, point source solutions and some applications to earthquake processes. In: Thoft-Christensen P (ed) Continuum mechanics aspects of geodynamics and rock fracture mechanics. D. Reidel Publ. Comp., Dordrecht-Holland/Boston, pp 201–244
- Teisseyre R (ed) (1995) Theory of earthquake premonitory and fracture processes. PWN, Warszawa, p 648
- Teisseyre R, Czechowski L, Leliwa-Kopystynski J (eds) (1993) Dynamics of the earth's interior. PWN, Warszawa, p 469
- Teisseyre R, Majewski E (eds) (2001) Earthquake thermodynamics and phase transformations in the earth's interior. Academic Press (Vol. 76 of International Geophysical Series), San Diego, New York, Boston, London, Sydney, Tokyo, p 670

Deep earthquake An earthquake characterized by a hypocenter located more than 100 km below the Earth's surface.

Hypocenter The point within the Earth where the earthquake rupture starts. The epicenter is the projection of the hypocenter onto the Earth's surface.

Local tsunami A tsunami that has little effect beyond 100 km from its source.

Magnitude m_B : The "broad-band" body-wave magnitude, generally based on measurements of the amplitude of P-waves with periods in the 2 to 20 s range.

M_S : The surface-wave magnitude. M_S is generally based on measurements of the amplitude of the surface (Love or Rayleigh) waves with periods of about 20 s. The US tsunami warning centers have applied a correction to the IASPEI formula that allows the estimation of M_S closer to the epicenter at a period of 20 s.

M_E : The "energy magnitude" scale, derived from velocity power spectra.

M_m : The mantle wave magnitude, based on the measurement of the amplitude of surface waves with periods of 50–400 s.

M_W : The moment magnitude or the "work magnitude" is based on the estimation of the scalar seismic moment, M_0 .

M_{wp} : The moment magnitude based on the initial long period P-waves.

M_L : The Local magnitude scale, based on the measurement of the maximum peak-to-peak amplitude observed on a Wood-Anderson seismogram, corrected for the decrease in amplitude with increasing epicentral distance. Generally based on the analysis of Sg, Lg or Rg surface waves oscillating with periods observed out to about 600 km. from the earthquake's epicenter.

pMag: A magnitude scale based on the average of the absolute values of the first three half cycles of the P-waves recorded at local distances.

Marogram A recording of sea-level variations obtained by tide gauges.

Regional tsunami A tsunami that has observable effects up to 1000 km from its source.

Seismic body waves Waves that propagate through the interior of an unbounded continuum. Primary waves (P-waves) are longitudinal body waves that shake the ground in a direction parallel with the direction of travel. Secondary body waves (S-waves) are shear waves that shake the ground in a direction perpendicular to the direction of travel. There are other types of arrivals (also known as phases) visible on seismographs corresponding to reflections of P- and S-waves from the earth's surface: The pP phase is a P-wave that

travels upwards from the hypocenter and reflects once off the surface and the PP phase is a P-wave that travels downwards from the epicenter and reflects once off of the surface. The definitions of the S-wave phases follow in the same manner.

Seismic moment The seismic moment M_0 , (expressed in units of force times distance; e. g. Newton-meters, or dyne-cm) is the moment of either couple of an equivalent double couple point source representation of the slip across the fault area during the earthquake. Mathematically, the Seismic Moment, $M_0 = \mu A d$, where μ denotes the shear rigidity, or resistance of the faulting material to shearing forces, A represents the area of the fault plane over which the slip occurs, and d represents the average co-seismic slip across A .

Seismic waves Elastic waves generated by movements of the earth's crust that propagate as radiated seismic energy, E_R .

Seismic surface waves Waves that propagate along the surface boundary of a medium, e. g. along the surface of the earth.

Shallow earthquake An earthquake characterized by a hypocenter located within 100 km of the Earth's surface.

Teletsunami A tsunami that has observable effects on coastlines more than 1000 km away from its source.

Tsunami A series of water waves generated by any rapid, large-scale disturbance of the sea. Most are generated by sea floor displacements from large undersea earthquakes, but they can also be caused by large submarine landslides, volcanic eruptions, calving of glaciers and even by meteorite impacts into the ocean.

Tsunami earthquake An earthquake that generates a much larger tsunami than expected given its seismic moment.

Tsunami warning system A tsunami warning system consists of a tsunami warning center such as the Pacific Tsunami Warning Center (PTWC), a formal response structure that includes Civil Defense authorities and Government Officials, and an education program that brings a minimum level of awareness and education to the coastal populations at risk.

Definition of the Subject

Tsunamis are among nature's most destructive natural hazards. Typically generated by large, underwater earthquakes near the Earth's surface, tsunamis can cross an ocean basin in a matter of hours. Although difficult to detect, and not dangerous while propagating in deep water, tsunamis can unleash awesome destructive power when

they reach coastal areas. With advance warning, populations dwelling in coastal areas can be alerted to move to higher ground and away from the coast saving many lives. Unfortunately, due to the lack of a tsunami warning system in the Indian Ocean, the Sumatra earthquake of Dec. 26, 2004 killed over 250 000 people with thousands of lives lost as far as away as East Africa many hours after the earthquake occurred. Had a tsunami warning system been in place many lives could have been saved [77].

As fast as tsunami waves are, seismic waves can travel at speeds more than 40 times greater. Because of this large disparity in speed, scientists rely on seismic methods to detect the possibility of tsunami generation and to warn coastal populations of an approaching tsunami well in advance of its arrival. The seismic P-wave for example, travels from Alaska to Hawaii in about 7 min, whereas a tsunami will take about 5½ h to travel the same distance. Although over 200 sea-level stations reporting in near-real time are operating in the Pacific it may take an hour or more, depending on the location of the epicenter, before the existence (or not) of an actual tsunami is confirmed. In other ocean basins where the density of sea-level instruments reporting data in near real-time is less, the delay in tsunami detection is correspondingly longer. In addition, global, regional, and local seismic networks, and the infrastructure needed to process the large amounts of seismic data that they record, are widespread. For these reasons, tsunami warning centers provide initial tsunami warnings to coastal populations based entirely on seismic data.

Introduction

A tsunami can be produced by any mechanism that causes a sudden displacement of the ocean's surface affecting a significant volume of water. Tsunamis can be generated by undersea earthquakes, landslides and volcanic explosions, calving of icebergs, and even meteorite impacts. However, the majority of tsunamis are generated by earthquakes. Not uncommon are earthquakes that trigger landslides so that both the displacement of the crust due to the earthquake, and the landslide, each contribute to the generation and size of the tsunami. Tsunamis are a devastating, natural, high fatality hazard [18]. In the absence of a proper tsunami warning system, a destructive tsunami will cause death and destruction as it encounters coastal areas while propagating across an ocean basin as it did in the case of the Sumatra tsunami of December 2004.

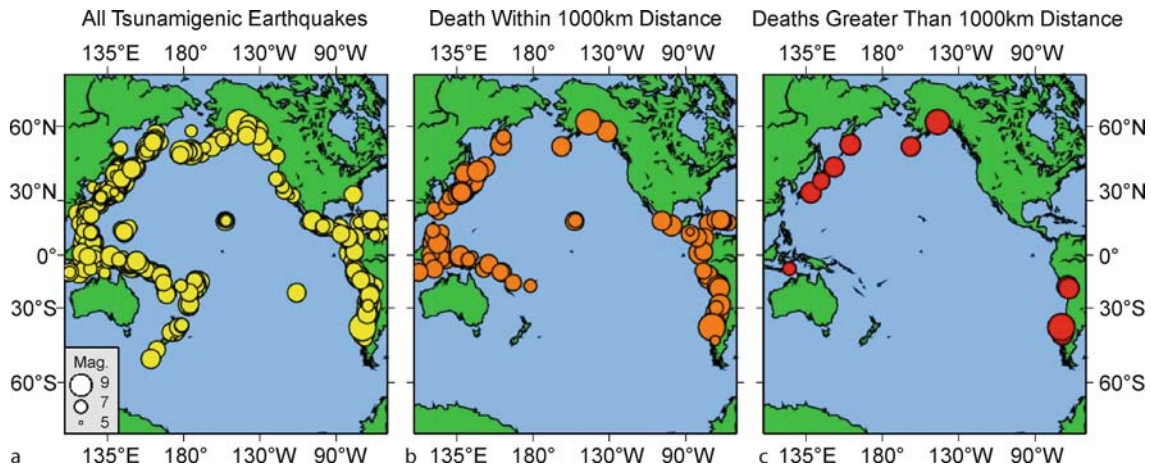
Although tsunamis propagate in deep water with speeds exceeding 900 km/h they are hard to detect in the open ocean. For instance, the first wave of the great Suma-

tra tsunami had a wave height of only one meter in deep water (> 500 m) [28], and a wavelength on the order of several hundred kilometers. Consequently, people aboard ocean vessels did not feel the accelerations caused by the Sumatra tsunami as it passed under them. However, as the speed, v , of a tsunami is governed by the simple relation $v = \sqrt{gh}$ where g is the acceleration of gravity (in m/s), and h is the thickness of the water column (in m), the tsunami will slow down as it propagates into shallow waters. At this point, the wave speed and wavelength decrease, causing the wave height to increase. Depending on the nature of the tsunami, and the shape and bathymetry of the coastal area, the tsunami wave height can be greatly amplified, thus magnifying its destructive power.

Because most tsunamis are generated by earthquakes, and seismic waves travel more than 40 times faster than tsunamis, the first indication that a tsunami may have been generated is the earthquake itself. Depending on the earthquake's location (undersea or inland), depth (shallow or deep) in the Earth's crust, and magnitude, a warning center may be required to issue an official message product. If the earthquake is a shallow, under sea earthquake, the severity of the message will depend upon the magnitude of the earthquake. The more rapidly and accurately the tsunami warning center can characterize the earthquake source, the quicker the initial evaluation of the tsunami-genic potential of the earthquake can be disseminated.

While some tsunamis are destructive, most are rather small, producing few if any casualties and little or no damage, although they are easily observable on marograms (Fig. 1). On the basis of how widespread their effects are, tsunamis can be classified as local (within 100 km of the epicenter), regional (up to 1000 km from the epicenter) or teletsunamis (greater than 1000 km from the epicenter). In the Pacific Basin there are warning centers designed to respond to tsunami threats on each of these scales.

The Richard H. Hagemeyer Pacific Tsunami Warning Center (PTWC) provides basin-wide warnings to the coastal areas of the Pacific basin. PTWC also functions as a local tsunami warning center for the Hawaii region. Other local warning centers include CPPT (French Polynesia Tsunami Warning Center) which is based in Tahiti, GFZ-Indonesia which is based in Jakarta provides local warnings for Indonesia, and Japan's JMA (Japan Meteorological Agency) which operates Japan's tsunami warning centers provides local tsunami warnings to Japan. Examples of regional warning centers include JMA which provides regional tsunami warnings to the Northwest Pacific and the West Coast and Alaska Tsunami Warning Center (WC/ATWC) which provides regional and local warnings to the US mainland coasts and the west coast of Canada.



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 1

Epicenters of tsunamigenic earthquakes occurring in the Pacific since 1 A.D. Of those earthquakes that do produce a tsunami a, most tsunamis cause no damage. Most events that cause casualties and/or damage do so within 1000 km of the epicenter b, leaving only a few great earthquake sources that generated tsunamis which caused casualties and/or damage more than a 1000 km from the epicenter, c. Data provided by the NOAA National Geophysical Data Center (NGDC), (www.ngdc.noaa.gov/hazard/tsu.html)

The tsunami warning centers themselves are not a complete tsunami warning system, they are simply the first of line of defense within the warning system. The warning system consists of three main components a) the tsunami warning centers, b) emergency management/civil defense authorities who receive tsunami warning center message products and c) a public in coastal areas that is educated in how to respond to tsunami emergencies. If any of these three components are lacking, the tsunami warning system can fail. Unfortunately, none of these components existed in the Indian Ocean at the time of the December 2004 Sumatra earthquake.

The greatest challenge for a tsunami warning system, particularly in the near field, is the slow (in terms of rupture speed) or “tsunami” earthquake. Tsunami earthquakes are so-called because they generate much larger than expected tsunamis given the size of the seismic moment of the earthquake [47]. In a well functioning tsunami warning system, residents in coastal areas are educated to immediately move inland and onto higher ground if they feel strong ground shaking and not wait for an official tsunami alert [24]. However, because a tsunami earthquake produces much less radiated high frequency body-wave energy than normal, even a large (in terms of moment magnitude) tsunami earthquake may not be strongly felt in the near field so that this strategy of having people self-evacuate upon feeling strong ground shaking will not work. This was, unfortunately, made dramatically clear by the Java earthquake of July 17, 2006. The tsunami generated by the Java earthquake killed ~ 500 people as many

residents in coastal areas near the earthquake did not feel strong shaking [83]. Tsunami warning centers need to be able to properly detect the occurrence of these tsunami earthquakes.

Tsunami Warning Center Operations

Tsunami warning center functions are much like those of a seismic observatory, i.e.: detecting, locating and characterizing the source of major earthquakes occurring around the world as fast as possible. Depending on the earthquake’s location (underwater vs. inland), depth below the surface, and magnitude, tsunami warning centers may issue an official message product to advise Civil Defense/Emergency Management authorities within the warning centers AOR (area of responsibility), of the occurrence of a large earthquake and its potential for generating a tsunami. The PTWC, located in Ewa Beach, Hawaii, provides advance warning of the generation of a destructive tsunami for the Pacific Ocean Basin, and on an interim basis, for the Indian, and Caribbean ocean basins.

After consultation with the member states of the Pacific Tsunami Warning System (PTWS), the PTWC has agreed to issue tsunami bulletins for the Pacific Ocean Basin according to the criteria shown in Fig. 2.

An Observatory (Earthquake) Message: The PTWC sends observatory messages to certain seismological observatories and organizations for any earthquake in or near the vicinity of the Pacific, Indian, or Caribbean ocean basins for seismic events when the magnitude is larger

Mw less than 6.5 (Mw: Moment Magnitude)	Earthquake Message Only
Mw 6.5 to 7.5	Tsunami Information Bulletin
Mw 7.6 to 7.8	Regional Tsunami Warning
Mw > 7.8	Expanding Warning / Watch
Confirmed Teletsunami	Pacific-Wide Warning

Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 2

PTWC Bulletin Criteria for the Pacific Basin

than about 5.5. This unofficial message contains only the earthquake's epicentral location, origin time, depth, magnitude, and a list of stations used in computing these parameters. These messages contain no evaluations regarding seismic or tsunami hazard, as the magnitude of the earthquake is far too small to have a significant tsunami generation potential.

A Tsunami Information Bulletin (TIB): The PTWC issues this message product for any earthquake in or near the vicinity of the Pacific Basin with a magnitude in the range $6.5 \leq M_W \leq 7.5$. A TIB states that a destructive tsunami is not expected outside the area of the epicenter. However, it does warn of the possibility of a destructive tsunami along coastlines within 100 km of the epicenter.

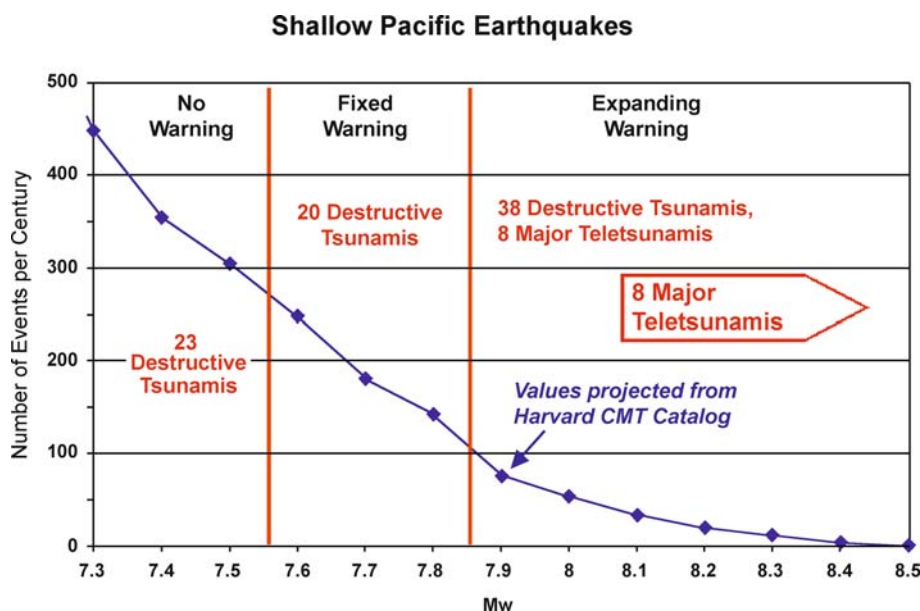
A Fixed Regional Warning Bulletin: The PTWC issues this message product for shallow underwater earthquakes (depth < 100 km) with a magnitude in the range $7.6 \leq M_W \leq 7.8$. This bulletin warns of the possibility of a regionally destructive tsunami within 1000 km of the epicenter. All regions within 1000 km of the epicenter are thus initially placed in a warning.

An Expanding Watch/Warning Bulletin: The PTWC issues this message for shallow underwater earthquakes with magnitude $M_W \geq 7.9$. This criteria is similar to the fixed watch/warning with the exception that this bulletin also warns of the possibility of a destructive tsunami traveling greater than 1000 km away from its source area. The use of the term "expanding" stems from the fact that the watch and warning regions expand across the Pacific as time progresses until the watch/warning is canceled. The extensions of the watch/warning area are referenced to the leading edge of the tsunami waves at the time the bul-

letin is issued. Areas within 3 to 6 h tsunami travel-time from the predicted current leading edge of the tsunami are placed in a watch. Areas within 3 h tsunami travel-time are placed in a warning. All other areas are placed in an advisory. Because of the expanding nature of the watch/warning, areas that were initially only placed in an advisory may eventually come to fall into the watch or warning region. If no potentially destructive tsunami is detected by sea-level stations, the watch/warning is canceled. On the other hand, if the data provided by sea-level stations provide evidence that a destructive tsunami is moving across the Pacific, the PTWC may upgrade to its most severe message, the Pacific-Wide Warning. A Pacific-Wide Warning is a tsunami warning for all coasts in the Pacific Basin. Before issuing a Pacific-Wide Warning the scientist on duty must confirm the presence of a potentially dangerous tsunami on sea-level instruments.

Figure 3 summarizes our response in retrospect, had these criteria been in place over the 20th century, after applying them to the earthquakes and tsunami that occurred during this period. The application of these criteria would have resulted in the issuance of a TIB, but no warning for 23 destructive locally generated tsunamis occurring over the last century in the Pacific Basin. At larger distances from the earthquake, the PTWC would have issued a *Fixed Regional Warning Bulletin* ahead of 20 destructive tsunamis generated in the last century, an *Expanding Watch/Warning Bulletin* for thirty-eight destructive tsunamis, as well as eight major Pacific Basin wide tsunami warnings within the same time period.

Coastlines close to the earthquake epicenter can experience tsunami waves within two to fifteen minutes after the earthquake; hence a local tsunami warning needs to be issued within a few minutes to be effective. This requires access to real-time data provided by a dense local network of seismic stations near the epicenter that allows both, the rapid location, and source characterization of the earthquake. In the case of the Hawaii region, the PTWC uses data from its own seismic network, and from the dense seismic network maintained by the USGS Hawaii Volcano Observatory (HVO) to rapidly detect Hawaii earthquakes. These data enabled PTWC to issue an information bulletin to the state of Hawaii, and to the Pacific Basin for the Kiholo Bay earthquake (M_W 6.7) within 3 min of the origin time of the earthquake [42]. However, without access to dense local seismic networks around the Pacific rim, the PTWC is unable to issue timely warnings for populations in the immediate vicinity of large earthquakes outside of Hawaii. As a result, the PTWC does not have the capability of functioning as a local warning center for areas outside of Hawaii.



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 3
Retroactive performance based on current bulletin criteria

The WC/ATWC of the US National Weather Service has access to dense local seismic networks on the US mainland, Puerto Rico, and Canada, and can therefore provide rapid warnings to the US and Canadian West Coast as well as to Puerto Rico and the Virgin Islands. Japan has a similar capability for its coastlines, and spurred on by the December 2004 Sumatra earthquake, several other nations such as Indonesia and New Zealand, for example, are rapidly developing and improving their seismic networks in an effort to improve their tsunami warning capabilities.

Seismic Methods

To rapidly detect, locate, and characterize the source of earthquakes occurring around the world, tsunami warning centers rely on the Global Seismic Network (GSN USGS/IRIS) which has many contributors in the US and worldwide. It is this unfettered access to real time seismic data supplied by a number of different networks that makes a basin-wide tsunami warning center possible. To rapidly deal with the threat posed by locally generated tsunamis to the state of Hawaii, PTWC processes seismic data from about 70 stations located in the Hawaiian Islands. The USGS HVO's dense network supplies most of this data. The US tsunami warning centers use the Earthworm software developed by the USGS to import and export seismic data [46].

PTWC duty scientists can receive automatic pages at any time, for any earthquake with magnitude M_W above ~ 5.5 . The system generating these pages combines Evan's and Allen's [23] teleseismic event detection algorithm, adapted for broadband data by Wither's [84], and Whitmore's [81] teleseismic picker and associator. In the Hawaiian Islands, the application of Hirshorn and Lindh's [40] algorithm notifies duty scientists for earthquakes with magnitudes larger than about 3.5 within 10 to 20 s of the earthquakes origin time. Other software automatically locates the event, and provides a first estimate of the earthquake's magnitude, and other source parameters in real time [3,39,40,45]. PTWC duty scientists then refine and supplement the software's automated real-time hypocenter location and magnitude estimates. Determining the earthquake's depth is particularly important as earthquakes occurring at depths greater than about 100 km generally don't cause tsunamis.

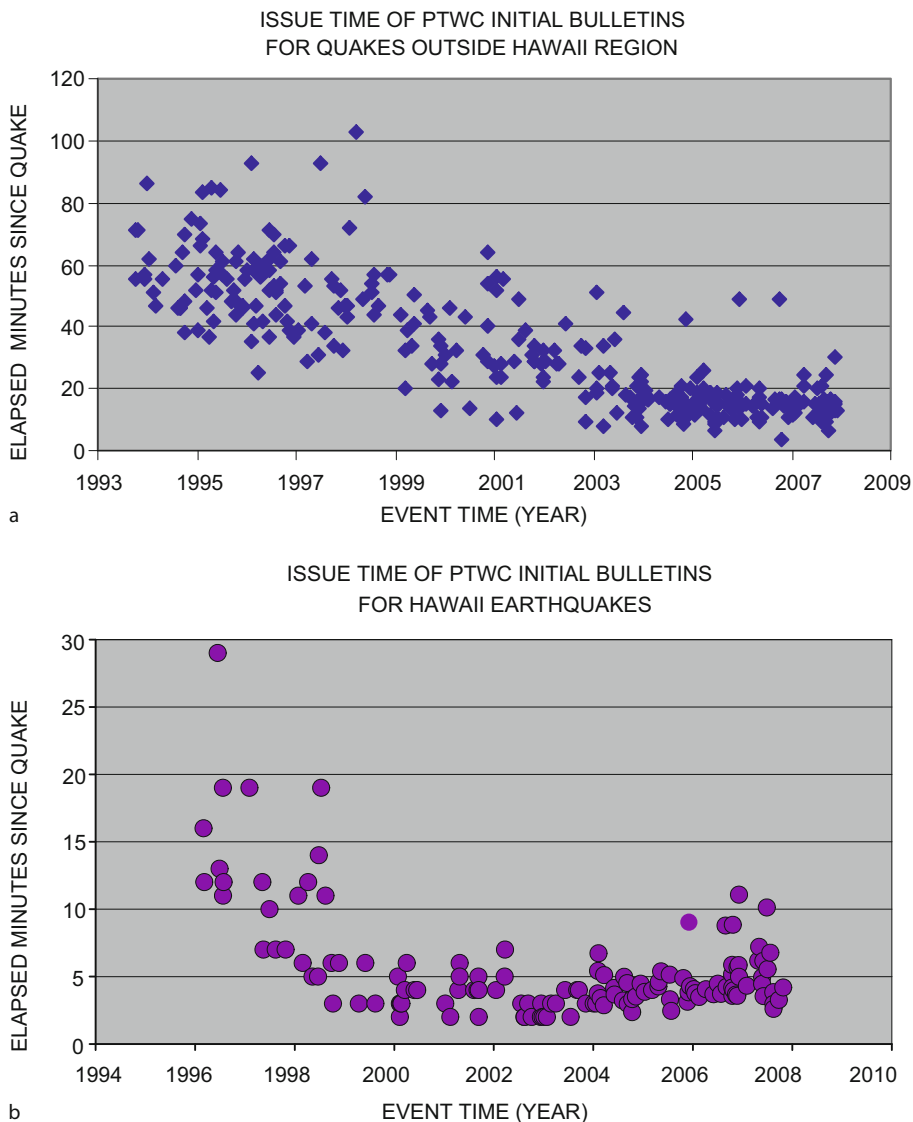
Earthquakes are located using P-wave arrival times recorded at a number of seismic instruments. As both the locations of seismic instruments and P-wave travel-times as a function of distance are well known, a process analogous to triangulation is used to locate the earthquake. The pickers and associators perform these functions automatically on a continuous basis.

While the depth of the earthquake can be estimated on the basis of P-times alone, a more robust result often requires the addition of depth phases such as pP which is

a P-wave that travels directly up to the earth's surface from the earthquake source and reflects once off of the Earth's surface before arriving at the seismometer. The duty scientists use pP arrival times to refine hypocentral depths of distant earthquakes (teleseisms). For earthquakes in Hawaii, observed at local distances, the S-wave arrival time would be useful for constraining earthquake depth. However, automatic picking of S-wave arrival times is not yet robust, and manual picking and incorporation of S-wave arrival times into the analyzes by a duty scientist would

take too long in the local earthquake case. In addition, as the deepest earthquakes in Hawaii have a hypocenter located about 50 km deep, (Fig. 6) which is comparable to the rupture length of an earthquake over about magnitude seven, the PTWC's local tsunami warning criteria are currently based on magnitude only.

Seismologists use a panoply of different magnitudes to characterize the seismic source. These different methods examine different parts of the seismic wave train, such as short and long period body waves (seismic waves that



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 4

Elapsed time from earthquake origin time to issuance of first official message product for a earthquakes outside the Hawaii Region and **b** for earthquakes within the Hawaii Region

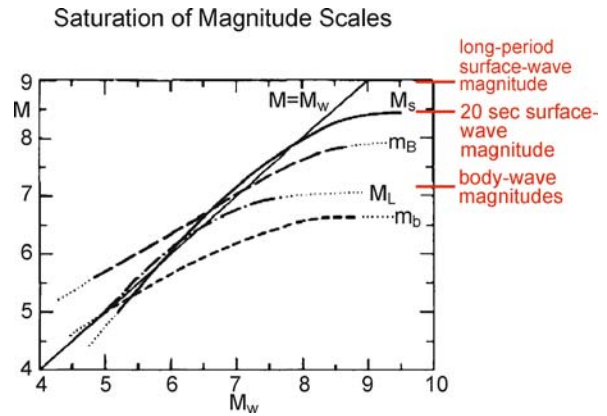
travel through the earth's interior like the P- and S-waves) and long period surface waves (slower seismic waves that are constrained to travel along the earth's surface). Most of these magnitude scales were developed to estimate the same quantity; the energy released by the earthquake as radiated seismic wave energy, E_R . Traditional magnitude measures such as M_L [66], and m_b (a shorter period variant of m_B Gutenberg [30,31] that examines high frequency body waves). The Gutenberg surface wave magnitude M_S [9], modified by Vanek [75] as well as the newer M_m Okal and Talandier [63], or mantle magnitude are derived from the surface waves. A relatively new and quick method, M_{wp} analyzes long period P-waves [73,74,82]. The M_{wp} magnitude is now the magnitude used in the decision process for deciding which if any official message product to issue, supplanting the M_S method which had been used for over 50 years. For large earthquakes, duty scientists also routinely estimate M_m , a very long period surface wave magnitude based on mantle waves with periods in the range 50–410 s [63]. The relationship between these magnitudes, each looking at different parts of the seismic wave spectrum of an earthquake, can be used to characterize the earthquake source [2,16,17].

When evaluating the tsunamigenic potential of an event, PTWC duty scientists also compute the quantity $\log_{10}(E_R/M_0)$, known as “Theta”, Θ , where M_0 is the seismic moment [1]. Newman and Okal [62] showed that Θ is anomalously small for tsunami earthquakes.

Since about the mid 1990's the two US Tsunami Warning Centers response times to potentially tsunamigenic teleseisms has decreased dramatically due to the much larger amounts of seismic data that they now receive, and to the switch from the slower M_S magnitude method to the faster M_{wp} moment magnitude method for their initial messages (Fig. 4a). For local events, using the real-time associator binder_agl [45], and the very fast $pMag$ scale [40], has brought the PTWC's response time down to less than 3 min (Fig. 4b).

Earthquake Source Parameters

A fundamental problem with traditional magnitude estimates such as M_L , m_b , and M_S , is that they are based on the amplitudes of relatively short period seismic waves with periods usually less than 3 s for m_b and M_L , and about 20 s for M_S . When the largest rupture dimension of the earthquake exceeds the wavelength of these seismic waves, which is about 50 km for the 20 s period surface waves used for M_S [48,49], these magnitude values will start to “saturate”. Saturation in this case means that these magnitude measures will underestimate the true size of the



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 5

Saturation of different classical magnitude scales with respect to non-saturating moment magnitude according to Kanamori [50]. Note that m_b refers to the original Gutenberg–Richter [32,33] body-wave magnitude scale based on amplitude measurements made on medium-period broadband instruments. It saturates at larger magnitudes when compared to the short-period based m_b .

earthquake when the periods of the amplitudes on which they are based are shorter than the corner period of the earthquake's seismic wave spectrum [2,16,17] (see Fig. 5). Another, equivalent explanation is that these magnitude methods, which look at waves with periods of a fraction of a second to a few tens of seconds, cannot sample enough of the energy released by an earthquake whose source duration (the length of time over which the rupture occurs) is many times larger than the periods used by these methods. As the earthquake becomes very large, one needs to examine longer period waves to avoid saturation.

K. Aki used a spectral representation to establish that earthquakes of varying size had spectra of similar shape, differing primarily in the low frequency amplitude, proportional to seismic moment, and the location of the “characteristic frequency” (corner frequency of the source spectrum) which he related to the characteristic length scale of an earthquake [2]. Subsequent studies by Brune [16,17], and Savage [68] also related the corner frequency to the dimensions of the fault plane.

To circumvent the saturation problem, Kanamori defined a new magnitude scale, the moment magnitude M_w , [48], in terms of a minimum estimate of the total co-seismic strain energy drop, W_0 , via Gutenberg and Richter's energy-magnitude relationship [33]. The M_w scale, more properly a magnitude that describes the total “work” required to rupture the fault, is computed from the seismic moment, M_0 , assuming 1) that the stress changes associated with large earthquakes are approximately con-

stant, and 2) that the stress release during an earthquake is about the same as the kinetic frictional stress during faulting. M_W and its agreement with the M_L and M_S magnitude scales in their unsaturated ranges was discussed by T.C. Hanks and H. Kanamori [34] while Kanamori [50] discusses additionally the average relationship of M_W with m_b and m_B , also in the range where these magnitudes saturate (see Fig. 5).

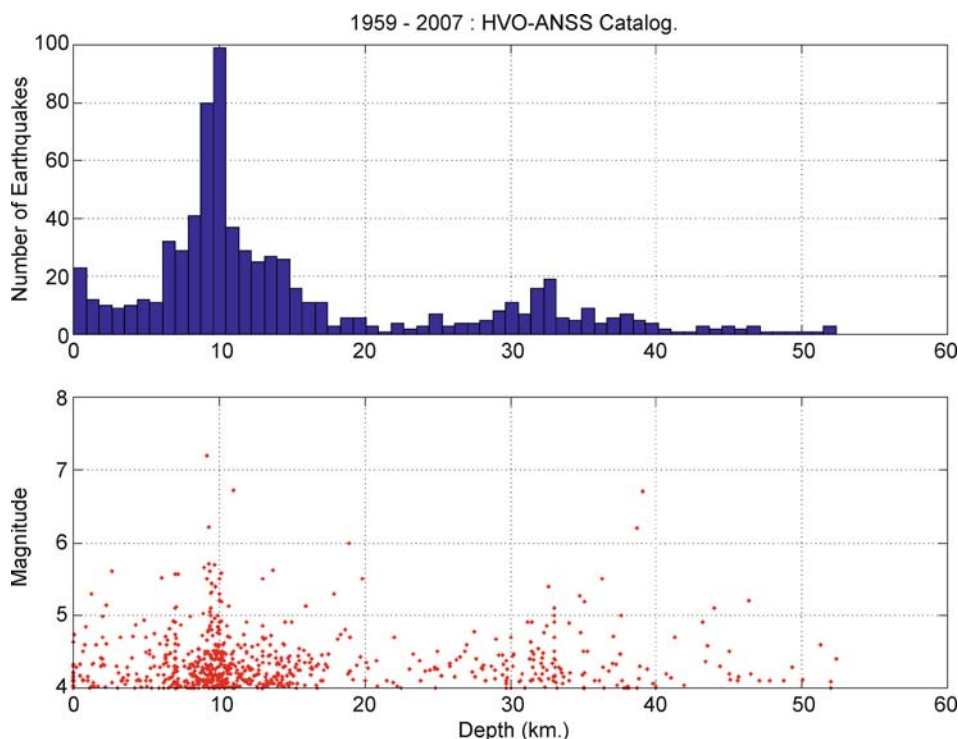
Traditional Amplitude Based Magnitudes at the PTWC

Local Earthquake Magnitude Methods Hirshorn and Lindh [40] developed a short period P-wave magnitude scale, called *pMag*, which is based on the average of the absolute values of the amplitudes of the first three half-cycles of the initial p-waves recorded, at local distances, on short period seismometers [40,45]. The *pMag* scale is based on the assumption that the decrease of locally recorded initial P-wave amplitudes with increasing hypocentral distance shares a common decay curve in a given geographic area, independent of the magnitude of the earthquake. Lindh and Hirshorn incorporated *pMag* into Carl John-

son's [46] local p-wave associator, binder_agl, enabling automatic pages, containing the hypocentral parameters and a the lower bound magnitude estimate provided by *pMag*, within about 10 to 20 s of an events origin time. At the PTWC, the System for Processing Local Earthquakes in Real Time (SPLERT) is based on this software.

PTWC also uses a very band-limited M_L scale, based on the maximum amplitudes measured on the horizontal components of short period seismograms recorded at local hypocentral distances from earthquakes that occur in Hawaii. These short period waves attenuate about 6 times less along the path between the hypocenter and recording stations for the "deeper" population of Hawaiian earthquakes (with hypocenters (Fig. 6) located between 20 and 50 km depth), then they do along the path from source to receiver for the "shallower" event population in the lower oceanic crust (about 10–20 km below the sea surface). For this reason, PTWC adds 0.8 magnitude units to the M_L values obtained for events with hypocentral depths ≥ 25 km.

Because of the bimodal depth distribution of Hawaii earthquakes (Fig. 6.) our M_L calculation requires only



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 6

This figure shows the bimodal depth distribution of Hawaii earthquakes of $M \geq 4$, taken from the USGS Hawaiian Volcano Observatory (HVO) ANSS Catalog. The *top figure* is a histogram of all events binned by hypocentral depth. The *bottom figure* shows the magnitude vs. hypocentral depth for each event

a good enough depth estimate to discriminate between these two populations.

Teleseismic Magnitude Methods PTWC's body wave magnitude method is called *bMag* and has similarities to the intermediate period broadband body wave m_B magnitude as defined by IASPEI. The IASPEI m_B [11,16] is based on Gutenberg's [30,31] and Gutenberg and Richter's m_B [32,33]. *bMag* uses a 90 s window of broadband vertical component seismogram starting 30 s prior to the arrival of the P-wave. This window is band-pass filtered between .3 s and 5 s. The largest amplitude and its period found in the 60 s after the first P-wave arrival are chosen for use in the magnitude formula. In PTWC's implementation, the formula used is the same as Gutenberg and Richter's [32,33] relation, adopted by IASPEI for m_B :

$$bMag = \log(A_{\max}/T_{\max}) + Q(\Delta, z),$$

where Δ is the epicentral distance ($15^\circ \leq \Delta \leq 90^\circ$), z is the hypocentral depth, A_{\max} is the maximum wave amplitude obtained from the band-pass filtered record, and T_{\max} is the period of the wave with the maximum amplitude. Gutenberg and Richter's [33] table of $Q(\Delta, z)$ is used to provide the distance and depth correction. The largest amplitude found in the 30 s prior to the P-wave arrival time is used as the basis for the signal to noise ratio. *bMag* differs from the IASPEI m_B in three respects,

1. m_B uses the largest amplitude wave in the P-wave coda up to the arrival of the PP phase,
2. m_B uses a slightly different distance range ($15^\circ \leq \Delta \leq 90^\circ$), and
3. for m_B , the seismogram is band-pass filtered using the band .3 s to 5 s.

bMag will saturate at lower magnitudes than M_S does, so it is of limited use for large earthquakes. However, *bMag*, is still useful for three main reasons

1. unlike M_S , *bMag* has a correction for the depth of the event's hypocenter,
2. it is useful for determining the magnitude of moderate earthquakes that occur as aftershocks of much larger earthquakes, e.g. when longer period energy is still present in the signal from an earlier, larger event, that can adversely affect magnitude methods based on longer periods, and
3. by comparison with magnitudes based on longer periods, such as M_w , it can also provide a way to detect slow or tsunami earthquakes.

In computing M_S (first proposed by Gutenberg [9] and later revised by Vanek et al. [75]) at the PTWC, we band-pass filter 14 min of the broadband velocity seismogram with a 7 s band, from 16 to 23 s, starting 3 min before the expected arrival time of the surface waves. We then apply the following equation, similar to the IASPEI [11,75] formula:

$$M_S = \log_{10}(A_{\max}/T_{\max}) + 1.66 \log_{10}(\Delta) + 3.3 + \text{correction}.$$

The *correction* term is 0 for epicentral distances, Δ , greater than 16° , and $0.53 - 0.033\Delta$ for Δ less than 16° . Note that the IASPEI implementation considers a much greater range in periods from 3 to 60 s for T_{\max} and has no need for the *correction* term. The *correction* term allows the US TWC's to compute the M_S magnitudes from stations as close as 600 km to the epicenter at a period of 20 s. This method is susceptible to saturation effects as the magnitude reaches the high 7's.

Although M_S is no longer used as the basis for issuing bulletins, it is still helpful in diagnosing deep earthquakes, and for comparing the amount of 20 s radiated energy to the amounts of radiated energy at other periods. Deep earthquakes do not excite large surface waves. Hence if $bMag > M_S$ the hypocenter is likely to be deep.

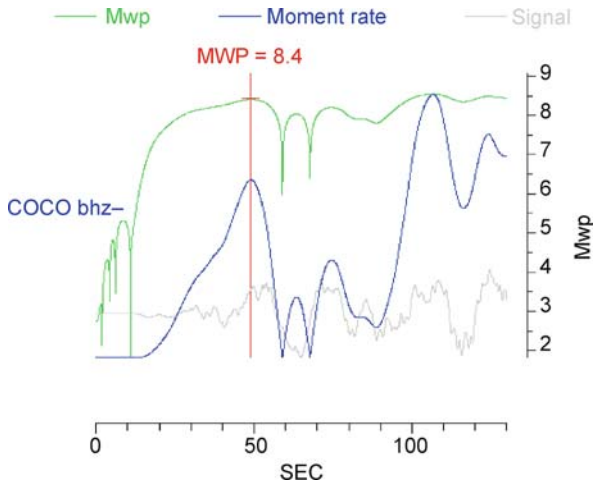
The M_{wp} Method

The broadband P-wave moment magnitude, M_{wp} , has replaced M_S as the magnitude upon which the US TWCs initial tsunami messages are based [73,74,82]. This is because M_{wp} uses P-waves, recorded at any epicentral distance, up to about 90° , when the observed initial P-waves are affected by refraction due to the earth's outer core. M_{wp} is obtained much quicker than M_S , which is based on the slower traveling surface waves, and because M_{wp} examines much longer period waves than the 20 s surface used by M_S making it less susceptible to the saturation effects discussed above. M_{wp} , as implemented at the PTWC, uses the first 120 s of the vertical component, broadband velocity seismogram, beginning at the P-wave arrival time (gray trace in Fig. 7).

The derivation of M_{wp} assumes that we can obtain the seismic moment, M_0 , from the far-field P- and/or pP-wave portion of the vertical broadband displacement waveform, $u_Z(x_r, t)$, using

$$M_0 = (4\pi\rho\alpha^3 r/F^P) \text{Max} \left| \int u_Z(x_r, t) dt \right|,$$

where ρ and α are the density and P-wave velocity averaged along the propagation path, r is the epicentral



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 7

The first 2 min of the broadband vertical velocity seismogram (gray) recorded by the GSN USGS/IRIS broadband station COCO, on Cocos Island, about 15 degrees south of the epicenter of the M_W 9.2 Sumatra Earthquake of Dec. 26, 2004. Note that this portion of the broadband velocity seismogram is not clipped. This instrument, a KS54000, has a flat frequency response to velocity to a period of about 350 s. The blue trace is the integrated displacement record (doubly integrated velocity), and the green trace is M_{wp} as a function of time

distance, and F^P is the earthquake source radiation pattern [73,74,82]. At the PTWC, we follow Tsuboi [73], approximating $\text{Max} \left| \int u_Z(x_r, t) dt \right|$ by the first significant or “big” peak in the absolute value of the integrated displacement record. We prefer to use velocity seismograms, $v(t)$, from STS-1 or KS54000 broadband seismometers as there is then no need to deconvolve the instrument response from the data. We simply scale the data by a gain factor, because we can assume that the instrument response function is flat in the frequency band of interest. For the STS-1, or the KS54000, which both have a flat velocity up to about 350 s, this works for all but the very largest or slowest earthquakes.

We first remove any pre-event offset from $v(t)$, ending before the P-waves from the earthquake arrive, integrate $v(t)$ twice, and then multiply the absolute value of each data point by $4\rho r\alpha^3$ to obtain $M_0(t)$ in N-m (the blue trace in Fig. 7). We then apply the standard IASPEI moment magnitude formula [34]:

$$M_W = (\log_{10} M_0 - 9.1)/1.5$$

to $M_0(t)$, to calculate $M_W(t)$ (the green trace in Fig. 7). To correct for the radiation pattern, F^P , we then add 0.2 to the average of the individual M_{wp} values, each obtained

at different azimuths and distances from the epicenter. This is because $\int (F^P)^2 d\Omega = 4/15$, where Ω is the azimuthal angle of the observation around the epicenter, and $\sqrt{4/15} = 0.52$. Therefore, we multiply the averaged M_0 by 2, which is equivalent to adding 0.2 to M_{wp} . Finally, we apply the Whitmore et al. [82] magnitude dependent correction, $M_{wp} = (M_{wp} - 1.3)/0.843$, to get a final value for M_{wp} .

Figure 8 compares these final M_{wp} values resulting from this procedure with the Harvard moment magnitude M_W estimated from their CMT [19] solutions.

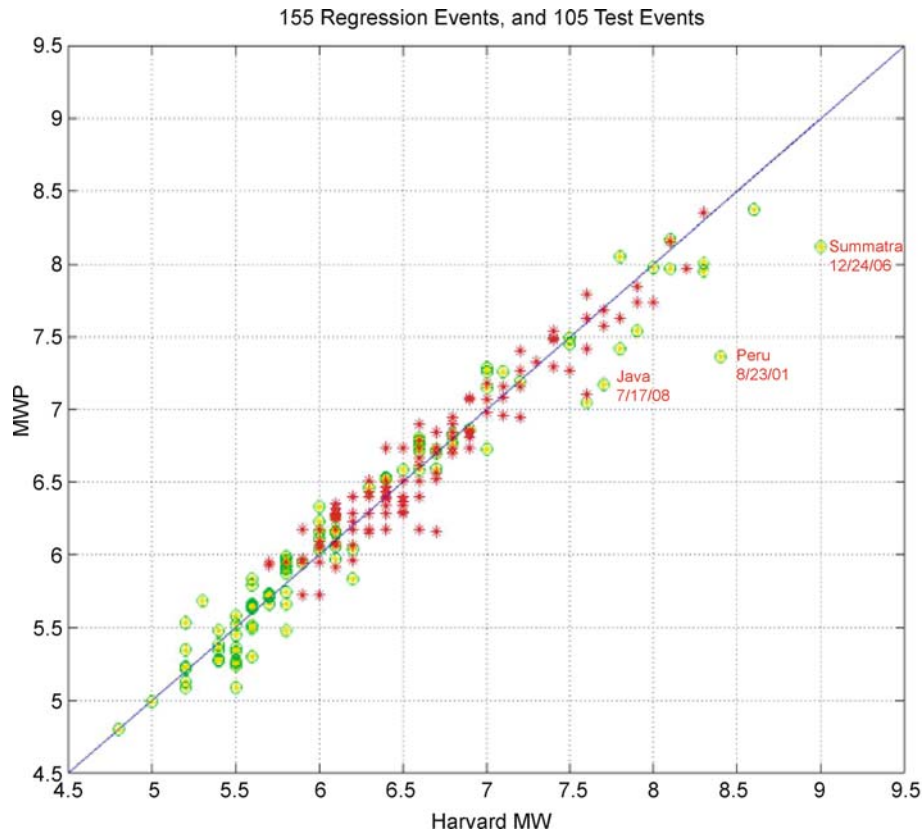
For some complex earthquakes, such as the M_W 8.4 [19] Peru earthquake of June 21, 2001, or the great M_W 9.2 [5,65,71] Sumatra earthquake of December 2004, M_{wp} (7.4 and 8.1, respectively) will underestimate M_W , when the first moment release is not the largest. In contrast, the PTWC’s final estimate of M_{wp} 8.4 for the M_W 8.6 [19] Nias event of March 28, 2005 was acceptable, as it was for 9 other earthquakes in the range $8.0 \leq M_W \leq 8.4$ (Fig. 8).

The PTWC also uses M_{wp} for large local earthquakes, occurring in the Hawaiian Islands [41,42]. M_{wp} is based on the far field formulation for P-wave displacements [73]. For earthquakes whose largest source dimension is small compared to the distances at which the P-waves are observed, this assumption is satisfied. Up to approximately M_W 7.5, M_{wp} calculated from these locally recorded, far field P-waves agrees well with the Harvard M_W values for the same events [41,42]. For example the PTWC calculated a value of M_{wp} 6.5 for the M_W 6.7 [19] Kiholo Bay event within two minutes of the initiation of rupture at the hypocenter [41,42].

The Mantle Magnitude (M_m) Method

Emile Okal and J. Talandier developed the M_m method in 1989 [63]. The mantle magnitude is related to the moment magnitude via the simple expression $M_W = M_m/1.5 + 2.6$. This work was inspired by the need to develop a magnitude method for tsunami warning centers that would not suffer the saturation problem of M_S [64]. Not only may M_S saturate as the magnitude becomes large (> 8) but slow earthquakes can cause M_S to be seriously deficient and $bMag$ even more so. Severely underestimating the magnitude of an earthquake can lead to a failure to warn. PTWC’s implementation of the M_m method is based on analyzing Rayleigh waves obtained on vertical component seismograms.

M_m being based on slow traveling long period surface waves, is available too late to be used in the decision process for issuing an initial bulletin. Notwithstanding, it does



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 8

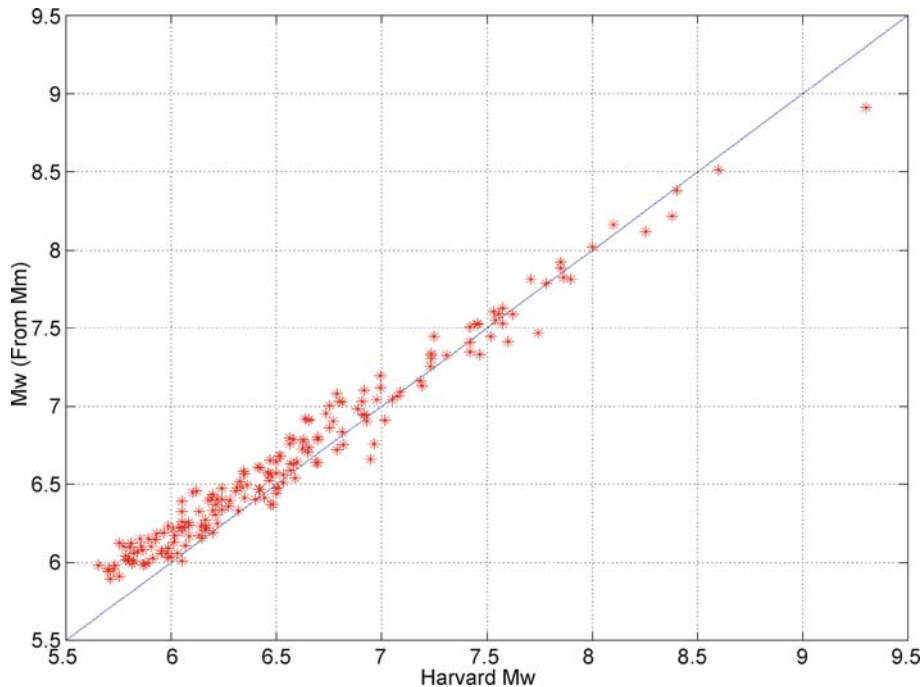
A Scatter plot of average M_{wp} versus the M_W [19] for a set of 260 earthquakes with magnitudes in the interval $4.8 \leq M_W \leq 9.2$ occurring from 1994 through October of 2007. Whitmore et al. [82] found that with the application of an empirical correction made to the results, satisfactory results could be obtained. Our linear, least squares fit to the 155 earthquakes (red stars) yielded a slope of 0.83, close to the Whitmore et al. [82] slope of 0.84. Since April of 2002, we have used this corrected relationship (green circles filled in yellow) to calculate M_{wp}

provide a useful check on the magnitude obtained from the M_{wp} method and if there is a discrepancy between M_{wp} and $M_W(M_m)$ on the order of 2–3 tenths or more in the 7+ magnitude range, the duty scientist may instead use the results of the M_m method in subsequent bulletins. The M_m method overcomes the limitation of saturation because it is a variable period magnitude. Multiple values of M_m are routinely computed for a number of fixed periods ranging from 50 to 270 s for each station. Because M_m may saturate at the smaller periods for great earthquakes, while at longer periods M_m will be unsaturated, Okal and Talandier's [63] procedure was to choose the largest M_m mitigating the effects of saturation.

M_m is more complicated than the other methods described here as it uses frequency domain deconvolution. This can cause problems due to deconvolution noise at low magnitudes, where the amplification of noise by the deconvolution process at long periods may result in spu-

rious magnitudes. Thus M_m works best with very long period broadband seismometers such as the KS54000's, KS36000's and STS-1's. While STS-2 seismometers tend to do well, however, the shorter period broadband seismometers tend to behave poorly at the longest periods [79]. Using the maximum M_m obtained for each station proved to be suboptimal due the heterogeneous distribution of instruments coupled with the total automation of the procedure at the PTWC. Weinstein and Okal [79] devised a sampling method that alleviates most of these difficulties in PTWC's implementation of M_m .

The December 2004 Sumatra earthquake showed that for earthquakes with an unusually long source duration (in this case ~ 600 s), even M_m at 270 s will saturate. Hence PTWC's M_m implementation will now automatically extend the period range to 410 s when the magnitude exceeds 8.0. At 410 s $M_W(M_m)$ is 8.9 [79] for the December 2004, Sumatra earthquake, still deficient, but a marked improve-



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 9
Scatter plot of $M_W(M_m)$ vs. Harvard/GCMT [19] M_W for over 200 Earthquakes

ment over the moment magnitude 8.5 obtained by PTWC and 8.2 obtained by the USGS (NEIC Fast Moment Tensor) on Dec. 26, 2004. M_m normally uses a 660 s window of the surface wave train, but when M_m exceeds 8.0, the window expands to 910 s. Given the mix of instruments and their distribution used at PTWC, and the effects of broadband deconvolution noise, Weinstein and Okal [79], found that the M_m method was not useful for $M_W < 6.0$.

Figure 9 compares $M_W(M_m)$ values obtained for more than 200 recent earthquakes with the respective M_W values of HRV/GCMT [19] for the same events. PTWC's implementation of M_m of still tends to over estimate M_W by about .15 magnitude units for $M_W < 7.0$.

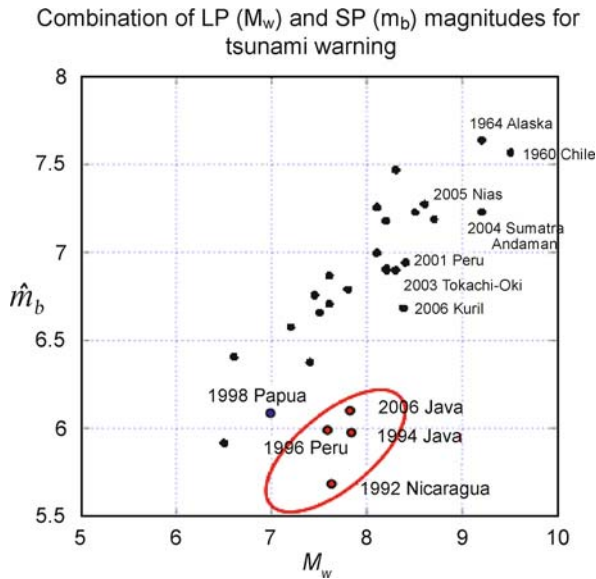
Rupture Slowness Estimation (Theta Program)

One way in which the occurrence of a tsunami earthquake may be indicated is if $bMag$ and/or M_S are significantly smaller than M_W obtained from the longer period P-waves, or longer period mantle waves. This is made clear in Fig. 10. Note the population of 4 tsunami earthquakes that fall well off the trend. The short period magnitudes may also simply be deficient simply due to the size of the earthquake. Measuring the "rupture slowness" of the earthquake can further aid the warning centers in deciding between the two possibilities. As can be determined from

Fig. 10, the body wave magnitude for July 2006, Java earthquake was deficient by nearly 1.5 magnitude units.

As mentioned earlier, a fundamental characteristic of a tsunami earthquake is the slowness of the rupture speed. Newman and Okal [62] showed that the log ratio of the radiated energy E_R [9,10], to the seismic moment M_0 , $\text{Log}_{10}(E_R/M_0)$ (also denoted by Theta, or " Θ ") is anomalously small for tsunami earthquakes. A number of factors can affect this ratio such as rupture velocity, stress-drop/apparent stress, fault plane geometry, maximum strain at rupturing, and directivity (bi-lateral vs. unilateral rupture). However, for shallow thrust, low stress-drop subduction zone earthquakes, unusually slow rupture velocity may have the largest influence on the value of Θ .

Newman and Okal [62] showed that for tsunami quakes, the value of Θ is usually about -6.0 or less. For an earthquake with a unilateral rupture with nominal speed (~ 3 km/s), theory suggests that Θ is about -4.9 [26,69,76]. Weinstein and Okal [79] extended the original dataset of Newman and Okal [62] by including an additional 118 earthquakes. The mean value of all Θ values is approximately -5.1 . However, when averaged by event, the distribution of Θ 's peaks precisely at -4.9 , in accordance with theoretical expectations. Given the standard deviation of 0.39 for all Θ 's (for the 118 earthquakes), Weinstein and Okal [79] found that values of Θ around



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 10

Comparison between short-period \hat{m}_b [43] and M_w for earthquakes with $M_w > 6$. Note the cluster of red dots representing tsunami earthquakes. This illustrates the diagnostic potential of short-period/long-period magnitude ratios to identify unusually slow earthquakes with high tsunami potential (Kanamori 2007, Talk at the PTWC in April of 2007)

— 6.0 or below are more than 2σ off the mean and hence clearly anomalous.

The PTWC uses broadband vertical component seismograms, obtained in the distance interval $25^\circ \leq \Delta \leq 90^\circ$, to compute Θ . A window of 75 s is used starting approximately 5 s before the P-wave arrival. This is done to insure that the first arrivals are not missed by the integration. This window is deconvolved with the instrument response and the radiated energy contained between .1 Hz and 2 Hz is computed.

In general it is thought that anomalously slow rupture speed is due to either low rigidity sediments in the fault or faulting through an accretionary prism [6,8,25,47,52,67]. In either case, the small shear rigidity associated with weak materials retards the rupture speed. As to why “slow” earthquakes produce more destructive than expected tsunamis, one can look at the well-known relation for moment magnitude:

$$M_0 = \mu A d,$$

where μ is the shear rigidity, A is the fault plane area and d is the average slip over the fault plane. Thus for two quakes with the same M_0 and all other properties equal except for μ and hence rupture speed, the slow quake requires

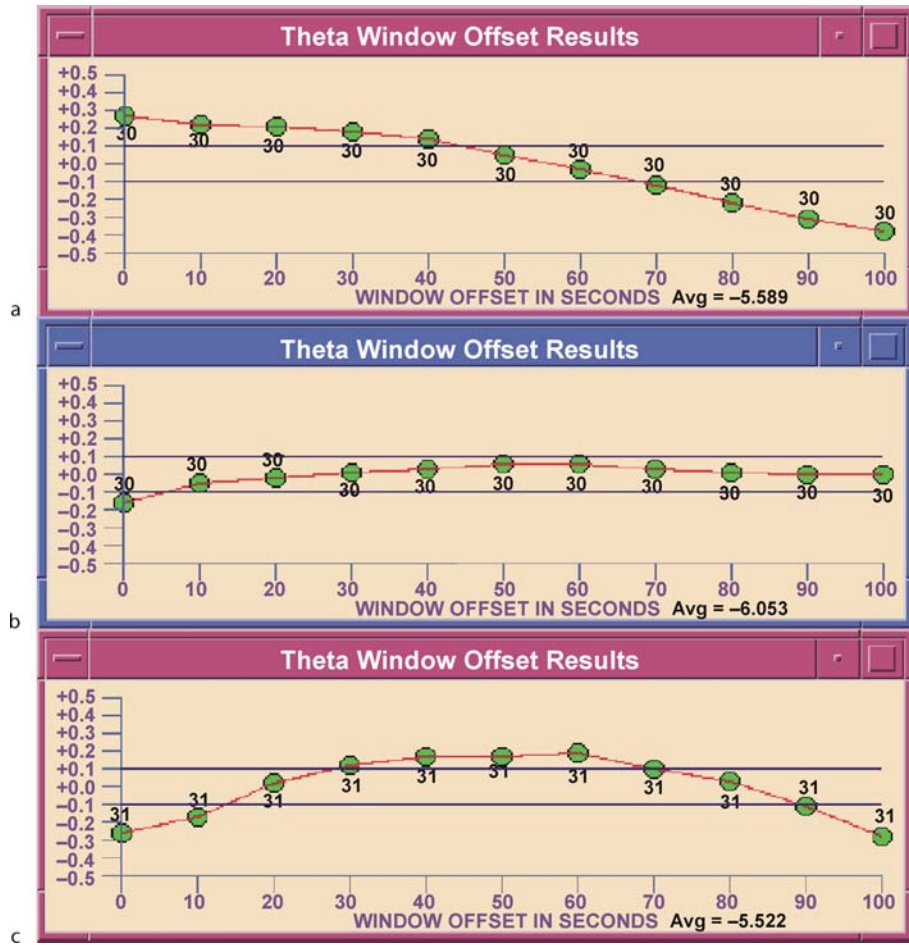
a correspondingly larger slip, d , in order to achieve the same moment as the earthquake with a nominal value of μ and hence normal rupture speed.

One problem with Θ , is that it can be misleading and occasionally yield false indications of rupture slowness. This was made apparent by the Peru earthquake of June 23, 2001. This earthquake began with a initial event that had a moment magnitude of approximately 7.4, followed almost 60 s later by a much larger event, which had a moment magnitude of almost 8.4. [7,27,55]. Due to the 60 s delay, the Θ computation used mainly P-wave coda from the first shock, and little if any energy from the main shock. As a result the PTWC initially obtained a Θ of -6.1 , using a moment based on M_m , making this earthquake appear very slow indeed. However, this result is spurious and was due to the complexity of the earthquake itself and not to actual slowness of the rupture.

Weinstein and Okal [79] found that by sliding the window over which Θ is computed forward in time, Θ would increase as the Θ window overlapped with the occurrence of the main event of the Peru 2001 earthquake. Indeed for a window offset of 70 s, Θ increases to -5.6 , which is a strong trend to slowness, but not a slow or tsunami quake. This was further borne out by the size of the tsunami, which while detected on sea-level instruments around the Pacific (more than 2 m peak-to-peak in Chile), was not destructive outside of Peru.

Weinstein and Okal [79] explored the windowing technique and found that in actuality, it was a more comprehensive method than the single determination of Θ (zero offset). Computing theta in a succession of windows separated in time by 10 s (each window spanning 70 s) up to 100 s post P-wave arrival yields a better method of detecting slowness (see Fig. 11). What Weinstein and Okal [79] found is that for true tsunami earthquakes, the variation of Θ with offset time was small, generally no more than 0.1 log units over the entire 100 s. It is this flat trend that is probably the best discriminant for tsunami earthquakes.

In effect, the curve resulting from the window-offset technique tells us something about the source duration of the earthquake. Gigantic earthquakes have long source durations, and slow earthquakes have anomalously long source durations for their seismic moment. Therefore Θ can be viewed as a measurement of how anomalous the source duration is in terms of whether the earthquake is anomalously slow, or simply anomalously large. It turns out that in the case of the Sumatra earthquake of December 2004, Θ has little variation, even when the integration window is increased to 200 s and the offset carried out to 300 s. The magnitude of Θ based on PTWC's $M_w(M_m)$ of



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 11

The variation of Θ with offset for a a "normal" earthquake, b a "slow earthquake" (Java, 2006), and c a complex earthquake (Peru, 2008) respectively. In these plots Θ is de-meaned (the mean is found on the *bottom right* of the plot in black) and the number next to the dots indicates how many stations were used in computing that value. These plots are taken from PTWC's operational software

8.5 was ~ -5.6 , a trend to slowness, but not slow. Using the M_W based on normal mode studies, Θ is ~ -6.1 (with a 200 s integration window!) and discussion continues to the current day as to whether or not the Sumatra earthquake of 2004 was slow, simply had aspects of a tsunami earthquake, or none at all [5,20,53,57,70].

Future Directions

Given the availability of high quality broadband seismic data, the tsunami warning centers can determine basic earthquake source parameters rapidly. However, the source characterization at the warning centers has rested largely on scalar measures of earthquake magnitude and slowness. The reasons for this are historical and practical. The warning centers have not always received the

quantity of seismic data they do now, and in the interest of speed, the calculation of scalar measures can be accomplished with the data at hand in a small amount of time. One issue the PTWC faced during the 2004 Sumatra earthquake was that no near real-time magnitude method existed at the time that would correctly estimate the size of the Sumatra earthquake. Since then, new techniques have been developed to determine the magnitude of great earthquakes. Among these are techniques based on P-wave broadband signals [12,13,15,22,35,36,59,60] and the W-phase [21,52,55,58].

Hara [35,36] and Lomax et al. [59], both use techniques that involve estimating the source duration from the P-wave coda. This estimate is obtained by applying a high pass filter to the velocity seismogram, squaring the result and smoothing it. This procedure results in a rela-

tively smooth curve or envelope function that tracks the variation in the velocity-squared time-series. The source duration estimate is then obtained by measuring the time from the beginning of the P-wave to the point when the envelope function falls below a certain percentage of its maximum value. However, these studies also show that the use of source duration alone is not a completely satisfactory basis for a moment magnitude estimate.

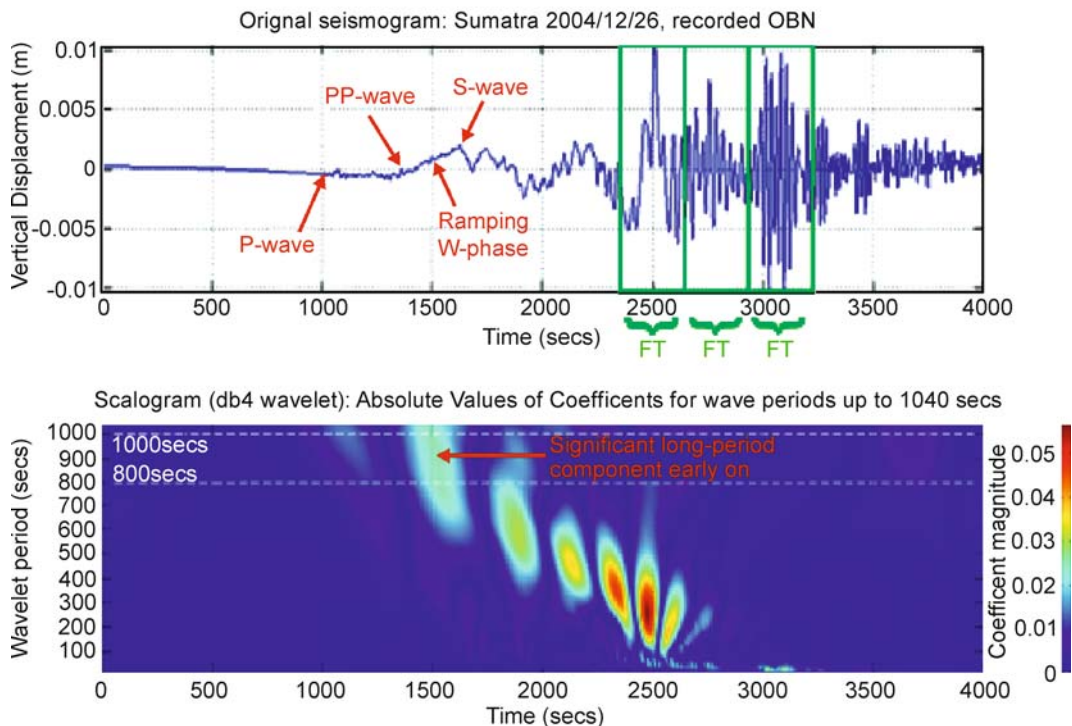
Lomax et al. [59] also measures the radiated energy of the P-wave in the interval between the P- and S-arrivals, and uses both the radiated energy and source duration estimate to formulate a magnitude scale based on the relation $M_0 \approx E^{1/2} \cdot T^{3/2}$ [76] where E is the high frequency radiated energy and T is the source duration estimate. Hara [35] uses the estimated source duration and the maximum displacement measured in the interval of the estimated source duration from a number of earthquakes to construct an empirical formula for the magni-

tude. Hara [36] showed that this technique also works well for tsunami earthquakes.

Lomax et al. [60] has derived a duration-amplitude procedure for determination of a moment magnitude, M_{wpd} , for large earthquakes within 20 min of the event origin time using teleseismic P-wave recordings. Their procedure determines apparent source durations, T_0 , from high frequency, P-wave records, and estimates seismic moments via integration of broadband seismograms over the interval t_p to $t_p + T_0$, where t_p is the P-wave arrival time. The characteristics of this method make it an extension of M_{wp} .

De Kool et al. [22] present a variation of the M_{wp} method which estimates the asymptotic behavior of the integrated displacement seismogram caused by the P-wave arrivals. Their results for M_{wp} show less scatter than do the PTWC's M_{wp} values, described above. In addition, they have automated their method.

Wavelet Scalogram of the Sumatra-Andaman Earthquake



Earthquake Source Parameters, Rapid Estimates for Tsunami Warning, Figure 12

Top: (Fig. 2a and b from [58]). Displacement seismogram of the 2004 Sumatra-Andaman earthquake recorded at OBN. Bottom: Scalogram of top seismogram. A diagram which displays the wavelet scale as a function of time is called a "scalogram". Bottom figure shows the scalogram for the 2004 event. Color intensity at any point in the picture corresponds to the coefficient magnitude of a wavelet with a particular period at a particular point of the time series. The y-axis has been translated from wavelet scale into corresponding wavelet time period. The long-period component arrived at about 1500 s. The wavelet transform can simultaneously achieve: (1) Accurate frequency representation for low frequencies, and (2) Good time resolution for high frequencies [58]

The method of Bormann and Wylegalla [13] and Bormann and Saul [14] calculate what they refer to as a cumulative body wave magnitude. They do this by summing up all of the peak velocity amplitudes for all pulses, which represent the rupturing of sub-faults, over the P-wave coda. For the December 2004 Sumatra earthquake they obtained an M_W of 9.3 in agreement with the estimate of Stein and Okal [71].

Since these methods are based on analyzing the P-waves in the P-S interval, they provide accurate moment magnitude estimates for great earthquakes within 20–25 min of the earthquake origin time. Therefore these estimates will come before estimates obtained by other methods like M_m .

The W-phase is a long period, up to 1000 s, wave that arrives before the S-wave (Fig. 12). It can be interpreted as a superposition of the fundamental, 1st, 2nd, and 3rd overtones of spheroidal modes or Rayleigh waves and has a group velocity of 8 km/s at 1000 s period, and 8.6 km/s at 100 s period [51,54]. Kanamori [54] has devised a magnitude scale based on the amplitude of the W-phase observed on deconvolved displacement records.

In addition to size and source duration, the warning centers are interested in more detailed properties of the source than can be obtained from the scalar measures we have just discussed, such as direction of rupture and the distribution of slip along the fault. This information is important as it can be used by tsunami wave-height forecast models to better their predictions.

In the near future, the tsunami warning centers will incorporate the results of centroid moment tensors and finite fault modeling. Finite Fault modeling involves the inversion of seismic waveforms to recover more detailed information about the source process including the slip distribution, rupture propagation speed and moment release history [7,27,37,38,62,77]. Weinstein and Lundgren [80] explored the potential of a simple teleseismic P-wave inverse method for the rupture history of an earthquake for use in a tsunami warning center context. The calculations proceed quickly enough that a slip distribution may be available just a few minutes after a suitable set of P-waveforms are obtained. Hence finite fault modeling results can be used in tsunami wave height forecast models to provide a timely initial estimate of tsunami wave heights.

The warning centers are also actively investigating the use of seismic arrays. Seismic arrays can be used to determine the direction along which the P-waves have propagated to the array. As the rupture propagates, this direction will change. By analyzing the seismic array data, this change in direction can be measured (by computing

back azimuths) and the history of energy/moment release as well as the extent and direction of the rupture propagations can be determined [44,56].

Acknowledgments

The authors greatly appreciate very thorough reviews by Kenji Satake, Anthony Lomax, Peter Bormann and our colleague Victor Sardina. Their comments greatly improved this manuscript. We also thank Paula Dunbar of the National Geophysical Data Center for helping us obtain the data we needed and Nathan Becker's GMT artistry for Fig. 1. We also thank the PTWC for use of its facilities in preparing this manuscript.

Bibliography

Primary Literature

1. Aki K (1966) Generation and propagation of G waves from the Niigata earthquake of June 16, 1964, Part 2. Estimation of earthquake moment, from the G wave spectrum. *Bull Earthquake Res Inst Tokyo Univ* 44:73–88
2. Aki K (1967) Scaling law of seismic spectrum. *J Geophys Res* 72:1217–1231
3. Allen RV (1978) Automatic earthquake recognition and timing from single traces. *Bull Seism Soc Am* 68:1521–1532
4. Allen RV (1982) Automatic phase pickers: their present use and future prospects. *Bull Seism Soc Am* 72:225–242
5. Ammon CJ, Ji C, Thio HK, Robinson D, Sidao N, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmlinger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman Earthquake. *Science* 6:113–1139
6. Bilek SL, Lay T (1999) Rigidity variations with depth along interplate megathrust faults in subduction zones. *Nature* 400:443–446
7. Bilek SL, Ruff LJ (2002) Analysis of the 23 June 2001 $M_W = 8.4$ Peru underthrusting earthquake and its aftershocks. *Geophys Res Lett* 8:21–1–21–4
8. Bilek SL, Lay T, Ruff LJ (2004) Radiated seismic energy and earthquake source duration variations from teleseismic source time functions for shallow subduction zone thrust earthquakes. *J Geophys Res* 109:B09308
9. Boatwright J, Choy GL (1986) Teleseismic estimates of the energy radiated by shallow earthquakes. *J Geophys Res* 91:2095–2112
10. Boatwright J, Choy GL, Seekins LC (2002) Regional estimates of radiated seismic energy. *Bull Seism Soc Am* 92:1241–1255
11. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) *IASPEI New Manual Seismological Observatory Practice*, vol 1, Chapter 3. GeoForschungsZentrum Potsdam, pp 1–94. (This can also go as in the text cited review paper)
12. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *Eos Trans AGU* 86(46):464
13. Bormann P, Wylegalla K, Saul J (2006) Broadband body-wave magnitudes m_B and m_{BC} for quick reliable estimation of the size of great earthquakes. *USGS Tsunami Sources Work-*

- shop 2006, poster, http://spring.msi.umn.edu/USGS/Posters/Bormann_et_al_poster.pdf
14. Bormann P, Saul J (2008) Earthquake magnitude. In: Meyers A (ed) *Encyclopedia of complexity and systems science*. Springer, Heidelberg
 15. Bormann P, Saul J (2008) The new IASPEI standard broadband magnitude m_B . *Seism Res Lett* 79:698–705
 16. Brune J (1970) Tectonic stress and seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
 17. Brune J (1971) Tectonic stress and seismic shear waves from earthquakes; Correction. *J Geophys Res* 76:5002
 18. Bryant E (2001) Distribution and fatalities. In: Bryant E (ed) *Tsunami: The underrated hazard*. School of Science and the Environment, Coventry University, Coventry. Cambridge University Press, Cambridge, pp 15–24
 19. Centroid Moment Tensor (2008) Catalog. <http://www.globalcmt.org>, accessed June 2008
 20. Choy GL, Boatwright J (2007) The energy radiated by the 26 December 2004 Sumatra–Andaman earthquake estimated from 10-minute P-wave windows. *Bull Seism Soc Am* 97: S18–S24
 21. Cummins PR (1997) Earthquake near field and W phase observations at teleseismic distances. *Geophys Res Lett* 24: 2857–2860
 22. De Kool M, Jepsen D, Purss, Matthew (2007) Rapid moment estimation of large earthquakes using a variation of the Mwp method. In press
 23. Evans JR, Allen S (1983) A teleseism-specific detection algorithm for single short-period traces. *Bull Seism Soc Am* 73:1173–1186
 24. Fryer G, Hirshorn B, McCreery S, Cessaro RK, Weinstein S (2005) Tsunami warning in the near field: The approach in Hawaii. *EOS* 86:S44B-04
 25. Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
 26. Geller RJ, Kanamori H (1977) Magnitudes of great shallow earthquakes from 1904 to 1952. *Bull Seismol Soc Am* 67: 587–598
 27. Giovanni MK, Beck SL, Wagner L (2002) The June 23, 2001 Peru earthquake and the southern Peru subduction zone. *Geophys Res Lett* 29:14-1–14-4
 28. Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. *Eos Trans AGU* 86(4):37–38
 29. Gutenberg B (1945) Amplitudes of surface waves and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:3–12
 30. Gutenberg B (1945) Amplitudes of P, PP, and S, and magnitudes of shallow earthquakes. *Bull Seism Soc Am* 35:57–69
 31. Gutenberg B (1945) Magnitude determinations of deep-focus earthquakes. *Bull Seism Soc Am* 35:117–130
 32. Gutenberg B, Richter CF (1956) Earthquake magnitude, intensity, energy and acceleration. *Bull Seism Soc Am* 46:105–145
 33. Gutenberg B, Richter CF (1956) Magnitude and energy of earthquakes. *Annali di Geofisica* 9:1–15
 34. Hanks T, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
 35. Hara T (2007) Measurement of the duration of high-frequency radiation and its application to determination of the magnitudes of large shallow earthquakes. *Earth Planets Space* 59:227–231
 36. Hara T (2007) Magnitude determination using duration of high frequency energy radiation and displacement amplitude: application to tsunami earthquakes. *Earth Planets Space* 59: 561–565
 37. Hartzell S, Heaton T (1986) Rupture history of the 1984 Morgan Hill, California, Earthquake from the inversion of strong motion records. *Bull Seism Soc Am* 76:649–674
 38. Hartzell S, Mendoza C (1991) Application of an iterative least-squares waveform inversion of strong-motion and teleseismic records to the 1978 Tabas, Iran earthquake. *Bull Seism Soc Am* 81:1:305–331
 39. Hirshorn B, Lindh A, Allen R (1987) Real Time Signal Duration Magnitudes from Low-gain Short Period Seismometers. *USGS OFR* 87:630
 40. Hirshorn B, Lindh G, Allen RV, Johnson C (1993) Real time magnitude estimation for a prototype early warning system (EWS) from the P-wave, and for earthquake hazards monitoring from the coda envelope. *Seis Res Lett* 64:48
 41. Hirshorn B (2004) Moment magnitudes from the initial P-wave for local tsunami warnings. *Seism Res Lett* 74:272–273
 42. Hirshorn B (2007) The Pacific Tsunami Warning Center Response to the Mw6.7 Kiholo Bay Earthquake and Lessons for the Future. *Seism Res Lett* 78:299
 43. Houston H, Kanamori H (1986) Source spectra of great earthquakes, teleseismic constraints on rupture process and strong motion. *Bull Seism Soc Am* 76:19–42
 44. Ishii M, Shearer PM, Houston H, Vidale JE (2005) Extent, duration and speed of the 2004 Sumatra–Andaman earthquake imaged by the Hi-Net array. *Nature* 435:933–936
 45. Johnson CE, Lindh A, Hirshorn B (1994) Robust regional phase association. *US Geol Surv Open-File Rept* 94:621
 46. Johnson CE, Bittenbinder A, Bogaert B, Dietz L, Kohler W (1995) Earthworm: a flexible approach to seismic network processing. *IRIS Newslett* XIV 2:1–4
 47. Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Inter* 6:246–259
 48. Kanamori H (1977) The energy release in great earthquakes: *J Geophys Res* 82:2981–2987
 49. Kanamori H (1978) Quantification of earthquakes. *Nature* 271:411–414
 50. Kanamori H (1983) Magnitude scale and quantification of earthquakes. *Tectonophysics* 93:185–199
 51. Kanamori H (1993) W Phase. *Geophys Res Lett* 20:1691–1694
 52. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediment. *Nature* 361:714–715
 53. Kanamori H (2006) Seismological Aspects of the December 2004 great Sumatra Andaman Earthquake. *Earthquake Spectra* 22:S1–S12
 54. Kanamori H, Rivera L (2008) Source inversion of W phase – Speeding up tsunami warning. *Geophys J Int* 175:222–238
 55. Kikuchi M, Yamanaka Y (2001) EIC Seismological Note Number 105, www.eic.eri-u-tokyo.ac.jp/EIC/EIC_news/105E.html
 56. Krüger F, Ohrnberger M (2005) Tracking the rupture of the Mw9.3 Sumatra earthquake over 1150 km at teleseismic distance. *Nature* 435:937–939
 57. Lay T, Kanamori H, Ammon C et al (2005) The Great Sumatra – Andaman Earthquake of 26 December 2004. *Science* 308:1127–1133
 58. Lockwood OG, Kanamori H (2006) Wavelet analysis of the seismograms of the 2004 Sumatra–Andaman earthquake and its application to tsunami early warning. *Geochem Geophys Geosyst* 7, Q09013, doi:10.1029/2006GC001272

59. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
60. Lomax A, Michelini A (2008) Mwpd: A duration-amplitude procedure for rapid determination of earthquake magnitude and tsunamigenic potential from P waveforms. *Geophys J Int* (accepted, in press)
61. Mendoza C (1996) Rapid derivation of rupture history for large earthquakes. *Seismol Res Lett* 67:19–26
62. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The E/M0 discriminant for tsunami earthquakes. *J Geophys Res* 103:26885–26898
63. Okal EA, Talandier J (1989) Mm: a variable-period mantle magnitude. *J Geophys Res* 94:4169–4193
64. Okal EA (1992) Use of mantle magnitude M_m for reassessment of the moment of historical earthquakes-I: Shallow events. *PA-GEOPH* 139:17–57
65. Park J, Song TA, Tromp J, Okal E, Stein S, Roullet G, Clevede E, Laske G, Kanamori H, Davis P, Berger J, Braitenberg C, Camp MV, Xiang'e L, Heping S, Houze X, Rosat S (2005) Earth's free oscillations excited by the 26 December 2004 Sumatra–Andaman earthquake. *Science* 308:1139–1144
66. Richter CF (1935) An instrumental earthquake magnitude scale. *Bull Seism Soc Am* 25:1–32
67. Satake K (1994) Mechanism of the 1992 Nicaragua tsunami earthquake. *Geophys Res Lett* 21:2519–2522
68. Savage JC (1972) Relation of corner frequency to fault dimensions. *J Geophys Res* 77:3788–3795
69. Scholz C (1982) Scaling laws for large earthquakes: Consequences for physical models. *Bull Seism Soc Am* 72:1–14
70. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? *Bull Seism Soc Am* 97:S296–S306
71. Stein S, Okal EA (2007) Ultralong period seismic study of the December 2004 Indian Ocean earthquake and implications for regional tectonics and the subduction process. *Bull Seism Soc Am* 97:279–295
72. Indian Ocean Earthquake and implications for regional tectonics and the subduction process. *Bull Seism Soc Am* 97: S279–S295
73. Tsuboi SK, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband P waveforms. *Bull Seism Soc Am* 83:606–613
74. Tsuboi S, Whitmore PM, Sokolowski TJ (1999) Application of Mwp to deep and teleseismic earthquakes. *Bull Seism Soc Am* 89:1345–1351
75. Vanek J, Zatopek A, Karnik V, Kondorskaya N, Riznichenko Y, Savarenski S, Solovov S, Shebalin N (1962) Standardization of magnitude scales. *Izv Acad Sci USSR Geophys Ser*, pp 108–111 (English translation)
76. Vassiliou MS, Kanamori H (1982) The energy release in earthquakes. *Bull Seism Soc Am* 72:371–387
77. Wald DJ, Helmberger DV, Hartzell S (1990) Rupture process of the 1987 Superstition Hills earthquake from the inversion of strong-motion data. *Bull Seism Soc Am* 80:1079–1098
78. Weinstein SA, McCreery C, Hirshorn B, Whitmore P (2005) Comment on “A strategy to rapidly determine the magnitude of great earthquakes” by Menke W, Levin V. *Eos* 86:263
79. Weinstein SA, Okal EA (2005) The mantle magnitude M_m and the slowness parameter Θ : Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 95:779–799
80. Weinstein SA, Lundgren PL (2008) Finite fault modeling in a tsunami warning center context. In: Tiampo KF, Weatherley DK, Weinstein SA (eds) *Earthquakes: Simulations, sources and tsunamis*. Birkhauser, Basel
81. Whitmore PM, Sokolowski TJ (2002) Automatic earthquake processing developments at the US West Coast/Alaska tsunami warning center. *Recent Research Developments in Seismology*, 1–13. Kervala, India: Transworld Research Network. ISBN 81-7895, 072-3
82. Whitmore PM, Tsuboi S, Hirshorn B, Sokolowski TJ (2002) Magnitude-dependent correction for M_{wp} . *Sci Tsunami Hazards J* 20:187–192
83. Widjo K et al (2006) Rapid survey on tsunami Java 17 July, 2006. http://nctr.pmel.noaa.gov/java20060717/tsunami-java170706_e.pdf. Accessed July 2008
84. Withers M, Aster R, Young C, Beiriger J, Harris M, Trujillo J (1998) A comparison of select trigger algorithms for automated global seismic phase and event detection. *Bull Seism Soc Am* 88:95–106

Books and Reviews

- Abercrombie R, McGarr A, Di Toro G, Kanamori H (eds) *Earthquakes: Radiated energy and the physics of faulting*. In: *Geophysical Monographs* 170. American Geophysical Union, Washington DC
- Aki K, Richards PG (1980) *Quantitative Seismology Theory and Methods*, 2 vol. W.H. Freeman Co, San Francisco
- Aki K, Richards PG (2002) *Quantitative Seismology*, 2nd edn. University Science Books, Sausalito
- Båth M (1981) Earthquake magnitude – recent research and current trends. *Earth Sci Rev* 17:315–398
- Bormann P (ed)(2002) *IASPEI new manual of seismological observatory practice*, vol 1 and 2. GeoForschungsZentrum Potsdam, p 1250
- Duda S, Aki K (eds) (1983) *Quantification of earthquakes*. *Tectonophysics* 93, Special issue 3/4:183–356
- Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seismol Soc Am* 65(5): 1073–1095
- Kanamori H (1994) *Mechanics of Earthquakes*. *Annu Rev Earth Planet Sci* 22:307–237
- Kanamori H, Rivera L (2006) Energy partitioning during an earthquake. In: Abercrombie R, McGarr A, Kanamori H, Di Toro G (eds) *Earthquakes: Radiated energy and the physics of faulting*. *Geophysical Monograph* 170, American Geophysical Union, Washington, DC, pp 3–13
- Kanamori H, Brodsky E (2004) The Physics of earthquakes. *Rep Prog Phys* 67:1429–1496
- Kanamori H, The diversity of the physics of earthquakes. *Proc Japan Acad Ser B* 80:297–316
- Lay T, Wallace TC (1995) *Modern global seismology*. Academic Press
- Okal EA (1992) A student's guide to teleseismic body wave amplitudes. *Seism Res Lett* 63(N2):169–180
- Richter CF (1958) *Elementary seismology*. WH Freeman Co, San Francisco
- Stein S, Wyssession M (2003) *An introduction to seismology, earthquakes, and earth structure*. Blackwell Publishing

Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in

YASUKO TAKEI

Earthquake Research Institute, University of Tokyo,
Tokyo, Japan

Article Outline

Glossary

Definition of the Subject

Introduction

General Theoretical Framework to Describe
the Dynamics of Solid–Liquid Composite Systems

Overview of Applications

Elastic Wave Propagation
in a Solid–Liquid Composite System

Future Directions

Acknowledgments

Bibliography

Glossary

Partially molten rock The partially molten state is a thermodynamic state between solidus and liquidus temperatures, where both solid and liquid phases co-exist. In the Earth's interior, partial melting of rocks occurs in the upper mantle and/or crust beneath volcanic areas.

Melt Liquid phase in partially molten rocks or completely molten rock above the liquidus temperature is called melt. Density of melt is about 10% lower than solid. Hence, melt phase in the partially molten rocks tend to ascend toward the Earth's surface.

Aqueous fluid H_2O -rich fluid. In a subducting oceanic plate, at the depths of several tens of km, aqueous fluids are released by the dehydration of hydrated minerals. Aqueous fluids, having much lower density and viscosity than melts, tend to ascend due to the buoyancy force.

Seismic tomographic image A number of seismometer networks have been placed on the surface of the Earth to record the seismic wave propagation from seismic sources at depths to the surface. Using the traveltime data obtained from these observations, three-dimensional seismic velocity structures in the Earth can be obtained, with a process called seismic tomographic imaging. By using P and S wave traveltimes, V_P and V_S structures, respectively, can be obtained.

Definition of the Subject

The dynamics of solid-liquid composite systems are of great relevance to many problems in the earth sciences, including how melts or aqueous fluids generated by partial melting or dehydration migrate through the mantle and crust toward the surface, how deformation and fracture in these regions are influenced by the existence of fluids, and also how these fluids can be observed in the seismic tomographic images. The mechanical and transport properties of the solid-liquid composite systems strongly depend on liquid volume fraction and pore geometry, such as pore shape, pore size, and a detailed porosity distribution. Therefore, the microstructural processes that control pore geometry influence macroscopic dynamics, and vice versa. This article introduces a general continuum mechanical theory to treat the macroscopic dynamics of solid-liquid composite systems with a special emphasis on how such interactions with pore geometry can be described. Although intensive experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, many problems still remain unsolved and the actual liquid content and pore geometries in the crust and mantle are poorly understood. Therefore, in the latter part of this article, by applying the general theoretical framework introduced in the former part to the seismic wave propagation, the determinability of porosity and pore geometry from seismic tomographic images is discussed in detail. Due to the recent advances in seismic tomography, we can obtain three-dimensional and highly resolved images of both V_P and V_S structures. From the V_P or V_S structure alone, neither porosity nor pore geometry can be determined independently. However, if both V_P and V_S structures are available, porosity and pore geometry can be determined independently, thus providing valuable information complementary to experimental and modeling approaches. A practical method to determine porosity and pore geometry from the V_P and V_S images is presented.

Introduction

Melt segregation from partially molten mantle or crust to the surface is the fundamental process of volcanism. In arc volcanism, aqueous fluids derived from the dehydrating subducting slab migrate through the mantle wedge and play an important role in the melting process. Modeling approaches to these fluid migration processes have been developed to explain various chemical and petrological observations, e.g. [13,20,35,37]. To make in situ observations of these fluids, seismic and/or electromagnetic tomographic images of the partially molten regions are

produced. These images are then interpreted using methods built on experimental and theoretical studies of the effects of fluids on seismic and electromagnetic properties, e. g. [2,23,24,44]. The effects of fluids on the tectonic and/or volcanic earthquake source process have also been of great interests; recently, observations of deep low-frequency earthquakes and tremors have been considered to be indicators of the presence or active migration of fluids in the crust and mantle, e. g. [9,26,29].

Fluids in the crust and mantle exist as solid-liquid composite systems in which fluid-filled pores are included in the solid matrix. Solid-liquid composite systems are characterized by high structural sensitivity. When the liquid volume fraction increases from zero to a few tens of %, the mechanical and transport properties of the system change greatly from those of a solid to those of a liquid. These properties are not simply determined by the liquid volume fraction but also strongly depend on the geometry of the liquid-filled pores. Here, the term pore geometry represents not only pore shape, but also pore size, orientation, and homogeneous or heterogeneous porosity distribution. Therefore, liquid content and pore geometry strongly influence the melt segregation and/or matrix deformation dynamics. Pore geometry is not constant but can change interactively with the macroscopic dynamics. Experimental and modeling approaches have been performed to investigate the interactive evolution of pore geometry and matrix deformation or fluid flow, e. g. [11,35,39]. However, many problems, including a poor understanding of the underlying physics, discrepancies between the experimental and modeling results, and the issue of scaling laboratory results to km scale, remain unsolved and detailed forward approaches to the interactions are the subjects of future studies.

Remarkable progress has been made in the seismic observations of the solid-liquid composite systems in the crust and mantle. Due to the recent advances in seismic tomography, we can now obtain three-dimensional and highly-resolved images of both V_p and V_s structures. Because the seismic velocity depends on both liquid volume fraction and pore geometry, neither can be estimated independently from V_p or V_s alone. However, when both V_p and V_s are available, the liquid volume fraction and pore geometry can be estimated independently. By using V_p and V_s images, an inverse approach to estimate the actual pore geometry and fluid content in the crust and mantle can be performed, yielding complementary information to the experimental and modeling approaches mentioned above. Therefore, in the present introduction of the dynamics of solid-liquid composite systems, elastic wave propagation is discussed in detail with a special focus on

the determinability of fluid content and pore geometry from seismic tomographic data.

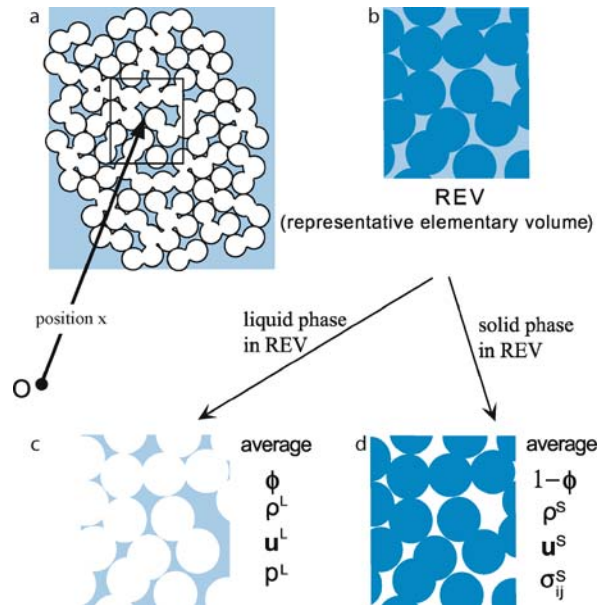
First, in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems”, I introduce a general continuum mechanical formulation of the macroscopic dynamics of solid-liquid composite systems, with a special emphasis on how structural sensitivity is described. The elastic version of this theory, called “linear poroelasticity”, was developed by Biot and coworkers, e. g. [5,6,31,48]. This theory, which assumes infinitesimal strain in the solid phase, is applicable to the propagation of elastic waves. A more general version applicable to large deformations was developed based on the fluid dynamic theory, and applied to the melt segregation dynamics in partially molten mantle, e. g. [7,20]. On the one hand, in most reviews of the theory of linear poroelasticity, the governing equations are introduced empirically and are difficult to compare to the mass and momentum conservation equations used in the usual continuum mechanical formulation, e. g. [48]. However, in those reviews, detailed explanations are given for the meaning of the macroscopic constitutive relation that describes the structural sensitive character of the solid-liquid composite systems. On the other hand, in most reviews of the general fluid-dynamical formulation the mass and momentum conservations can be easily confirmed, but the physical meanings of the macroscopic constitutive relation and the structural-sensitive character are difficult to understand. Therefore, by taking advantage of both fields of study, I introduce here a general fluid-dynamical formulation with a detailed explanation of the macroscopic constitutive relation. The general formulation includes as a special case the theory of linear poroelasticity. The governing equations introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” include several structural-sensitive parameters that are given as functions of liquid volume fraction and pore geometry. Provided an evolution equation of pore geometry is obtained in future studies, it will be possible to investigate the interaction between pore geometry and macroscopic dynamics by solving the governing equations including the constitutive relation together with the evolution equation.

In Sect. “Overview of Applications”, a brief summary of the various applications of the general theory introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” is presented. In Sects. “Derivation of Wave Equations” to “Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems”, the theory is applied to the elastic wave propagation in a solid-liquid composite system; we linearize the general formulation and derive the wave

equations (Sect. “Derivation of Wave Equations”). Based on the detailed descriptions of V_p and V_s obtained from the wave equations, the effects of liquid volume fraction, pore geometry, and liquid compressibility on the velocities are summarized systematically (Sects. “Porosity and Pore Shape” to “Determinability of Porosity and Pore Shape from Elastic Wave Velocities”). To assess the determinability of pore geometry, the usual forward modeling based on a priori assumed pore geometries, e.g. [15,17,18,27,41] are not enough, and a systematic treatment of general pore geometries is required, which is enabled by the introduction of the concept of equivalent aspect ratio (Sect. “Porosity and Pore Shape”). The determinability of porosity and pore geometry from seismic tomographic data is discussed in Sect. “Application to Seismic Tomographic Images”. In Sect. “Dispersion and Attenuation of Waves in Solid-Liquid Composite Systems”, so as to clarify the limitation in the application of the theoretical results, I briefly discuss the attenuation and dispersion of elastic waves. In this article, the term ‘liquid’ is used with the same meaning as ‘fluid’. Hence, ‘liquid’ in this paper includes ‘gas’.

General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems

A schematic illustration of a solid-liquid composite system considered in this article is shown in Fig. 1a. Length scales less than the grain size are referred to as “microscopic”; length scales greater than the grain size are referred to as “macroscopic”. Solid-liquid composite systems are characterized by large differences in mechanical properties between solid and liquid phases, and hence the stress and velocity fields that develop under external forces are usually highly heterogeneous at the microscopic scale. However, when studying macroscopic dynamics such as mantle-scale melt segregation and propagation of seismic waves with wavelengths much larger than the grain size, it is not practical to solve both the microscopic and macroscopic processes simultaneously. In this section, I review the theoretical framework to treat the macroscopic dynamics separately from the microscopic processes. In this theory, macroscopic dynamics of solid-liquid composite systems are described within the framework of continuum mechanics, using the macroscopic variables obtained by averaging the microscopic fields. The averages within the solid and liquid phases are taken separately, so that the theory can be applied to the phenomena involving a relative motion between the two phases. Although microscopic variables do not explicitly appear in the governing equations, several parameters included in these equa-



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 1

a A solid-liquid composite system with a representative elementary volume (REV) at position x . **b** REV consisting of solid (thick blue) and liquid (light blue) phases. **c** Liquid phase in REV and the averaged quantities representing the macroscopic mechanical state of the liquid. **d** Solid phase in REV and the averaged quantities representing the macroscopic mechanical state of the solid

tions are sensitive to the microstructures and thus the microscopic processes do affect the macroscopic dynamics through these parameters. In this review, a special emphasis is given to these structurally sensitive parameters.

Macroscopic Variables

We consider a region called REV (representative elementary volume), which is small enough to consider as a point in the macroscopic scale but large enough to contain a number of solid grains. Macroscopic quantities are defined by the average of the corresponding quantities in REV, where the averages within the liquid phase and those within the solid phase are taken separately. The averaging procedure can be defined by using a window function W and phase function A . By considering the REV as a cuboid with edge length L_i , the window function $W(\mathbf{x})$ takes the value 1 at $-L_i/2 \leq x_i \leq L_i/2$ ($i = x, y, z$) and value 0 otherwise. The phase function $A(\mathbf{x})$ takes the value 1 if the position \mathbf{x} is in the liquid phase, and value 0 in the solid phase; A and $1 - A$ quantify the properties of the liquid and solid phases, respectively. Let $a(\mathbf{x}, t)$ be a microscopic

field of the physical quantity a , which is a function of the position \mathbf{x} and time t . The phasic average of a in the liquid phase (a^L) or solid phase (a^S) is defined by

$$\begin{aligned} a^L(\mathbf{x}, t) &= \frac{1}{V\phi} \int W(\boldsymbol{\xi} - \mathbf{x}) A(\boldsymbol{\xi}, t) a(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}} \\ a^S(\mathbf{x}, t) &= \frac{1}{V(1-\phi)} \int W(\boldsymbol{\xi} - \mathbf{x}) (1 - A(\boldsymbol{\xi}, t)) \cdot a(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}}, \end{aligned} \quad (1)$$

where $V(=L_x L_y L_z)$ represents the volume of REV and ϕ represents the liquid volume fraction, defined as

$$\phi(\mathbf{x}, t) = \frac{1}{V} \int W(\boldsymbol{\xi} - \mathbf{x}) A(\boldsymbol{\xi}, t) dV_{\boldsymbol{\xi}}. \quad (2)$$

The volume integrals in Eqs. (1)–(2) are taken over whole of the solid-liquid system.

The dynamic state of the solid-liquid composite system at a spatially-fixed position \mathbf{x} is described by the following 7 macroscopic variables defined by Eqs. (1) and (2):

$\phi(\mathbf{x}, t)$...liquid volume fraction (nondimensional)

$\rho^L(\mathbf{x}, t)$...density of liquid (kg/m³)

$\rho^S(\mathbf{x}, t)$...density of solid (kg/m³)

$\mathbf{u}^L(\mathbf{x}, t)$...displacement of liquid (m)

$\mathbf{u}^S(\mathbf{x}, t)$...displacement of solid (m)

$p^L(\mathbf{x}, t)$...liquid pressure (Pa), with compression positive

$\sigma_{ij}^S(\mathbf{x}, t)$...solid stress (Pa), with tension positive.

It is rigorous to define \mathbf{u}^L and \mathbf{u}^S by using mass-weighted average [7]. However, for simplicity, the density heterogeneity within each phase is assumed to be small and the mass-weighted average is approximated by the phasic average.

Governing Equations

The seven variables introduced in Sect. “Macroscopic Variables” are governed by the following seven equations:

mass conservation of liquid

$$\frac{\partial(\phi\rho^L)}{\partial t} + \nabla \cdot (\phi\rho^L \dot{\mathbf{u}}^L) = \Gamma \quad (A)$$

mass conservation of solid

$$\frac{\partial\{(1-\phi)\rho^S\}}{\partial t} + \nabla \cdot \{(1-\phi)\rho^S \dot{\mathbf{u}}^S\} = -\Gamma \quad (B)$$

intrinsic constitutive relation of liquid

$$\frac{\delta\rho^L}{\rho^L} = \frac{1}{k_L} \delta p^L \quad (C)$$

intrinsic constitutive relation of solid

$$\frac{\delta\rho^S}{\rho^S} = \frac{1}{k_S} \delta p^S \quad (D)$$

macroscopic constitutive relation of solid framework

$$\epsilon_{ij} = S_{ijkl} (\sigma_{kl}^S + p^L \delta_{kl}) - \frac{1}{3k_S} p^L \delta_{ij} \quad (E)$$

linear momentum conservation of liquid

$$\phi\rho^L \ddot{\mathbf{u}}^L = -\nabla(\phi p^L) + \phi\rho^L \mathbf{g} + \mathbf{I} \quad (F)$$

linear momentum conservation of solid

$$(1-\phi)\rho^S \ddot{\mathbf{u}}^S = \nabla \cdot \{(1-\phi)\sigma^S\} + (1-\phi)\rho^S \mathbf{g} - \mathbf{I}, \quad (G)$$

where \mathbf{I} (N/m³) in Eqs. (F) and (G) represents the interaction between the solid and liquid phases (the force applied to the liquid from the solid is taken positive), and is explicitly written as

$$\mathbf{I} = -\frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) + p^L \nabla \phi. \quad (3)$$

The Γ (kg/s/m³) in Eqs. (A) and (B) represents the net mass flux from the solid to the liquid, k_L and k_S (Pa) in Eqs. (C), (D), and (E) represent the intrinsic bulk moduli of the liquid and solid, respectively, $p^S = -(\sigma_{xx}^S + \sigma_{yy}^S + \sigma_{zz}^S)/3$ (Pa) in Eq. (D) represents the solid pressure (compression positive), ϵ_{ij} in the left hand side of Eq. (E) represents the framework strain (extension positive), whose definition is given in Sect. “Equation (E)”, S_{ijkl} (Pa^{−1}) in Eq. (E) represents the elastic compliance tensor of the solid framework, \mathbf{g} (N/kg) in Eqs. (F) and (G) represents the gravitational acceleration vector, and η_L (Pa s) and k_ϕ (m²) in Eq. (3) represent the liquid viscosity and permeability, respectively. For $\alpha = S, L$, $\dot{\mathbf{u}}^\alpha = D\mathbf{u}^\alpha/Dt$ and $\ddot{\mathbf{u}}^\alpha = D\dot{\mathbf{u}}^\alpha/Dt$ represent velocity and acceleration vectors, respectively, where $D/Dt = \partial/\partial t + \dot{\mathbf{u}}^\alpha \cdot \nabla$. Also, $\nabla \cdot \sigma = \partial\sigma_{ij}/\partial x_j = \sigma_{ij,j}$. The summation convention for repeated subscripts is employed.

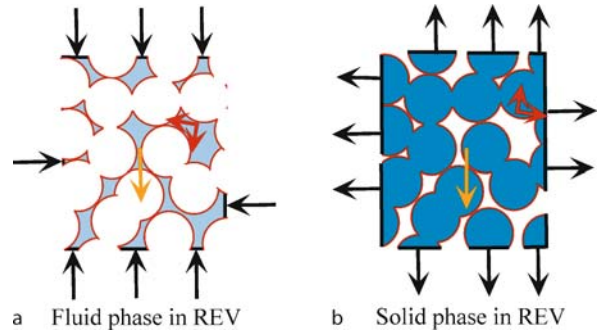
These equations, except for Eq. (E), are rigorously derived by averaging the mass and linear momentum conservation equations and constitutive relations required in the microscopic scale [7]. The physical meanings of Eqs. (A)–(D) are made clear by analogy with the standard fluid dynamic equations. The term Γ included in Eqs. (A)–(B) is zero unless melting/solidification or dissolution/precipitation occurs. The parameters that are sensitive to the microstructures are S_{ijkl} and k_ϕ included in Eqs. (E)–(G). Therefore, in the following part of this

section, detailed discussions of these three equations are given. In Sect. “Semi-Intuitive Derivations of Equations (E), (F), and (G)”, derivations of Eqs. (E)–(G) are given in a semi-intuitive manner to clarify the physical meaning of these equations and the structural-sensitive parameters. Several important aspects of these equations are also discussed in Sects. “Relation to the Effective Medium Theory”–“Fundamental Assumption for Stress Heterogeneity”.

Semi-Intuitive Derivations of Equations (E), (F), and (G)

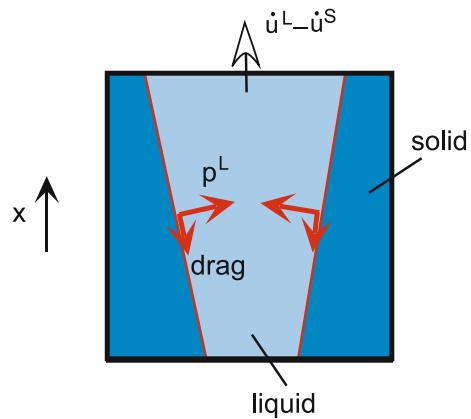
Equations (F) and (G) Equations (F) and (G) describe the linear momentum conservations for the liquid and solid phases, respectively, in the REV. The left hand sides (LHS) of these equations represent the acceleration terms. The right hand sides (RHS) represent the total forces applied on each system, including the body force and the surface force. The surface force is applied through the boundary surface of each system; this boundary is made of the boundary on the surface of REV (black boundary in Figs. 2a and 2b) and the boundary with the other phase (red boundary in Figs. 2a and 2b). The former boundary (black) exists within each phase and the latter boundary (red) exists on the phase boundary. In the RHS of Eq. (F) or (G), the second term represents the body force due to gravity, and the first and the third terms represent the surface forces applied through the black and red boundaries, respectively. The surface force through the red boundary is the interaction between solid and liquid, and hence the third term in the RHS of Eq. (F), I , and that of Eq. (G), $-I$, are of the same magnitude and of opposite sign.

Equation (3) shows that interaction I consists of two terms, corresponding to the contributions from the traction components tangential and normal to the phase boundary. The first term, corresponding to the tangential component, represents the viscous drag force proportional to the relative velocity of the two phases. The permeability k_ϕ included in the proportionality constant depends on the detailed geometry of the liquid-filled pores. The second term in the RHS of Eq. (3) represents the contribution from the normal component of traction, which is determined by the liquid pressure. An intuitive explanation for the proportional dependence of this force on the porosity gradient is given in Fig. 3, in which a simple pore geometry is assumed. A more rigorous derivation of this term for general pore geometries is presented in Sect. “Fundamental Assumption for Stress Heterogeneity”, where the assumption needed to derive this term is clarified and its validity is discussed. In the dynamics of solid-liquid com-



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 2

a Forces applied on the liquid phase in REV. **b** Forces applied on the solid phase in REV. Forces applied on each system consist of body force (orange arrow) and surface force (black and red arrows). The body force is due to gravity. The surface force is applied through the boundary surface of each system, which is divided into the boundary on the surface of REV (black boundary) and the boundary with the other phase (red boundary). The surface force applied through the red boundary is shown by dividing into normal and tangential components to the interface (red arrows). Solid stress is taken to be tension positive and liquid pressure is taken to be compression positive

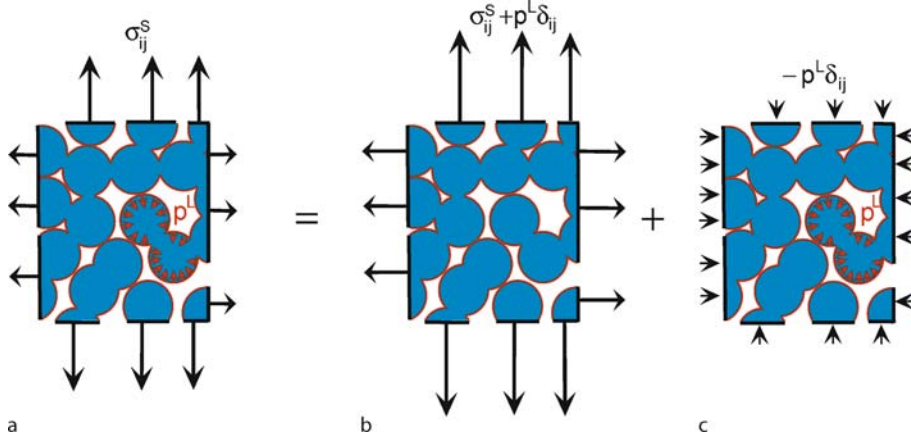


Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 3

Interaction between the solid and liquid phases through the phase boundary (red boundary) is schematically shown for a simple pore geometry with a porosity gradient. The forces applied from solid to liquid are shown. The total force of the force component normal to the phase boundary does not vanish if porosity gradient is not zero, and that tangential to the phase boundary does not vanish if average velocity of liquid relative to solid is not zero

posite systems, the motion of each phase is significantly affected by the interaction with the other phase through the phase boundary.

Note that Eqs. (F) and (G) are derived by implicitly assuming a connectivity of each phase (black boundary in



Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 4

a A stress state of the solid phase in REV, generally represented by the solid stress σ_{ij}^S (tension positive) and liquid pressure p^L (compression positive), is expressed by the superposition of b effective stress state, defined by solid stress $\sigma_{ij}^S + p^L \delta_{ij}$ and liquid pressure 0, and c uniform stress state, defined by solid stress $-p^L \delta_{ij}$ and liquid pressure p^L . Equation (E) in the text states that the framework strain under a given stress state is obtained by the superposition of the framework strain under the effective stress state and that under the uniform stress state

Figs. 2a and 2b). Therefore, we need to be careful in applying the present theory to end-member systems of suspensions of solid or isolated inclusions of liquid, in which one phase is dispersed in the other phase without connectivity.

Equation (E) The framework strain ϵ_{ij} in the LHS of Eq. (E) is defined by using the macroscopic displacement of the solid \mathbf{u}^S as

$$\epsilon_{ij} = \frac{1}{2} \left(\frac{\partial u_i^S}{\partial x_j} + \frac{\partial u_j^S}{\partial x_i} \right), \quad (4)$$

and hence ϵ_{ij} represents the macroscopic deformation of the solid framework. Equation (E) provides a macroscopic constitutive relationship between the framework strain and the macroscopic (averaged) stress state of the framework. The coefficient S_{ijkl} included in this equation is not only determined by the solid intrinsic properties, but shows a large structural sensitivity, which plays a significant role in the dynamics of solid-liquid composite systems. Equation (E) is one of the key equations characterizing the two-phase dynamics. I present here a semi-intuitive derivation of Eq. (E).

As shown in Fig. 4a, a macroscopic stress state of the solid framework is generally described by the solid stress σ_{ij}^S applied through the boundary on the surface of REV (black boundary) and liquid pressure p^L applied through the boundary with the liquid phase (red boundary). The framework strain under this stress state is calculated by the superposition of the framework strain under stress

$\sigma_{ij}^S + p^L \delta_{ij}$ applied through the black boundary while leaving the red boundary as a free surface (Fig. 4b) plus the framework strain under uniform pressure p^L applied to all boundaries (Fig. 4c). The assumption of the superposition is valid when the response of the solid framework to the applied stresses is linear. The term $\sigma_{ij}^S + p^L \delta_{ij}$ is called the effective stress. Figure 4b shows that the elastic compliance tensor S_{ijkl} specifying the effect of effective stress on the framework strain describes the mechanical properties of the ‘skeleton’ (solid framework obtained by replacing the regions occupied by the liquid phase with a vacuum), which is not only determined by the intrinsic properties of solid but also strongly depends on the porosity and pore geometry. For a given solid-liquid composite system, S_{ijkl} is estimated by using experimental and/or modeling approaches, some of which are presented in a later section. The framework strain under a given pressure p^L applied to all boundaries (Fig. 4c) is identical to the strain of REV completely filled with solid ($\phi = 0$) under uniform pressure p^L , and hence is determined only by the intrinsic bulk modulus of the solid phase (the last term in the RHS of Eq. (E)).

In the derivation mentioned above, the solid phase is assumed to deform elastically. For an isotropic system, Eq. (E) is written as

$$\epsilon_{ij} = \frac{1}{3k_{sk}} \left(\frac{\sigma_{kk}^S}{3} + p^L \right) \delta_{ij} + \frac{1}{2\mu_{sk}} \left(\sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right) - \frac{1}{3k_s} p^L \delta_{ij}, \quad (E_e)$$

where k_{sk} and μ_{sk} represent the bulk and shear moduli, respectively, of the skeleton. Similarly, when the solid phase deforms viscously, the macroscopic constitutive relation for an isotropic system is written as

$$\dot{\epsilon}_{ij} = \frac{1}{3\xi_{sk}} \left(\frac{\sigma_{kk}^S}{3} + p^L \right) \delta_{ij} + \frac{1}{2\eta_{sk}} \left(\sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right), \quad (E_v)$$

where ξ_{sk} and η_{sk} represent the bulk and shear viscosities, respectively, of the skeleton. Because viscous deformation is usually large in amplitude, the volumetric deformation due to the intrinsic compressibility of the solid phase, which corresponds to the 3rd term in the RHS of Eq. (E_e), is neglected in Eq. (E_v). Similar to k_{sk} and μ_{sk} , ξ_{sk} and η_{sk} are not only determined by the intrinsic property of the solid phase but also strongly depend on the porosity and pore geometry. Although the intrinsic compressibility of the solid phase is neglected in Eq. (E_v), the volumetric component of the framework strain rate controlled by ξ_{sk} cannot be neglected. This is because even when the solid phase is made of incompressible material, the solid framework can change its volume by changing the porosity. This demonstrates the essential difference between the skeleton property and the intrinsic property of the solid. A more general description of the viscous constitutive relation is given by neglecting the last term in the RHS of Eq. (E) and replacing ϵ_{ij} and elastic compliance tensor S_{ijkl} by $\dot{\epsilon}_{ij}$ and viscous compliance tensor S_{ijkl}^V , respectively.

Relation to the Effective Medium Theory

The structural sensitivity of the skeleton properties S_{ijkl} plays an important role in the two-phase dynamics. To predict quantitatively the microstructural effects on the skeleton properties, an effective medium theory has been developed. However, the applicability of these theoretical results to Eq. (E) is not self-evident, because the definition of the macroscopic strain given by Eq. (4) is different from that of the average strain used in the effective medium theory. In the effective medium theory, it is well-known that the stress and strain fields of the solid phase are highly heterogeneous at the microscopic scale, so that the local stress can be largely different from the macroscopic stress. However, in the fluid dynamical formulation of the two-phase dynamics, this point is not emphasized and a confusion between microscopic and macroscopic stresses sometimes occurs (Sect. “Fundamental Assumption for Stress Heterogeneity”). Therefore, it is important to establish a connection between

the fluid dynamical formulation and the effective medium theory.

When the skeleton properties are calculated in the effective medium theory, the effective properties of the solid-liquid composites are calculated under either drained conditions, in which the liquid pressure is kept constant, or under dry conditions, in which the liquid phase with zero bulk and shear moduli is kept under undrained conditions. In both cases, the space which remains after subtracting the solid phase from REV is considered to be filled with the pore phase. The microscopic displacement field in the pore phase can be obtained from the undrained solution by setting the bulk and shear moduli of the liquid to zero. The displacement field in the solid phase is continuously connected with that in the pore phase and there is no relative motion between the solid and pore phases. The effective properties under the dry (or drained) condition is given by

$$\epsilon_{ij}^B = S_{ijkl}^{\text{dry}} \sigma_{kl}^B, \quad (5)$$

where ϵ_{ij}^B and σ_{ij}^B represent the average strain and stress, respectively, over both solid and pore phases,

$$\begin{cases} \sigma_{ij}^B &= (1-\phi)\sigma_{ij}^S \\ \epsilon_{ij}^B &= (1-\phi)\epsilon_{ij}^S + \phi\epsilon_{ij}^P, \end{cases} \quad (6)$$

the superscript “B” means bulk, and

$$\epsilon_{ij}^\alpha = \frac{1}{2} \left\langle \frac{\partial u_i}{\partial \xi_j} + \frac{\partial u_j}{\partial \xi_i} \right\rangle_\alpha \quad (\alpha = S, P) \quad (7)$$

represents the phasic average of strain in the solid or pore phase, e.g. [21,50]. To clarify the relationship between S_{ijkl} and S_{ijkl}^{dry} , the relationship between ϵ_{ij} and ϵ_{ij}^B needs to be specified.

Let $\langle a \rangle_\alpha$ ($\alpha = S, L$) be the phasic average of the quantity a defined in Eqs. (1). Let $\Sigma_{\text{black}}^\alpha$ ($\alpha = S, L$) be the boundary of the phase α on the surface of REV (black boundary in Figs. 2 and 4) and let Σ_{red} be the phase boundary in REV (red boundary in Figs. 2 and 4). Let n_i be the outward unit normal to these boundaries, where the positive direction at Σ_{red} is outward to the liquid phase. From the definition of phasic average $\langle \cdot \rangle_S$, we obtain

$$\begin{aligned} & \frac{\partial \langle u_i \rangle_S}{\partial x_j} \\ &= \frac{\partial}{\partial x_j} \left[\frac{1}{V(1-\phi)} \int W(\xi - \mathbf{x})(1 - A(\xi))u_i(\xi) dV_\xi \right] \end{aligned}$$

$$\begin{aligned}
&= \frac{\partial \phi}{\partial x_j} \frac{1}{V(1-\phi)^2} \int W(\xi - \mathbf{x})(1 - A(\xi))u_i(\xi) dV_\xi \\
&\quad + \frac{1}{V(1-\phi)} \frac{\partial}{\partial x_j} \left[\int W(\xi - \mathbf{x})(1 - A(\xi))u_i(\xi) dV_\xi \right] \\
&= \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
&\quad + \frac{1}{V(1-\phi)} \int -\frac{\partial W(\xi - \mathbf{x})}{\partial \xi_j} (1 - A(\xi))u_i(\xi) dV_\xi \\
&= \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
&\quad - \frac{1}{V(1-\phi)} \int \frac{\partial \{W(\xi - \mathbf{x}) \cdot (1 - A(\xi))u_i(\xi)\}}{\partial \xi_j} dV_\xi \\
&\quad + \frac{1}{V(1-\phi)} \int W(\xi - \mathbf{x})(1 - A(\xi)) \frac{\partial u_i(\xi)}{\partial \xi_j} dV_\xi \\
&\quad + \frac{1}{V(1-\phi)} \int W(\xi - \mathbf{x}) \frac{\partial (1 - A(\xi))}{\partial \xi_j} u_i(\xi) dV_\xi \\
&= \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} + \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_s \\
&\quad + \frac{1}{V(1-\phi)} \int W(\xi - \mathbf{x}) \frac{\partial (1 - A(\xi))}{\partial \xi_j} u_i(\xi) dV_\xi \\
&= \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} + \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_s \\
&\quad + \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{red}}} u_i(\xi) n_j(\xi) dS_\xi, \quad (8)
\end{aligned}$$

where the volume integral of factor $\partial \{W(\xi - \mathbf{x})(1 - A(\xi))u_i(\xi)\} / \partial \xi_j$ in the RHS of the 4th equation is converted to the surface integral of $W(\xi - \mathbf{x})(1 - A(\xi))u_i(\xi)$ on the outermost boundary of the solid-liquid system, which is zero because $W(\xi - \mathbf{x})$ is zero outside the REV.

Because the displacement field u_i of the solid phase can be continuously connected to u_i of the pore phase, by using Gauss's theorem, the integral over Σ_{red} in the third term on the RHS of the last equation of (8) can be rewritten in terms of the volume integral in the pore phase. Because the boundary of the pore phase is given by Σ_{red} and Σ_{black}^L , we thus obtain

$$\begin{aligned}
&\frac{\partial \langle u_i \rangle_s}{\partial x_j} \\
&= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_s + \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
&\quad + \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{red}} + \Sigma_{\text{black}}^L} u_i(\xi) n_j(\xi) dS_\xi \\
&\quad - \frac{1}{V(1-\phi)} \int_{\Sigma_{\text{black}}^L} u_i(\xi) n_j(\xi) dS_\xi
\end{aligned}$$

$$\begin{aligned}
&= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_s + \frac{\langle u_i \rangle_s}{1-\phi} \frac{\partial \phi}{\partial x_j} \\
&\quad + \frac{\phi}{(1-\phi)} \frac{1}{V\phi} \int W(\xi - \mathbf{x}) A(\xi) \frac{\partial u_i}{\partial \xi_j} dV_\xi \\
&\quad - \frac{1}{(1-\phi)} \frac{\partial (\phi \langle u_i \rangle_s)}{\partial x_j} \\
&= \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_s + \frac{\phi}{1-\phi} \left\langle \frac{\partial u_i}{\partial \xi_j} \right\rangle_p - \frac{\phi}{1-\phi} \frac{\partial \langle u_i \rangle_s}{\partial x_j}, \quad (9)
\end{aligned}$$

where u_i at the surface of REV (Σ_{black}^L and Σ_{black}^S) is considered to be the same between the solid and pore phases, because there is no relative motion between these two phases. From Eqs. (4) and (9), we obtain

$$\begin{aligned}
\epsilon_{ij} &= (1-\phi) \epsilon_{ij}^S + \phi \epsilon_{ij}^P \\
&= \epsilon_{ij}^B. \quad (10)
\end{aligned}$$

Therefore, a simple conversion formula can be obtained linking S_{ijkl}^{dry} and S_{ijkl}

$$S_{ijkl}^{\text{dry}} = (1-\phi) S_{ijkl}. \quad (11)$$

We further discuss the peculiarity of Eq. (E) that it cannot be derived by the phasic average of the microscopic constitutive relationship. Because the microscopic stress and strain fields in the solid phase are related to each other by the intrinsic constitutive relationship of the solid phase, we may take the average of this relationship over the solid phase in REV, to obtain

$$\epsilon_{ij}^S = S_{ijkl}^{\text{intrinsic}} \sigma_{kl}^S, \quad (12)$$

where $S_{ijkl}^{\text{intrinsic}}$ represents the solid intrinsic properties. By comparing Eq. (E) to Eq. (12), it is apparent that the significant difference between S_{ijkl} and $S_{ijkl}^{\text{intrinsic}}$ comes from the significant difference between ϵ_{ij} and ϵ_{ij}^S in that ϵ_{ij} represents not only ϵ_{ij}^S but also ϵ_{ij}^P . From the definition of ϕ ,

$$\begin{aligned}
\frac{\partial \phi}{\partial x_i} &= \frac{\partial}{\partial x_i} \left[\frac{1}{V} \int W(\xi - \mathbf{x}) A(\xi) dV_\xi \right] \\
&= \frac{-1}{V} \int W(\xi - \mathbf{x}) n_i(\xi) dS_\xi. \quad (13)
\end{aligned}$$

Then, from Eqs. (4), (8), and (13), ϵ_{ij} can be rewritten as

$$\epsilon_{ij} = \epsilon_{ij}^S + \frac{\int_{\Sigma_{\text{red}}} ((u_i(\xi) - \langle u_i \rangle_s) n_j(\xi) + (u_j(\xi) - \langle u_j \rangle_s) n_i(\xi)) dS_\xi}{2V(1-\phi)}. \quad (14)$$

Thus, a significant difference between ϵ_{ij} and ϵ_{ij}^S implies a significant contribution from the second term in the RHS

of (14). This means that the microscopic displacement field of the solid phase at the boundary with the liquid can be systematically different from the average displacement of the solid phase. An example of such systematic deviation can be seen in the compaction of the solid framework, where a contraction of the solid phase observed macroscopically is compensated by a displacement of the solid-liquid phase boundary into the pore space.

Fundamental Assumption for Stress Heterogeneity

A fundamental assumption implicitly used in formulating the dynamics of solid-liquid composite systems is that the microscopic stress field is homogeneous in the liquid phase but can be heterogeneous in the solid phase. This point is rarely stated explicitly. Here, I discuss this assumption first with respect to the liquid phase and then with respect to the solid phase.

The second term in the RHS of Eq. (3) represents the total force due to the normal traction component applied to the liquid phase through the phase boundary (red boundary). This term is derived as

$$\begin{aligned}
 & -\frac{1}{V} \int_{\Sigma_{\text{red}}} p(\xi) n_i(\xi) dS_\xi \\
 &= -\frac{1}{V} \int_{\Sigma_{\text{red}} + \Sigma_{\text{black}}^L} p(\xi) n_i(\xi) dS_\xi \\
 & \quad + \frac{1}{V} \int_{\Sigma_{\text{black}}^L} p(\xi) n_i(\xi) dS_\xi \\
 &= -\frac{1}{V} \int_{\text{REV}} A(\xi) \frac{\partial p(\xi)}{\partial \xi_i} dV_\xi + \frac{1}{V} \int_{\Sigma_{\text{black}}^L} p(\xi) n_i(\xi) dS_\xi \\
 &= -\phi \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_L + \frac{\partial (\langle p \rangle_L \phi)}{\partial x_i} \\
 &= -\phi \frac{\partial \langle p \rangle_L}{\partial x_i} + \frac{\partial (\langle p \rangle_L \phi)}{\partial x_i} \\
 &= p^L \frac{\partial \phi}{\partial x_i},
 \end{aligned} \tag{15}$$

where the relationship

$$\left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_L = \frac{\partial \langle p \rangle_L}{\partial x_i} \tag{16}$$

is assumed to obtain the 4th equation. The validity of this assumption is checked as follows. In the same manner as Eq. (8), we obtain

$$\frac{\partial \langle p \rangle_L}{\partial x_i} = \frac{-\langle p \rangle_L}{\phi} \frac{\partial \phi}{\partial x_i} + \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_L - \frac{1}{V \phi} \int_{\Sigma_{\text{red}}} p(\xi) n_i(\xi) dS_\xi. \tag{17}$$

From Eqs. (13) and (17),

$$\frac{\partial \langle p \rangle_L}{\partial x_i} = \left\langle \frac{\partial p}{\partial \xi_i} \right\rangle_L - \frac{1}{V \phi} \int_{\Sigma_{\text{red}}} (p(\xi) - \langle p \rangle_L) n_i(\xi) dS_\xi \tag{18}$$

is obtained. Equation (18) shows that the phasic average $\langle \rangle_L$ is not exchangeable with the differential operator, and that Eq. (16) is not valid if the 2nd term in the RHS of Eq. (18) is non-negligible, that is, if the liquid pressure at the boundary with the solid phase is systematically different from the average. In solid-liquid composite systems, the stress heterogeneity within the liquid phase relaxes quickly and the liquid pressure is usually considered to be uniform in REV. Therefore, the 2nd term in the RHS of (18) is considered to be negligible. This also confirms the validity of an assumption implicitly used in Figs. 3 and 4: because the liquid pressure is homogeneous at the microscopic scale, then from the continuity of stress, the normal compressive stress of the solid phase at the solid-liquid interface is equal to the macroscopic liquid pressure p^L .

The inexchangeability between the differential operator and phasic average is described by Eq. (14) for the solid phase and by Eq. (18) for the liquid phase. Although the second term in the RHS of Eq. (14) was considered to be non-negligible, the corresponding term in Eq. (18) was considered to be negligible. The different treatments applied to the two phases are based on the fact that the stress heterogeneity within the liquid phase relaxes quickly but that the stress heterogeneity within the solid phase does not relax (elastic solid phase) or relaxes much more slowly (viscous solid phase). As can be seen from Fig. 4b, under nonzero effective stress, the traction applied to the surface of each solid grain is significantly different between the areas in contact with the liquid phase and the areas in contact with the neighboring grains, indicating that the microscopic stress field in each grain is highly heterogeneous. Such a heterogeneous stress field causes a systematic deviation of the microscopic displacement at the boundary with the liquid, making the second term in the RHS of Eq. (14) non-negligible.

For the solid phase, it is therefore important to recognize a possible difference between local and macroscopic stresses. However, there seems to be a confusion in some studies considering the effect of the solid-liquid interfacial tension γ_{sl} . When γ_{sl} is taken into account, the stress continuity condition required at the solid-liquid interface is replaced by the Laplace condition, e.g. [38]. Hence, the solid stress used in the Laplace condition is the local stress, which is locally determined by p^L , γ_{sl} , and interfacial mean curvature, regardless of the macroscopic solid stress σ_{ij}^S . I emphasize this point because in the previous studies,

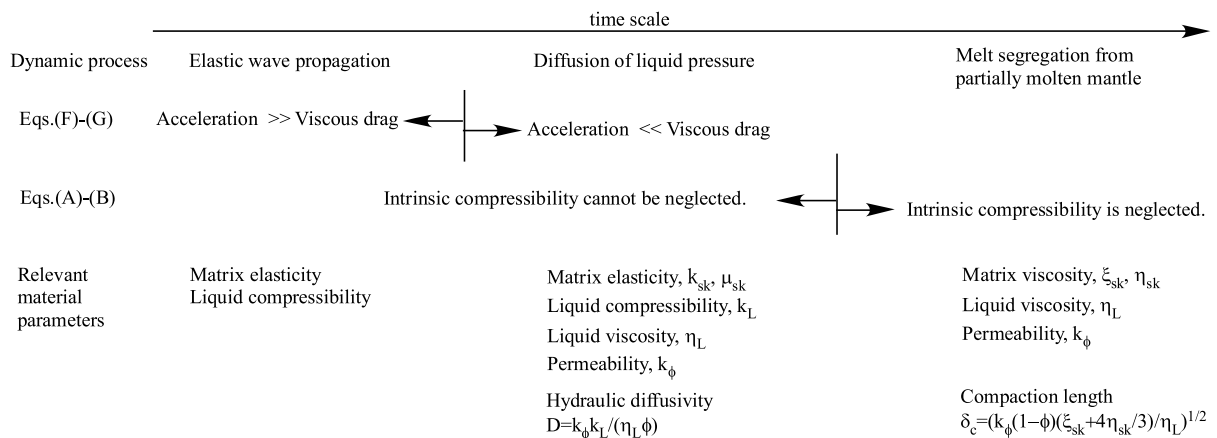
several questionable results were obtained due to the improper use of σ_{ij}^S in the Laplace condition, e. g. [38,46].

Overview of Applications

The theoretical framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems” is applicable to various geophysical and geological phenomena occurring in solid–liquid two-phase systems. Typical applications are summarized in Fig. 5 by classifying processes into three categories based on their time scales. The propagation of elastic waves is investigated using the wave equations derived from Eqs. (A)–(G). The application to elastic wave propagation is discussed in detail in the following sections. The time-dependent evolution of a perturbation in liquid pressure in the porous media, and the interaction between liquid pressure and matrix deformation or fracture, have long been of interest in investigations of the possible occurrence of earthquakes due to dehydration, e. g. [51] and earthquake triggering, e. g. [16]. These processes have much longer time scales than the periods of elastic waves, and the acceleration terms in Eqs. (F) and (G), which play an essential role in the wave equations, are negligible compared to the viscous drag force included in the interaction term I. Then, as is well known, the evolution equation for liquid pressure is derived in the form of a diffusion equation, where the diffusivity is given by $D = k_\phi k_L / (\eta_L \phi)$ (hydraulic diffusivity).

Melt segregation from a partially molten mantle has been of great interest in volcanology, petrology, and

geochemistry. This process occurs over much longer time scales than the processes mentioned above, and involves large viscous deformations. Accordingly, the intrinsic compressibilities of the constituent materials are neglected. By applying the governing equations (A)–(D), (E_v), and (F)–(G) to a one-dimensional column of partially molten mantle with a homogeneous porosity distribution at $t = 0$ ($\phi = \phi_0$ at $z \geq -H$ and $\phi = 0$ at $z < -H$), the initial stage of melt segregation was solved analytically [20]. In most of the column, the buoyancy force $(1 - \phi)(\rho^S - \rho^L)g$ is in equilibrium with the viscous drag force $\eta_L \phi (\dot{u}_z^L - \dot{u}_z^S) / k_\phi$, while within the compaction length δ_c from the bottom ($-H \leq z < -H + \delta_c$), the buoyancy force balances with the compaction resistance of the solid framework. Therefore, when the compaction length $\delta_c = \sqrt{k_\phi(1 - \phi)(\xi_{sk} + 4\eta_{sk}/3)/\eta_L}$ is smaller than H , which is the case for the mantle, the segregation velocity is determined by the permeability k_ϕ . The permeability control is more clear in the steady-state model, in which the steady-state porosity structure develops to satisfy the balance between melt production rate and melt segregation rate, which is often assumed for the decompressional melting of the upwelling mantle below a ridge [30]. However, several instabilities inherent to the solid–liquid systems that result in a time-dependent evolution of the porosity distribution have also been reported. These include the propagation of melt as solitary waves or porosity waves, e. g. [3,34], and the unstable evolution of the perturbation in melt fraction under pure shear deformation of the solid matrix [39], in which both melt migration and matrix deformation are involved. These phe-



Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 5

Dynamic processes in solid–liquid composite systems. Typical assumptions adopted in applying the governing equations (A)–(G) in the text to these processes are shown, with the material properties relevant to the processes

nomena are described by Eqs. (A)–(D), (E_v), and (F)–(G), where, in the solitary or porosity waves, the nonlinearity caused by the dependence of permeability on melt fraction ($k_\phi \propto \phi^{n>1}$) plays an essential role and, in the latter instability, the dependence of η_{sk} on ϕ plays an essential role. The possible occurrence of these instabilities has been of great interest, because the melt ascending velocity is significantly affected by the spatial distribution of porosity. The melt velocity and its spatial distribution determine the degree of chemical interaction between the melt and host rocks, and thus influence the major and trace element compositions. The microstructural dependences of permeability and viscous properties of the partially molten rocks affecting these instabilities are poorly understood and are the subject of future studies. The basic framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” can be further extended to take into account additional factors, such as the chemical interaction between the fluid and host rocks, e. g. [1,36] or interfacial tension e. g. [32].

Elastic Wave Propagation in a Solid-Liquid Composite System

Derivation of Wave Equations

In this section, the general theoretical framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems” is applied to the elastic wave propagation in a solid-liquid composite system. The governing equations are linearized by considering the infinitesimal strain and displacement of a macroscopically homogeneous medium, and the wave equations are derived. The linearized equations are shown to be equivalent to the basic equations used in the theory of linear poroelasticity.

When we consider a macroscopically homogeneous solid-liquid composite system, the spatial and temporal variations in ϕ , ρ^L , and ρ^S are caused by the displacements \mathbf{u}^L and \mathbf{u}^S . Therefore, if \mathbf{u}^L and \mathbf{u}^S are infinitesimally small, such terms as $\dot{\mathbf{u}}^L \cdot \nabla(\phi\rho^L)$ and $\dot{\mathbf{u}}^S \cdot \nabla((1-\phi)\rho^S)$ are negligible as higher-order terms. Under these approximations, by substituting Eqs. (C) and (D) into (A)/ $\rho^L + (B)/\rho^S$, and taking the time integration, we obtain

$$-\phi \nabla \cdot (\mathbf{u}^L - \mathbf{u}^S) = \frac{\phi}{k_L} p^L + \frac{1-\phi}{k_S} p^S + \nabla \cdot \mathbf{u}^S. \quad (19)$$

By using Eq. (E_c), $\nabla \cdot \mathbf{u}^S = \epsilon_{kk}$ in Eq. (19) can be expressed in terms of stresses. Then, Eq. (19) and Eq. (E_c)

are written as

$$\begin{aligned} & \phi \nabla \cdot (\mathbf{u}^L - \mathbf{u}^S) \\ &= \phi \left(\frac{1}{k_S} - \frac{1}{k_L} \right) p^L + (1-\phi) \left(\frac{1}{K_b} - \frac{1}{k_S} \right) (p^S - p^L) \end{aligned} \quad (20)$$

$$\begin{aligned} \epsilon_{ij} &= \frac{(1-\phi)}{2N} \left(\sigma_{ij}^S - \frac{\sigma_{kk}^S}{3} \delta_{ij} \right) \\ &\quad - \frac{(1-\phi)}{3K_b} (p^S - p^L) \delta_{ij} - \frac{1}{3k_S} p^L \delta_{ij}, \end{aligned} \quad (21)$$

where K_b and N represent the bulk and shear moduli, respectively, of the skeleton and are related to k_{sk} and μ_{sk} introduced in Eq. (E_c) as $K_b = (1-\phi)k_{sk}$ and $N = (1-\phi)\mu_{sk}$. In the theory of linear poroelasticity, K_b and N , rather than k_{sk} and μ_{sk} , are commonly used. Similarly, by substituting Eq. (3) into Eqs. (F) and (G), and neglecting the effect of gravity, we obtain

$$\phi \rho^L \ddot{\mathbf{u}}^L = -\phi \nabla \cdot p^L - \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) \quad (22)$$

$$(1-\phi) \rho^S \ddot{\mathbf{u}}^S = (1-\phi) \nabla \cdot \sigma^S + \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S). \quad (23)$$

Equations (20)–(23) are equivalent to the basic equations used in the theory of linear poroelasticity. Compared to the framework introduced in Sect. “General Theoretical Framework to Describe the Dynamics of Solid-Liquid Composite Systems”, the number of governing equations is reduced from 7 to 4, because the variables ρ^L and ρ^S are eliminated and ϕ in Eqs. (20)–(23) can be treated as constant. In the theory of linear poroelasticity, Eq. (20) is called the constitutive relation for the relative motion between the two phases, and this relation is usually introduced empirically, e. g. [48]. The present derivation from the more general framework shows that Eq. (20) is based on the requirement of mass conservation and intrinsic constitutive relations. When the acceleration terms are negligible, Eqs. (22) and (23) are further rewritten as

$$\phi (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) = -\frac{k_\phi}{\eta_L} \nabla p^L \quad (24)$$

$$\nabla \cdot \sigma^B = 0, \quad (25)$$

where Eq. (24) represents Darcy's law, and the bulk stress σ_{ij}^B represents $\sigma_{ij}^B = (1-\phi)\sigma_{ij}^S - \phi p^L \delta_{ij}$.

By eliminating pressures and stresses from Eqs. (20)–(23), we obtain

$$\begin{aligned} (1-\phi) \rho^S \ddot{\mathbf{u}}^S &= P \nabla (\nabla \cdot \mathbf{u}^S) - N \nabla \times \nabla \times \mathbf{u}^S \\ &\quad + Q \nabla (\nabla \cdot \mathbf{u}^L) + \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S) \end{aligned} \quad (26)$$

$$\phi \rho^L \ddot{\mathbf{u}}^L = Q \nabla(\nabla \cdot \mathbf{u}^S) + R \nabla(\nabla \cdot \mathbf{u}^L) - \frac{\eta_L \phi^2}{k_\phi} (\dot{\mathbf{u}}^L - \dot{\mathbf{u}}^S), \quad (27)$$

where P , Q , and R are given by

$$\begin{cases} P = K_b + \frac{4}{3}N + \frac{(1-\phi - \frac{K_b}{k_S})^2 k_S}{1-\phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \\ Q = \frac{\phi(1-\phi - \frac{K_b}{k_S}) k_S}{1-\phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \\ R = \frac{\phi^2 k_S}{1-\phi - \frac{K_b}{k_S} + \phi \frac{k_S}{k_L}} \end{cases} \quad (28)$$

By taking the curl of Eqs. (26) and (27), and using the expressions $\Omega_S = \nabla \times \mathbf{u}^S$ and $\Omega_L = \nabla \times \mathbf{u}^L$, we obtain wave equations for the shear component;

$$\begin{cases} (1-\phi) \rho^S \ddot{\Omega}_S = N \nabla^2 \Omega_S + \frac{\eta_L \phi^2}{k_\phi} (\dot{\Omega}_L - \dot{\Omega}_S) \\ \phi \rho^L \ddot{\Omega}_L = -\frac{\eta_L \phi^2}{k_\phi} (\dot{\Omega}_L - \dot{\Omega}_S) \end{cases} \quad (29)$$

By taking the divergence of Eqs. (26) and (27), and using the expressions $e_S = \nabla \cdot \mathbf{u}^S$ and $e_L = \nabla \cdot \mathbf{u}^L$, we obtain wave equations for the longitudinal component;

$$\begin{cases} (1-\phi) \rho^S \ddot{e}_S = P \nabla^2 e_S + Q \nabla^2 e_L + \frac{\eta_L \phi^2}{k_\phi} (\dot{e}_L - \dot{e}_S) \\ \phi \rho^L \ddot{e}_L = Q \nabla^2 e_S + R \nabla^2 e_L - \frac{\eta_L \phi^2}{k_\phi} (\dot{e}_L - \dot{e}_S) \end{cases} \quad (30)$$

The elastic wave propagation in a solid-liquid composite system was first formulated by Biot [5,6]. The wave equations (29)–(30) are almost the same as those obtained by Biot [5,6], except for the acceleration terms, which are slightly different. This is because the interaction \mathbf{I} given in Eq. (3) does not take into account the effect of the relative acceleration between the solid and liquid phases. If an additional term proportional to the relative acceleration is added to the RHS of Eq. (3), then the same equations as Biot [5,6] can be derived. The proportionality constant attached to the relative acceleration is called tortuosity; tortuosity represents the deviation from the straight pore channel.

The elastic waves obtained by solving Eqs. (29)–(30) are dispersive and dissipative due to the relative motion between the solid and liquid phases. However, as shown below in Sect. “Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems”, the characteristic frequency for the dispersion and attenuation is much higher than the seismic frequency range. Therefore, in predicting

the seismic wave velocities, the solutions obtained at the low-frequency limit are of importance. The wave solutions at the low-frequency limit do not involve relative motions, because the velocity terms, $\dot{\Omega}_L - \dot{\Omega}_S$ and $\dot{e}_L - \dot{e}_S$, if any, dominate the acceleration terms, and hence are not dispersive nor dissipative. The longitudinal and shear wave velocities at the low-frequency limit are obtained as

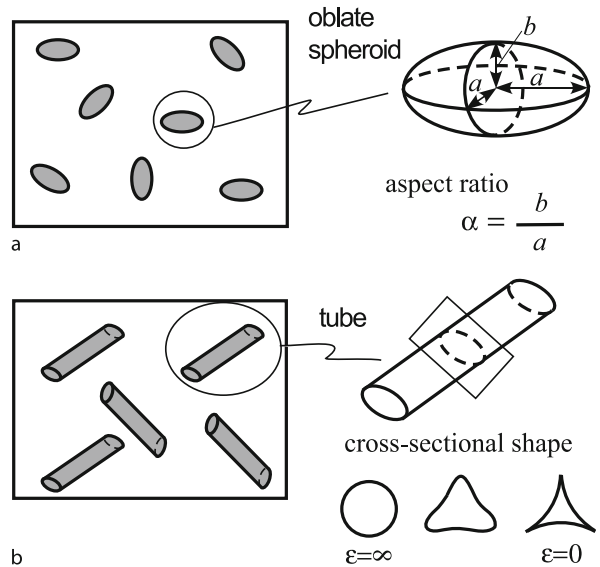
$$V_P = \sqrt{\frac{K_b + \frac{4}{3}N + \frac{k_S(1-K_b/k_S)^2}{1-\phi-K_b/k_S+\phi k_S/k_L}}{\bar{\rho}}} \quad (31)$$

$$V_S = \sqrt{\frac{N}{\bar{\rho}}}, \quad (32)$$

where $\bar{\rho} = (1-\phi)\rho^S + \phi\rho^L$ represents the average density of the medium. Without relative motion, neither permeability nor tortuosity affect the velocities. Therefore, Eqs. (31)–(32) are exactly the same as the results of Biot [5,6].

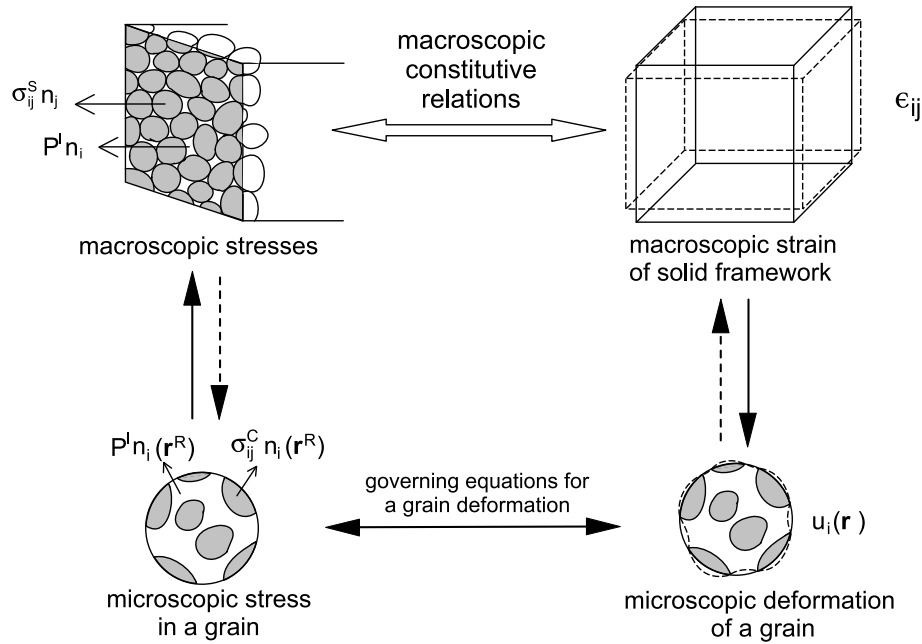
Porosity and Pore Shape

Because the bulk and shear moduli of the skeleton, K_b and N , included in Eqs. (31)–(32) depend not only on porosity but also on pore geometry, various models assuming various pore geometries have been developed to predict K_b and N quantitatively, e.g. [15,17,18,27,41]. The oblate spheroid model (Fig. 6a), tube model (Fig. 6b),



Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 6

Inclusion models of a solid-liquid composite system. **a** Oblate spheroid model. **b** Tube model



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 7

Granular model of a solid-liquid composite system, showing the procedures to derive macroscopic constitutive relation, Eq. (E), based on the microscopic deformation of each grain [41]

granular model (Fig. 7), and crack model are four representative models in which analytical results can be obtained for K_b and N (Table 1). All but the granular model are inclusion models in which the liquid phase is modeled by inclusions contained in a continuum solid phase, and K_b and N are derived based on the effective medium theories. In the granular model, the constitutive equation (E) is derived directly. The results from the different theories can be compared based on Eq. (11). Although the connectivity of the liquid phase is not guaranteed in the inclusion models, when considering waves in the low-frequency regime where relative motion between solid and liquid does not occur, connectivity of the liquid phase is not important.

Here, by assuming a random orientation and homogeneous distribution of the pores, the macroscopic properties are assumed to be isotropic. Then, K_b and N are derived as functions of the porosity ϕ and aspect ratio α (α =short radius/long radius) for the oblate spheroid model, as functions of the porosity ϕ and parameter ε for the tube model, where ε represents the cross-sectional tube shape (Fig. 6b), as functions of the contiguity φ for the granular model, and as functions of the crack density parameter κ for the crack model. The contiguity φ used in the granular model is defined by the ratio of the grain-to-grain contact area relative to the total surface area of each grain;

thus, $\varphi = 0$ when there is no grain-to-grain contact, and $\varphi = 1$ when there is no liquid or pore phase. The crack density parameter κ used in the crack model is defined by $\kappa = n_\kappa a_\kappa^3$, where a_κ represents the radius of the circular crack and n_κ represents number density. Walsh [47] showed that in the limit of small aspect ratio, K_b and N obtained from the oblate spheroid model depend only on the crack density parameter $\kappa = 3\phi/(4\pi\alpha)$, and the results of the oblate spheroid model and crack model become equivalent. Therefore, the crack model can be included in the oblate spheroid model as a special case of small aspect ratio.

In the granular model, the dependence of contiguity φ on porosity ϕ first needs to be assessed in order to specify the dependences of K_b and N on the porosity ϕ . For a random packing of elastic spheres, which is often used as a model of soil, the total area of the elastic contacts of the spheres increases with raising confining pressure, and φ and ϕ are derived as functions of confining pressure [8]. However, when the temperature is higher than a few hundreds °C, such an elastic model is not realistic. In the deep crust and mantle, solid grains are single crystals and the contiguity is related to the area of (liquid-free) grain boundaries. Because the grain boundary and crystal-liquid interface both have interfacial energies, there ex-

Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Table 1
Microstructural Models for Solid-Liquid Composites

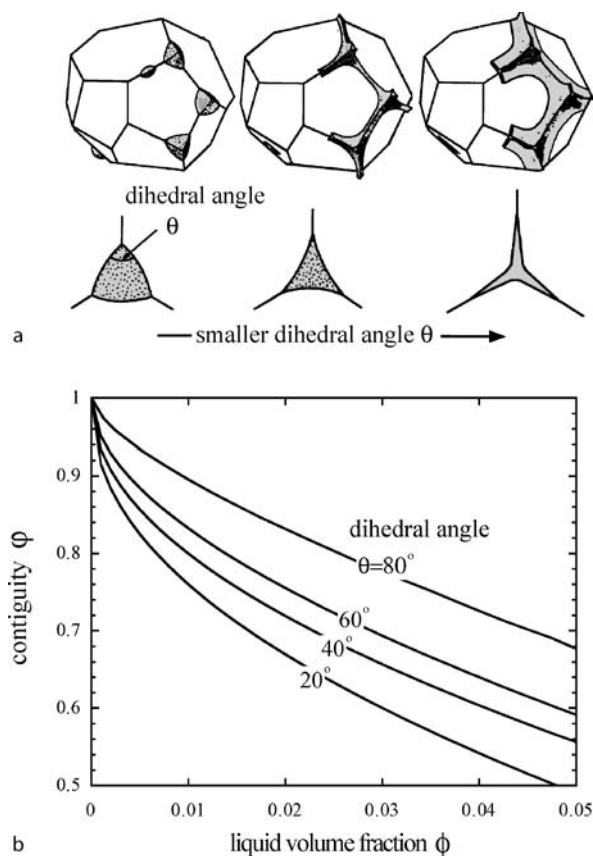
Model	Structural parameters		References
Oblate spheroid	Porosity ϕ	Aspect ratio α	e. g., [4] ¹
Tube	Porosity ϕ	Tube geometry ε	[17]
Granular	Contiguity φ		[41]
Equilibrium geometry ²	Porosity ϕ	Dihedral angle θ	[41]
Crack	Crack density parameter κ		[27]

1: Typographical errors in the former studies were corrected.

2: Equilibrium geometry model is a special case of granular model.

ists an equilibrium shape of the liquid phase which minimizes the total interfacial energy of the system. The equilibrium shapes were actually observed in the high T and high P experiments for various rock + melt and rock + aqueous fluid systems e. g. [10]. Therefore, the relationship between contiguity φ and porosity ϕ is derived by assuming the equilibrium shape of the liquid phase. The granular model under this assumption is called the equilibrium geometry model (Table 1). Under a given liquid volume fraction ϕ , the equilibrium shape is controlled by the dihedral angle θ , which is determined by the grain boundary energy γ_{ss} and crystal-liquid interfacial energy γ_{sl} as $\gamma_{ss}/\gamma_{sl} = 2 \cos(\theta/2)$ (Fig. 8a). The theoretical results of von Bargen and Waff [45] show that under a given ϕ , the equilibrium contiguity is smaller for smaller θ (Fig. 8b). By substituting φ obtained as functions of ϕ and θ into the results of the granular model, K_b and N in the equilibrium geometry model can be derived as functions of ϕ and θ . Most rock + melt systems have θ between 20–40° and most rock + aqueous fluid systems have θ between 40–100° [10]. When $\theta \leq 60^\circ$, a connected liquid network develops along the grain edges at $\phi > 0$. Although the tube model considers such grain edge tubules, the parameter ε in the tube model cannot be quantitatively related to the dihedral angle θ .

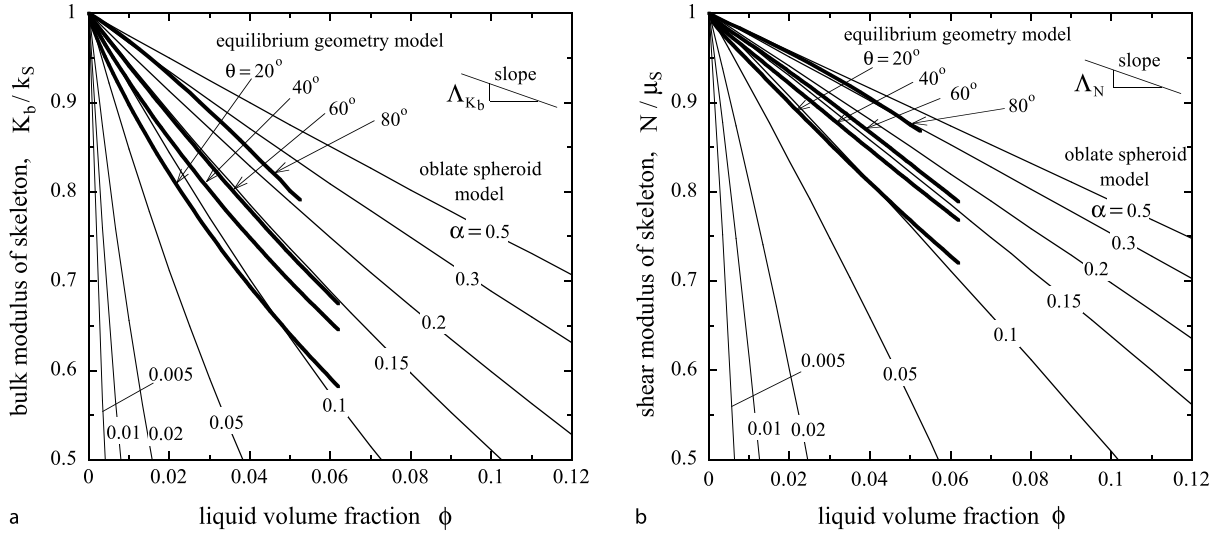
Now, solid-liquid composite systems are characterized in terms of the two parameters: porosity ϕ and pore shape x ($= \alpha, \varepsilon$, or θ). Because a random orientation and homogeneous distribution of the pores are assumed for simplicity, and since K_b and N do not depend on pore size, pore geometry is parameterized only by the shape. Porosity ϕ and pore shape x are generally not dependent but can vary independently governed by different physics. For example, in a texturally equilibrated system, the dihedral angle θ is determined by thermodynamic conditions such as temperature, pressure, and chemical compositions, whereas ϕ can vary mechanically through flow-in or flow-out of the liquid. In a rock + water system stressed



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 8

a Equilibrium geometry of the liquid phase characterized by dihedral angle θ . **b** Contiguity φ versus liquid volume fraction ϕ calculated theoretically for the equilibrium geometry with dihedral angle θ

under undrained condition, the pore shape can vary by fracture, while ϕ is kept constant. In most solid-liquid systems in the Earth, neither the porosity nor the pore shape are known.



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 9

a Bulk modulus K_b and **b** shear modulus N of skeleton versus porosity ϕ , for the oblate spheroid model with various aspect ratios α and for the equilibrium geometry model with various dihedral angles θ

Because the parameter describing the pore shape is different in different models, it seems difficult to investigate the effects of pore shapes systematically. However, introduction of the concept of “equivalent aspect ratio” enables us to treat various pore shapes systematically. When the porosity ϕ is small, K_b and N are closely approximated by linear functions of ϕ ,

$$\begin{cases} \frac{K_b}{k_s}(\phi, x) = 1 - \phi \Lambda_{K_b}(x) \\ \frac{N}{\mu_s}(\phi, x) = 1 - \phi \Lambda_N(x) \end{cases} \quad (x = \alpha, \varepsilon, \theta), \quad (33)$$

where k_s and μ_s represent the intrinsic bulk and shear moduli, respectively, of the solid, and the proportionality coefficients Λ_{K_b} and Λ_N are functions of pore shape x (Fig. 9). In other words, the effects of porosity and pore shape on K_b and N can be separated in such simple forms as given in Eq. (33), in which the pore shape given by x is characterized in terms of two parameters Λ_{K_b} and Λ_N . If a tube model with ε (or an equilibrium geometry model with θ) has almost the same values of Λ_{K_b} and Λ_N as the oblate spheroid model with α ,

$$\begin{cases} \Lambda_{K_b}(x) \simeq \Lambda_{K_b}(\alpha) \\ \Lambda_N(x) \simeq \Lambda_N(\alpha) \end{cases} \quad (x = \varepsilon, \theta), \quad (34)$$

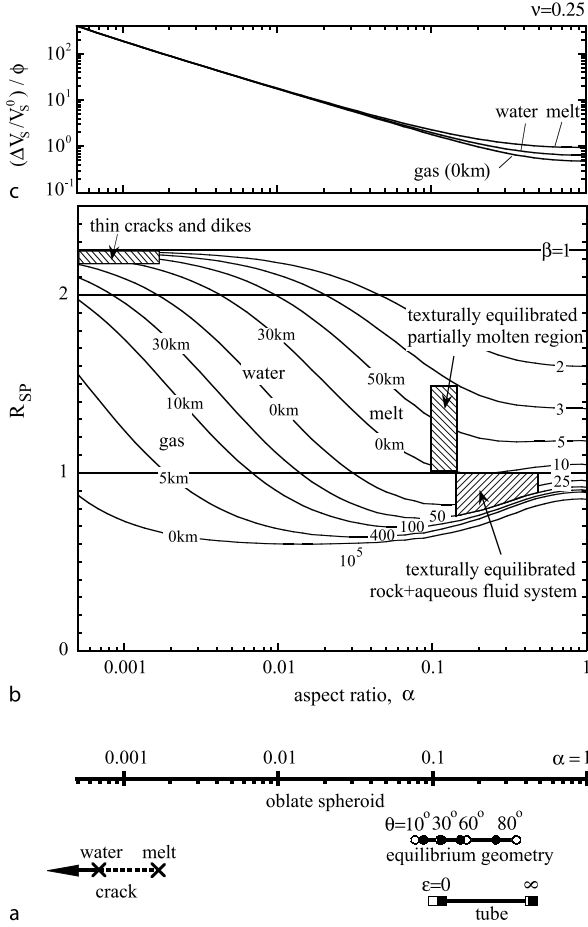
these two models yield almost the same values of K_b and N for a given ϕ . This also means that V_P and V_S calculated with Eqs. (31)–(32) are almost the same in these two mod-

els. Therefore, the value of α satisfying Eqs. (34) is called the equivalent aspect ratio of x ; this aspect ratio guarantees the equivalence between the different pore shapes in predicting V_P and V_S .

Figure 10a shows the relationships between the different models in terms of equivalent aspect ratio. The solid and open symbols represent the equivalent aspect ratios determined from Λ_{K_b} and Λ_N , respectively. The difference between these symbols is small, indicating that, in a practical sense, one value of equivalent aspect ratio satisfying both equations in (34) can be determined. Figure 10a shows that the tube model with $\varepsilon = 0$, equilibrium geometry model with $\theta = 30^\circ$, and oblate spheroid model with $\alpha = 0.1$ are all equivalent. Rigorously speaking, Λ_{K_b} and Λ_N depend on the intrinsic Poisson's ratio ν of the solid phase. The results shown in Fig. 10a are calculated for $\nu = 0.25$. Fortunately, however, the effects of ν are almost the same in all models, and the equivalent aspect ratio can be determined almost independently of ν . The present method to determine the equivalent aspect ratio from Eqs. (33)–(34) is applicable to general isotropic solid-liquid systems. By using the equivalent aspect ratio, general pore geometries can be treated systematically.

Determinability of Porosity and Pore Shape from Elastic Wave Velocities

It can be shown that when the porosity ϕ is small, the effects of ϕ on the skeleton properties K_b and N can be



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 10

a Equivalence of the equilibrium geometry model, tube model, and crack model to the oblate spheroid model is shown by the equivalent aspect ratio α . Solid and open symbols are the equivalent aspect ratios determined from Λ_{K_b} and Λ_N , respectively. **b** R_{SP} , representing the ratio between V_S and V_P perturbations, $(\Delta V_S/V_S^0)/(\Delta V_P/V_P^0)$, versus pore aspect ratio α , for various fluid compressibilities $\beta = k_S/k_L$. **c** Proportionality constant between V_S perturbation $\Delta V_S/V_S^0$ and porosity ϕ versus pore aspect ratio α , for various liquid types (gas, water, and melt)

closely approximated by linear functions of ϕ . In the same manner, when ϕ is small, the effects of ϕ on the velocities V_P and V_S can be closely approximated by linear functions of ϕ . Let $\Delta V_P = V_P^0 - V_P$ and $\Delta V_S = V_S^0 - V_S$ be reductions in V_P and V_S , respectively, caused by liquid-filled pores, where $V_P^0 = \sqrt{(k_S + 4\mu_S/3)/\rho^S}$ and $V_S^0 = \sqrt{\mu_S/\rho^S}$ represent the intrinsic elastic wave velocities of the solid phase. Let $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ be perturbations in V_P and V_S , respectively. By substituting Eqs. (33) into Eqs. (31)–(32) and neglecting higher-order terms in ϕ (ϕ^n

with $n \geq 2$), we obtain

$$\begin{cases} \frac{\Delta V_P}{V_P^0} = \left[\frac{(\beta-1)\Lambda_{K_b}}{(\beta-1)+\Lambda_{K_b}} + \frac{4}{3}\gamma\Lambda_N \right] \frac{\phi}{2} - \left(1 - \frac{\rho^L}{\rho^S}\right) \\ \frac{\Delta V_S}{V_S^0} = \left[\Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right) \right] \frac{\phi}{2}, \end{cases} \quad (35)$$

where $\beta = k_S/k_L$ and $\gamma = \mu_S/k_S$. Without loss of generality, Λ_{K_b} and Λ_N can be treated as functions of the equivalent aspect ratio α . Equations (35) demonstrate that the velocity perturbations are affected by the five non-dimensional parameters ϕ , α , β , γ , and ρ^L/ρ^S . Because a possible variation in γ in response to a variation in the intrinsic Poisson's ratio ν of the solid phase is small, γ can be fixed to 0.6 ($\nu = 0.25$). Also, as shown below, the effect of ρ^L/ρ^S on the perturbations is small. Therefore, in a practical sense, the velocity perturbations are controlled by the three non-dimensional factors: liquid volume fraction ϕ , pore aspect ratio α , and liquid compressibility β . If only one of $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ is known, ϕ cannot be determined without knowing α (and β). However, if both $\Delta V_P/V_P^0$ and $\Delta V_S/V_S^0$ are known, significant constraints can be placed on ϕ , α , and/or β . A practical method to obtain these constraints is presented below.

First, we introduce R_{SP} , representing the ratio of the perturbations in V_S and V_P . From Eq. (35), R_{SP} is written as

$$\begin{aligned} R_{SP} &= \frac{\Delta V_S/V_S^0}{\Delta V_P/V_P^0} \\ &= \frac{\Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right)}{\frac{(\beta-1)\Lambda_{K_b}}{(\beta-1)+\Lambda_{K_b}} + \frac{4}{3}\gamma\Lambda_N - \left(1 - \frac{\rho^L}{\rho^S}\right)}. \end{aligned} \quad (36)$$

R_{SP} can be closely related to the V_P/V_S ratio frequently used in seismology: when $R_{SP} < 1$, the perturbation (reduction positive) in V_P is larger than that in V_S and hence the V_P/V_S ratio decreases; when $R_{SP} > 1$, the perturbation in V_S is larger than that in V_P and thus the V_P/V_S ratio increases. Because R_{SP} is independent of the liquid volume fraction ϕ , this factor provides a useful insight into the effects of pore shape α and liquid compressibility β on the velocity perturbations. R_{SP} is sometimes written as $d \ln V_S / d \ln V_P$. Figure 10b shows R_{SP} versus pore aspect ratio α for various compressibility β of pore fluids. Values of β are estimated as 5–10 for rock + melt systems, 10–40 for rock + water systems, and 50–10⁵ for rock + ideal

Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Table 2

Liquid bulk modulus k_L

depth, km	P , GPa	T , °C	k_L , GPa		
			gas ¹	water ²	melt ³
0	10^{-4}	20	1.3×10^{-4} ($\beta = 10^5 - 10^6$) ⁴	2.2 (18–50)	7–25 (4–10)
5	0.15	75	0.2 (200–600)	3.1 (13–40)	
10	0.3	150	0.4 (100–300)	1.8 (22–66)	
35	1	500	1.3 (30–100)	4.5 (9–25)	
70	2				20–40 (3–6)

- 1: Adiabatic bulk modulus estimated by $1.3 P$.
- 2: Isothermal bulk modulus estimated at each (P, T) condition. Data from Schäfer [33].
- 3: Data from Stolper et al. [40]. Data at $P = 2$ GPa are estimated from $\partial k_L / \partial P = 6 - 7$.
- 4: Numerals in the parentheses show $\beta = k_S / k_L$ evaluated for $k_S \simeq 40 - 120$ GPa

gas systems in the 0–50 km depth range (Table 2). Values of ρ^L / ρ^S are estimated as 0.92, 0.33, and 0 for rock + melt, rock + water, and rock + ideal gas systems, respectively. The effect of ρ^L / ρ^S on R_{SP} is small, and practically the same figure as Fig. 10b can be obtained by simply assuming $\rho^L / \rho^S = 1$ ([42] Fig. 4). Figure 10b shows that for a given pore shape α , R_{SP} increases with decreasing liquid compressibility β . Figure 10b also shows that for a fixed liquid compressibility β , R_{SP} varies significantly with the variation of pore shape α . When β is fixed to 25, for example, R_{SP} is smaller than 1 for moderate values of pore aspect ratio ($\alpha > 0.03$), larger than 1 for small aspect ratio (< 0.03), and larger than 2 for very small aspect ratio (< 0.0016). Therefore, R_{SP} provides a good seismological indicator of pore shape.

When $\Delta V_P / V_P^0$ and $\Delta V_S / V_S^0$ are obtained from seismological observations or laboratory experiments, R_{SP} is calculated by taking the ratio of these two. If we know whether the liquid phase is melt, water, or gas, Fig. 10b can be used for estimating the equivalent aspect ratio α from R_{SP} under known β . Without any additional information about the liquid phase, α is estimated from R_{SP} under an assumed β . Figure 10c shows the proportionality coefficient between $\Delta V_S / V_S^0$ and ϕ (the 2nd equation of 35) versus α for various liquid types. By applying α estimated from R_{SP} to Fig. 10c, the liquid volume fraction ϕ can be determined from $\Delta V_S / V_S^0$.

Figures 10a–10c are a complete summary of the effects of liquid volume fraction ϕ , pore shape α , and liq-

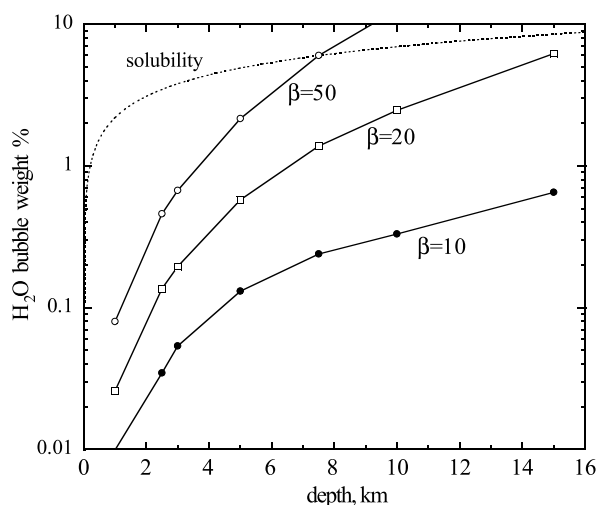
uid compressibility β on V_P and V_S . Using these figures, we can combine and confirm our understandings from the previous forward approaches, such as the different effects of gas, water, and melt, on the V_P / V_S ratio [27,49]. Also, Figs. 10b–10c represent a practical method in the inverse approach to constrain ϕ , α , and/or β from the observation of V_P and V_S . Figures 10a–10c are based on Eqs. (33), (35), and (36) in which the higher-order terms in ϕ ($\phi^{n \geq 2}$) are neglected. A possible variation of R_{SP} with ϕ caused by the higher-order terms in ϕ can be obtained by calculating $(\Delta V_S / V_S^0) / (\Delta V_P / V_P^0)$ directly from Eqs. (31) and (32) without using Eq. (33). Using results from the latter calculations, it is confirmed that the present method based on the linearized equations is valid for $\Delta V_S / V_S^0 \leq 0.35$ for the equilibrium geometry model, which corresponds to $\phi \leq 0.15$, and valid for $\Delta V_S / V_S^0 \leq 0.2$ for the oblate spheroid, tube, and crack models. The applicable range is large for the equilibrium geometry model because the higher-order effects neglected in obtaining Eq. (33) cancel those neglected in obtaining Eq. (35). Compared to the equilibrium geometry model, the applicable range is smaller for the oblate spheroid, tube, and crack models. To use the present method outside the applicable range, a small modification is required ([42], Appendix B).

Application to Seismic Tomographic Images

The information on pore geometry obtainable from seismic tomographic data is the equivalent aspect ratio. Whether the actual geometry is an oblate spheroid or tube, for example, cannot be determined from seismological data. A reasonable interpretation of the derived equivalent aspect ratio requires a knowledge of probable pore geometries in the Earth. Two end-member images of pore geometry have been inferred based on experimental and field observations: one is the equilibrium geometry characterized by a dihedral angle, and the other is thin dikes and veins. In the equilibrium geometry, the pore size is much smaller than the grain size and the permeability is small. For rock + water systems, because the dihedral angle is usually larger than 60° , the equilibrium geometry may have nearly zero permeability. Dikes and veins can develop on much larger scales than the grain size and hence may have much larger permeability than the equilibrium geometry. Therefore, the particular geometry realized between these two end-members significantly affects the liquid migration velocity of the buoyancy ascent. As discussed in Sect. “Overview of Applications”, whether the liquid phase is in the equilibrium shape or not can provide us with valuable information about the degree of interaction between pore geometry and macroscopic dynam-

ics. It is therefore desirable to distinguish between these two end-members using seismic tomographic data. Because the equivalent aspect ratio for the equilibrium geometry is about a factor of 100 larger than that for the thin cracks and dikes (Fig. 10a), the expected values of R_{SP} are significantly different between these two end-member geometries (hatched regions in Fig. 10b); for rock + water systems, R_{SP} is <1 for the equilibrium geometry and >2 for thin cracks and dikes; for rock + melt systems, R_{SP} is 1–1.5 for the equilibrium geometry and >2 for thin cracks and dikes. Therefore, R_{SP} can be used as a seismological indicator of the textural equilibrium, and the information on pore geometry obtained from this indicator can provide us a valuable constraint on actual fluid migration processes in the Earth. Low-velocity regions observed in the mantle wedge beneath Northeastern Japan subduction zone show a systematic change in R_{SP} with depth. Nakajima et al. [24] applied the method introduced in Sect. “Determinability of Porosity and Pore Shape from Elastic Wave Velocities” to seismic tomographic data and inferred a systematic change in pore geometry from an equilibrium geometry at a depth of ~ 90 km to thin cracks and dikes at a depth of ~ 65 km.

Beneath volcanic areas, low-velocity regions with lower V_P/V_S ratio than the surrounding regions are sometimes observed at depths of several km [22]. Rock + melt systems usually have β smaller than 10. This means that R_{SP} is larger than 1 regardless of α (Fig. 10b) so that the



Earth's Crust and Upper Mantle, Dynamics of Solid–Liquid Systems in, Figure 11

The amount of H_2O bubble (weight %) in melt at which the compressibility of the mixture is equal to $\beta = 10, 20$, or 50 , is shown as a function of depth

observed low V_P/V_S ratio cannot be explained by the melt-filled pores. At these shallow depths, however, the melt phase can be a mixture of melt and H_2O vapor, because H_2O initially dissolved in the melt in the deeper reaches of the subduction zone starts to exsolve. Here, we briefly discuss such situation, which is not considered in Fig. 10b and Table 2. Because of the high temperature (900 – 1000°C) of melt, the water phase in the melt is much more compressible than the estimates in Table 2. Figure 11 shows the fraction of water (wt%) above which β of the mixture exceeds 10, 20, or 50. It is shown that at a depth of 3–4 km, $\beta \geq 20$ occurs for the melt containing 0.2–0.5 wt% water as vapor phase, which is realistic in the subduction zone. Therefore, from Fig. 10b, $R_{SP} < 1$ can occur at large α and can explain the observed reduction in the V_P/V_S ratio. If the H_2O vapor in the melt phase can be detected by the low V_P/V_S ratio, we can obtain valuable constraints on the water content of melts and on the evolution of a magma chamber in the crust. However, because melt viscosity is considered to be high in shallow magma chambers, it is important to be careful about wave dispersion, as discussed below.

Dispersion and Attenuation of Waves in Solid–Liquid Composite Systems

In deriving the elastic wave velocities in Sect. “Derivation of Wave Equations”, it was implicitly assumed that the frequencies of the seismic waves are lower than the characteristic frequencies of several relaxation processes inherent to solid-liquid composite systems. If this assumption is not valid, the relaxation processes affect the wave velocities due to dispersion. To assess the applicability of the theoretical results presented in Sect. “Determinability of Porosity and Pore Shape from Elastic Wave Velocities”, I present here a brief discussion of such relaxation processes. To show the mutual relationship between dispersion and attenuation, I consider the relaxation mechanism predicted from Eqs. (29)–(30). This mechanism was first studied by Biot [5,6] and is hereafter referred to as the Biot mechanism. The behavior obtained for the Biot mechanism describes the fundamental characteristic of relaxation. Several other relaxation mechanisms inherent to the solid-liquid composites also affect wave propagation and these are further summarized below.

Let ω and k be the angular frequency and wave number, respectively. By substituting the traveling wave solutions $\Omega_S = \Omega_S^0 e^{-i(\omega t - kx)}$ and $\Omega_L = \Omega_L^0 e^{-i(\omega t - kx)}$ into Eq. (29), the dispersion relation, under which non-trivial solutions exist, is obtained as

$$\left(\frac{k}{\omega}\right)^2 = \frac{\rho_U}{N} \cdot f(\omega), \quad (37)$$

where the complex function $f(\omega) = f_1(\omega) + if_2(\omega)$ is explicitly written as

$$\begin{cases} f_1(\omega) = 1 + \frac{\Delta}{1 + \left(\frac{\omega}{\omega_c}\right)^2} \\ f_2(\omega) = \frac{\Delta \cdot \left(\frac{\omega}{\omega_c}\right)}{1 + \left(\frac{\omega}{\omega_c}\right)^2}, \end{cases} \quad (38)$$

and $\rho_U = (1 - \phi)\rho^S$, $\rho_R = (1 - \phi)\rho^S + \phi\rho^L$, $\Delta = (\rho_R - \rho_U)/\rho_U$, and $\omega_c = \eta_L\phi/(k_\phi\rho^L)$. The phase velocity V and attenuation Q^{-1} are defined by $k/\omega = V^{-1}(1 + i/(2Q))$. By assuming Q^{-1} to be small ($f_2 \ll f_1$), we obtain

$$\begin{cases} V = \sqrt{\frac{N}{\rho_U \cdot f_1(\omega)}} \\ Q^{-1} = \frac{f_2(\omega)}{f_1(\omega)}. \end{cases} \quad (39)$$

The phase velocity V and attenuation Q^{-1} given by Eqs. (39) are shown in Fig. 12 as functions of the normalized frequency ω/ω_c . Both dispersion and attenuation occur near $\omega/\omega_c = 1$ and the total amplitude of dispersion and peak value of Q^{-1} are equal;

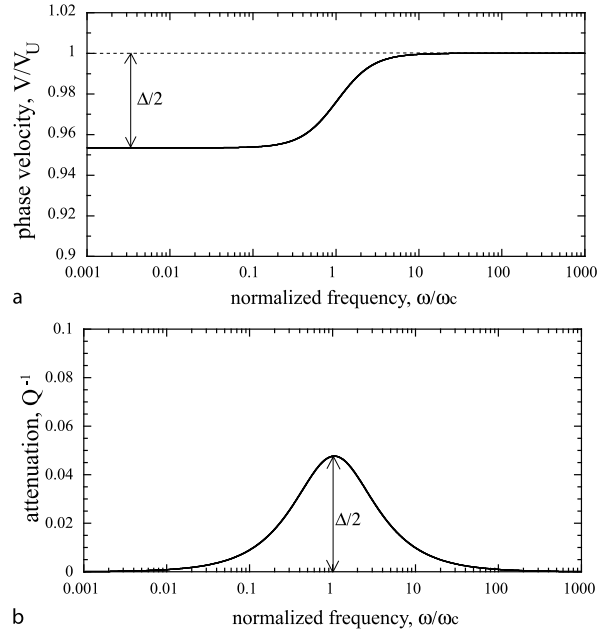
$$Q_{\text{MAX}}^{-1} = \frac{\Delta}{2} = \frac{V_U - V_R}{V_U}, \quad (40)$$

where V_U (unrelaxed velocity) represents V at $\omega/\omega_c \gg 1$ and V_R (relaxed velocity) represents V at $\omega/\omega_c \ll 1$.

Although obtained for the Biot mechanism, Eqs. (38) and (39) describe the fundamental characteristics of dispersion and attenuation regardless of the individual mechanism. These equations simply mean that the dispersion and attenuation are caused by a relaxation process. When a response $J(t)$ (e. g., strain) of a medium to a constant unit force (in the form of a Heaviside function $H(t)$) applied at $t \geq 0$ is not instantaneous but shows a time delay expressed in the form of

$$J(t) = J_U [1 + \Delta \cdot (1 - e^{-\omega_c t})] \cdot H(t), \quad (41)$$

this phenomenon is called relaxation. In other words, Eq. (41) gives a phenomenological model of relaxation. The relaxation process is characterized by relaxation strength Δ and relaxation time scale ω_c^{-1} . The time derivative of Eq. (41), $\dot{J} = J_U[\delta(t) + \omega_c \Delta e^{-\omega_c t} H(t)]$, yields the impulse response, where the Fourier transform of \dot{J}/J_U is equal to $f(\omega) = f_1(\omega) + if_2(\omega)$ with $f_1(\omega)$ and $f_2(\omega)$ given by Eq. (38). Therefore, Eq. (38), which is called Debye equation [25], is a phenomenological model of relaxation in the frequency domain.



Earth's Crust and Upper Mantle, Dynamics of Solid-Liquid Systems in, Figure 12

a Frequency-dependent phase velocity (dispersion) and **b** Q^{-1} (attenuation) described by Debye equations (38)

Solid-liquid composite systems show several relaxation mechanisms caused by the liquid phase. One example is the Biot mechanism exemplified above, in which the density of the system relaxes from ρ_U to ρ_R . At $\omega/\omega_c \gg 1$, due to a dominant effect of liquid inertia, a wave field cannot cause motion in the liquid phase. Hence, ρ_U is associated only with the solid mass. At $\omega/\omega_c \ll 1$, due to a dominant effect of viscous drag force, relative motion does not occur between the solid and liquid phases. Hence, ρ_R is associated with the total mass. The characteristic frequency ω_c is estimated as 200 kHz for water-saturated sandstone ($k_\phi = 10^{-12} \text{ m}^2$, $\eta_L = 10^{-3} \text{ Pa s}$, $\rho^L = 10^3 \text{ kg/m}^3$, and $\phi = 0.2$). Because k_ϕ in the mantle is usually smaller than the value for sandstone, ω_c is usually much higher than the seismic frequency. The Biot mechanism for the longitudinal waves also has the value of ω_c much higher than the seismic frequency. Therefore, the assumption of the relaxed state is valid for the Biot mechanism.

Another relaxation mechanism is squirt flow [19]. When the pore shape is not spherical, changes in pore pressure induced by the elastic waves depend on the aspect ratio and orientation of each pore. Therefore, liquid pressure becomes heterogeneous at the microscopic scale, and this pressure heterogeneity is relaxed by viscous flow of the liquid between pores (squirt flow). This causes a relax-

ation of the skeleton moduli from N_U and K_{bU} to N_R and K_{bR} . The relaxation strength of N , $(N_U - N_R)/N_R$, is generally larger than that of K_b , $(K_{bU} - K_{bR})/K_{bR}$, and therefore the effect of squirt flow is larger on V_S than on V_P . In Sect. “General Theoretical Framework to Describe the Dynamics of Solid–Liquid Composite Systems”, the liquid pressure was assumed to be homogeneous within REV and hence the velocities obtained in Sect. “Derivation of Wave Equations” represent relaxed velocities. When α is large, the relaxation strength is small and the difference between the relaxed and unrelaxed velocities is not significant. As α becomes smaller, the relaxation strength becomes larger, and hence the relaxed velocities can be used only when liquid pores are not isolated and the frequency of the waves are much lower than $\omega_c = k_S \alpha^3 / \eta_L$ [28]. For water ($\eta_L = 10^{-3}$ Pa s) and basaltic melt ($\eta_L = 1 - 10^3$ Pa s), ω_c approaches the seismic frequency when α is smaller than $10^{-2} - 10^{-3}$. If the pore orientation is not random, even under the relaxed state of squirt flow, the pore pressure is different from that estimated for the random orientation. Therefore, when V_P and V_S derived in Sect. “Determinability of Porosity and Pore Shape from Elastic Wave Velocities” are applied to $\alpha < 10^{-2} - 10^{-3}$, the connectivity of the pores, characteristic frequency of squirt flow, and the randomness of pore orientation should all be checked.

As exemplified in the previous section, the presence of H_2O vapor in the melt phase can relax the liquid compressibility from that of a pure melt (β_U) to that of a water–melt mixture (β_R). Figure 11 is obtained by assuming that the characteristic frequency ω_c is much higher than the seismic frequency. However, ω_c decreases with increasing viscosity of the melt. For andesitic and rhyolitic melts, the characteristic frequency is close to or lower than the seismic frequency range [12].

Future Directions

One practical problem limiting the determinability of porosity and pore geometry from seismological data lies in the difficulty of accurately estimating ΔV_P and ΔV_S . For example, a low-velocity anomaly in the upper mantle is generally caused by the superposition of high-temperature anomaly and partial melting. In order to determine porosity and pore geometry from these data, ΔV_P and ΔV_S associated only to the existence of liquid-filled pores (hereafter referred to as poroelastic effect) should be estimated by accurately correcting the data for the temperature effect. Recent experimental studies on the elastic properties of melt-free olivine polycrystals performed in the seismic-frequency range have demonstrated that at $T > 1000^\circ\text{C}$ and at such low frequency, the temperature effect consists of

both anharmonic and anelastic effects [14]. Unlike the anharmonicity, anelasticity cannot be measured by the usual experimental methods using ultrasonic waves. Because experimental data on anelasticity are still limited and the detailed mechanism of anelasticity with or without melt is poorly understood, correction of the data for the temperature effect is difficult. Also, in the crust, accurate estimation of ΔV_P and ΔV_S for the poroelastic effect is difficult because the effect of anelasticity has not been assessed under crustal conditions and also because the lithological heterogeneity is considered to be larger than in the mantle. Therefore, the separation of poroelastic, temperature, and lithological effects affecting the velocity perturbations is an important subject of future study. Recently, not only the V_P and V_S structures but also the three-dimensional Q_P and/or Q_S (seismic attenuation) structures and two-dimensional or three-dimensional electrical conductivity structures have become available. These structures provide additional information on liquid-filled pores, temperature anomaly, and/or lithological heterogeneity. The Q_P and/or Q_S structures, for example, are important in constraining the magnitude of the anelastic effect [24]. Although the separation of individual factors is difficult to determine from velocity structures alone, additional information from Q and/or electrical conductivity structures will be very valuable.

Acknowledgments

The original and more simplified form of this article was published in Japanese [43]. I especially thank Tokyo Geographical Society, for the permission to use a modified version of figures and limited text. I thank S. Nagumo for helpful discussions. I also thank B. K. Holtzman and B. Chouet for reading the manuscript and providing helpful comments.

Bibliography

1. Aharonov E, Whitehead JA, Kelemen PB, Spiegelman M (1995) Channeling instability of upwelling melt in the mantle. *J Geophys Res* 100:20433–20450
2. Baba K, Chave AD, Evans RL, Hirth G, Mackie RL (2006) Mantle dynamics beneath the East Pacific Rise at 17°S : Insights from the mantle electromagnetic and tomography (MELT) experiments. *J Geophys Res* 111:B02101. doi:10.1029/2004JB003598
3. Barcion V, Richter FM (1986) Nonlinear waves in compacting media. *J Fluid Mech* 164:429–448
4. Berryman JG (1980) Long-wavelength propagation in composite elastic media 2: Ellipsoidal inclusions. *J Acoust Soc Am* 68:1820–1831
5. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 1, Low-frequency range. *J Acoust Soc Am* 28:168–178

6. Biot MA (1956) Theory of propagation of elastic waves in a fluid-saturated porous solid, 2, Higher frequency range. *J Acoust Soc Am* 28:179–191
7. Drew DA (1983) Mathematical modeling of two-phase flow. *Annu Rev Fluid Mech* 15:261–291
8. Duffy J, Mindlin RD (1957) Stress-strain relations and vibrations of a granular medium. *J Appl Mech* 24:585–593
9. Hasegawa A, Yamamoto A (1994) Deep low-frequency micro-earthquakes in or around seismic low-velocity zones beneath active volcanoes in northeastern Japan. *Tectonophysics* 233:233–252
10. Holness MB (1997) Surface chemical controls on pore-fluid connectivity in texturally equilibrated materials. In: Jamveit B, Yardley B (eds) *Fluid flow and transport in rocks*. Chapman and Hall, London, pp 149–169
11. Holtzman BK, Groebner NJ, Zimmerman ME, Ginsberg SB, Kohlstedt DL (2003) Stress-driven melt segregation in partially molten rocks. *Geochem Geophys Geosyst* 4:8607, doi:10.1029/2001GC000258
12. Ichihara M (1997) Mechanics of viscoelastic liquid containing bubbles; implications to the dynamics of magma. Ph D thesis, Univ. of Tokyo (in Japanese)
13. Iwamori H (1994) ^{238}U , ^{230}Th , ^{226}Ra - and ^{235}U , ^{231}Pa disequilibria produced by mantle melting with porous and channel flows. *Earth Planet Sci Lett* 125:1–16
14. Jackson I, Fitz Gerald JD, Faul UH, Tan BH (2002) Grain-size-sensitive seismic wave attenuation in polycrystalline olivine. *J Geophys Res* 107(B12):2360, doi:10.1029/2001JB001225
15. Kuster GT, Toksöz MN (1974) Velocity and attenuation of seismic waves in two-phase media, 1, Theoretical formulations. *Geophysics* 39:587–606
16. Masterlark T, Wang HF (2002) Transient stress-coupling between the 1992 landers and 1999 Hector Mine, California, earthquakes. *Bull Seism Soc Am* 92:1470–1486
17. Mavko GM (1980) Velocity and attenuation in partially molten rocks. *J Geophys Res* 85:5173–5189
18. Mavko G, Mukerji T, Dvorkin J (1998) *The Rock Physics Handbook*. Cambridge University Press, New York
19. Mavko GM, Nur A (1975) Melt squirt in the asthenosphere. *J Geophys Res* 80:1444–1448
20. McKenzie D (1984) The generation and compaction of partially molten rock. *J Petrol* 25:713–765
21. Mura T (1987) *Micromechanics of defects in solids*, 2nd edn. Martinus Nijhoff Publishers, Dordrecht
22. Nakajima J, Hasegawa A (2003) Tomographic imaging of seismic velocity structure in and around the Onikobe volcanic area, northeastern Japan: implications for fluid distribution. *J Vol Geotherm Res* 127:1–18
23. Nakajima J, Matsuzawa T, Hasegawa A, Zhao D (2001) Three-dimensional structure of Vp, Vs, and Vp/Vs beneath the northeastern Japan arc: Implications for arc magmatism and fluids. *J Geophys Res* 106:21843–21857
24. Nakajima J, Takei Y, Hasegawa A (2005) Quantitative analysis of the inclined low-velocity zone in the mantle wedge of northeastern Japan: A systematic change of melt-filled pore shape with depth and its implications for melt migration. *Earth Planet Sci Lett* 234:59–70
25. Nowick AS, Berry BS (1972) *Anelastic relaxation in crystalline solids*. Academic Press, New York
26. Obara K (2002) Nonvolcanic deep tremor associated with subduction in Southwest Japan. *Science* 296:1679–1681
27. O'Connell RJ, Budiansky B (1974) Seismic velocities in dry and saturated cracked solids. *J Geophys Res* 79:5412–5426
28. O'Connell RJ, Budiansky B (1977) Viscoelastic properties of fluid-saturated cracked solids. *J Geophys Res* 82:5719–5735
29. Ohmi S, Obara K (2002) Deep low-frequency earthquakes beneath the focal region of the Mw 6.7 2000 Western Tottori Earthquake. *Geophys Res Lett* 29:1807, doi:10.1029/2001GL014469
30. Ribe N (1985) The deformation and compaction of partially molten zone. *Geophys J R astr Soc* 83:487–501
31. Rice JR, Cleary MP (1976) Some basic stress diffusion solutions for fluid-saturated elastic porous media with compressible constituents. *Rev Geophys* 14:227–241
32. Riley GN, Kohlstedt DL (1991) Kinetics of melt migration in upper mantle-type rocks. *Earth Planet Sci Lett* 105:500–521
33. Schäfer K (ed) (1980) *Landolt-Börnstein Numerical Data and Functional Relationships in Science and Technology, New Series IV, vol 4, High-Pressure Properties of Matter*. Springer, Berlin
34. Scott DR, Stevenson DJ (1984) Magma solitons. *Geophys Res Lett* 11:1161–1164
35. Spiegelman M, Kelemen PB (2003) Extreme chemical variability as a consequence of channelized melt transport. *Geochem Geophys Geosyst* 4:1055, doi:10.1029/2002GC000336
36. Spiegelman M, Kelemen PB, Aharonov E (2001) Causes and consequences of flow organization during melt transport: The reaction infiltration instability in compactible media. *J Geophys Res* 106:2061–2077
37. Spiegelman M, McKenzie D (1987) Simple 2-D models for melt extraction at mid-ocean ridges and island arcs. *Earth Planet Sci Lett* 83:137–152
38. Stevenson DJ (1986) On the role of surface tension in the migration of melts and fluids. *Geophys Res Lett* 13:1149–1152
39. Stevenson DJ (1989) Spontaneous small-scale melt segregation in partial melts undergoing deformation. *Geophys Res Lett* 16:1067–1070
40. Stolper E, Walker D, Hager BH, Hays JH (1981) Melt segregation from partially molten source regions: The importance of melt density and source region size. *J Geophys Res* 86:6261–6271
41. Takei Y (1998) Constitutive mechanical relations of solid-liquid composites in terms of grain-boundary contiguity. *J Geophys Res* 103:18183–18203
42. Takei Y (2002) Effect of pore geometry on Vp/Vs: From equilibrium geometry to crack. *J Geophys Res* 107(B2):2043, doi:10.1029/2001JB000522
43. Takei Y (2005) A review of the mechanical properties of solid-liquid composites. in Japanese, *J Geography* 114(6):901–920
44. Tsumura N, Matsumoto S, Horiuchi S, Hasegawa A (2000) Three-dimensional attenuation structure beneath the northeastern Japan arc estimated from spectra of small earthquakes. *Tectonophysics* 319:241–260
45. von Bagen N, Waff HS (1986) Permeabilities, interfacial areas and curvatures of partially molten systems: Results of numerical computations of equilibrium microstructures. *J Geophys Res* 91:9261–9276
46. Waff HS (1980) Effects of the gravitational field on liquid distribution in partial melts within the upper mantle. *J Geophys Res* 85:1815–1825
47. Walsh JB (1969) New analysis of attenuation in partially melted rock. *J Geophys Res* 74:4333–4337

48. Wang HF (2000) Theory of linear poroelasticity with applications to geomechanics and hydrogeology. Princeton University Press, Princeton, New Jersey
49. Watanabe T (1993) Effects of water and melt on seismic velocities and their application to characterization of seismic reflectors. *Geophys Res Lett* 20:2933–2936
50. Watt JP, Davies GF, O'Connell RJ (1976) The elastic properties of composite materials. *Rev Geophys* 14:541–563
51. Wong T-F, Ko S, Olgaard DL (1997) Generation and maintenance of pore pressure excess in a dehydration system 2, Theoretical analysis. *J Geophys Res* 102:481–4852

Ecological Complexity

BRIAN A. MAURER

Department of Fisheries and Wildlife, Graduate Program in Ecology, Evolutionary Biology, and Behavior, Michigan State University, East Lansing, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction: The Nature of Ecological Complexity](#)

[Describing Ecological Complexity](#)

[Survey of Different Solutions](#)

[to the Problem of Describing Ecological Complexity](#)

[Future Directions: Complexity and Complementarity](#)

[Cross References](#)

[Bibliography](#)

Glossary

Ascendancy The tendency, in the absence of disturbances, for an ecosystem to increase in size or total throughput and to have more constrained pathways for within system flows.

Bioinformatics The storage, processing, and analysis of very large arrays of biological data.

Dispersal limitation Limitation of the number of species within an ecological community due to decreased probabilities of some species entering a local ecological community by dispersal.

Ecological community The collection of individual organisms of different species that are found within the boundaries of an ecosystem.

Ecological drift Random changes in the relative abundances of species within a community due to stochastic population processes.

Ecosystem an arbitrary ensemble of macroscopic matter that captures, stores, and uses energy to circulate and rearrange matter within the system.

Emodied energy (emergy) Potential energy stored in chemical bonds within an ecological entity (organism, population, community, etc.).

Food web A network describing the flows of energy and matter within an ecosystem.

General metabolic equation Phenomenological description of how mass, temperature, and resource concentration affect the metabolic rate of an organism or an ensemble of organisms.

Metabolic scaling The exponential relationship of average body mass with the rates of many metabolic processes.

Metacommunity A collection of many local communities aggregated are larger spatial scales.

Tranformity The total amount of solar energy required to form a unit of biological material.

Definition of the Subject

Living systems are collections of entities at multiple scales (e.g., cells, organisms, populations) that undergo a wide variety of interactive processes. Simply by the sheer magnitude of the different possible behaviors of such systems, the problem of describing and understanding ecological processes is daunting. In addition to the size problem, ecological systems have large numbers of unique parts, the behavior of which can vary considerably in space and time. Because of the complexity inherent in ecological ensembles, devising adequate methods to describe, analyze, and predict their behavior is a major challenge to science. Ecological systems combine both idiosyncratic, unpredictable outcomes with strong constraints on system structure that makes them paradoxically both deterministic and unpredictable at the same time. Because of this, has been no universal theory to guide research on ecological phenomena. What is needed is the development of complementary approaches that emphasize some important aspect of the greater whole of an ecosystem, but are amendable to synthesis with other approaches that emphasize other aspects. Thus, ecological complexity presents a unique challenge to science that will require a wide variety of approaches and conceptual infrastructures.

Introduction: The Nature of Ecological Complexity

Ecological systems are large ensembles of macroscopic matter that capture energy, store it, and use it to circulate and rearrange matter within the system. The term “use energy” aptly describes the material flows within an ecological system because information-based processes determine both the material structure of the system and its function. Because of the vast number of possible config-

urations ecological systems can take on, determining fundamental properties and cause-effect relationships within such systems is fraught with many conceptual and empirical challenges.

Energy flow through ecological systems occurs via a hierarchical arrangement of matter starting at microscopic scales and continuing up to encompass all life on earth. At the microscopic scale, complex molecules located within the bodies of plants and microbes intercept high-energy photons. The captured energy drives a variety of molecular pathways that ultimately produce potential energy stored in chemical bonds of complex polymers of glucose. The energy capture system is located in small organelles called chloroplasts, which are part of large, highly organized molecular ensembles called cells. Chemical work done by the transformation of potential energy in chemical bonds maintains cellular cohesion and reproduction. The information regarding the configuration of cellular structure and execution of chemical work is stored on large highly organized polymers of nucleotides and ribose (DNA and RNA).

Chemical work generated by cells is expended in a variety of ways beyond maintenance of intracellular cohesion and information processing. In prokaryotes, individual cells form networks of interactions among themselves and their environments. These interactions involve secretions of chemicals produced within cells, energy-driven movement of cells within complex media, and complex interactions with other cells of the same or different species. In multicellular organisms, cellular interactions proceed along canonized sequences of cell division and proliferation determined by intracellular information contained on DNA. Cell ensembles follow developmental pathways driven by energy derived from stored cellular potential energy. Cell proliferation and diversification ultimately results in organisms, which obtain energy either through photosynthesis or by ingestion of potential energy stored in the cells of other organisms.

Organisms form various types of ensembles that give rise to a variety of potential ways of describing them. Because the same organism can be part of several different types of ensembles, a great deal of confusion in terminology and conceptualization ensues. Furthermore, the spatial and temporal structure of these ensembles makes most attempts at identifying system boundaries for them at best arbitrary. Terms like population, species, community, ecosystem, and biome lack rigorous, consistent definitions, and often mean different things in different situations. The necessity for arbitrary system definitions makes the study of ecological complexity above the organismal level particularly challenging.

Describing Ecological Complexity

Three general types of phenomena constitute the complex spatial and temporal structure of ecological systems: (1) energetic phenomena resulting from energy transfer and storage; (2) kinetic phenomena describing changes in amounts of substances or entities; and (3) informational phenomena describing information transfer and storage within and among ecological entities. These phenomena are at the same time complementary and interacting. Explicit description of one aspect often requires implicit assumptions regarding other aspects. Presumably, the same sequence of events in an arbitrarily defined ecological system could be described from the perspective of any one of these three aspects, although it may be more straightforward to develop one type of description for a specific phenomenon.

Ecological Energetics

Because of the energy gradient necessitated by the second law of thermodynamics, maintenance of any ecological system requires a constant influx of energy. Thus, any change in the structure or function of a system requires an exchange of energy, requiring the eventual replacement of expended potential energy. Failure to replenish stored energy results in dissolution of the system.

Although simple in concept, it is difficult in practice to describe an arbitrary ecological system as a purely energetic phenomenon. Because most of the energy contained in such systems is stored potential energy, the mass of the compounds storing that energy must be measured. It is very difficult to separate out compounds that do not store energy from those that do, so typically, either the total mass of the system or the mass of some important element (most often carbon) is measured. Since energy exchanges occur on the microscopic scale, this necessary simplification is justifiable, but also creates further difficulties. Describing the energetic aspects of the system as mass storages and flows makes concurrent definition of energy losses difficult. Energy is lost as heat, so in order to resolve the measurement scales, either mass must be expressed in energetic units, or heat loss must become an implicit assumption of the description. This creates a mismatch between the units of measure used to describe the system (g) and the units that measure fluxes (kcal).

Despite the conceptual difficulty between relating the microscopic description of system energetics (units of energy) with the macroscopic description (units of mass), for many practical problems the macroscopic description is sufficient. For example, net primary productivity (usually expressed as grams of carbon per unit time per unit area)

shows striking geographical patterns that match the geophysical properties of the earth closely. With recent developments in remote sensing, however, the thermal properties of ecosystems can be measured directly. Furthermore, many important organismal activities are also measurable directly using thermal units. It is often difficult to reconcile mass-based descriptions of the energetic properties of ecological systems with thermal-based descriptions.

Ecological Kinetics

As implied by the preceding discussion, ecological systems exhibit a kind of duality in that they are both energetic systems as well as material systems. When the focus of study is on the material properties of ecological systems, it is often convenient to express state changes of the system in material units. Often the units of interest are concentrations. In a kinetic system description, the energetic mechanisms underlying material flows are not explicitly included.

One of the most useful types of kinetic descriptions models the spatial and temporal properties of ensembles of organisms. Because organisms are fundamental entities in the organization of living matter, many practical applications focus on the kinetics of organisms belonging arbitrary ensembles such as populations and communities. These descriptions, however, require the energetic basis of organismal kinetics to be simplified or only implicitly expressed. Most often, this difficulty arises when incorporating mass transfer from organisms of one species (say a prey organism) to an individual of a different species (e.g., a predator). The actual mechanism of the mass transfer involves significant and complex mechanical and chemical processing, which in turn, requires expenditure of large amounts of stored energy. This complexity cannot be captured by a kinetic description in which the units of measure are organisms.

Further complications for kinetic descriptions arise because, most of the time, the fundamental process of change is discrete, involving addition and subtraction of individual units (organisms). Times between additions and subtractions vary in length. In addition, a large amount of uncertainty may exist regarding the sequence of events leading to these discrete changes. All of these complications create challenges for particular formalisms.

Ecological Information Content and Exchange

There are two very different ways of describing the information content of an ecological system. First, the information contained in cellular DNA profoundly shapes both kinetics and energetics in ecosystems. Fundamental biochemical pathways may differ profoundly among

different kinds of organisms. For example, there are at least three different photosynthetic pathways exhibited by plants, each of which has profound consequences for energy capture efficiency under different ecological conditions. Secondly, ecological systems often exhibit configurations far from thermodynamic equilibrium. The difference between the system state and its thermodynamic equilibrium indicates ecological systems exist in very low probability states resulting from continual energetic expenditures. These low probability states imply a high degree of “organization”, that is, system configurations that are perpetuated by the self-maintenance and self-replicating nature of the constituent energetic and kinetic processes.

With the emergence of molecular biology, description of genetic information content in ecological systems became possible. Although it is sometimes possible to establish direct links between genetic information and specific protein products, it is difficult to connect the spatial and temporal productions of multiple proteins across taxa and relate them to basic ecosystem functions such as matter cycling and energy flow. Furthermore, it is not apparent how to relate genetic information to kinetic descriptions of ecosystems. Since most kinetic descriptions focus on organisms, the link between genetic information content of an organism and its performance in specific environments is difficult to establish and incorporate into modeling efforts.

Genetic information exchange occurs within sexually reproducing species or by other mechanisms such as conjugation in bacteria. However, not all information contained in an ecosystem is genetic. There is a wide variety of chemical and physical signals used by organisms to exchange information about their proximate environment. Pheromones, sounds, visual stimuli, and toxins are examples of such extra-genetic information exchange. This specialized information profoundly influences ecosystem kinetics and energetics.

Another quite different aspect of ecological information is the degree to which an ecological system departs from some low-information content state. The most extreme of such states is thermodynamic equilibrium, where system components have dispensed all of their kinetic and potential energy. Such a state is problematic and of limited utility, since the entire earth is far from thermodynamic equilibrium. Alternatively, a low information state of an ecosystem would exist if its components were apportioned equally among N possible states or configurations. In such a situation, the information about any single state provides information on all other states. Borrowing from information theory, the system would be in a state of maximum

“entropy”, where entropy is a measure of the homogeneity of the system. The more homogeneous a system is, the less information is needed to describe it. Therefore, in this information theoretic sense, entropy is the opposite of information. N equiprobable states correspond to maximum entropy (S), or low information content. Using a maximum entropy approach [1], the lowest information state has an entropy of

$$S_1 = K \log(N),$$

where K is an arbitrary constant. If the probabilities (p_i) of each state differ, then maximum entropy is less than $\log N$, and is given as

$$S_2 = -K \sum_i p_i \log p_i.$$

In ecological applications, K is usually given an arbitrary value of 1. The difference between S_1 and S_2 represents the difference between the actual information content of the system and the lowest possible information content (equiprobability). Difficulties arise when attempting to define a useful set of states that can be related to ecological kinetics or energetics. States that are amenable to kinetic analysis, such as population abundance, are not readily definable in terms of entropy or information. For example, it is not clear how to establish the lowest information configuration of a population.

Survey of Different Solutions to the Problem of Describing Ecological Complexity

It was widely accepted by the end of the nineteenth century that assemblages of plants and animals presented a unique challenge to science. A simple Newtonian mechanics solution to the problem of describing the huge variety of life forms on earth was not possible. At a time when physics was being revolutionized, biologists were often left to enumerating ecological phenomena without reference to any attempts to develop a comprehensive theory. Applying methods of the preceding generations of nineteenth century naturalists, ecology became a field that simply cataloged ecological phenomena without reference to testing or developing new theory. The only guiding theory was Darwin's poorly understood ideas regarding evolution. Ecologists did not test hypotheses about evolution; rather, they provided verbal descriptions of natural systems that were at best confirmatory.

The practice of ecology as descriptive natural history began to change early in the twentieth century. Mathematical probability and statistical physics were maturing, and new tools became available for application to the problem

of ecological complexity. By the end of the twentieth century, a myriad of approaches had developed, each of which made major strides towards finding a comprehensive solution to ecological complexity. The advantages and drawbacks of these different solutions are examined in what follows.

Physical Biology

One of the first attempts to develop a comprehensive theoretical structure for ecology was Lotka's concept of “physical biology” [2]. Lotka's ideas had far-reaching implications that several scientific generations of ecologists used in developing specific types of solutions to the problem of complexity. Many of his ideas were co-opted for specific problems, and after passing through several generations of ecologists, the original context of his ideas were lost. Lotka's vision, however, persists as a fundamental attempt to integrate ecological complexity with the broader framework of theoretical physics. Although both physics and ecology have undergone radical changes since his time, Lotka's insights permeate many of the theoretical approaches to ecological complexity developed over the decades since he formulated them.

Using fundamental insights from thermodynamics and physical chemistry, Lotka recognized that kinetic and energetic descriptions of biological systems were complementary. He viewed the kinetic problem as the formulation and solution of a set of differential equations representing the different material states that comprised an evolving biological system. The parameters for these equations described both the interactions among the material states as well as any influences from outside the material system. Most of his examples dealt with populations of organisms, but he also envisioned kinetic descriptions of element cycling and storage. To make the kinetic equations amenable to analysis, Lotka studied the behavior of the material system in the neighborhood of its multi-dimensional equilibrium. This simplification allowed the linearization of the system of differential equations around the equilibrium. Descriptions of transient dynamics in the region of equilibrium were limited to smooth approaches towards or away from equilibrium, periodic oscillations dampening towards equilibrium or increasing in amplitude away from equilibrium, or “stable limit cycles”, where oscillations persist in the system indefinitely.

The energetic description of biological systems Lotka found more difficult to specify. He envisioned the need for a “statistical mechanics” to describe the macroscopic consequences of the myriad of interactions that material entities in biological systems undergo. Here again he consid-

ered the problem from the perspective of a population of individual organisms encountering one another and undergoing interactions that caused the organisms to undergo changes. In these interactions, organisms could be viewed as having “kinetic energy” that was exchanged via the type of interaction the organisms engaged in (e.g., a prey organism transfers its kinetic energy to a predator). Ultimately, each ensemble of organisms (e.g., populations of predators and prey) increased or decreased in size. The rate of this increase, Lotka thought, could be related to general boundary constraints imposed by the environment to create macroscopic laws isomorphic with physical laws like the ideal gas law.

Neither information theory nor genetics were available at the time Lotka was conducting his early work on physical biology. Although he was aware, for example, of Fisher’s early work in genetics, he was unable to elaborate much on the nature of ecological information. The closest he came to addressing that problem was in his comments regarding what he called “intra-species” evolution. Lotka appreciated that within any given species, individuals were not identical and could not be treated as simple particles with only a few properties relevant to the kinetics and energetics of the systems that contained them. He understood the significance of this variability to defining system level properties, but was unable to move much beyond outlining the general nature of this problem.

These early insights provided the conceptual basis for a number of different approaches to ecological complexity that emerged in the later decades of the twentieth century. Ultimately, bringing these differing approaches back together into a unifying framework, such as Lotka envisioned, would fuse a strong theoretical foundation for understanding ecological complexity.

Kinetics of Steady State Systems

Because many ecological systems are persistent as recognizable entities over the time periods typically studied by ecologists, the relatively stable configuration lends itself to description as a steady state ensemble. Borrowing from the ecological idea of an environmentally determined “carrying capacity” for each component of the ensemble, a kinetic description of the system’s deviations away from stable states (i.e., equilibria) is derivable by linearization of the unknown underlying kinetics. That is, for each state variable X_i representing a component of the ecological ensemble, a steady state value for that variable (X_i^*) is assumed. Given a description of the (unknown) kinetics for component i as $dx_i/dt = f_i(\mathbf{X})$, where \mathbf{X} is a vector containing variables for each component in the ensemble, ki-

netics around the collective equilibria for all variables (\mathbf{X}^*) is obtained by a Taylor series expansion as

$$f_i(\mathbf{X}) = f_i(\mathbf{X}^*) + \sum_j \partial f_i / \partial X_j (X_j - X_j^*) + \mathbf{O}_i(\mathbf{X} - \mathbf{X}^*),$$

where j indexes over all system components, and \mathbf{O}_i contains higher order terms of deviations away from equilibria. Since $f_i(\mathbf{X}^*) = 0$, and the higher order terms are assumed to be close to zero, evaluating the partial derivatives at \mathbf{X}^* leads to a linearization of the near equilibrium kinetics [3,4,5,6]. Eigenanalysis of the resulting linear system indicates a limited range of potential behaviors for the system. If all eigenvalues are real, the system converges towards equilibrium or expands away from it. Imaginary eigenvalues indicate either convergent or divergent oscillations, or, if the real parts of all imaginary eigenvalues are zero, sustained periodic oscillations.

The linearized solution to ecological dynamics is most appropriate over relatively short time spans. Because organisms are somewhat resistant to change, kinetics of ecological systems are expected to maintain inertia in the face of changing environments. However, the interpretation of this inertia is somewhat different from standard physical models. In physical models, a force applied to a mass must overcome the inertia of that mass for the system state to change. In a sense, the system is pushed into a new state by expenditure of energy. Ecological systems generally change from within. The environment acts as a rate control on internal processes within such systems, so that changes in the state of the system are accomplished by shifting the balance among competing internal processes. The inertia of the system corresponds to the finite time lag over which internal processes reconfigure. Over appropriately scaled time spans, the system can be thought of as fluctuating around a relatively static state. However, over longer time spans, the static state disappears, limiting the usefulness of the linearized description.

Systems Analysis

The linearization of ecological kinetics has an implicit assumption of closure. The state variables, \mathbf{X} , are assumed to completely specify the system. Unmeasured quantities that may impinge on that description do so primarily through the constants obtained through the Taylor expansion. However, ecological systems are almost never perfectly closed, and it is therefore necessary to recognize how inputs to and outputs from them influence the internal workings of the system. Following a general trend in science to consider phenomena as coherent systems [7], ap-

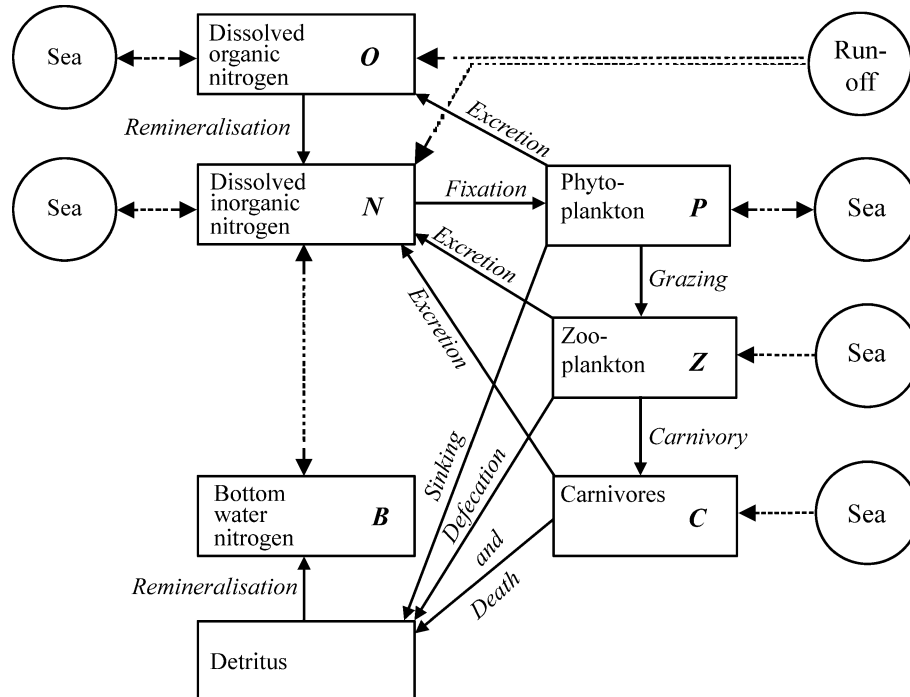
plication of systems analysis to ecological complexity extends the range of describable behaviors.

Systems theory recognizes that complex behavior arises both from the internal workings of the system and its connections to other systems with which it is associated. Given a set of state variables, the effects of variables on one another can generate “feedback” loops. That is, any given state variable can affect its own kinetics by increasing effects of other variables on itself (positive feedback) or decreasing them (negative feedback). Each state variable is connected either directly or indirectly with many other variables, so the balance between positive and negative feedbacks determines the overall impact of the variable on itself. Steady state is achieved when positive and negative feedbacks are nearly equivalent across all state variables. Because the balance between positive and negative feedbacks can never be precise, the expectation is that a system that maintains integrity over time will show bounded oscillations.

Ecologists typically choose state variables as the amount of a particular element (say nitrogen) or as the amount of potential energy (usually expressed as the amount of organic carbon). Each state variable is repre-

sented as the amount of matter or potential energy located in different intermediate locations in the transfer of mass from one group of organisms to another. Different types of organisms obtain matter/energy in different ways. Producers generate potential energy from sunlight and absorb matter (hereafter called nutrients) from their surroundings. Consumers obtain energy and nutrients from other organisms. Nutrients are recycled through the physical environment by decomposition of dying and dead organisms or their parts. By tracing the route by which nutrients and energy are exchanged in an ecosystem, it is possible to describe a complex network, where nodes represent organisms obtaining nutrients and energy from the same sources. There are several approaches to describing these networks, or food webs.

Kinetic descriptions use differential or difference equations to describe changes in nutrients or energy within nodes (i. e. the state variables X). Including terms for other state variables in the functional form of the differential equation for a given state variable is typically used to represent feedbacks. Information or hypotheses regarding specific interactions among organisms determine the functional form of equations. For example, the amount



Ecological Complexity, Figure 1

Schematic representation of nitrogen fluxes in sea lochs in Scotland. Circles represent influxes into the system. Rectangles represent the various state variables describing nitrogen accumulation in different populations. Arrows represent nitrogen movements among components. System kinetics are modeled using a variety of functional relationships among inputs and state variables. From [8]

of nitrogen obtained by a consumer may saturate beyond a certain nitrogen concentration. Solutions to these sets of equations are most often obtained by numerical methods. Sensitivity analyses that vary key parameters or functional forms gauge the dependence of numerical solutions on different assumptions made about the system. The final step in system modeling is statistical comparison of model output to data collected from the system.

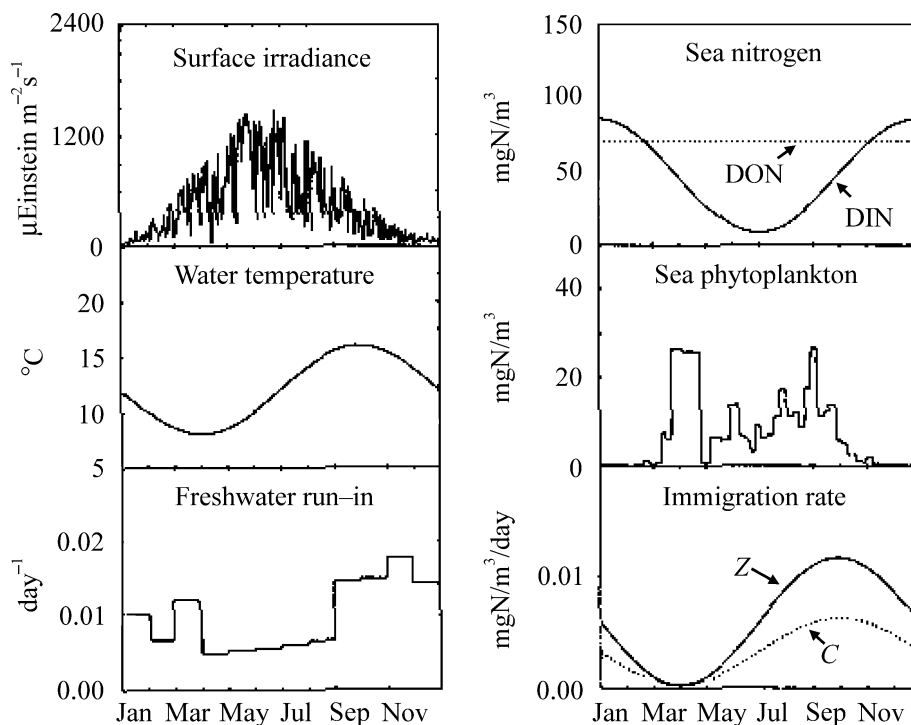
The following example highlights the details of the systems modeling approach. Scottish sea-loch ecosystems are formed along the coast by variation in tidal cycles, which isolate bodies of seawater behind topographic features [8]. Freshwater enters the sea-lochs through runoff from the land and is mixed with saline water at high tide. The network of nutrient flows in these ecosystems is described by a schematic representation of the major nutrient storages and fluxes (Fig. 1). The kinetics of the system are represented by a system of differential equations such as

$$dX_i/dt = f_i(\mathbf{X}) \quad i = 1, 2, \dots, 7,$$

where \mathbf{X} is a vector containing the X_i 's. Rather than linearize the system, each function is determined by the na-

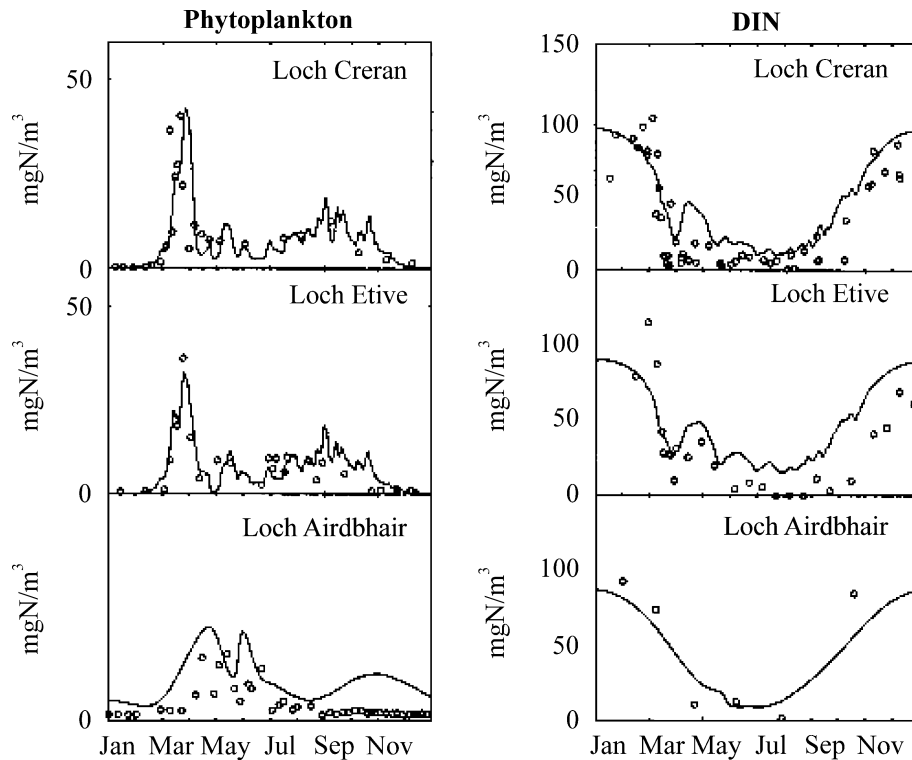
ture of the processes responsible for nutrient fluxes. The system of equations includes terms that connect the state variables \mathbf{X} to influxes from outside the ecosystem. These influxes are "driving functions", that is, they are inputs into the system that are responsible for system change. The system of equations are then numerically solved by dividing time into discrete intervals, and computing the resulting difference equations. The discrete intervals are made to be small relative to the span of time over which the system is being modeled, so the result is somewhat like a numerical integration. At each time step, input from driving functions is determined by data or functions thought to describe temporal variation in those inputs (Fig. 2). Validation of model output is done by comparison to data from specific systems (Fig. 3). Sensitivity analysis is often conducted by varying different parameters or functional forms in the model and determining how such changes influence model output. Once the model has been validated, it is used to forecast system behavior in response to changes in the driving functions.

Another systems-based approach to ecological complexity focuses on the topology of the network describing material flows in ecosystems (see ► [Ecological Topology](#)



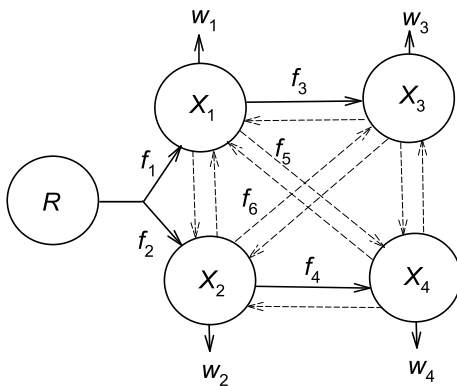
Ecological Complexity, Figure 2

Driving functions used to model exogenous inputs of nitrogen into sea lochs in Scotland. These inputs are incorporated into a variety of functions that relate inputs with nitrogen uptake rates in different organisms. From [8]



Ecological Complexity, Figure 3

Model output for monthly concentrations of nitrogen in two state variables (phytoplankton and dissolved inorganic nitrogen) for three different sea lochs in Scotland compared with field measurements. Note the general agreement between model output and field measurements. From [8]



Ecological Complexity, Figure 4

Network diagram for a relatively simple food web with four nodes. R represents exogenous input into the web, w_i 's represent outfluxes. Under conservation of matter and energy, $R = \sum w_i$. Solid arrows denote flows within the network, while dashed lines represent possible flows. As is typical with most food webs, the network is not over-connected, with flows to higher trophic levels (X_3 and X_4) dominating the network. Adding flows f_5 and f_6 to the web would retain these properties while making the web more complex

and Networks. A food web comprises the network of connections among different components of an ecosystem. These links can be represented as a directed graph, with edges identifying one-way flows of energy and biomass from one node (usually a species population) to others (Fig. 4). Rather than representing the amount of materials flowing through the network, analysis of food webs focuses on the nature of connectedness among food web components. Food webs are not maximally connected, that is, not all nodes in a food web are connected to all others. There is a definite hierarchical arrangement of nodes, with producers (photosynthetic/chemosynthetic organisms) at the basal level and predators at the apices. Edges rarely describe flows moving opposite the direction from base to apex. Many insights from general networking theory can be applied to the analysis of food webs. Simple food webs often have unexpected properties [9]. Furthermore, relatively simple models can describe many of the properties observed in data accumulated on trophic connections among species in nature [10,11,12]. These simplifications, however, are not fundamentally derived from the ther-

modynamics of ecosystems, but rather represent heuristic tools for describing ecosystem organization and the response of that organization to changing conditions.

The most wide-ranging attempt to deal with ecosystems as thermodynamic systems has been H.T. Odum's concept of systems ecology [13]. Beginning from fundamental principles of energy exchange co-opted from Lotka [2], Odum constructed an elaborate framework capable of expressing nearly every ecological phenomenon, from energy exchanges among molecules to the flow of goods and services through economies. The key concept uniting such diverse systems was the idea that all entities in a complex system must ultimately be composed of "embodied" energy, that is, stored energy available to do work. With few exceptions, energy in ecological systems originates from solar radiation captured by photosynthesis, hence, at least in concept, any quantity in an ecological system could be expressed in terms of the amount of solar energy expended to get the quantity to its current state [14,15]. This amount, referred to as the "transformity" of an entity, requires knowledge of the amount of solar energy that flows from its source through a complex ecosystem to its current storage location. For example, Odum [15] calculated that a unit of biological matter containing one joule (J) of energy, had a transformity of approximately 4000 solar emjoules/J. The dimension "emjoule" refers to the embodied energy contained in an entity. Odum's sweeping vision of ecological organization and its consequences for human energy transforming systems (e.g., economies) has yet to be fully realized.

Nonlinear Dynamics

In the late 1970's, the search for general solutions to the dynamics of ecological systems took an unexpected turn. Beginning with May's [16,17] demonstration that simple population models produced unexpectedly complicated behavior under certain conditions, it became apparent that departures from the "linearized" version of ecological kinetics could produce behavior that mimicked persistent fluctuations observed in ecological communities [18]. Further exploration in both ecology and other applications indicated that solutions to arbitrary sets of non-linear kinetic equations under certain conditions could generate a complex topology in the multidimensional state space representing the set of state variables describing a system. The solution set to such equations, often called a "strange attractor", has the paradoxical properties of being a deterministic attractor that produces a unique set of solutions for any arbitrary starting point in the

state space. The attractor represents a complicated "folding" and "compression" of the state space onto a fractal-like solution manifold. An original high dimensional system is mapped onto a surface with a smaller fractal dimension.

Dynamics resulting from systems whose solution is a strange attractor are often referred to using the misleading label of "chaos". Chaos as defined in this way results in microscopic (i.e., the original state variables) uncertainty due to dependence on initial conditions coupled with macroscopic (i.e., the attractor surface) certainty. The nature of the macroscopic attractor can be derived from general properties of the ensemble of local state change vectors. Chaotic attractors have the property that correlations in the time evolution of points that begin in the same region of state space decline as the trajectories resulting from those points follow their unique paths on the attractor surface. This type of behavior departs from the expectation of a linearized system, where points close to one another converge along similar trajectories to the final steady state (which is often a point attractor).

A significant challenge in identifying the existence of deterministic uncertainty is the fact that any ecological system also contains other types of uncertainty, or "randomness". Uncertainty can arise from the inability to specify completely all factors responsible for system change in the kinetic description of the system. Uncertainty can also arise from the indeterminate nature of the behavior of the state variables. In an open ecological system, both types of uncertainty are likely to occur because the specified components of the system are embedded in a more inclusive system that impinges on events occurring within the specified system boundaries. For example, a community of organisms at the same trophic level may be represented by a set of differential equations, yet there may be other species not included explicitly in the kinetics represented by those equations that nonetheless cause changes in the species being modeled. There are a number of approaches that attempt to differentiate between kinetics due to "random motion" in state space from those governed by the "pull" of a strange attractor [19,20].

The importance of uncertainty in the face of arbitrary kinetic specifications of an ecological system has led to many attempts to develop stochastic process models that might account for this uncertainty. In addition, the data themselves may have some degree of measurement error associated with them. Estimating parameters for such models using existing types of data is complicated by the existence of "measurement error", that is, uncertainty due to the relationship between measured quantities and the underlying state variables [21].

Statistical Mechanics

As the size and complexity of a system being modeled increases, the more uncertainty enters into the resulting kinetic descriptions. This is reminiscent of Lotka's [2] vision of a statistical mechanics for ecological systems. During the middle of the twentieth century, a number of attempts to define statistical descriptions of ecological systems that might arise from their complicated dynamics arose [22,23,24,25,26,27]. In the late 1970's and early 1980's, these attempts were abandoned in favor of a "mechanistic" approach to ecology, shored up by increasingly sophisticated experimental approaches [28, 29,30]. This necessary phase in the development of ecology produced a wealth of evidence for the importance of local interactions among species. However, most studies were limited to a few years in length, and few lasted long enough to observe the long-term outcomes of manipulations. Those that lasted longer showed complicated behaviors, suggesting that events occurring outside the experimental system modified the results of experimental manipulations in unexpected ways [5].

Renewed interest in the statistical mechanical approach arose when it was determined that relatively simple assumptions applied to stochastic population change could reproduce a number of statistical patterns observed in large ensembles of species in space and time [31,32,33]. The basic approach begins with the description of the random variable X_{ij} , defined as

$$\begin{aligned} X_{ij} &= 1 && \text{if species } j \text{ has exactly } i \text{ individuals,} \\ &&& \text{if species probability } P_{ij} \\ X_{ij} &= 0 && \text{otherwise, with probability } (1 - P_{ij}). \end{aligned}$$

If the occurrence of species in a community of S species is independent of other species, the number of species that have i individuals at a given steady state is the sum over $j = 1, 2, \dots, S$ of the individual X_{ij} . The expected value of this sum, S_i^* is

$$S_i^* = \sum_j P_{ij}$$

and the variance is

$$\sigma_i^2 = \sum_j P_{ij}(1 - P_{ij})$$

The set of expected values S_i^* , under the assumption that species are approximately equivalent in their demography and migration rates, produces a species relative abundance distribution (SAD).

There are a wide variety of methods to obtain estimates of the probabilities P_{ij} [34,35,36,37,38,39,40,41]. The simplest formulation assumes that species are ecologically equivalent (ecological symmetry) and that the environment in which they interact is homogeneous. Under these assumptions, a stochastic birth-death process ensues which leads to fixation of a single species in a completely isolated local community [42]. That is, $P_{ik} \rightarrow 1$ for species k and $P_{ij} \rightarrow 0$ for all $j \neq k$. The random fluctuations that result from this stochastic birth-death process is referred to as *ecological drift* [31].

For species richness to be greater than unity, the community must be open to influx of individuals of other species [42]. The source of individuals available for immigration into the community is called a *metacommunity*. There are a variety of ways to define a metacommunity, and different definitions result in different expressions for the expected relative frequencies of species (P_{ij}). The fundamental result under ecological symmetry is that the SAD becomes a function of the relative frequencies of species in the metacommunity and the rate of migration into local communities [31]. The form of this function depends on the underlying structure of the metacommunity (e.g., [41]). If migration into a local community is relatively low, the community is *dispersal limited*. A balance between ecological drift and dispersal limitation then determines local species diversity. The challenge then becomes to obtain appropriate descriptions of metacommunities for an ecosystem of interest [43] and deduce the sampling distribution for local communities under the appropriate metacommunity description [41].

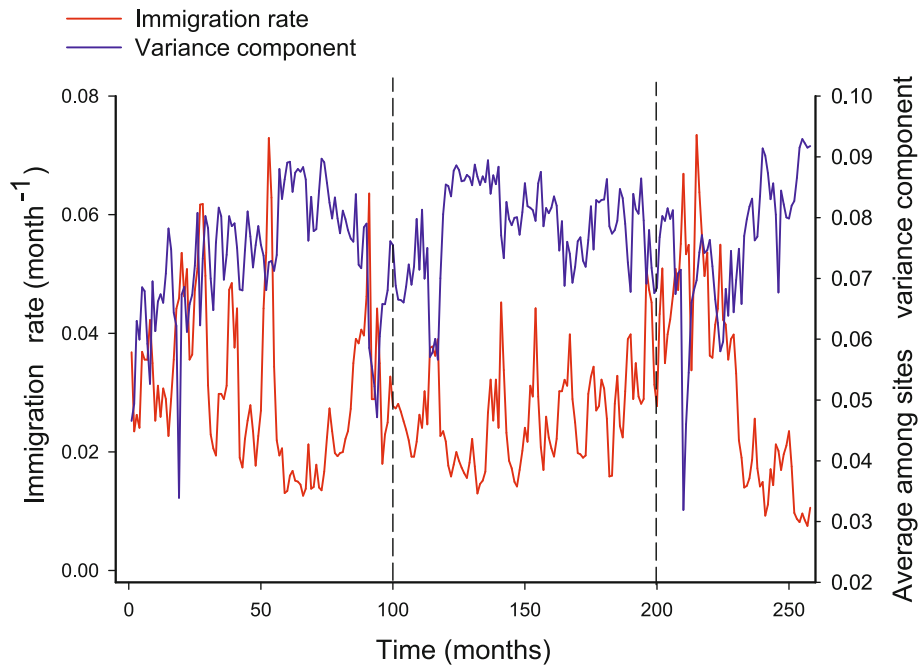
Solutions for SAD's become more complicated under ecological asymmetry. For example, consider the situation where species are ecologically symmetric except for their immigration probabilities. For a sample in a local community containing J individuals, rescaling the P_{ij} 's so that they sum to one, the average probability of a species being found in a community of size J is $1/S$. The rescaled probability for species j (dropping the redundant index i) is p_j , and the variance of its limiting distribution under migration rate m_j is

$$\begin{aligned} \text{var}(p_i) &= [(p_i/J)(J\{1 - p_i\} - m_i\{1 - p_i\})]/ \\ &\quad [m_i(J - 1) + 1 - m_i] \end{aligned}$$

The average variance in relative frequency across species is

$$\sigma_S^2 = \sum_j p_j(1 - p_j)/S$$

Although the migration rate of individual species may not be known, it is possible to find the average migration rate



Ecological Complexity, Figure 5

Temporal patterns of variation in spatial variation in abundance and average immigration rate in a community of desert rodents living in a Chihuahuan desert ecosystem. Vertical dashed lines represent periods of transition among relatively distinct sets of species in response to changes in the local environment. Data provided by S.K. M. Ernest, Utah State University

across species (m^*) as a function of the p_j 's, which is

$$m^* = 2J\sigma_S^2 / (J^2\sigma_S^2 - 2J\sigma_S^2 + 1 - 1/S)$$

For an open community constantly undergoing fluctuations, the average migration rate provides an index of the degree to which the community fluctuations are influenced by dispersal limitation (Fig. 5).

Information Theory

Recall that there are two different types of information that can be used to describe an arbitrary complex ecological system: genetic information and what we might refer to as “configurational” information. This second type of information describes the departure of an aspect of the ecosystem, such as nutrient flows, from some hypothetical “most probable” state. Thermodynamic equilibrium is ultimately most probable, but in the face of import of energy and matter into an ecosystem, transient kinetics can maintain the system far from thermodynamic equilibrium for indefinite lengths of time, causing it to occupy a large range of low probability states. To distinguish among these alternative, low probability states, some baseline is needed for comparison. As mentioned above, if equiprobability of

individual states holds, then there is very little information needed to describe the entire system. This condition implies symmetry exists among components. More accurately, if there are N possible states that the system can occupy, each of which has a relative frequency of p_i , the “entropy” of the system (see ► [Entropy Maximization and Species Abundance](#)) is

$$S = -k \sum_i p_i \log p_i$$

which is maximized when $S = k \log N$, in other words, all relative frequencies are equal. MacArthur [44] introduced S as a measure of ecological system “stability”. Using heuristic arguments, he reasoned that stability in this sense increased either by increasing the number of system states (N) or more evenly distributing relative frequencies among system states. For many reasons, MacArthur’s interpretation has been questioned, but the basic application of information theory in this context is sound. The question has become, what exactly does S represent in ecological systems? The answer hinges on how one defines the N states being considered. In a kinetic system description, the amounts of the state variables X_i are summed and used to calculate S . N represents the number of state variables (e. g., the number of species in the ecosystem or some ar-

bitrary part of it). For a kinetic description, when k is set to unity, S is the *ecological diversity* of the ecosystem. If the state variables are counts of organisms (or other ecological units), setting $k = \sum_i X_i$ gives the likelihood function for a multinomial distribution describing the probability density of the counts. This fact can be used to test hypotheses about the distribution of counts among state variables (assuming stationarity). In an energetic ecosystem description, the states of interest are the flow rates between state variables [45,46]. The p_i 's are the fraction of total system throughput flowing through a specified edge connecting two state variables.

Based on phenomenological considerations, Ulanowicz [45,46] argued that because S represented the degree to which flows in an ecosystem are organized, that is, the degree to which they depart from symmetry of flows among all ecosystem components, it expressed a fundamental quality that changes directionally in an ecosystem over time. Letting k = total system throughput (a measure of system size), S was called the *ascendancy* of an ecosystem [45,46]. Barring any outside disturbances, according to Ulanowicz's phenomenology, ascendancy increases because over time, an ecosystem evolves to maximize the rate of energy processing [2,13], which involves adding additional energy pathways to the ecosystem (effectively increasing N , the number of system units, i.e., energy pathways and/or species). However, there is also a tendency for the pathways that maximize energy flow to increase their share of energy at the expense of other pathways, preventing the system from reaching the maximum possible ascendancy, namely $(\log N) \sum_i X_i$. Ulanowicz argued that all of E.P. Odum's [47] principles of ecosystem succession where specific realizations of the overall tendency of ecosystems to maximize ascendancy.

It is interesting to note that in these sense Ulanowicz used ascendancy, it plays a descriptive role not unlike fitness in Fisher's fundamental theorem of natural selection [48]. In both cases, biological processes operating among a large number of entities lead to increasing levels of organization through competitive dominance among entities balanced by inevitable tradeoffs within the entities. In most cases, this leads to a breaking of symmetry among entities, increasing the amount of information needed to describe the system.

Bioinformatics

Because the amount of information that ecosystems store is vast, unique challenges face scientists attempting to describe that information. This stored information exists as sequences of cellular DNA and RNA, distributions of

gene products in space and time, and higher-level chemical and biophysical transmissions among organisms. Because of the sheer magnitude of this information, processing, storing, and describing data presents unique technological challenges. Solutions to these problems require very large computational capacity, algorithmic efficiency, and a mathematical infrastructure capable of extracting desired information from large databases. *Genomics* is the application of these technologies to large databases on gene sequences, while *ecoinformatics* does the same with large databases on geographic variation of ecologically relevant information. Much of this information involves remote sensors deployed on satellite platforms in orbit around the earth.

The ability to access such vast amounts of biological information opens the door to ask questions that have been unanswerable or even unthinkable in the past. For example, human capacity to modify the earth's ecological systems (see ► [Human-Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development](#)) has increased rapidly as both human populations and resource use have increased nearly exponentially in the past century. These global changes have induced widespread changes in the chemistry of the earth's atmosphere and oceans. Understanding how and why such changes occur is impossible without the ability to process the vast amounts of data required to describe and monitor global ecosystems. Obtaining such an understanding is critical to the long-term ability of humans to persist while maintaining an acceptable quality of life for every person.

Bioinformatics in many ways is a science in its formative stages. Although technologies are evolving rapidly, there are yet to emerge accompanying theoretical constructs that will allow the formulation and testing of hypotheses about how large-scale ecosystems interact and function. These interactions and functions operate on a vast range of scales from molecular to planetary. It is not entirely clear that present mathematical and technological developments are adequate to describe such multiscaled complexity. Because of this, bioinformatics represents a fundamental conceptual frontier into which science is just beginning to explore.

Metabolic Scaling

There has been a long tradition in biology to examine constraints on biological form and function that result from physical limitations [49,50]. Towards the end of the twentieth century, it was discovered that size, in particular, played a fundamental constraining role in the transport of energy through individual organisms [51,52,53]. The scal-

ing is allometric, that is, if M is a measure of size (usually mass) and E is the flow of energy through an individual organism, then

$$E = aM^b.$$

The empirically derived exponent, b , is usually close to 0.75. The coefficient of this scaling can be considered a function of other effects, such as temperature [54] and resource concentration [55,56,57], rendering a *general metabolic equation*. Although phenomenological in its derivation, the equation can be thought of as a macroscopic, statistical-mechanical description of energetic processes involving organisms. The mass effect represents a fundamental design constraint on fractal-like networks that distribute mass, energy, and information within cells and organisms [58]. In organisms, those networks have a minimum size corresponding to scales below which only passive diffusion can move molecular products carrying energy and information. Below this minimum threshold, movement of energy and information is random (Brownian) in nature. Here, temperature exerts its effect on metabolism. Molecular collision is required for almost any cellular action, and under Brownian motion, collisions will occur more rapidly as temperature increases. Many enzymatic-regulated cellular processes operate optimally under a relatively narrow range of temperatures. Within that range of temperatures, the range of energy states that characterize a particular enzymatic reaction can be modeled using a Boltzmann-like function, where the energy state needed to produce the enzyme product is the “activation energy” of the reaction. Metabolism, however, involves many such molecular processes, but at the macroscopic scale of the organism, one can envision an “activation energy of metabolism” that results from the various cellular processes. Hence, the macroscopic description of a metabolic process, E , can be written as

$$E = a'M^b \exp(-E_a/kT)$$

where E_a is the activation energy of the metabolic process, T is temperature (in Kelvin) and k is Boltzmann's constant. The exponential term, in statistical mechanics, represents the fraction of molecules that reach energy level E_a . However, at the macroscopic scale, this formulation can only be approximate, since so many individual reactions are involved in metabolism. Finally, the rate of supply of resources to an organism saturates at high resource density, hence, a Michaelis–Menten type function has been used to represent this final component of metabolism [55,56,59].

The application of these considerations to ecological processes has yielded a number of important insights into

how energy processing in living systems constrains the form and function of organisms and the ensembles they form [60,61]. The metabolic theory in some ways brings both energetic and kinetic descriptions of ecological complexity together in a synthesis that sheds light on both types of processes. Although in its infancy, the metabolic approach to ecology is a promising approach to ecological complexity that is solidly grounded in basic physical principles and links a wide range of ecological phenomena into a coherent theoretical construct.

Future Directions: Complexity and Complementarity

Summing together what is known about ecological complexity and how it is described underscores some fundamental insights into how living systems produce order along energy gradients on the earth. The most basic principle is that the order observed in nature is a direct consequence of thermodynamic principles in open systems. The abundance of high-energy solar radiation on the earth's surface is captured as potential energy that is degraded in relatively small steps to assemble highly complicated physical networks of material and information. These networks perform recursive physical operations that open a vast number of possible routes by which solar energy can travel as it is transformed into biological work and heat.

Because of the huge number of possibilities that life presents, it is difficult, or even impossible to provide a complete description of ecological systems. To this end, scientific approaches to ecological complexity must be necessarily incomplete. The fact that there are at least three ways of describing complexity, as discussed above, is a result of this incompleteness. However, since the same set of ecological phenomena can, at least in principle, be described in several different ways, it may be possible to envision progress by looking for theoretical and empirical complementarities among different descriptions of the same system. The metabolic approach to ecology described in the last section is one such approach. For example, by linking body mass with individual production, it is possible to link kinetic descriptions of population dynamics with energy processing by the constituent organisms [60,62]. This in turn can be used to examine geographic patterns in biological diversity [63].

Ecology is a science on the verge of discovering the fundamental conceptual framework that will unite it with other sciences to portray a more complete picture of how life develops in the one system currently available to scientific study. The principles that will be formulated in the future will provide the necessary paradigms and empirical

approaches that will be needed when, if ever, living systems other than the earth are discovered.

Cross References

This article serves as the introduction to the Ecological Complexity Section. The three additional articles in the section elaborate on portions of this article. They are:

- [Ecological Topology and Networks](#)
- [Entropy Maximization and Species Abundance](#)
- [Human–Environment Interactions, Complex Systems Approaches for Dynamic Sustainable Development.](#)

Bibliography

Primary Literature

1. Jaynes ET, Rosenkrantz RD (1983) E.T. Jaynes: Papers on probability, statistics, and statistical physics. Kluwer, Boston, Dordrecht, Holland; Boston Hingham
2. Lotka AJ (1925) Elements of physical biology. Williams and Wilkins, Baltimore
3. Pielou EC (1977) Mathematical ecology. Wiley, New York
4. Puccia CJ, Levins R (1985) Qualitative modeling of complex systems: An introduction to loop analysis and time averaging. Harvard University Press, Cambridge
5. Maurer BA (1999) Untangling ecological complexity: The macroscopic perspective. University of Chicago Press, Chicago
6. Schaffer WM (1981) Ecological abstraction – the consequences of reduced dimensionality in ecological models. *Ecol Monogr* 51:383–401
7. von Bertalanffy L (1969) General systems theory. George Braziller, New York
8. Gurney WSC, Nisbet RM (1998) Ecological dynamics. Oxford University Press, New York
9. Holt RD, Lawton JH (1994) The ecological consequences of shared natural enemies. *Annu Rev Ecol Syst* 25:495–520
10. Pimm SL (1991) The balance of nature? Ecological issues in the conservation of species and communities. University of Chicago Press, Chicago
11. Cohen JE, Briand F, Newman CM (1990) Community food webs: Data and theory. Springer, Berlin
12. Jonsson T, Cohen JE, Carpenter SR (2005) Food webs, body size, and species abundance in ecological community description. *Adv Ecol Res* 36:1–84
13. Odum HT (1994) Ecological and general systems: An introduction to systems ecology. University Press of Colorado, Niwot
14. Odum HT (1988) Self organization, transformity, and information. *Science* 242:1132–1139
15. Odum HT (1995) Self organization and maximum empower. In: Hall CAS (ed) Maximum power. University Press of Colorado, Niwot, pp 311–330
16. May RM (1975) Biological populations obeying difference equations: Stable points, stable cycles, and chaos. *J Theor Biol* 51:511–524
17. May RM (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
18. Turchin P (2003) Complex population dynamics: A theoretical/empirical synthesis. Princeton University Press, Princeton
19. Parker TS, Chua LO (1989) Practical numerical algorithms for chaotic systems. Springer, Berlin
20. Barahona M, Poon CS (1996) Detection of nonlinear dynamics in short, noisy time series. *Nature* 381:215–217
21. Dennis B, Ponciano JM, Lele SR, Taper ML, Staples DF (2006) Estimating density dependence, process noise, and observation error. *Ecol Monogr* 76:323–341
22. Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. *J Anim Ecol* 12:42–58
23. Kendall DG (1948) On some modes of population growth leading to R.A.Fisher's logarithmic series distribution. *Biometrika* 35:6–15
24. Kerner EH (1957) A statistical mechanics of interacting biological species. *Bull Math Biophys* 19:121–146
25. Kerner EH (1959) Further considerations on the statistical mechanics of biological associations. *Bull Math Biophys* 21:217–255
26. Preston FW (1948) The commonness and rarity of species. *Ecology* 29:254–283
27. Preston FW (1962) The canonical distribution of commonness and rarity. *Ecology* 43:185–215, 410–432
28. Hairston NG (1989) Ecological experiments: Purpose, design, and execution. Cambridge University Press, Cambridge
29. Resetarits WJ, Bernardo J (1998) Experimental ecology: Issues and perspectives. Oxford University Press, New York
30. Underwood AJ (1997) Experiments in ecology: Logical design and interpretation using analysis of variance. Cambridge University Press, New York
31. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton
32. Bell G (2001) Neutral macroecology. *Science* 293:2413–2418
33. Bell G (2000) The distribution of abundance in neutral communities. *Am Nat* 155:606–617
34. Etienne RS (2005) A new sampling formula for neutral biodiversity. *Ecol Lett* 8:253–260
35. Etienne RS (2007) A neutral sampling formula for multiple samples and an 'exact' test of neutrality. *Ecol Lett* 10:608–618
36. Etienne RS, Alonso D (2007) Neutral community theory: How stochasticity and dispersal-limitation can explain species coexistence. *J Stat Phys* 128:485–510
37. Etienne RS, Olff H (2005) Confronting different models of community structure to species-abundance data: A bayesian model comparison. *Ecol Lett* 8:493–504
38. Alonso D, Etienne RS, McKane AJ (2006) The merits of neutral theory. *Trends Ecol Evol* 21:451–457
39. Alonso D, McKane AJ (2004) Sampling hubbell's neutral theory of biodiversity. *Ecol Lett* 7:901–910
40. Volkov I, Banavar JR, Hubbell SP, Maritan A (2003) Neutral theory and relative species abundance in ecology. *Nature* 424:1035–1037
41. Volkov I, Banavar JR, Hubbell SP, Maritan A (2007) Patterns of relative species abundance in rainforests and coral reefs. *Nature* 450:45–49
42. Maurer BA, McGill BJ (2004) Neutral and non-neutral macroecology. *Basic Appl Ecol* 5:413–422
43. Leibold MA, Holyoak M, Mouquet N, Amarasekare P, Chase JM, Hoopes MF, Holt RD, Shurin JB, Law R, Tilman D, Loreau M, Gonzalez A (2004) The metacommunity concept: A framework for multi-scale community ecology. *Ecol Lett* 7:601–613

44. MacArthur RH (1955) Fluctuations of animal populations and a measure of community stability. *Ecology* 36:533–536
45. Ulanowicz RE (1986) Growth and development: Ecosystems phenomenology. Springer, New York
46. Ulanowicz RE (1997) Ecology, the ascendent perspective. Columbia University Press, New York
47. Odum EP (1969) The strategy of ecosystem development. *Science* 164:262–270
48. Fisher RA (1958) The genetical theory of natural selection. Dover Publications, New York
49. Thompson DAW (1942) On growth and form. The University Press, Cambridge
50. Pennycuik CJ (1992) Newton rules biology: A physical approach to biological problems. Oxford University Press, Oxford
51. Calder WA (1984) Size, function, and life history. Harvard University Press, Cambridge
52. Peters RH (1983) The ecological implications of body size. Press Syndicate of the University of Cambridge, Cambridge
53. Schmidt-Nielsen K (1984) Scaling, why is animal size so important? Cambridge University Press, Cambridge, New York
54. Gillooly JF, Brown JH, West GB, Savage VM, Charnov EL (2001) Effects of size and temperature on metabolic rate. *Science* 293:2248–2251
55. Real L (1977) The kinetics of functional response. *Am Nat* 111:239–300
56. Real L (1979) Ecological determinants of functional response. *Ecology* 60:481–485
57. Marquet PA, Labra FA, Maurer BA (2004) Metabolic ecology: Linking individuals to ecosystems. *Ecology* 85:1794–1796
58. West GB, Brown JH, Enquist BJ (1999) The fourth dimension of life: Fractal geometry and allometric scaling of organisms. *Science* 284:1677–1679
59. Maurer BA (1990) Dipodomys populations as energy-processing systems – regulation, competition, and hierarchical organization. *Ecol Model* 50:157–176
60. Brown JH, Gillooly JF, Allen AP, Savage VM, West GB (2004) Toward a metabolic theory of ecology. *Ecology* 85:1771–1789
61. Brown JH, Gillooly JF, West GB, Savage VM (2003) The next step in macroecology: From general empirical patterns to universal ecological laws. In: Blackburn TM, Gaston KJ (eds) *Macroecology: Concepts and consequences*. Blackwell, Oxford, pp 408–423
62. Ernest SKM, Enquist BJ, Brown JH, Charnov EL, Gillooly JF, Savage VM, White EP, Smith FA, Hadly EA, Haskell JP, Lyons SK, Maurer BA, Niklas KJ, Tiffney B (2003) Thermodynamic and metabolic effects on the scaling of production and population energy use. *Ecol Lett* 6:990–995
63. Allen AP, Gillooly JF (2006) Assessing latitudinal gradients in speciation rates and biodiversity at the global scale. *Ecol Lett* 9:947–954

Books and Reviews

- Allen TFH, Starr TB (1982) *Hierarchy: perspectives for ecological complexity*. University of Chicago Press, Chicago
- Botkin DB (1990) *Discordant harmonies: a new ecology for the twenty-first century*. Oxford University Press, New York
- Brown JH (1999) *Macroecology: progress and prospect*. *Oikos* 87:3–14
- Brown JH, West GB (2000) *Scaling in biology*. Oxford University Press, Oxford, New York

- Gaston KJ, Blackburn TM (2000) *Pattern and process in macroecology*. Blackwell Science, Oxford, Malden
- Gotelli NJ, Graves GR (1996) *Null models in ecology*. Smithsonian Institution Press, Washington
- Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology – Is Mother-Nature a strange attractor. *Annu Rev Ecol Syst* 24:1–33
- Holling CS (1992) Cross-scale morphology, geometry, and dynamics of ecosystems. *Ecol Monogr* 62:447–502
- Levin SA (1999) *Fragile dominion: complexity and the commons*. Perseus Books, Reading
- Milne BT (1998) Motivation and benefits of complex systems approaches in ecology. *Ecosystems* 1:449–456
- Peterson DL, Parker VT (1998) *Ecological scale: theory and applications*. Columbia University Press, New York
- Roughgarden J (1979) *Theory of population genetics and evolutionary ecology: An introduction*. Macmillan, New York
- Weber BH, Depew DJ, Smith JD (1988) *Entropy, information, and evolution: New perspectives on physical and biological evolution*. MIT Press, Cambridge
- Yodanis P (1989) *Introduction to theoretical ecology*. Harper and Row, New York

Ecological Systems

JORDI BASCOMPTE

Integrative Ecology Group, Estación Biológica de Doñana, CSIC, Seville, Spain

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Information Theory and Diversity](#)
[Networks](#)
[Complex Dynamics](#)
[Spatiotemporal Dynamics](#)
[Thresholds](#)
[Future Directions](#)
[Bibliography](#)

Glossary

- Food webs** Networks depicting who eats whom in an ecological community.
- Compartments** Groups of highly interacting nodes with few connections to nodes from other groups.
- Scale-free networks** Very heterogeneous networks in which the bulk of nodes have a few links, but a few nodes have a very large number of links.
- Mutualistic networks** Two-mode networks depicting the mutually beneficial interactions between plants and their pollinators or seed dispersers.

Connectivity correlation A measure of network structure that represents the correlation between the number of interactions of a node and the average number of interactions of the nodes it interacts with. A negative connectivity correlation would represent a modular network.

Species strength A measure of the importance of a species in terms of the total weight of its connections.

Network motifs Patterns of interconnections significantly over-represented in complex networks. These may be regarded as the simple building blocks of complex networks.

Trophic cascades Changes in population abundance that propagate through more than one trophic link in the food chain.

Ecosystem shifts Sudden qualitative changes in the state of an ecosystem (i. e., from clear to turbid waters in a lake) following a continuous tuning of a variable such as nutrient load.

Deterministic chaos A periodic, random-like time series generated by low dimensional, non-linear, deterministic models.

Lyapunov exponent A measure of the degree of divergence of initially close trajectories in the phase space that is characteristic of deterministic chaos.

Coupled map lattice Dynamical system with discrete time, discrete space, and continuous state. It was first used by the physicist Kunihiro Kaneko in relation to spatiotemporal chaos and later on used in ecology as a model of spatiotemporal systems.

Interacting particle system Stochastic spatial models with discrete time, discrete space, and finite states. They have been used as spatially extended models of populations and epidemics, and have been widely analyzed by Richard Durrett and Simon Levin.

Metapopulation A population of populations maintained in a dynamical balance between local extinctions and recolonizations from nearby local populations.

Extinction thresholds Critical values in the amount of habitat destroyed at which a metapopulation goes extinct.

Definition of the Subject

Ecological systems are paradigmatic examples of complex systems. Just think about the thousands of species interacting in complex ways within rich communities such as tropical rainforests or coral reefs. The most pressing questions ecologists face deal with concepts such as stability, resilience, thresholds and non-linearities which are at the

core of the sciences of complexity. How robust are these cathedrals of biodiversity? At which rate will they disassemble as a consequence of global change? For example, one of the long-standing questions in ecology is the relationship between complexity and stability. This contribution will present a brief review of some of the applications of the complexity sciences into the realm of ecological systems and discuss the implications for our understanding of ecosystems. Predicting the consequences of global change on biodiversity and the services it provides will need an interdisciplinary approach in which concepts from the sciences of the complexity may be very useful. Not only complexity sciences are important for ecology, but ecological research has also provided concepts and ideas to the science of complexity, for example in the context of deterministic chaos.

Introduction

Ecology is a relatively new science. It focuses on the relationships of species among themselves and with their environments. Because of the huge number of entities and their multiple interactions and feedbacks, the study of ecosystems is amenable to some macroscopic approaches.

Although ecology has been an eminently descriptive science, important theoretical contributions were made almost from the beginning starting with the pioneering work by Lotka [68], Volterra [125], and Nicholson and Bailey [89]. These first contributions analyzed the dynamics of simple models describing two coupled populations such as a predator and its prey or two competing species. This early theoretical work defined the steady state solutions of these systems and their stability. The lessons from this exercise were to understand the possible dynamic outcomes from species interactions. For example, predators and their prey may become engaged in cycles. Some of these cycles were quite similar to cycles observed in nature such as the textbook example of the Canadian lynx and its main prey, the snowshoe hare. The competition models, on the other hand, were used to understand under what circumstances two species will coexist. These type of models are usually more useful when they do not describe appropriately the reality pointing towards important missing variables. It is the case with the Nicholson–Bailey model [89] of a host-parasitoid interaction, a type of specific predator-prey interaction in which an insect such as a wasp lays its eggs in, at or near the body of an arthropod such as a caterpillar. It is nature's own version of the celebrated movie *Alien*. Nicholson was fascinated by the coexistence of these insects whose abundances tend to oscillate in the field. However, the model was unstable and

led to the extinction of one of the species. Thus, something else, such as the consideration of space, was needed as we will see below.

Models of one or a few species were later on replaced by other type of models representing entire communities. The dominant question revolved around the relationship between the complexity and stability of ecological communities [74]. Also, single population models were analyzed in the context of nonlinearities, as for example in relation to deterministic chaos [75]. Another emphasis was in stochastic models where ecologists explored how time to extinction scales with population size [31,43,62], and how species coexistence depends on fluctuating environments [23,24].

Similarly, another extension of single-population, or two coupled population models were in the direction of addressing spatial degrees of freedom, that is, incorporating a spatial dimension and exploring how this new dimension made species coexistence easier.

Field ecologists, on the other hand, took other avenues but with similar goals. What regulates populations? What shapes the structure of communities? The first type of question emphasized the role of density-dependence versus external variables in explaining population change through time. The interest of this work is twofold since it may guide a biologically-informed pest control as William Murdoch and colleagues have advocated [83,84]. At the community level, the question was to understand the suite of mechanisms allowing the high levels of biodiversity that can be found in coral reefs and tropical rainforests. Joseph Connell, for example, analyzed the role of competitive interaction in structuring the marine intertidal [26]. In particular, he analyzed how patterns of recruitment, mortality and competition affected the distribution of barnacles [26]. More generally, he addressed how the high diversity in coral reefs and tropical forests is related to external perturbations in his famous intermediate disturbance hypothesis [27]. This states that the highest diversity levels are found neither in the absence of perturbations (competitive exclusion eliminates some species), nor with perturbations too frequent or intense (the bulk of species can not survive). Highest diversity levels are found at intermediate levels of perturbation.

Robert Paine emphasized the role of predation in controlling biodiversity in the intertidal [95]. His seminal work led to the concept of keystone species. He experimentally excluded the starfish in plots of the intertidal. The starfish is an important predator. It mainly preys on a species of algae which is competitively superior, keeping a control on its abundance and allowing the coexistence of several algae species. When Paine removed the starfish,

the competitively superior algae out-competed the other algae species and the system became quite simplified. The predator had a strong interaction with its main prey and that had implications for the whole ecosystem. The importance of some species is much higher than what one would have predicted based on their abundance. The keystone concept is extremely important in ecology. It has shown beyond any doubt the potential ecological impact of a single species, and thus that we need to consider the roles of individual species in order to manage ecological communities [95].

What precedes is a very simplified and biased review of milestones in ecology and does not claim to be representative of the wonderful work that has been achieved in its century of history. It just tries to provide some general background on the ideas of diversity, complexity, nonlinear dynamics and threshold behavior that will be illustrated in the following sections as examples of applications of the paradigm of complex systems to the problems of ecology and the preservation of natural resources in the face of human-induced perturbations. Next, I will explore this suite of studies and how they have shed light on our understanding of ecological processes.

Information Theory and Diversity

A great contribution in ecology was made from the perspectives of general systems. Ramon Margalef was pioneering the use of Information Theory as a way to describe ecological systems [70]. He was inspired by the work of Norbert Wiener, who introduced the concept of cybernetics [4,126]. The key idea was to emphasize the feedbacks between components of the ecosystem as a way to understand the control of one system by another. A classic example of negative feed-back is that between a predator and its prey. Predator and prey regulate each other's population as a thermostat would regulate a room's temperature. Margalef's book *Perspective in Theoretical Ecology* [71] was a classic in that regard. Margalef felt completely comfortable in the context of cybernetics because it perfectly described his view of ecology, a view defined as "the study of systems at a level in which individuals or whole organisms may be considered elements of interaction, either among themselves, or with a loosely organized environmental matrix" [70]. Information Theory was applied to ecology mainly as a way to characterize the diversity of an ecosystem measured as the number of different species and their relative abundances. Diversity would be maximum when each individual was from a different species, and minimum when all individuals were from the same species. Margalef used to talk about the museum

and the field of agriculture to refer to these two extreme cases.

In the context of the theory of information, an ecosystem is like a channel that transfers information. The amplitude of this channel is measured by the Shannon Entropy, which is a measure of disorder or uncertainty. In our context, let's say that we randomly pick up an individual. What is the uncertainty that this individual belongs to a specific species? Let's assume that an ecosystem has s different species, each one with an abundance n_1, n_2, \dots, n_s , so that the total number of individuals is $N = \sum_{i=1}^s n_i$. The probability that the randomly picked individual belongs to species i is then $p_i = n_i/N$, and one can define the diversity of the community as

$$H = - \sum_{i=1}^s p_i \log_2 p_i. \quad (1)$$

MacArthur [69] and many others advocated the use of this type of measure to describe diversity and many different uses of these indices have been applied since then, for example in trophic studies of animals' diets, or in the quantification of energy flows in food webs [123].

Networks

Another significant contribution to a general system view of ecology was the concept of food webs, networks that represent who eats whom in ecological communities. These graphical representations of communities were first drawn by ecologists such as Lindenman and Odum [32,67,90]. Odum [90] used his engineering training to represent the interrelationships of ecological systems. As in the case of Margalef, he was emphasizing the interrelationships more than the nodes. He also had a broad and rich background that allowed him to think about ecosystems with fresh views. And he insisted on the concept of energy as one of the most important currencies in ecology. Odum used energy diagrams in the hope of seeing general patterns across systems regardless of taxonomic differences [118].

Food webs have constituted one of the classic subjects in ecology, with changing emphasis through the years. In the 1970s, and as a consequence of the seminal paper by Robert M. May [74], people started looking at food web structure due to the evidence that structure greatly affects food web dynamics.

Stability and Complexity

May [74] used Gardner and Ashby's previous result to determine under what circumstances a random food web will

be stable. This work was based on matrix algebra and was very successful at starting a rich research agenda. Roughly speaking, May was using Lotka–Volterra models with random interactions among species, and analyzed the probability of this model to be linearly stable. Given a certain connectance C measuring the fraction of non-zero interactions among species, May used previous results on random matrices to show that the system will be stable if

$$\alpha < SC^{-\frac{1}{2}}, \quad (2)$$

where S is the number of species and α is the average interaction strength among species. As noted from the previous inequality, the probability of a community to remain stable decreases as either the number of species or connections increases. This result essentially tells us that there are some constraints to randomly built communities in order to remain stable. Complexity begets instability, which contrasted with classical arguments by MacArthur, Elton and Margalef that suggested that complex ecosystems are more stable than simple ones. The question, thus, was to explore what properties of food webs counterbalance this tendency towards instability.

As a consequence of May's [74] paper, ecologists became interested in the modularity or compartmentalization of these ecological networks. The reason, at least in part, was the discussion at the end of that influential paper. May, after showing that complexity begets instability, performed some numerical experiments with non-random networks. He concluded his paper by noting that *"such examples suggest that our model multispecies communities, for a given average interaction strength and web connectance, will do better if the interactions tend to be arranged in "blocks" – again a feature observed in many natural ecosystems"*. Thus a whole research program was set on compartments. Part of the research focused on exploring the theoretical implications of compartments [98]; part was trying to explore whether real food webs are compartmentalized [58,100,103]. For a review on the studies on food webs see [25,94,99,117]. This body of work emphasized invariant properties of food webs, their structure, the frequency distribution of interaction strengths, and simple models able to generate food webs with a similar structure as the observed in nature. I will not review this interesting literature here, instead I will emphasize the latest round of research in food webs that echoes similar work in complex networks. Recently, tools from the study of complex networks have been successfully applied to food webs. Food webs are now seen as another example of a complex network, with several papers comparing their structure with that shown by other types of networks such as the Internet, protein networks, or social networks [2].

Scaling in Ecological Networks

A first descriptor of network structure is the connectivity distribution, defined as the probability distribution of the number of interactions per node. The idea is to pick randomly a node in the network and represent the probability of this node interacting with one, two, \dots , n other nodes. The relevance of this descriptor of network structure stems from two facts. First, from its relationship with early graph theory by Paul Erdős and Alfred Rényi [34] and recent build up models that generated a good correspondence between several models of network formation and their consequent connectivity distribution. Second, because the paper by Albert et al. [1] clearly related the shape of the connectivity correlation with the network robustness to error and attack. Albert et al. [1] found that the Internet has a connectivity distribution that follows a scale-free distribution defined by a power-law of the type:

$$p(k) \propto k^{-\gamma}, \quad (3)$$

where $p(k)$ is the probability of a node having k links and γ is a critical exponent. In a log-log plot, this relationship is defined by a straight line of slope $-\gamma$ for all the range of k values. That is, Eq. (3) is a relationship not defined on a particular scale. This would not be the case, for example, for an exponential distribution that has a specific scale, the average number of links per node [108].

Barabasi and Albert [6], building on a previous result by Simon [110] showed that a process of network build up where new nodes link preferentially with already well-connected nodes (a type of “rich gets richer process”) is a simple recipe to generate scale-free networks. On the contrary, a random model such as the classical random graph by Erdős and Rényi [34] generates distributions with Poisson distributions or exponential distributions if the number of nodes keeps growing. The important point is that in the latter case, the resulting network is much more homogeneous in the sense that all nodes have a similar number of interactions.

A randomly built network with connectivity distributions with thin tails, such as the Erdős–Rényi random graphs, are very fragile to the random deletion of nodes. After a certain fraction has been removed, the networks fragment. This fragmentation threshold will be revisited later on in the context of spatial processes, where space is represented as a regular (or irregular) network of points. Thus, random networks are very fragile [1]. On the other hand, scale-free networks are much more robust to the random deletion of nodes. One has to remove a high fraction before the network gets fragmented. The reason is that the few highly connected nodes (the hubs) play a major



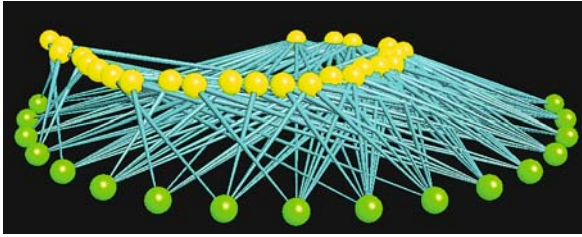
Ecological Systems, Figure 1

The mutually beneficial interactions between plants and their animal pollinators (picture) and seed dispersers have played a major role in the generation of earths' biodiversity. Picture courtesy of Mark Chappell

role in keeping the entire network together. Since these hubs are quite rare, it is very unlikely to remove them by chance. However, as shown by Albert et al. [1], these hubs are the Achilles' heel of the network. If one now starts removing the most connected nodes, the whole network collapses. Thus, scale-free networks are very robust to the random loss of nodes but very fragile to the loss of the hubs.

The work by Albert et al. inspired ecologists who turned to food webs in search of their connectivity distributions. Solé and Montoya [112] first analyzed a few food webs and found evidence for a scale-free distribution, while Camacho et al. [20] compared different distributions and found the best fit to be to an exponential distribution. Dunne et al. [30] generalized these previous results by using a broader data set and testing several functions. Their conclusion was that even when there were a few food webs described by fat tail distributions, the bulk of the food webs had tails following an exponential distribution.

Jordano et al. [53] extended the argument by focusing on a different type of ecological network, the one describing the mutually beneficial interactions between plants and their animal pollinators or seed dispersers (Fig. 1). These are two-mode networks with a much higher level of resolution than traditional food webs. While food webs have a high level of lumping so that a node contains several taxonomic species, mutualistic networks have a level of resolution almost always corresponding to a taxonomic species. These networks describe the coevolutionary process in species-rich communities [8]. This study analyzed 53 communities and concluded that in the bulk of cases,



Ecological Systems, Figure 2

The mutualistic interactions such as the one depicted in Fig. 1 form complex networks of species interdependence. The architecture of these networks greatly affect their robustness to the extinction of one of the species. The picture represents a plant-pollinator network in the Arctic. Plants and insects are represented as green and yellow nodes, respectively

connectivity distribution for both plants and animals was best fitted by a truncated power law, a distribution of the following form:

$$p(k) \propto k^{-\gamma} e^{-k/k_c}. \quad (4)$$

The main difference in relation to Eq. (3) is the existence of a critical connectivity level k_c beyond which the connectivity distribution decays faster than expected for a power-law. These mutualistic networks are still very heterogeneous but not as heterogeneous as predicted for a scale-free distribution (Fig. 2) [53].

There are several non-exclusive factors that may account for the existence of these truncated power-law distributions. Jordano et al. [53] focused on what they termed *forbidden links*, that is, the existence of interactions that are not possible due to size or phenology uncoupling. For example, an insect cannot pollinate a plant species if it is a migrant that arrives after the flowering period of the plant. Or a bird species will not disperse a tree species if their seeds are larger than the width of the bird's beak. By combining analytic thinking and natural history, Jordano et al. [53] were able to account for a large fraction of the non-observed interactions in two well-studied communities.

Of course the fact that forbidden links exist and that their existence can lead to a truncation of an otherwise power-law does not exclude additional mechanisms. Forbidden links and similar mechanisms such as filtering information (i. e., a new node can sample only a subset of the network) constrain the preferential attachment mechanism. Other processes also lead to truncated power-law distributions without any constraint on such a process. For example, the same preferential attachment process taking place on a bipartite network leads to a truncated power-law distribution if there is any asymmetry between the two

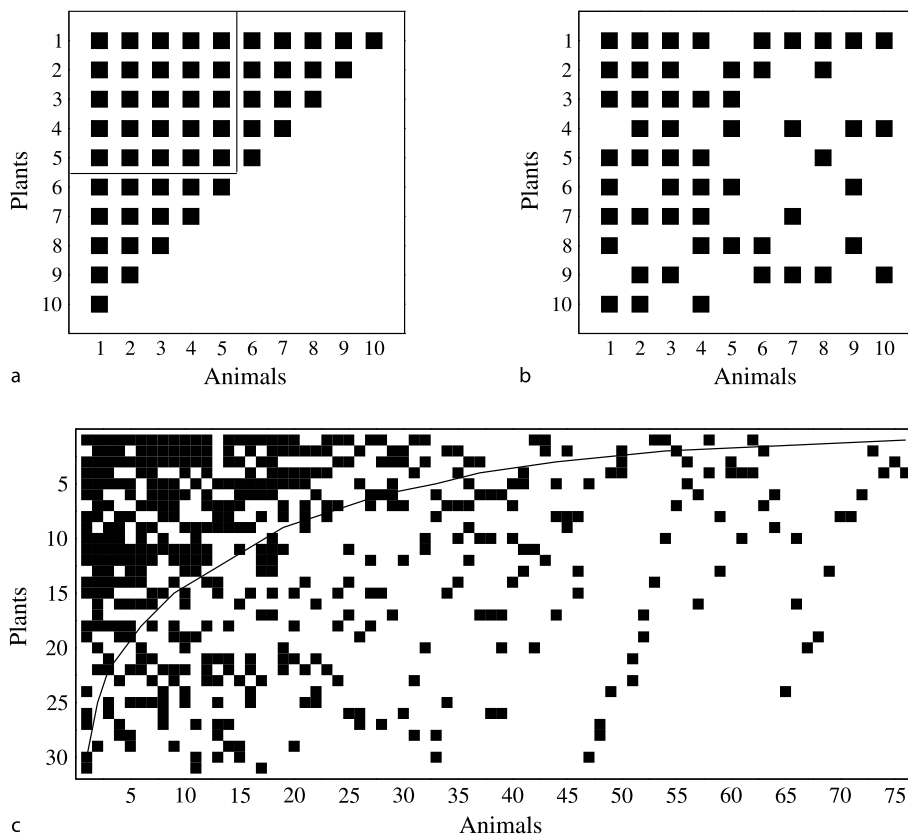
sets such as one set (e. g., plants) growing faster than the other set (e. g., animals) [39,41].

From the point of view of the robustness to species extinction of these mutualistic networks, the truncated power-law distribution confers more robustness than an exponential distribution to the random extinction of species but less dependence to the extinction of the hubs than for a power-law distribution.

Network Structure: Modules

The connectivity distribution is just a first description of network structure. In the general field of complex networks, scientists looked at deeper measures of network structure such as connectivity correlation or modularity. This was mainly analyzed for genetic networks and the Internet. For example, the connectivity correlation measures the average correlation between the number of links of a node and the average number of links of the nodes it interacts with. Maslov and Sneppen [73] found that both the internet and protein networks had a negative connectivity correlation, which means that hubs tend to interact with poorly connected nodes. This corresponds to an organization in compartments, which may buffer from the propagation of mutations or other perturbations [73]. Melián and Bascompte [78] applied this idea to food webs and found that they are more cohesive than the Internet or protein networks. Generalist species tend to interact among themselves. This may make these communities less robust to the propagation of a perturbation such as a contaminant, but more resistant to the extinction of a species. There is more than a single way to be robust [78]. This description of food webs is complementary to a cohesive modular organization where several k -subwebs, that is, groups of species with at least k interactions among other species in the subweb, are linked to a densest central subweb which induces cohesion to the entire food web [79].

This work on the structure of food webs links to the early attempts to characterize compartments mentioned above [74,100,103]. In this regard, research on food webs [74,98] pioneered the search for network structure that 30 years later would be so important in complex networks. The search for compartmentalization in food webs has not found too much evidence, partly because a lack of high quality data, partly because a lack of appropriate statistical tools to unambiguously define and characterize modules. More recently, Krause et al. [58] used software available to sociologists and found the strongest evidence of compartmentalization in three out of five food webs studied. In the context of the physics of complex networks, recent work has addressed the role of compartments in the



Ecological Systems, Figure 3

The nested assembly of mutualistic networks. Two-mode interacting networks are represented as a matrix with plants in rows and animals in columns. A square indicates that the plant in this row and the animal in this column interact. Panels **a**, **b** and **c** represent a totally nested, random and real network. The line in **c** represents the isocline of perfect nestedness. Modified from [14]

structure of complex networks such as the world-wide air traffic [42]. Several algorithms to quantify modularity are now available. Olesen et al. [92] have used these algorithms to detect modularity in pollination networks. These modules are interpreted as the basic units of coevolution, that is, small groups of highly interacting plants and animals. The modularity analysis is useful in this context in showing the denser areas of the network. These denser areas have the potential to be coevolutionary hotspots or vortices [119].

Another concept ecologists have used to further describe the structure of mutualistic networks provides from island biogeography, nestedness. In the mutualistic context, a matrix of plant-animal interactions is nested if specialists interact with species that form perfect subsets of the species with which generalists interact (Fig. 3) [14]. This is a pervasive community organization that has been described for other ecological interaction such as those be-

tween cleaning fish and their hosts [41], or parasites [60]. Nestedness implies a central core of interactions where generalist plants and generalist animals interact among themselves. This originates a dense core of interactions with a high level of redundancy and the possibility for the system to respond to perturbations. This is somehow in agreement with the cohesive organization of food webs found through the connectivity correlation and k -subweb distribution seen above. On the other hand, a nested mutualistic pattern implies an asymmetric pattern of specialization since specialists tend to interact with generalists. The latter tend to be more abundant and less fluctuating, and thus these community patterns confer mechanisms for the persistence of rare species [14]. Ecologists are now starting to explore the implications of these universal community patterns from the point of view of community responses to perturbations such as habitat loss [5,37] or the invasions of foreign species [19,80,82,91].

Weighted Ecological Networks

These heterogeneous, asymmetric network patterns in mutualistic networks are also observed in weighted networks. In this case, species strength, the weighted equivalent of species degree, grows faster than linear with species degree [7]. This pattern had been previously found for the world-wide airport network, but not for the scientific collaboration network [7]. The strength of highly connected species is even higher than expected based on their degree because specialists tend to interact exclusively with the most generalized species [14], and so depend completely on them. Thus, specialists contribute disproportionately to increase the overall strength of the generalists they depend upon. The nested structure of these mutualistic networks accounts for this pattern. The predominance of weak interactions and the asymmetry in pairwise interaction when a plant, for example, depends highly on an animal, tends to increase the conditions for the persistence and stability of species-rich communities as indicated by analytical results of a simple community model [16].

The role of weak interaction strengths on community stability has also been analyzed in studies of weighted food webs. This research agenda may be traced back to the seminal work by Robert Paine [95,96] who in his classic experiments on the intertidal noted in the introduction, found that the strength of interactions between predators and their prey are defined by a few strong interactions in a matrix of weak interactions. This pattern has then been observed over and over in other food webs and using other measures of interaction strength [15,35,103,123,127]. Simple dynamic models have shown that this frequency distribution of interaction strengths increases the stability of communities [57,77]. However, the frequency distribution of interaction strength is only a first descriptor that does not tell us how these interaction strengths are combined in the basic components of the food web. Thus, Neutel et al. [88] found that weak interactions tend to be distributed in long loops. This avoidance of strong interactions in long loops induces the stability of the whole community [88]. Similarly, Bascompte et al. [15] found that the co-occurrence of two strong interactions in a tri-trophic food chain occurs less often than expected by chance, and that in the few cases in which this occurs, it tends to be accompanied by strong omnivory (predator preying on two consecutive levels of the food chain) more often than expected by chance. These results have implications for the likelihood of trophic cascades, that is, changes in species abundance that transmit at least through two consecutive levels of a food chain. An example of a trophic cascade would be a decrease in sharks through overfishing, a subsequent in-

crease in abundance of big fish that constitute their prey, and a concomitant decrease in the abundance of smaller, herbivorous fish the former prey on. Two strong interaction strengths have the potential to induce trophic cascades after the overfishing of top predators, but when accompanied by strong omnivory the magnitude of this cascade is severely reduced. Thus, in the light of the dynamic results of a biologically parametrized bioenergetic model, one can conclude that, other things being equal, the reduced frequency of two consecutive interaction strengths and their association to strong omnivory reduce the likelihood of trophic cascades [15]. However, this global pattern does not assure that the marine food web is buffered from the effects of overfishing, since overfishing does not affect randomly picked species but tends to focus on large, top predators. In the Caribbean food web, fishing selectively targets a biased sample of species belonging to upper trophic levels [86,97]. These species include ten heavily fished shark species from seven families that account for almost one-half of the strongly interacting food chains in the Caribbean food web [15]. The likelihood of trophic cascades after the overfishing of these predators is thus high. These cascades can contribute to the depletion of herbivorous fishes at the base of the chain such as parrotfishes that are important grazers of the algae. The reduction of these herbivorous fishes can accelerate the transition from corals to algae, an example of bistable steady state that will be considered later on in the context of ecosystem shifts as examples of phase transitions (see below).

First quantifications of interaction strength through energy fluxes was using information theory, the same framework we have described in the previous section in the context of species diversity [123]. New metrics to characterize weighted food webs build on this preliminary study [18].

Network Motifs and Trophic Modules

A final parallelism between complex networks and ecological food webs has to do with network motifs, patterns of interconnections that are over-represented in complex networks and that can be considered as the simple building blocks of complex networks [81]. Interestingly enough, there is a significant difference between research on network motifs and its equivalent research in ecology. The approach in complex networks is eminently structural, while that of their ecological counterpart is eminently dynamical. For example, the first papers on network motifs quantified their representation in entire networks and compared their frequency with that predicted by appropriate

null models. Only later on there were some studies exploring the dynamics of these different motifs [101]. On the other hand, ecology has been studying the dynamics of simple trophic modules such a tri-trophic food chain without looking at how frequent are these simple modules in entire food webs [9]. What remains to be done now is to scale-up from these isolated modules to the entire food web.

Complex Dynamics

Let us now move from structural complexity to dynamical complexity. There is a strong relationship between ecology and complexity sciences in the context of population dynamics. As a matter of fact, one of the more seminal contributions to deterministic chaos came from theoretical ecology. Once more, the great talent of Robert May was behind this contribution [79]. May was looking at one of the simplest models one could think of in theoretical ecology. It describes the dynamics of populations with non-overlapping generations. Let's assume that the density of insects in a generation t is N_t . Let's first normalize this value by dividing it by the highest density ever observed N_{\max} so that $x_t = N_t/N_{\max}$. If we imagine a deterministic model with density dependence, one can write what the density of insects at the next generation will be

$$x_{t+1} = \mu x_t(1 - x_t), \quad (5)$$

where μ represents the population rate of increase. Robert May analyzed the temporal dynamics of system (5) as he was increasing the growth rate μ . For very low values, the population goes extinct. When $\mu > 1$ the population reaches a steady state. If μ is further increased, at $\mu > 3$ the steady state becomes unstable and a cycle of period two becomes stable. For even higher μ -values, there are other period-doubling bifurcations and so the population oscillates with cycles of higher frequency. Finally, when μ reaches a critical value the dynamics never repeats itself, the system shows deterministic chaos [79].

Finding this period-doubling route to chaos in an ecological model opened a research agenda that found that this scenario has universal properties, e.g., the relationship between the successive critical μ_k values at which a new bifurcation k appears are independent of the model. More than that, even in experimental systems one could observe the same universal laws [37]. Specifically, Feigenbaum showed that for a large enough μ -value, the following relationship takes place:

$$\delta = \lim_{k \rightarrow \infty} \frac{\mu_k - \mu_{k-1}}{\mu_{k+1} - \mu_k} = 4.6692 \dots, \quad (6)$$

The co-discovery of deterministic chaos in the logistic equation (together with parallel work in meteorology and mathematics) had a huge importance, not only in the field of ecology, but beyond. It is one of the few examples in which the flow of ideas has gone from ecology to physics. Since this important discovery, a rich research program of research in ecology revolved around the role of deterministic chaos in ecological systems, both theoretically [76] as well as empirically [51,106].

Chaos in the Real World

William Schaffer and Mark Kot [106] were among the pioneers in the search for deterministic chaos in real ecological systems such as the cycles of the Canadian lynx or the monthly records of measles in big cities. Ecology was facing the possibility that complex temporal series were not the result of hundreds of stochastic variables, but of a few variables in deterministic, yet non-linear dynamical systems [52]. This is not just a technical issue. If complex dynamics in the populations of diseases or pests were deterministic we could understand the underlying rules. However, the evidence for chaos has been more evasive due to the shortness of temporal series and their high amounts of noise.

Arguably, the first serious attempt to quantify chaos in nature was the paper by Hassell et al. [46], who analyzed the temporal series of 28 arthropod insects from both the lab and the field. Detecting chaos depends very much on the way it is attempted, and ecologists have been very imaginative in their search for chaos in a noisy world. In this first study, Hassell et al. fitted their temporal series to a previously studied non-linear population model. Twenty-six populations had temporal series which best fit model parameters within the parameter region corresponding to steady states, and only one example corresponded to the region of deterministic chaos. This was thought to be of little empirical support for chaos to begin with.

Schaffer and Kott [106] used a different approach and different data sets, and their work supported the notion that chaos may be common in ecology. They used the same techniques physicists were using, such as attractor reconstruction and estimation of its fractal dimension. Several time series such as the Canadian lynx had attractors reminiscent of the strange attractors that are the hallmark of deterministic chaos. Similar results were obtained for the measles records in Baltimore [3,106].

Sugihara and May [116] used non-linear forecasting techniques to distinguish deterministic chaos from noise (both correlated and uncorrelated). This clever approach

is based on dividing the temporal series in two halves and using the first half as a source of known data, and considering the second half as the unknown future. The correlation coefficient between predicted and observed values is plotted versus prediction time. Deterministic systems show an exponential decay in correlation, while this is constant for noisy systems. Ellner and Turchin [32] used non-linear techniques to estimate the largest Lyapunov exponent, i. e., the parameter telling at which rate two nearby trajectories in the phase space will diverge. Their strategy to reduce the high levels of noise present in the original time series was to first estimate the map that best fits their temporal series, and then use it to calculate the Lyapunov exponent from the dynamics of the map. Tilman and Wedin [120] built a Poincaré map by plotting the biomass of annual plants one year versus the same biomass the next year. The slope of this map for a vegetation model previously used dictates whether the system is or is not chaotic. In summary, there was a serious effort to find chaos in real ecosystems by using a broad spectrum of techniques. There was evidence of chaos, but also evidence against its presence. This line of research is almost extinguished, but a recent paper by Sibly et al. [109] touched on it by estimating return rates after a perturbation for a very large number of temporal series of groups of mammals, birds, fish, and insects. They found that in the bulk of cases the return rates were quite below the threshold for chaos, which corresponds to stable populations. However, there is a clear case for the potential of chaos in population dynamics, a beautiful example that comes from an interaction between analytical and lab work.

Costantino and colleagues [28,29] combined an experimental setting where a population of the flour beetle *Tribolium* was growing in milk bottles, and a population dynamic model that, although simple, incorporated the basic dynamics of the species life cycle. This was a discrete time model that described the three phases of the beetle life cycle, namely feeding larvae, non-feeding larvae, and adults. Cannibalism is very common in this species and mathematically induces a strong nonlinear term that is partly responsible for the presence of deterministic chaos. This team first proceeded by parametrizing their model with the temporal data they obtained in the lab. From the experimental point of view, they manipulated the recruitment rate into the adult stage by adding or removing adults at the time of the census. As this recruitment rate was increased, there was a sequence of period-doubling route to chaos shown both by their lab census as well as by their model with the adequate parameter values.

In sum, maybe chaos is not common in nature, but nature certainly has the potential to show chaos.

Spatiotemporal Dynamics

Deterministic chaos was perhaps the first pedagogical example of the potential for non-linear dynamics in ecology. The lesson was that other dynamical behaviors beyond steady states and cycles are compatible with a deterministic, density-dependent model. The next finding of the potential of non-linear dynamics to generate complex phenomena was provided by the study of dynamic systems extended in space. Imagine a discrete lattice of sites simulating the patchy distribution of some available habitat. Within each one of these habitat patches a local population can be described by a dynamic model such as the logistic map. However, we allow now for the fact that a fraction of the individuals born in a patch disperse to neighboring sites. The resulting spatiotemporal dynamics can be described by a coupled map lattice (CML), a dynamical system with discrete time, discrete space and continuous state, first used by the Japanese physicist Kunihiko Kaneko working with problems of diffusion and spatiotemporal chaos [54,55]. This approach allows one to easily study the combined action of two processes: local dynamics (described by an appropriate *map* or discrete time model such as the logistic Eq. (5), and the coupling through dispersal of these local maps. A CML can be written in the following way:

$$x_{t+1}(i) = (1 - D)F\{x_t(i)\} + \frac{D}{k} \sum_{j=1}^k F\{x_t(j)\}, \quad (7)$$

where $x_t(i)$ is the density of population at site i and time t , D is the fraction of individuals leaving its patch and k is a certain neighborhood around a local patch to where individuals can move. Coupled map lattices such as (7) have been extensively used in ecology [10,112]. For example, coupled map lattices have shown how dispersal may affect the temporal dynamics [49], and the length of the transients [50]. Particularly relevant is the finding of spatial self-organizing patterns such as spiral waves in the abundance of populations [47,48,115]. This phenomenon is qualitatively similar to the one found for excitable media where symmetry breaking takes place around a pacemaker [85]. From the point of view of ecology this suggests that simple rules can create long-range spatial patterns. Similar spatial models belong to the class of interacting particle systems, where not only space and time are discrete, but also the state of a cell is on one of a few discrete values. As opposed to CMLs, the latter are stochastic models for which there is a rich body of mathematical work addressing, among others, how spatial pattern arises in ecology [31]

Coupled map lattices are useful to link patterns at different spatial scales. The problem of pattern and scale is at the core of ecology [63]. For example, one can describe the population dynamics at a lattice site. In the case of a chaotic map, this dynamics will be quite unstable, with strong fluctuations. This unstable characteristic was one of the arguments by which some field ecologists argued that chaos would not be common in nature [17]. If populations oscillate so heavily, at some point population density will be low enough for stochastic events to lead the population to extinction. However we can now see the same unstable dynamics from a larger spatial scale, let's say that we plot the total abundance in 2×2 lattice sites, 4×4 lattice sites, and so on. How unstable will the chaotic dynamics look? Surprisingly, the dynamics will now appear very constant. Solé and Bascompte [12,111] coined the term *chaotic stability* to refer to the fact that chaos and its instability at a local scale can induce a strong stability at a global scale. Thus, the criticism of chaos based on its instability does not apply when space is considered.

One way to understand the previous result is to remember the strong dependence on initial conditions of a chaotic system. Because of this, two nearby local populations will start to oscillate out of phase, and so ups and downs will soon cancel each other. Technically, this can be analytically proved following reference [105]. These authors found a relationship between the largest Lyapunov exponent and the spatial coherence length. The coherence length describes how far away two points oscillate in a correlated way. As a matter of fact, the inverse Lyapunov exponent can be regarded as a correlation time, i.e., the time horizon beyond which two initial trajectories fluctuate totally independent from each other. In the vicinity of the onset of chaos, the following relationship between the largest Lyapunov exponent (λ) and the spatial coherence length (ξ) holds: $\xi \approx \lambda^{-1}$. Thus, the more chaotic a system is, the faster spatial correlation decays with distance. An application of this idea in ecology was proposed by [10], providing a clear mechanism for an early suggestion by [120] on the difficulty of detecting chaos at larger spatial scales even when present at local scales.

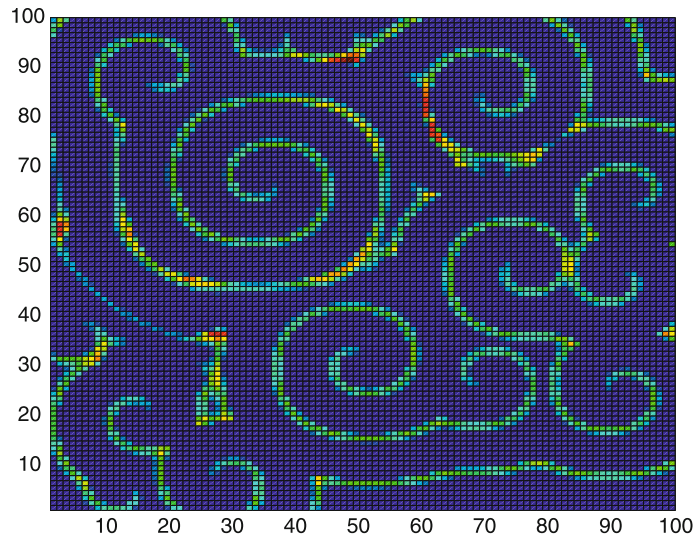
Another interesting application of coupled map lattices is to the problem of pattern formation, which is a celebrated one in several fields such as developmental biology and excitable media [85]. In the context of ecology, pattern formation in space is related to the problem of species coexistence. This is another example where introducing spatial degrees of freedom changes entirely our picture of ecological systems. At the beginning of this contribution I mentioned that Lotka and Volterra had derived a mathematical model for two competing species. The lesson from

that model was that the coexistence of the two species is only compatible with low values of interspecific competition. One interesting avenue has been analyzing the mathematical conditions under which species coexist with different environmental fluctuations [23,24]. Another aspect has been considering the spatial component. If one considers the spatial extension of a competitive model, local exclusion is compatible with global coexistence if stochasticity plays a role in the dynamics. From a spatially homogeneous setting where both species were initially present, one ends up with clusters of patches where species one is the survivor and alternative clusters where the second species wins [114]. These steady state spatial patterns are equivalent to Turing patterns in models of development.

An even more striking example of spatial self-organization is the existence of spiral waves such as the ones observed in excitable media as some chemical reactions or electric activity in the heart [38,85]. If one extends the host parasitoid model by Nicholson and Bailey into a coupled map lattice, one can observe the spontaneous emergence of traveling waves in the density of one of the species (Fig. 4) [47,113,115]. This is relevant from several points of view. From the point of view of complex systems, this shows how the interplay between local non-linear dynamics and short-range dispersal can generate large scale self-organized spatial patterns as first shown by the great mathematician Alan Turing [63,121]. This opens a new way to interpret large scale patterns in ecology, traditionally adduced to reflect environmental causes [15]. Interestingly, these spiral waves are related to the persistence of interacting species: once more, despite the local instability, the system is globally stable and all species coexist more easily than predicted by non-spatial models.

Coupled models have also shown how dynamics can be affected by dispersal. For example, reaction-diffusion models of population dynamics have illustrated the dispersal-induced route to chaos, that is, the change in the type of dynamics from steady states to cycles and to chaos as the dispersal rate is increased [59,93]. Similarly, coupled map lattices have shown how the transient time in non-spatial models becomes now extremely large [50], which had also been noted by Kaneko from a physical perspective. This has important implications in ecology. We implicitly assume that the steady state is the relevant dynamics, but if transients are as long as thousands of years, transient dynamics may be much more relevant for ecology than long-term steady states [50].

As for the case of deterministic chaos, a myriad of papers looked for these self-organizing patterns in nature, and good evidence for pattern formation come from ex-



Ecological Systems, Figure 4

Spiral waves in a coupled map lattice model of interacting populations. The figure corresponds to an iteration and each lattice size codes the abundance of hosts in a host-parasitoid system. These self-organizing spatial patterns are very much related to the persistence of populations

amples of rodents in northern Europe [104], host-parasitoid interactions [72], and outbreaks of the moth *Zeiraphera diniana* in the Alps [19]. Once more, theory was ahead and lead field ecologists to search for examples of complexity in real nature. This was expanding our horizons and moving from a classical view of ecosystems where all complex processes were associated with external variables, to another scenario where internal processes were able to account for much of the complexity observed in real nature.

Thresholds

One important application of the sciences of complexity is the concept of phase transition from statistical mechanics. This is very important because in ecology we are used to thinking in terms of linear relationships between a cause and its effects. Oftentimes, as we tune a parameter we find the occurrence of a critical point in which a sudden qualitative change takes place. A previously stable solution becomes unstable and two new stable solutions emerge, as we have seen in the period-doubling route to chaos. A symmetry-breaking process takes place and the system chooses one of two possible solutions. A mechanical analogy would be a ball rolling on a surface with two minima. This describes the dynamics of a phase transition.

The paradigmatic example of a second-order phase transition in physics is the Ising model. This model de-

scribes the behavior of a set of magnets on a square lattice of length side N . The state of each lattice site i at time t is defined by the spin $S_t(i)$. Each spin can be in the states upwards (1) or downwards (-1) and interacts with its four nearest neighbors to minimize energy, that is, to have parallel alignment. The global magnetization is $M = \sum_K S_i$, and the idea is to plot this measure as a function of the temperature. For high temperatures, noise dominates, and the distribution of spins is random, i. e., $M = 0$. At very low temperatures, the system is ordered and all spins point towards the same direction (either upwards or downwards). M becomes maximum. As we progressively decrease the temperature, a sudden transition takes place at a critical temperature T_c . The magnetization per spin $m = M/N$ behaves close to T_c (for $T < T_c$) as $m \approx |\tau|^\alpha$ where $\tau = (T - T_c)/T_c$ [107].

A relevant parameter to characterize spatially distributed systems is the correlation length ξ . For $T > T_c$ we already said the system is random and correlation lengths are small. Close to T_c , ξ scales as $\xi \approx |\tau|^{-\nu}$, ν being another critical exponent. Below the critical point, the model exhibits long-range order. Clusters exist on every length scale. That is, the system is scale-free. The correlation length ν , the size of the maximum cluster, and the variance in sizes diverges to infinity as we approach the critical point.

Percolation theory has had a nice application as a null model in landscape ecology [122]. A useful example is its

application to the problem of habitat fragmentation. This is an extraordinarily complex problem due to the accelerating rates of habitat destruction everywhere and the well-known fact that habitat transformation is the number one cause of biodiversity decline. Imagine a spatial lattice as the one described above. A direct application of percolation consists in envisioning a situation in which each site is originally pristine, i. e., occupied by vegetation, and one proceeds by destroying an increasing fraction of randomly placed sites. As for the Ising example, the size of the largest patch starts declining smoothly at the beginning. A new destruction event just reduces the size of the single large patch by one site. But close to the percolation threshold, an additional destruction implies that the previously continuous cluster breaks down in small pieces. To separate the effects of habitat loss from those of habitat fragmentation, one can use the following order parameter [13]:

$$\Omega = \frac{S_{\max}}{\sum_{k=1}^N \Theta(k)}, \quad (8)$$

where S_{\max} is the size of the largest cluster, and $\Theta(k)$ is one if site k is available, and zero if it is destroyed. As can be easily seen, when all available sites belong to the same cluster, the previous equation is one. From a biological point of view, it means that we are only facing habitat loss. However, when habitat loss induces habitat fragmentation, the value of the order parameter drops suddenly. This is because we have now several disjoint clusters of vegetation, and thus only a small fraction of the available sites belong to the largest cluster. Interestingly enough, the order parameter drops really fast near the percolation threshold, so its value is one below a critical level of habitat destruction, and becomes almost zero after that threshold.

Extinction Thresholds

The above non-linear changes in landscape structure as more habitat is destroyed have implications for the persistence of a species inhabiting such a landscape. Species inhabiting heterogeneous landscapes living in a dynamical balance between local extinctions and recolonizations from neighborhood patches are called metapopulations [44,66]. If we plot the regional abundance of a metapopulation (i. e., the fraction of sites occupied) versus the fraction of sites destroyed, one observes the presence of an extinction threshold defined as a critical destruction value at which the metapopulation goes extinct despite a fraction of the habitat is still available [61]. In spatially explicit systems with local dispersal, the rate at which a metapopulation's regional abundance decreases is

faster than in the case of spatially implicit models. That is, the effects of habitat loss are higher as more habitat has already been destroyed [13]. The reason has to do with the previously reported non-linear changes in the landscape. Essentially thus, habitat destruction models are equivalent to models of phase transitions in statistical mechanics. This theory has served to better understand the consequences of habitat destruction on metapopulations. It has been very pedagogical in suggesting how changes in ecological systems are not smooth, but rather non-linear.

The presence of extinction thresholds in metacommunity dynamics is equivalent to the eradication thresholds in epidemiological models. Nee [87] already noted the equivalence between these two types of models and their critical points. These critical points can be simplified to expressions where species-specific parameters such as colonization or transmission rates cancel out. Eradication thresholds can be phrased as the points at which the destructive process reaches the amount of resource used when all the resource is available. For example, consider the following metapopulation model originally proposed by Levins [66]:

$$\frac{dv}{dt} = cv(1 - v - D) - ev. \quad (9)$$

The previous equation assumes an infinite number of habitat sites and describes the temporal dynamics in the fraction of sites occupied by a metapopulation (v). It contains a positive term describing the increase in occupied sites due to the colonization of empty sites. c is the species-specific colonization rate, and D is the fraction of sites permanently destroyed. The second term in Eq. (9) refers to the loss of previously occupied sites due to local extinction, e being the extinction rate. Note that the previous model can also be used as a toy model of an infectious disease, in which case v would be the fraction of hosts infected, c would be the transmission rate, D would be the fraction of hosts vaccinated and e would be the clearance rate. Model (9) has a positive steady state as long as the fraction of sites destroyed (hosts vaccinated) is lower than the threshold $D_c = 1 - e/c$. As noted by Nee [87], this threshold is equivalent to the steady state in the absence of destruction: $v^* = 1 - e/c$. Thus, one can predict D_c without knowing the parameters c and e by measuring the amount of habitat occupied when all habitat is pristine. This simple rule was coined as the Levins rule by Hanski et al. [45]. In epidemiological models, the infectious disease disappears when the fraction of hosts vaccinated is equal to the fraction of non-infected hosts when all hosts are suscep-

tible [11,87]. Similar ideas can be applied to the dynamics of transposable elements, a type of intragenomic parasites [11].

Stochastic spatial models such as the contact process have allowed us to derive mathematical conditions under which a species will persist or die out with almost certainty, that is, similar thresholds for species persistence [31]. In this review we have encountered both spatially explicit models such as the CMLs and the interacting particle system, as well as spatially implicit models such as Model (9). There are analytical approaches such as moment-closure that allow to bridge between these two extremes [64].

Ecosystem Shifts

Thresholds such as the one here illustrated for the case of habitat loss are common in ecology. One can find several examples in which a variable is smoothly changed with no apparent consequence in the macroscopic properties of the system until a threshold in which an abrupt transition in the state of the system takes place. These are known as ecosystem shifts [107,111]. One classical example of an ecosystem showing two alternative steady states is that provided by shallow lakes [22,107]. There are documented examples where a lake has shifted between an initial state characterized by clear water and submerged vegetation to a state characterized by turbulent water, a high concentration of phytoplankton, and an absence of submerged vegetation. This lack of vegetation is associated with a reduction in diversity, since several species of fish and other taxa use vegetation as food and refugia [107]. This transition in shallow lakes occurs as a consequence of human-induced eutrophication, and constitutes a global problem affecting also small seas such as the Baltic. The tuning parameter in this example would be the amount of fertilizers dumped into the lake. As one starts increasing this parameter nothing seems to occur for a while. Until the critical point is reached and the new state suddenly takes place. Once more, there is no apparent correlation between the last push and its amplified consequence. Shallow lakes show a profound hysteresis in response to nutrient load [107]. This means that the system is irreversible, and that environmental actions to recover the pristine state may be costly. Now one needs to almost clean the lake completely to revert the change to the pristine state.

The case of a lake is by no means the only example. Other examples involve the transitions from corals to macroalgae, or from herbaceous vegetation to bare desert. In the first case, extensive areas of coral reef have been

replaced by a system dominated by algae. Corals hosts countless numbers of other species. Currently, ecologists have started documenting these transitions and looked for their explanations. Several non-exclusive explanations involve increased nutrient loading and overfishing that have decreased the abundance of herbivorous fish, thus freeing algae from their control and allowing them to take over and replace corals [52]. Overfishing of sharks may have also contributed to the depletion of herbivorous fish through trophic cascades [15].

The transition from a vegetated state to a desert one is also one of concern. Vegetated and desert seem to be two alternative stable states. This example has further implications in the context of global change and can feed-back into further increases of temperature without the layer of vegetation. Knowing that these transitions are irreversible due to the hysteresis cycle is worrisome as it has also profound implications in the context of human migration in search of available water.

Future Directions

The previous cases of ecosystem shifts suggest the role for non-linearities in conservation biology. The take home message is that we can not think anymore in linear terms. There is not necessarily a proportional relationship between cause and consequence. This calls for caution when assessing the consequences of global change and other types of human-induced perturbations. For example, the consequences of habitat destruction may even be worse than expected and further destruction values may cause the system to cross a threshold where extinctions may take place at a much higher rate. The existence of ecosystem shifts suggests that ecological systems may behave in qualitatively similar terms as other simpler physical systems. This is good news in the sense of being able to use a well-developed theoretical framework to make predictions of ecological systems. Near the critical points the system's macroscopic properties may be described by simple models [111].

On the other hand, due to the abrupt changes that take place in the critical points, it is very important to develop early indicators of the proximity of a system to such thresholds. For example, Kleinen et al. [56] analyzed changes in the power spectrum of temporal series and concluded that there is a reddening of the signal in the vicinity of the critical point. Similarly, there is increasing evidence that a clear early-warning signal of an ecosystem shift is an increase in the variance of the temporal series [21,124]. Further studies will be very useful in developing new and easy to measure early-warning signs. This may be very im-

portant in predicting major shifts in the state of ecosystems and the services they provide.

To sum up, ecological systems are wonderful examples of complex systems with multiple states, phase transitions, and non-linear dynamics. They provide opportunities to further apply concepts and tools from the physics of complex systems. And in the face of the multiple risks from global change, there is an urgent need to do so.

Funding was provided by the European Heads of Research Councils, the European Science Foundation, and the EC Sixth Framework Programme through a EURYI (European Young Investigator) Award.

Bibliography

Primary Literature

- Albert R, Jeong H, Barabasi AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378–382
- Amaral LA, Scala A, Barthélemy M, Stanley HE (2000) Classes of small-world networks. *Proceedings of the National Academy of Science of the United States of America* 97:11149–11152
- Anderson RM, May RM (1991) *Infectious diseases of humans. Dynamics and control*. Oxford University Press, Oxford
- Ashby WR (1954) *Design for a Brain*. Chapman and Hall, London
- Ashworth L, Aguilar R, Galetto L, Aizen MA (2004) Why do pollinator generalist and specialist plant species show similar reproductive susceptibility to habitat fragmentation? *J Ecol* 92:717–719
- Barabasi AL, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
- Barrat A, Barthélemy M, Pastor-Satorras V, Vespignani A (2004) The architecture of weighted complex networks. *Proceedings of the National Academy of Sciences of the United States of America* 101:3747–3752
- Bascompte J, Jordano P (2007) Plant-animal mutualistic networks: the architecture of biodiversity. *Annual Rev Ecol Evol Syst* 38:567–593
- Bascompte J, Melián CJ (2005) Simple trophic modules for complex food webs. *Ecology* 86:2868–2873
- Bascompte J, Rodríguez MA (2000) Self-disturbance as a source of spatiotemporal heterogeneity: the case of the tall-grass prairie. *J Theor Biol* 204:153–164
- Bascompte J, Rodríguez-Trelles F (1998) Eradication thresholds in epidemiology, conservation biology and genetics. *J Theor Biol* 192:415–418
- Bascompte J, Solé RV (1995) Rethinking complexity: modelling spatiotemporal dynamics in ecology. *Trends Ecol Evol* 10:361–366
- Bascompte J, Solé RV (1996) Habitat fragmentation and extinction thresholds in spatially explicit metapopulation models. *J Anim Ecol* 65:465–473
- Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant-animal mutualistic networks. *Proceedings of the National Academy of Sciences of the United States of America* 100:9383–9387
- Bascompte J, Melián CJ, Sala E (2005) Interaction strength combinations and the overfishing of a marine food web. *Proceedings of the National Academy of Sciences of the United States of America* 102:5443–5447
- Bascompte J, Jordano P, Olesen JM (2006) Asymmetric coevolutionary networks facilitate biodiversity persistence. *Science* 312:431–433
- Berryman AA, Millstein JA (1989) Are ecological systems chaotic: and if not why not? *Trends Ecol Evol* 4:26–28
- Bersier L-F, Banašek-Richter C, Cattin M-F (2002) Quantitative descriptors of food-web matrices. *Ecology* 83:2394–2407
- Bjørnstad ON, Peltonen M, Liebhold AM, Baltensweiler W (2002) Waves of larch Budmoth outbreaks in the European Alps. *Science* 298:1020–1023
- Camacho J, Guimerà R, Amaral LAN (2002) Robust patterns in food web structure. *Physical Review Letters* 88:228102
- Carpenter SR, Brock WA (2006) Rising variance: a leading indicator of ecological transition. *Ecology Letters* 9:308–315
- Carpenter SR, Kitchell JF (1996) *The trophic cascade in lakes*. Cambridge University Press, Cambridge
- Chesson P (1994) Multispecies competition in variable environments. *Theor Popul Biol* 45:227–276
- Chesson PL, Ellner S (1989) Invasibility and stochastic boundedness in monotonic competition models. *J Math Biol* 27:117–138
- Cohen JE (1978) *Food Webs and Niche Space*. Princeton University Press, Princeton
- Connell JH (1961) Influence of interspecific competition and other factors on distribution of barnacle *Chthamalus stellatus*. *Ecology* 42:710–723
- Connell JH (1978) Diversity in tropical rain forests and coral reefs- high diversity of trees and corals is maintained only in a non-equilibrium state. *Science* 199:1302–1310
- Costantino RF, Cushing JM, Dennis B, Desharnais RA (1995) Experimentally induced transitions in the dynamic behaviour of insect populations. *Nature* 375:227–230
- Costantino RF, Desharnais RA, Cushing JM, Dennis B (1997) Chaotic dynamics in an insect population. *Science* 275:389–391
- Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: the role of connectance and size. *Proceedings of the National Academy of Sciences, USA* 99:12917–12922
- Durrett R, Levin SA (1994) Stochastic spatial models: a user's guide to ecological applications. *Philos Trans R Soc London B* 343:329–350
- Egerton FN (2007) Understanding food chains and food webs, 1700–1979. *Bull Ecol Soc Am* 88:50–69
- Ellner SP, Turchin P (1995) Chaos in a noisy world: new methods and evidence from time-series analysis. *Am Nat* 145:343–375
- Erdős P, Rényi A (1959) On Random Graphs. *Pub Math* 6:290–297
- Fagan WF, Hurd LE (1994) Hatch density variation of a generalist arthropod predator: population consequences and community impact. *Ecology* 75:2022–2032
- Feigenbaum M (1978) Quantitative universality for a class of non-linear transformations. *J Stat Phys* 19:25–52
- Fortuna MA, Bascompte J (2006) Habitat loss and the structure of plant-animal mutualistic networks. *Ecol Lett* 9:281–286

38. Glass L, Mackey MC (1990) *From Clocks to Chaos. The Rhythms of Life*. Princeton University Press, Princeton
39. Guimarães PR Jr, de Aguiar MAM, Bascompte J, Jordano P, dos Reis SF (2005) Random initial conditions in small Barabási-Albert networks and deviations from the scale-free behavior. *Phys Rev E* 71:037101
40. Guimarães PR Jr, Machado G, de Aguiar MAM, Jordano P, Bascompte J, Pineiro A, dos Reis SF (2007) Build-up mechanisms determining the topology of mutualistic networks. *J Theor Biol* 249:181–189
41. Guimarães PR Jr, Sazima C, Furtado dos Reis S, Sazima I (2007) The nested structure of marine cleaning symbiosis: is it like flowers and bees? *Biol Lett* 3:51–54
42. Guimerà R, Amaral LAN (2005) Functional cartography of complex metabolic networks. *Nature* 433:895–900
43. Gurney WSC, Nisbet RM (1998) *Ecological Dynamics*. Oxford University Press, Oxford
44. Hanski I (1999) *Metapopulation Ecology*. Oxford University Press, Oxford
45. Hanski I, Moilanen A, Gyllenberg M (1996) Minimum viable metapopulation size. *Am Nat* 147:527–541
46. Hassell MP, Lawton JN, May RM (1976) Patterns of dynamics behaviour in single-species populations. *J Anim Ecol* 45:471–486
47. Hassell MP, Comins HN, May RM (1991) Spatial structure and chaos in insect population dynamics. *Nature* 353:255–258
48. Hassell MP, Comins HN, May RM (1994) Species coexistence and self-organizing spatial dynamics. *Nature* 370:290–292
49. Hastings A (1993) Complex interactions between dispersal and dynamics: lessons from coupled logistic equations. *Ecology* 74:1362–1372
50. Hastings A, Higgins K (1994) Persistence of transients in spatially-structured ecological models. *Science* 263:1133–1136
51. Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology: is mother nature a strange attractor? *Annual Rev Ecol Syst* 24:1–33
52. Hughes TP (1994) Catastrophes, phase-shifts, and large-scale degradation of a Caribbean coral reef. *Science* 265:1547–1551
53. Jordano P, Bascompte J, Olesen JM (2003) Invariant properties in coevolutionary networks of plant-animal interactions. *Ecol Lett* 6:69–81
54. Kaneko K (1984) Period-doubling of Kink-antiking patterns, quasi-periodicity in Antiferro-like structures and spatial intermittency in coupled map lattices: towards a prelude to a Field Theory of Chaos. *Prog Theor Phys* 72:480–486
55. Kaneko K (ed) (1992) Special issue on Coupled Map Lattices. *Chaos* 2
56. Kleinen T, Held H, Petschel-Held G (2003) The potential role of spectral properties in detecting thresholds in the Earth system: application to the thermohaline circulation. *Ocean Dyn* 53:53–63
57. Kokkoris GD, Troumbis AY, Lawton JH (1999) Patterns of species interaction strength in assembled theoretical competition communities. *Ecol Lett* 2:70–74
58. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments revealed in food-web structure. *Nature* 426:282–285
59. Kuramoto Y (1984) *Chemical Oscillations, Chaos and Turbulence*. Springer, Berlin
60. Lafferty K, Dobson A, Kurtis A (2006) Parasites dominate food web links. *Proceedings of the National Academy of Sciences of the United States of America* 103:11211–11216
61. Lande R (1987) Extinction thresholds in demographic models of territorial populations. *Am Nat* 130:624–635
62. Lande R, Engen S, Saether BE (2003) *Stochastic Population Dynamics in Ecology and Conservation*. Oxford University Press, Oxford
63. Levin SA (1992) The problem of pattern and scale in ecology. *Ecology* 73:1943–1967
64. Levin SA, Durrett R (1996) From individuals to epidemics. *Philos Trans R Soc London B* 351:1615–1621
65. Levin SA, Segel LA (1976) Hypothesis for the origin of planktonic patchiness. *Nature* 259:659
66. Levins R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull Entomol Soc Am* 15:237–240
67. Lindenman RL (1942) The trophic dynamic aspect of ecology. *Ecology* 23:399–418
68. Lotka AJ (1925) *Elements of Physical Biology*. Williams and Wilkins, Baltimore
69. MacArthur RM (1964) Environmental factors affecting bird species diversity. *Am Nat* 98:387–397
70. Margalef R (1958) Information theory in ecology. *Gen Syst* 3:36–71
71. Margalef R (1968) *Perspectives in theoretical ecology*. Chicago University Press, Chicago
72. Maron JL, Harrison S (1997) Spatial pattern formation in an insect host-parasitoid system. *Science* 278:1619–1621
73. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. *Science* 296:910–913
74. May RM (1972) Will a large complex system be stable? *Nature* 238:413–414
75. May RM (1974) Biological populations with non-overlapping generations: stable points, stable cycles, and chaos. *Science* 186:645–647
76. May RM, Oster GF (1976) Bifurcations and dynamics complexity in simple ecological models. *Am Nat* 110:573–599
77. McCann K, Hastings A, Huxel GR (1998) Weak trophic interactions and the balance of nature. *Nature* 395:794–798
78. Melián CJ, Bascompte J (2002) Complex networks: two ways to be robust? *Ecol Lett* 5:705–708
79. Melián CJ, Bascompte J (2004) Food web cohesion. *Ecology* 85:352–358
80. Memmott J, Waser NM (2002) Integration of alien plants into a native flower-pollination visitation web. *Proceedings of the Royal Society Series B* 269:2395–2399
81. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alton U (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824–827
82. Morales CL, Aizen MA (2006) Invasive mutualisms and the structure of plant-pollinator interactions in the temperate forests of the north-west Patagonia, Argentina. *J Ecol* 94:171–180
83. Murdoch WW (1994) Population regulation in theory and practice. *Ecology* 75:271–287
84. Murdoch WW, Chesson J, Chesson PL (1985) Biological control in theory and practice. *Am Nat* 125:344–366

85. Murray JD (1989) *Mathematical Biology*. Springer, Heidelberg
86. Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. *Nature* 423:280–283
87. Nee S (1994) How populations persist. *Nature* 367:123–124
88. Neutel A, Heesterbeek JAP, Ruiters PC (2002) Stability in real food webs: Weak links in long loops. *Science* 296:1120–1123
89. Nicholson AJ, Bailey VA (1935) The balance of animal populations. 1st Proceedings of the Zoological Society of London 3:551–598
90. Odum HT (1956) Primary production in flowing waters. *Limnol Oceanogr* 1:102–117
91. Olesen JM, Eklundsen L, Venkatasamy S (2002) Invasion of pollination networks on oceanic islands: importance of invader complexes and epidemic super generalists. *Divers Distrib* 8:181–192
92. Olesen JM, Bascompte J, Dupont YL, Jordano P (2007) The modularity of pollination networks. submitted to Proceedings of the National Academy of Sciences of the United States of America
93. Pascual M (1993) Diffusion-induced chaos in a spatial predator-prey system. Proceedings of the Royal Society of London B 251:1–7
94. Pascual M, Dunne JA (eds) (2006) *Ecological Networks: Linking Structure to Dynamics in Food Webs*. Oxford University Press, Oxford
95. Paine RT (1969) A note on trophic complexity and community stability. *Am Nat* 103:91–93
96. Paine RT (1992) Food-web analysis through field measurements of per capita interaction strength. *Nature* 355:73–75
97. Pauly D, Christensen V, Dalsgaard J, Froese R, Torres F Jr (1998) Fishing down marine food webs. *Science* 279:860–863
98. Pimm SL (1979) The structure of food webs. *Theor Popul Biol* 16:144–158
99. Pimm SL (1982) *Food Webs*. University of Chicago Press, Chicago
100. Pimm SL, Lawton JH (1980) Are food webs divided into compartments? *J Anim Ecol* 49:879–898
101. Prill RJ, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization. *PLoS Biology* 3(11):e343
102. Raffaelli D, Hall SJ (1992) Compartments and predation in an estuarine food web. *J Anim Ecol* 61:551–560
103. Raffaelli D, Hall S (1995) Integration of Patterns and Dynamics. In: Polis G, Winemiller K (eds) *Food Webs*. Chapman and Hall, New York, pp 185–191
104. Ranta E, Kaitala V (1997) Travelling waves in vole population dynamics. *Nature* 390:456–456
105. Rasmussen DR, Bohr T (1987) Temporal chaos and spatial disorder. *Phys Lett A* 125:107, 125:107–110
106. Schaffer WM, Kot M (1986) Chaos in ecological systems: the coals that Newcastle forgot. *Trends Ecol Evol* 1:58–63
107. Scheffer M, Carpenter SR, Foley J, Folke C, Walker B (2001) Catastrophic shifts in ecosystems. *Nature* 413:591–596
108. Schroeder MM (1991) *Fractals, Chaos, Power Laws*. Freeman, New York
109. Sibly RM, Barker D, Hone J, Pagel M (2007) On the stability of populations of mammals, birds, fish and insects. *Ecol Lett* 10:970–976
110. Simon H (1955) On a class of skewed distribution functions. *Biometrika* 42:425–440
111. Solé RV, Bascompte J (2006) *Self-organization in complex ecosystems*. Princeton University Press, Princeton
112. Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. Proceedings of the Royal Society of London Series B 268:2039–2045
113. Solé RV, Valls J (1991) Order and chaos in a two-dimensional Lotka–Volterra coupled map lattice. *Phys Lett A* 153:330–336
114. Solé RV, Bascompte J, Valls J (1992) Stability and complexity of spatially extended two-species competition. *J Theor Biol* 159:469–480
115. Solé RV, Valls J, Bascompte J (1992) Spiral waves, chaos and multiple attractors in lattice models of interacting populations. *Phys Lett A* 166:123–128
116. Sugihara G, May RM (1990) Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature* 344:734–741
117. Sugihara G, Schoenly K, Trombla A (1989) Scale invariance in food web properties. *Science* 245:48–52
118. Taylor PJ (2005) *Unruly Complexity: Ecology, Interpretation, Engagement*. University of Chicago Press, Chicago
119. Thompson JN (1994) *The Coevolutionary Process*. University of Chicago Press, Chicago
120. Tilman D, Wedin D (1991) Oscillations and chaos in the dynamics of a perennial grass. *Nature* 353:653–655
121. Turing A (1952) On the chemical basis of morphogenesis. *Philos Trans R Soc Series B* 237:37–72
122. Turner MG, Gardner RH, Dale VH, O'Neill RV (1989) Predicting the spread of disturbances across heterogeneous landscapes. *Oikos* 55:121–129
123. Ulanowicz RE, Wolf WF (1991) Ecosystem flow networks: loaded dice. *Math Biosci* 103:45–68
124. van Nes EH, Scheffer M (2003) Alternative attractors may boost uncertainty and sensitivity in ecological models. *Ecol Model* 159:117–124
125. Volterra V (1926) Variations and fluctuations of the number of individuals in animal species living together. Translation In: Chapman RN (1931) *Animal Ecology*. Wiley, New York, pp 409–448
126. Wiener N (1948) *Cybernetics or control and communication in the animal and the machine*. Wiley, New York
127. Wootton JT (1997) Estimates and tests of per-capita interaction strength: diet, abundance, and impact of intertidally foraging birds. *Ecol Monogr* 67:45

Books and Reviews

- Hastings A, Hom CL, Ellner S, Turchin P, Godfray HCJ (1993) Chaos in ecology: is mother nature a strange attractor? *Annual Rev Ecol Syst* 24:1–33
- Levin SA (1999) *Fragile Dominion: Complexity and the Commons*. Perseus Publishing, Reading
- May RM (1973) *Stability and Complexity in Model Ecosystems*. Princeton University Press, Princeton
- Prigogine I, Stengers I (1984) *Order Out of Chaos*. Bantam Doubleday Dell Publishing Group, Toronto
- Stewart I (1989) *Does God play dice? The mathematics of chaos*. Basil Blackwell, Oxford

Ecological Topology and Networks

ÖRJAN BODIN

Stockholm Resilience Centre, Stockholm University,
Stockholm, Sweden

Article Outline

Glossary

Definition of the Subject

Introduction

Network Analysis

Food Webs

Network Perspective of Fragmented Landscapes

Summary

Future Directions

Bibliography

Glossary

Network A network is (1) a set of system entities (nodes), which may be interconnected through (2) links.

Graph A graph is the mathematical notation of a network, i. e., it is an abstraction of a network.

Nodes Nodes are the separate entities of a network, i. e., the system components that may be interconnected. Sometimes they are called vertices.

Links Links are the realizations of the possible connections among the nodes of a network. Sometimes they are called edges.

Topology Topology, in this context, represents the pattern of interconnections, i. e., the links, of a network. A topological analysis is hence concerned with analyzing the structural characteristics of the set of links defining the network.

Network analysis Network analysis refers to all kinds of quantitative approaches used in analyzing the topology of networked systems. In doing so, methods from the mathematical branch of graph theory are often applied.

Definition of the Subject

Ecology is, simply speaking, the science of ecosystems, i. e., sets of interacting species constrained by the physical environment. Due to earth's enormous species diversity, species' patterns of interactions quickly become very complex and thus difficult to oversee, although it is clear that the interaction patterns themselves often have profound effects on the behaviors and functioning of the ecosystems. To enable systematic pattern analyses, it is often favorable to represent these patterns of interactions as *net-*

works where the nodes are some sort of biological entities and the links represent some sort of interaction between these entities. The entities can e. g., represent species, but they could also represent individuals or groups of organisms. The links could e. g., represent trophic interactions ("who eats whom"). By representing interacting species as a network, analytical focus is set on the actual pattern, or *topology*, of the interactions themselves. Topological analysis of ecological systems has a relatively long history in ecology which can be exemplified by the long-lasting scientific debate, spurred by Sir Robert May in the 1970s when he, against prevailing interpretation, suggested that an increased number of species interaction would actually lead to decreased ecosystem stability. More recently, an increased interest among various scientific disciplines on network approaches in complex systems research has re-energized the topological perspective of ecosystem research.

Introduction

In nature, species interact in many different ways; no species exists in isolation. Understanding these interactions and how these affect individual organisms, species and whole ecosystems are, therefore, key to a systemic understanding of the natural environment. In the early twentieth century, Lotka and Volterra [1,2] paved the way for theoretical and mathematical approaches in understanding predator-prey interaction and the resulting dynamics of species populations. In the 1950s, the Odum brothers [3] revolutionized ecology by emphasizing the need for a systemic perspective of the natural environment [4]. In order to take the field of ecology beyond a mainly descriptive science, and to find solution to challenges facing the natural environment, they argued that better understanding of the large-scale properties of the environment is needed [4].

Ecosystems, although over time defined and/or perceived in many different ways, are basically systems of interacting species limited by constraints arising from the physical environment. The Odum brothers originally modeled ecosystems as sets of components (e. g., species) and flows of energy (the common denominator) cycling through these components. Thus, they essentially laid the foundation for seeing the environment as a networked system consisting of nodes and links. What constitutes a node depends on the question at hand; it could be a species, a group of similar species, an individual organism, groups of organisms, physical objects, etc. The links, i. e., the relations between the system components of interest, also depends on the chosen question; they could e. g., be flows

of energy going from prey to predators in food webs (e. g., [5]), or flows of genes spread through species dispersals among localized populations (e. g., [6]). Networks can be viewed as maps that outline these local interactions and preserve their importance at the system level [7]. The perspective of networks carries the advantage of simultaneously addressing the members of the systems as well as the patterns of their interactions. This modeling approach, emphasizing localized interaction between separated parts of the system, captures some of the fundamental characteristics of a complex adaptive system [8].

As mentioned earlier, the ecological impact of various characteristics of patterns of interactions has been studied and debated for a considerable period of time. Of focal interest here is, however, research approaches paying special attention to the actual pattern, or *topology*, of the interactions themselves. For example, how may the distribution of links among the nodes in food webs affect ecosystem stability in terms of risk of species extinctions? Alternatively, how may the average topological distance between pairs of habitat patches in a fragmented landscape affect metapopulation dynamics? These, and other topological aspects of ecological systems, are the subject of this review.

In order to analyze networked systems, formal methods addressing the topology are needed. Such methods have, fortunately, been developed in various scientific disciplines such as sociology, computer sciences and mathematics over a considerable amount of time (e. g., [9]). These methods aim at quantifying different topological aspects of concern, some of which are briefly described in the following section. Recently, the interest for topological analyses of various kinds of systems has grown tremendously; a development that e. g., has generated an abundance of new insights and approaches in topological analyses. This development has, among other things, contributed to re-energize the topological perspective in ecological research.

Network Analysis

A system of interconnected entities, i. e., a network, is mathematically represented as a graph. Graphs consist of nodes and links. Nodes are the terminal points or intersection points of the graph (sometimes called vertices). Links represent connections between nodes and represent the structure of the network over which interaction occurs. There is an entire branch of mathematics called graph theory that deals with the analysis of such graphs. Furthermore, network-oriented analyses are undertaken in several other disciplines; thus methodological, technical and theoretical developments of relevance for networked systems

are taking place across disciplines. An example of an interdisciplinary endeavor is the fast-growing organization INSNA (International Network for Social Network Analysis, see <http://www.insna.org>) consisting of sociologists, mathematicians, physicists, computer scientists and others that are mainly occupied with studying patterns of social interactions (i. e., social networks). Thus, the term network analysis will be used here when referring to all kinds of quantitative approaches in analyzing the patterns of interconnections in networked systems.

Modeling a Networked System

In conducting topological analyses, the first step would be to represent the system under study as a network, i. e., to define what entities will constitute the nodes, and what kind of relations among the nodes will constitute the links. Modeling a system as a network is in some cases straight forward, because it is more or less obvious what entities will make up the nodes, and what kind of relations would constitute the links. For example, in studying pattern of friendships among students in a classroom each student would constitute a node, and the reported friendship between any two students would be represented by a link. In other cases, it might be less obvious how to define a node. In studying the dynamic interactions in an ecosystem, should the nodes be represented by individual species, groups of similar species, or even by individual organisms? The method by which one chooses a suitable level of abstraction will depend on the research questions. The problem is, as always in modeling, to choose a level of abstraction that is as simple and aggregated as possible, but still fine-grained enough to capture the essential characteristics of the system in order to help answer the questions at hand.

The issue of choosing an appropriate level of abstraction likewise applies when defining the links. Often many different kinds of relations exist between the different entities (nodes) of the system under study; thus the question is which type of relations should be modeled? Furthermore, in many cases the strengths of the relations are of interest. The strength of a relation could be thought of as the intensity, and/or the frequency of interactions, the flow of energy, material, information etc between the nodes in question. Although the focus here is primarily on analysis of the topological properties of a networked system (where differences in the strength of the links are not considered), the issue of link strength is important when e. g., constructing the network. If the strength, or intensity, of the type of relation under consideration varies, it might be useful to define a threshold; and relations between any pair of nodes with strengths below such threshold would

thus be ignored. Therefore, when defining the threshold value, the resulting topology is affected since links below the threshold are omitted. Also, in this context it is important to mention that many commonly used measures defining various network characteristics can, in addition to purely topological characteristics, also take links strengths into account.

Finally, the links of the network representing the studied system could be bidirectional or unidirectional. For example, in food webs the predator consumes the prey, which is an example of a unidirectional flow of energy (which in a network context translates to a unidirectional link). If, however, two species prey on each other, the flow of energy is bidirectional.

Topological Characteristics

In this section, three important characteristics of networks, and some of the associated and commonly used measures defining these characteristics, are briefly reviewed. These are degree distribution, modularity and centrality. How these are of relevance in ecology will be discussed and exemplified in coming sections. Naturally, there are many characteristics that define a network other than those presented here. The applicability of various existing network measures in network-oriented ecological studies is almost a research topic in itself e. g., [10,11,12].

Degree Distribution and Small-World Properties In most real-world networks, links are very unevenly distributed among the nodes (see e.g., [13] for a review). If, for example, all links were distributed randomly, the degree distribution would follow a Poisson distribution. There are, however, lots of examples of real networks not following this distribution; thus there must be processes other than chance alone that are in play when networks are formed and shaped over time (e. g., [13]). Of recent interest are the so-called scale free networks following a power-law degree distribution, meaning that most nodes have few links, but that some rare nodes possess very many links (these nodes are often called hubs). Such networks are quite robust (in terms of the risk for severe network fragmentation) to random node removals (since most of the nodes have very few links), but they would be very vulnerable to a targeted removal of the hubs [14].

Somewhat related to the degree distribution is the concept of *small world* networks [15]. A small world network displays a high level of clustering, meaning that two nodes that both possess links to a common third are much more likely to be directly linked to each other as compared to the likelihood that any other arbitrary pair of nodes should be directly linked. In spite of this high degree of clustering,

in a small world network the average topological distance between any arbitrary chosen pair of node remains relatively short, thus implying that there are still many links that cross boundaries and therefore link together different clusters. A small world network does not necessarily follow any particular degree distribution, but it has been shown that scale-free networks often display small-world characteristics see (e. g., [13]).

Modularity Within a network there may be groups (or *modules*) of nodes that, from a topological perspective, distinguish themselves from the rest. For example, it could be that these groups of nodes are more internally than externally interlinked, i. e., the distinction of groups is based on a high density of interconnecting links among each group's members. In this way, a group would have a relatively high frequency of direct (or indirect) relations within the group compared to outside the group. Examples of such group-assessment methods are LS-sets and lambda sets (see [16] and references therein).

A specific example of another type of group definition is the clique ([17] and references therein). In a clique every member is connected to every other member; thus this definition of a group does not define members based on their relative cohesion versus non-member – instead it uses an absolute criterion for defining a group. The definition of a clique can be extended to account for directional and weighted links as well.

Generally, methods of assessing groups can be classified as either generating hierarchically nested groups, or as generating groups that can partially overlap. In the former case, an individual node cannot be a member of more than one hierarchical branch. Furthermore, a decomposition algorithm used to identify hierarchical branches can be either divisive or agglomerative, where agglomerative algorithms have been more commonly used [18]. If the focus is to find cores of strongly and/or intensely interlinked groups, agglomerative methods are preferred since they will identify groups using a bottom-up approach, starting with the most densely connected subgroups. Divisive algorithms, on the other hand, uses a top-down approach and iteratively divide the network in smaller and smaller branches.

A different way of distinguishing groups within a network is to group nodes according to the set of relations they have with one another (called equivalence, see e. g., [17]). Here, two members of a group would have similar sets of relations to others, i. e., they occupy similar positions in the network. This kind of group can e. g., be useful to locate functionally similar groups of species in a food web (as will be described below).

Centrality A fundamental topological characteristic of a node in a network is its level of *centrality*. The concept of centrality is devoted to analyzing the position of nodes' in the network. The underlying assumption is that some positions are more favorable than others in terms of the influence the nodes occupying them can exert on others (for an introduction and review of the literature, see e. g., [17]). There are, however, numerous ways to exert influence, and accordingly many different measures of centrality have been developed – each focusing on different topological aspects. Here some of these are presented:

1. Degree centrality. This is the number of links a node possesses. In a network with directed links, one could distinguish between in-degree and out-degree centrality.
2. Betweenness centrality (see [19]). This measure assesses how much “in-between” a particular node is, based on how many shortest pathways (connecting other nodes) that goes through this particular node.
3. Closeness centrality [20]. This measure assesses how close (from a topological perspective) a particular node is to the rest of the nodes in the network.

Food Webs

In nature, species interact in many ways. One of these types of interactions is predation, i. e., when a species eat another species. Predation can also occur within species (cannibalism) when larger individuals eats smaller one. In a food web, the nodes represent a set of interacting species and the links represent the structure of energy flows, i. e., the links represent who eats who. Typically, the links in a food web are directed, meaning that a specific predator species consumes one or several specific prey species, and not vice versa (even if bi-directed size-dependent predation does occur, which may e. g., pose challenges in modeling food webs, see e. g., [21]).

The strength of the link between species A and species B can be assessed in various ways. First, it can be modeled as binary, i. e., either species A eats species B or not, and the analytical focus is thus set on the food-web's topology alone. The strength could, however, be weighted according to the change in the abundance of species A following the removal of species B (functional edges, see [5]). It could also be assessed based on the flow of energy between species A and species B. Other assessment methods are also available (for an overview, see [22]).

When assessing the link strength as in the last two examples, research focus is not set solely on the topology of

the food-web. This increases the complexity of the analysis, since the researcher has to (1) somehow assess the link strength, and (2) take the assessed strength into account in studying the food-web structures. In some cases the increased level of complexity may be redundant (i. e., a topological analysis would suffice given the particular research question at hand), but in other cases the more complex approach might be needed to find structural network characteristics of interest (e. g., [5]). An example of the latter is a study by Krause et al. [23], where it was shown that the modular structure for some of the studied food webs was only detectable when taking link strength into account. Another example is given by Scotti and colleagues [24], where they showed that the ranking order of the most central nodes were highly affected when link weights were taken into account.

Much of early research, up until the late 1980s, in food web topology was in fact highly affected by problems associated with assessing link strength. Much of the empirical data available at the time were lacking information of the weakest links, i. e., the links representing low levels of energy flows (corresponding to gut content <5%, see e. g., [25] for a review). However, weak links abound in empirical food webs (e. g., [26,27,28]). Later research has subsequently showed that when weak links are taken into account, structural patterns of food webs change drastically (e. g., [29]). Up to this point, theoretical development was in many cases actually based on incomplete data, often leading to false conclusions on the structural properties of food webs. One example was the false assumption of the independence between a species average number of links and the total number of species in the food web. Instead, the number of links per species increases as the number of available species increase [30].

Finally, of significant interest with respect to the topological research approach of complex ecosystem as described here, is the insight that the actual pattern of linkages, i. e., the food web topology alone (regardless of link strengths) appears to be extremely important in defining ecosystem characteristics (e. g., [31]). In this context, weak interactions representing relatively small rates of energy flows can have a large impact on systemic properties whereas interactions with large flows can reversely have a relatively small impact (see [32] for a review of a number of important contributions in this research).

Degree Distribution

In terms of degree distributions, it appears that food webs, like many other types of networks, often experience a skewed link distribution where a few nodes pos-

sess many links [33]. This is completely different from what would be expected if the links were distributed randomly. This makes food webs, in accordance with other systems experiencing, a skewed link distribution, quite robust with respect to a random removal of nodes (i.e., species), but quite vulnerable to a targeted removal of the most connected nodes [33]. The distribution of links in food webs is generally different from the previously mentioned scale-free degree distribution in that the tail of the distribution towards very high degree centralities is truncated (see e.g., [7] and references therein). Furthermore, since food webs often tend to experience a relatively high density of links as compared to many other kind of networks, the degree distribution accordingly gets more uniform.

In using a method designed to find subgroups (identical to a method named *k*-core used to find cohesive groups in social network, see [34]), Melián and Bascompte [35] found that of the five studied food webs, all had one dense and cohesive core subgroup. In these dense subgroups all species have a least *k* links to other group members. Thus, it is not only the distribution of links that is skewed; the species with most links appear to be connected to other highly connected species. Such a positive correlation between a species number of links and its neighbors average number of links (connectivity correlation, see e.g., [36]), has been reported in other studies as well (e.g., [37]).

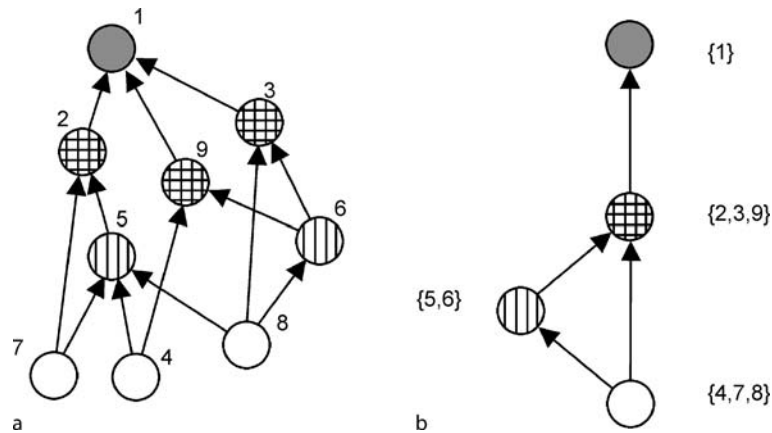
Sparsely connected food webs do also experience the small world effect [15], but this gets much less clearly articulated as the number of links increases [38,39,40]. Furthermore, it appears that the average distance between any two species in a food web is even smaller than in a small world network [33,41]. Thus, food webs could be considered as being rather tight when, on average, only two intermediate links are needed to jump between any two species in the network, and that >95% of the species are within three degrees of separation from each other [41].

In food webs with a high number of links, the small world effect often disappears. If, however, a food web's minimal spanning tree is analyzed, some interesting insights can be found. A minimal spanning tree is essentially a simplified version of the original food web where only the links that are minimizing the distances from all species to the basal species are kept, the rest are removed (e.g., [25]). Garlaschelli and colleagues [38] found that the minimal spanning trees experienced both the small-world effect and adhered to a scale-free degree distribution irrespective of the size or the link density of the set of seven different food webs studied. This suggests that there may exist some core structural principles upon which all food webs are built upon.

Modularity

With the significantly increased resolution in empirical food web data collection starting in the early 1990s, it has become increasingly feasible to look at the possibilities for modular structures in food webs (see [25] for a review). Modularity can, however, be defined – topologically – in (at least) two different ways. Modules can be defined as groups of species that are structural similar to each other, or it can be defined as groups of species that are interacting more intensively and/or frequently among themselves than with other species outside the group (hereafter called compartments). The former definition is exemplified in the concepts of trophic role or trophic level, although neither of these (and other similar) terms has been fully defined and collectively accepted in the ecological research community (see e.g., [10]). From a topological perspective, an interesting approach is to define, or single out, groups of species with similar trophic roles by analyzing the structure of the food web. Luczkovich and colleagues [10] used the network analytical approach based on regular equivalence, developed within the field of social network analysis (e.g., [17]), to partition species that play the same structural roles into different isotrophic classes (see Fig. 1). The approach, as described earlier, is based on finding positionally equivalent species in the food web, and provides a way to mathematically formalize trophic position, trophic group and trophic niche.

Research studying compartmentalization in food webs, according to the latter definition of modularity based on the frequency and/or intensity of interactions, has gained interest in recent years. One reason for the interest in compartmentalization lies in its potential effect on ecosystem stability and robustness (or using different terminology, resilience) (see e.g., [42]). A food web consisting of just one single coherent group of species would likely help to propagate both harmful and beneficial effects throughout the ecological community, whereas the presence of internally coherent but externally isolated or only weakly connected subgroups would reduce the possibilities for such large-scale propagations. This has led to the hypothesis that intermediate modularity may be the most beneficial structural composition of food webs (see e.g., [43]). Such structures are formed in networks where coherent groups of species with strong and/or frequent internal interactions are only weakly coupled to other groups. If this hypothesis holds true, one would expect to find modular structures in food webs since, over time, only the most robust food webs would be expected to prevail despite disturbances in the past.



Ecological Topology and Networks, Figure 1

a A food web where nodes (species) are patterned according to their regular equivalence, and **b** the same food web, but here are regular equivalent species grouped together. Thus, a simplified image of the food web is created where the relations among groups of species occupying analogous positions is presented. (Source: [10])

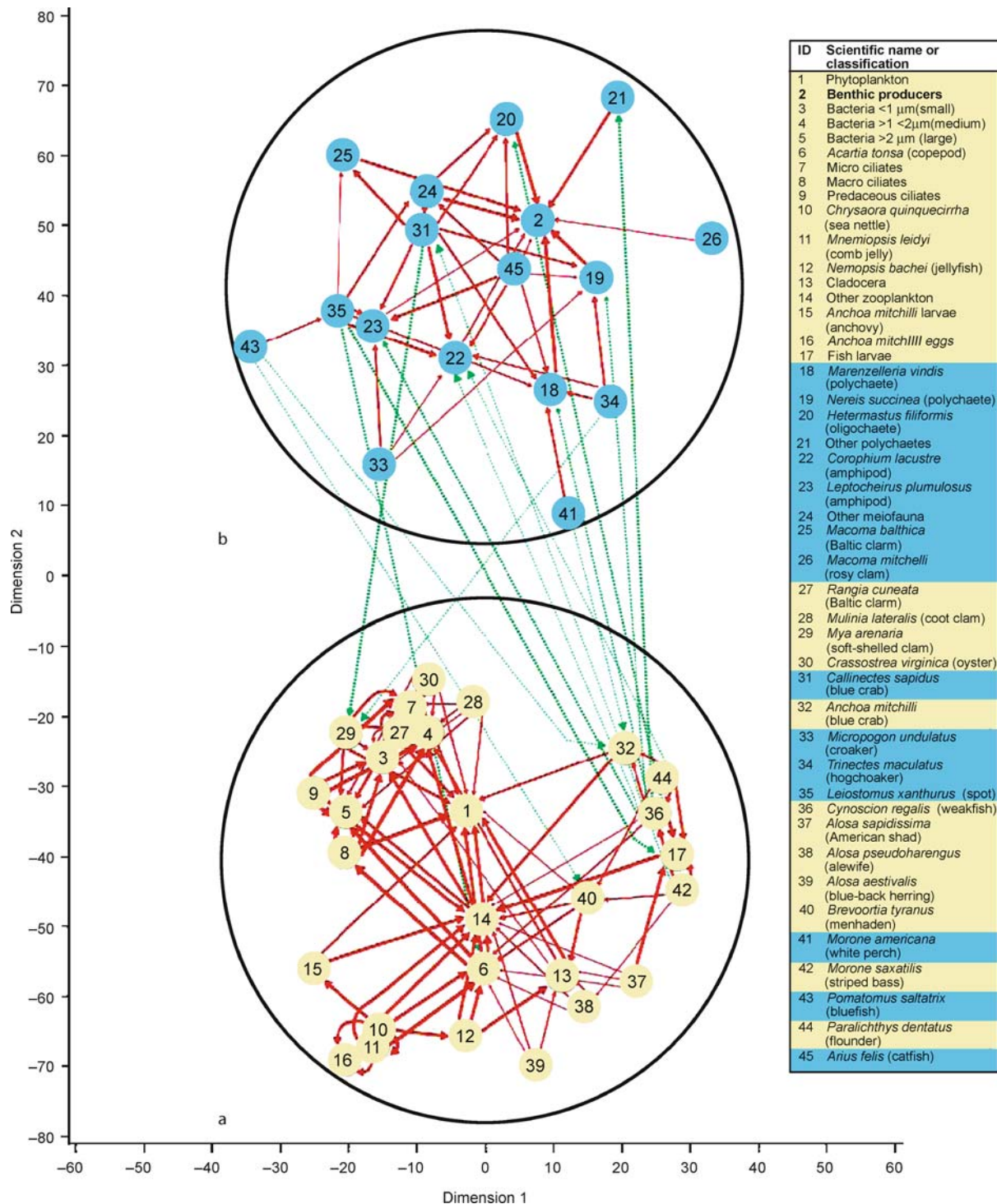
Food webs appear to be more coherent (i. e., less compartmentalized) than many other kind of biological networks (see e. g., [37,44]). This tendency seems to be derived, at least partly, from the relatively high link density that many food webs exhibit (e. g., [39]). It is obviously harder to distinguish subgroups when every species is connected to many others. In spite of this, recent studies have been able to find evidence for compartmentalization in food webs. Girvan and Newman [45] used e. g., a divisive hierarchical decomposition algorithm to detect ecologically relevant subgroups in a coastal food web. In addition, Krause and colleagues [23] were able to detect modularity in three out of five complex food webs (see Fig. 2 for an example) using a group assessment methods developed for analysis of groups in social networks [46].

Interestingly, Krause and colleagues barely found any compartments when the weakest links were left unconsidered, or when the food web consisted of very few species. In addition, the ability to detect compartments increased by taking link strengths into account. However, in another recent study by Dunne and colleagues [39], the presence of compartmentalization was not very pronounced. They found only five of sixteen food webs exhibiting a clear tendency for compartmentalization. Interestingly, but perhaps not surprisingly considering the previous argument, these were the sparsest food webs among the studied set.

Hence, although it seems clear that food webs often experience compartmentalization, existing evidence of a clear trend of intermediate level of compartmentalization in respect of stability is mainly from anecdotal evidence and models [7]. In the previously mentioned study by Krause and colleagues [23], they simulated the effect

of selectively removing species, one at the time, in one of the two distinct compartments in the Chesapeake Bay food web (each compartment corresponded to benthic and pelagic species respectively) alternatively. The resulting loss of links was found to be significantly higher in the compartment from where the species were removed. Thus this simulation supported the idea that compartmentalization reduces the spread of harmful effects. Direct evaluation of the impact of intermediate levels of compartmentalization for food web stability will, however, require long term food web data, for many different systems, with varying levels of compartmentalization which have experienced disturbance at some point during their monitoring.

Similar arguments as the ones proposing the beneficial effect of intermediate modularity are to be found in the discussion of the importance of weak links. Not only did the increased resolution of food web data reveal the abundance of weak links leading to the realization that food webs are much more complex then previously thought. Weak links may also help to stabilize oscillatory dynamics resulting from strongly interacting species [47], but see also May [48] for a different view on stability. In coupling oscillatory pairs, or groups, of species to other more stable groups of species, weak links act to reduce the amplitude of variations, thus helping to create a more stable system. Furthermore, if hypothesizing that interaction strength among species within coherent compartment is high, it may seem natural to assume that links bridging different compartments would be mostly weak. The presence of weak links thus suggests that the food web exhibits a modular structure (see e. g., [7]), although support for such statement is scant.



Ecological Topology and Networks, Figure 2

Graphical display of the results for the Chesapeake Bay food web with 45 taxa and weighted by interaction strength. Units are relative distances based on the inverse of the density of interactions. Within-compartment distances were decreased by a factor of 6.2 for aesthetic purposes. Circles indicate compartment boundaries and numbers identify taxa (yellow, within compartment a; blue, within compartment b). Arrows indicate interactions between taxa (solid red, within compartment; dashed green, between compartments; thickness indicates rank of associated interaction strength) and point from predator to prey (Source: [23])

Network Centrality and Keystone Species

The concept of keystone species in ecology is different, but related to, the concept of dominant species. A dominant species is a species that is high in abundance, and exert a large impact on the ecosystem where they are situated (see e. g., [33]). Keystone species, on the other hand, differ from dominant species in that their effects on the ecosystem are much larger than would be predicted from their abundance alone (e. g., [49]). It has been suggested that a keystone species is a species with a disproportionately high number of links in the food web [50]. Dunne et al. [33] have also put forward the idea of a structural keystone, i. e., a species that exerts influence on the basis of its structural position within the food web, and not only on the basis of the number of links it has to others.

The recent interest in using network analytical approaches in studying food webs have caused a renewed interest in the debate on whether increased ecosystem complexity (levels of diversity and connectivity) leads to less stability. Although many food webs do not experience a truly scale-free degree distribution, they nonetheless appear to be much more sensitive to a targeted removal of highly connected species than to a random removal (as described above). This is an effect of the skewed degree distribution, where the removal of the most connected species (i. e., the species with most links – highest degree centrality) will increase the average path length (i. e., increase the distance between consumer species and basal resources), and may also lead to network fragmentation. If a food web fragments into smaller isolated pieces, species might become excluded from their energy sources (their prey and/or basal resources); leading to secondary extinctions. From a purely topological perspective, a secondary extinction occurs when a non-basal species loses all of its prey species (e. g., [33]). Hence, in considering the risk of network fragmentation leading to secondary extinctions, and degree centrality is obviously of great relevance in assessing individual species' importance. Furthermore, the number of secondary extinctions following random species removal is often highly non-linear. Dense food webs are relatively insensitive to species loss up to a certain point. Beyond this point, the number of secondary extinctions sharply increases [33].

Dunne and colleagues [33] have, however, demonstrated that degree centrality alone cannot always predict the number of secondary species extinctions following species loss. In some cases the removal of species with a relatively few links led to a disproportional large number of secondary extinctions (hence leading to their notion of structural keystone species). Such structural keystone

species might play an important role in maintaining the minimum spanning tree, i. e., they may possess exclusive, non-redundant links that, more so than other links, help keep the food web together (see discussion below on non-redundant links and structural holes).

The insight that a species structural position in a food web may have a strong impact on the ecosystem is not entirely new. There are several examples in the literature of cascading effects taking place on the level of entire ecosystems following the extinction of certain keystone species. One of the most famous examples is the large scale destruction of underwater kelp forests following the extinction of sea otters due to intensive hunting for the fur trade [51,52,53]. Sea otters predate sea urchins, which in turn predate algae such as kelp. When sea otters disappeared, urchin abundance increased drastically leading to the destruction of the kelp forests. The destruction of kelp forests, in turn, produced cascading effects on other species dependant on the kelp resulting in a species-poor ecological community. Later, conservation measures designed to protect sea otters were implemented, which resulted in a recovery of the original ecological community composition [54]. Hence, as seen from a topological perspective, this example shows how a species (the sea otter), in occupying a structurally influential position in a food web (here, the Pacific coastal marine food web), become a keystone species.

Related to the discussion on structural keystone species is the insight that not only direct links, but also indirect links, are important in food webs because, in combination, they determine a species contribution in maintaining the overall structure of the food web. This was e. g., clearly demonstrated by the sea otter example presented above. The influence of indirect links is caused by the fact that these connect otherwise disconnected species. For example, a species that in itself possesses only very few links, but has network neighbors that possess many links, can still exert a very large impact if it is the only species that, indirectly, links its neighbor's neighboring species. This example illustrates the relevance of betweenness centrality [19], described earlier, in studying aspects of species' structural influence in food web. A similar example of a centrality measure is the concept of structural holes in social network analysis [55], where a distinction is made based on whether a particular link is redundant or not. A redundant link (or path) has an equivalent counterpart in the sense that there are other links that also connect its corresponding pair of nodes, whereas removing a non-redundant link would disconnect the pair of nodes. Species possessing non-redundant links clearly exert a negative influence on a food webs ability to withstand species loss or

extinctions since, if removed, they would disconnect consumer species from basal and fundamental energy sources. Non-redundant links can be particularly important when they connect otherwise disconnected groups of species.

Finally, there are numerous other measures of centrality available; each designed to assess a particular characteristic of a node's structural importance in a network. Research exploring the applicability of such, and other, topological centrality measures in advancing the concept of keystone species in food webs has just started (see e.g., [56,57,58]). In all, species abundance, link distributions, network position, and the strength of the links, clearly play a role in determining the impact a certain species exerts on the ecosystem. Thus, the relationships among network positions, keystone species, and highly interactive species (i.e., species with many connections) are not entirely straightforward but characterized by complex dependencies and dynamics. Some of these will be discussed below.

Dynamics of Secondary Extinctions

The topological approach to analysis of food web stability, with respect to their vulnerability to secondary extinctions following species removals, clearly lacks a dynamic part. I.e., the analysis is based solely on the topological changes of the food web following species removals. Hence, secondary extinctions are assessed only on the assumption that species that become isolated from all their prey species go extinct, otherwise they remain. Although this assumption is reasonable in many cases, it does not take into account indirect dynamical effects resulting from species loss (see e.g., [59]). For example, two competing prey species may coexist due to predator mediation. If the predator goes extinct, one of the prey species may then out-compete the other. This leads to the counter-intuitive result that a prey species may in fact depend on its predator for its prevalence. Such effects are very difficult to assess in purely topological analyses, as are effects of varying link strengths (see e.g., [60]). Also, the time it takes for a ghost species (a species doomed to become extinct, see [61]) to go extinct may be important. If, for example, the time it takes for a certain species to go extinct is significantly longer than the rate upon which new individuals are moving into the area, the extinction might never take place [62].

However, the strength in using topological analyses lies in its relative simplicity. A dynamical perspective typically involves solving N coupled differential equations (N -species Lotka-Volterra models) numerically in studying food webs with N species. This requires access to large computer processing capacity even for rather small food

webs. Secondly, in order to correctly parameterize the set of differential equations, detailed empirical data on the strength of all species interactions are needed.

Plant/Pollinator (Mutualistic) Networks

Until now, focus has been on patterns of predation among species. In nature, species relations are, however, often mutualistic. In such cases both species benefit from their interactions. A typical example of mutualism is given by the relations between plants and pollinators (e.g., insects, birds, mammals etc.). Such mutualism may be just as ecologically important as trophic interactions [63]. As for food webs, the set of mutualistic interactions among a given set of species, such as plants and pollinators, can favorably be described as an ecological network. Thus, the topological characteristics of mutualistic networks have spurred interest among scholars. Jordano and colleagues [64] have e.g., studied a large set of mutualistic network of plants and the animals that pollinate them or disperse their seeds. They found that the degree distribution follows a truncated power law. That is, up to a certain degree, the relation between the number of links per species versus the number of species follows a power law (i.e., the scale free characteristic as discussed previously). Beyond that critical degree, the number of species decline much faster. Thus, these networks are, as for many food webs, vulnerable to a targeted removal of highly connected species, but not as vulnerable as other kinds of networks that strictly follows the power law.

Furthermore, Bascompte and colleagues [65] have studied 27 plant-frugivore networks and 25 plant-pollinator networks, and they found that most of these network consisted of a highly connected core and asymmetrically connected peripherals. This topological characteristic was denoted *nestedness*, which is similar to the concept of core-peripheral structures in social network analysis (see e.g., [66]). The core consists of generalist species where, ideally, every species is connected to every other generalist species; whereas the peripherals are specialized species that, ideally, only link to generalist species. This topological structure clearly departs from the compartmentalization that is often found in food webs as discussed previously.

Network Perspective of Fragmented Landscapes

In the early days of theoretical ecology, space was largely omitted from the analysis. The only dimension was time, and two-species interaction described by continuous time equations made up the bulk of the work ([32], which present a comprehensive review of the development in

theoretical ecology up to present time). As time went on, observations from the real world began to challenge theory development; it seemed something was missing in the equations. It was later shown that this missing ingredient was space. For example, Alan Turing showed, in the early 1950s, that an initially homogenous system could emerge into a heterogeneous and ordered spatial structure – a finding that influenced e.g., theoretical ecology (see also [32,67]). Following the introduction of space into ecology, species movements were hereafter acknowledged as being of crucial importance.

In the late 60s, Levin [68] introduced the concept of metapopulation. A metapopulation is a set of spatially separated local populations which may, individually, undergo local extinctions, but where the aggregated population is maintained by re-colonization events resulting from dispersing organisms. Thus, even though a local population (confined to a spatially distinct patch in the landscape) may go extinct, dispersing organisms from other local populations can re-colonize the empty patch. The fairly simplistic model of metapopulation dynamics introduced by Levin was later extended by Lande [69] who incorporated more details. He also applied his metapopulation models to study the dynamics of the northern spotted owl in northern California. By assessing the fraction of occupied patches, and the fraction of suitable patches, he could define an extinction threshold. If the amount of suitable patches drops below a certain threshold, the whole metapopulation would go extinct.

Levin's and Lande's models are spatially implicit, i.e., they take space into account, but they do not explicitly incorporate geographic location in the models. They assume global mixing, i.e., a mean field approximation of dispersals where all patches are equally likely to receive dispersing organisms. Thus, these models are not able to model the effect of habitat fragmentation, i.e., when habitat patches gets more and more separated, or even disconnected, from each other (from the perspective of a dispersing organism). To enable that kind of analysis, spatially explicit metapopulations models are needed (see e.g., [70,71,72]).

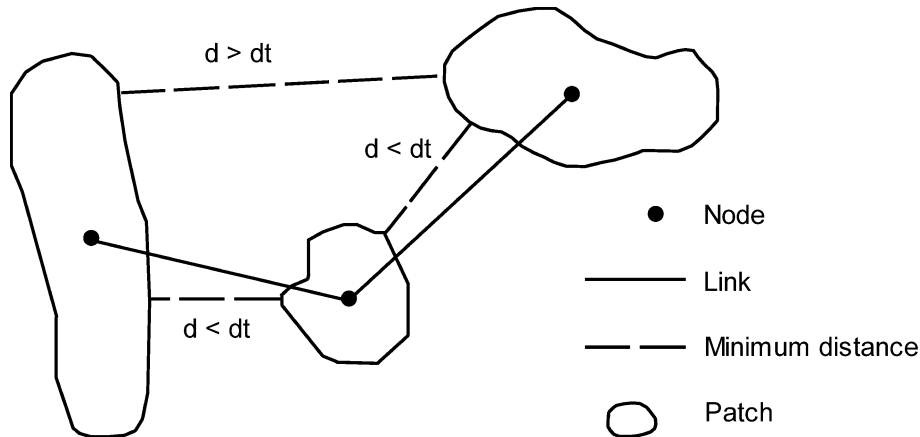
Hanski and Ovaskainen [73] used a matrix notation to describe the flux of dispersing organisms between individual patches at the landscape level in their spatially explicit metapopulation model. They then defined the leading eigenvalue of the matrix as the metapopulation capacity, and that single parameter summarized how the spatial structure of the landscape influenced the metapopulation's dynamics. Of particular interest here is the matrix representation. This matrix can be interpreted as a graph where each matrix element represents the link strength between

each pair of nodes (patches). Thus, it describes a network of interconnected patches. Keitt and colleagues [6] explicitly applied the graph-theoretical perspective in studying the level of connectivity of fragmented patches of ponderosa pine and mixed-conifer forest at the landscape level. They modeled how these patches might be experienced in terms of connectivity, or isolation, by dispersing juvenile Mexican spotted owls. Connectivity is, in this context, defined as the degree to which the scattered habitat patches in the landscape facilitate the dispersals of organisms [74].

The landscape network modeled by Keitt and colleagues consisted of nodes (representing individually and spatially distinct habitat patches in the landscape) and links (representing the possibility for dispersal between individual patches, see Fig. 3). Hence, such a network represents the landscape's spatial structure of connectivity (from an organism's point of view). It encapsulates the potential of an organism to traverse the landscape by moving from patch to patch [11]. Using this network perspective on fragmented landscapes, topological analysis of interacting local species populations is made possible. The network approach merges recruitment and dispersal processes with spatial patterns of habitat patches, thus enabling process-oriented analyses of landscape connectivity [75]. In comparison with food web analysis, where focus was on different interacting species, the network approach is here applied in studying how patterns of dispersals between separated habitat patches in the landscape may affect a target species metapopulation dynamics.

Modeling a Landscape as a Network

The starting point in studying topology of fragmented landscapes is to construct the network representation of the patches and their dispersal linkages. The first step in this process is to define what type of land is habitat, and what is not. In its simplest form, a binary image representing the landscape is then created, where all kinds of land cover assessed as being hospitable by the target species is marked, and the rest left unmarked and thus considered as inhospitable (called the landscape matrix, not to be confused with the matrix defined by Hanski and Ovaskainen [73]). The input data in constructing such binary image could be a satellite images or aerial photographs of the landscape under study. This binary simplification may, however, be too simplistic in analyzing heterogeneous landscapes (as will be discussed below), but there are available methods that can incorporate this heterogeneity into the model. Using the binary landscape image, a patch is defined as a coherent area of hospitable land



Ecological Topology and Networks, Figure 3

Construction of a landscape graph (d = distance, dt = threshold distance). This example landscape contains three separate habitat patches and a single component. The three patches belong to a single component because there exists a path along the links (solid lines) that connects all three patches. If either one of the links shown were removed, there would be two components. If all links were removed, then there would be three components, each consisting of a single patch (Adapted from: [6])

(in practice, all adjacent pixels that are marked as hospitable make up a single patch). From a network perspective, a patch is from now on seen as a node.

Emanating from the binary image, the dispersal links of the patches are then assessed. The simplest way to assess the links is to define a threshold distance [6]. Two habitat patches that are separated by a distance less than the threshold distance are considered connected, otherwise they are considered mutually disconnected and thus no link exists between these. The threshold distance should be defined based on the target species dispersal ability, i. e., how far the species is able to move in the inhospitable matrix. The threshold distance for a species depends on a number of different factors such as e. g., body mass, movement strategy, visual ability, and the type of land between the habitat patches.

Refinement of the dichotomous distance-based link assessment procedure is often preferred. Instead of the binary link/no link assessment; the link strength could be weighted proportionally, taking more factors than distance into account. The area of the patches could e. g., be factored in using a negative dispersal kernel (cf. [71], see also [76]). Thus, two large patches could be considered as connected, while another pair of smaller patches (although separated by the same distance) could be considered as disconnected (Eq. 1). Furthermore, directionality of dispersals could be accounted for. For example, consider a pair of patches where one is large and the other is small. Here, it might be reasonable to assume that organisms' movement towards the large patch is more probable than the opposite. Although link strengths are taken into consider-

ation, it would simplify the analysis if the weakest links are disregarded (by introducing a cut-off value C , see Eq. (1)).

$$S_{ij} = b \sqrt{A_j} e^{\alpha D_{ij}} \quad (1)$$

$$S_{ij} = 0 \quad (\text{if } S_{ij} \text{ was less than } C)$$

(see further details in [76])

S_{ij} = Metric proportional to the dispersal flux rate from patch i to patch j

A_j = Area of patch j (m^2)

D_{ij} = Distance between patch i and patch j (m)

α = Constant ($1/m$)

C = Cut-off value

b = Constant ($1/m$)

Until now, a binary habitat/matrix landscape has been assumed. This approximation may be too simplistic in heterogeneous landscapes where the permeability of different types of land covers differs considerably. If the matrix consists of an array of different land cover types, the ability to move between any arbitrary pair of patches depends on the exact composition of land types in-between, and not only on the geographic distance separating these patches (e. g., [77]). Fortunately, there are methods available that take the differing permeability of a heterogeneous matrix into account when calculating the least cost paths (or effective distance) of moving between any two arbitrary locations in the landscape [78,79,80,81]. Thus, more sophisticated link-assessment methods than just applying a threshold distance could favorably be used in analyzing heterogeneous landscapes. The key point is, however, that after all links have been assessed, the resulting network

representation will incorporate the topological characteristics of the fragmented landscape, and further analyses of the network do not depend on the chosen link-assessment method.

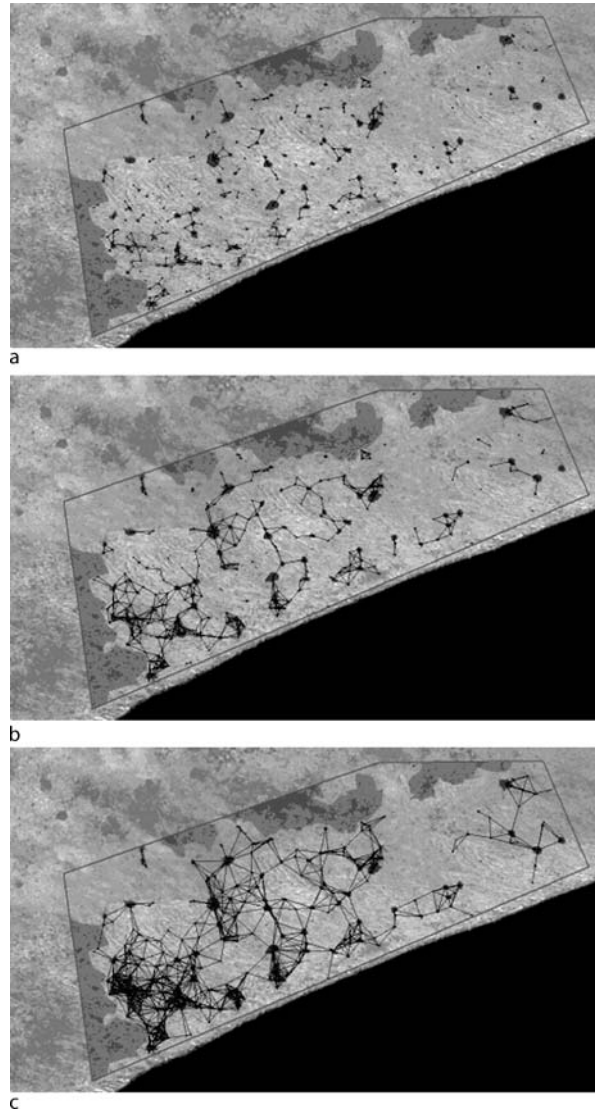
Assessing Fragmentation and Modularity

In spite of being fragmented, a landscape may very well be perceived as connected by species being able to move between patches of habitats [74]. In studying landscape connectivity, the overall research question is therefore to what extent the landscape supports species movement, i. e., how well connected are the scattered habitat patches as perceived by an organism on the move? Of central importance is the level of traversability, i. e., how far can dispersing organisms reach by moving from patch to patch in the landscape?

In terms of network terminology, traversability can be expressed as if the network itself is divided into separate subnets, or whether all nodes are confined within one single coherent network [6]. A subset of nodes where a path between all member nodes exists is called a network *component*. Thus, by definition, no paths exist between nodes of different components. Therefore, if a network is made up of two, or more, components, it is not possible to traverse the whole network by moving from node to node. Ecologically, an organism located in an arbitrary habitat patch can, directly or indirectly, move to any other patch within the same component, but it can never disperse to a patch belonging to another component (the landscape is compartmentalized). Thus, a component can host a set of connected but localized populations – a metapopulation – but the metapopulation itself would be isolated from other metapopulations confined to other components [11,75].

The number of components in a network representing the landscape's spatial structure of connectivity for a dispersing species would, therefore, directly relate to the actual level of fragmentation (or connectivity) in the landscape. A large number of components correspond to a heavily fragmented landscape, whereas a network with just one component implies a well-connected landscape where a large and persistent metapopulation may thrive. A second metric of connectivity is the percentage of patches confined to the largest component in the network. If the fraction of patches that belong to the largest component is high, the fraction of the landscape that is within reach for a single metapopulation would be high.

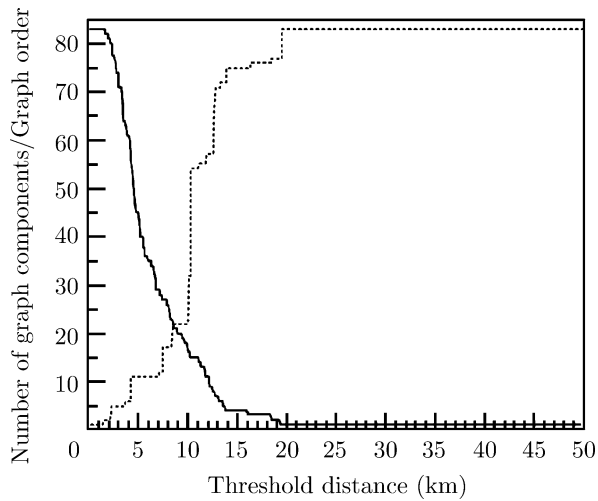
In recalling the concept of threshold distance, a question that arises is how the number of components relates to the threshold distance. Obviously, a very large threshold distance would imply that all patches would be reach-



Ecological Topology and Networks, Figure 4

The figures show an agricultural landscape in southern Madagascar with three different overlaid networks representing the connectivity among scattered forest patches assuming dispersal abilities set to a 500 m, b 1000 m, and c 1500 m. The forest patches are identifiable by the distinct dark spots, situated within a matrix consisting of cultivated land (light grey). Patches range in size from <1–95 ha and are fairly evenly distributed in the landscape. In the western and northern part of the studied area, the shaded/darker grey zones indicate areas classified as potential source areas (Source: [76])

able from each other, whereas a very small threshold distance would imply that virtually no patch is connected to any other patch (see Fig. 4). Thus, by measuring the number of components as a function of the threshold distance (edge thinning, see [78]), one can assess at what



Ecological Topology and Networks, Figure 5

Number of components (solid line) and size of largest component (dotted line) as a function of threshold distance (Source: [78])

threshold distance the landscape starts to be perceived as connected (see Fig. 5). The transition interval, i. e., where the number of components rapidly decreases as a function of an increased threshold distance, has been suggested as a scalar measure of the landscape's structural scale of

connectivity [82]. Species with dispersal capabilities below this transition interval would experience the landscape as severely fragmented, whereas species with dispersal capabilities above the interval would experience the landscape as connected.

Measuring the number of component seems to be a useful method in assessing spatial compartmentalization of the landscape. In some circumstances, it may however be too crude as it lacks the ability to estimate a more continual degree of landscape compartmentalization (i. e., where compartments are not necessarily isolated from each other, but merely separated to a certain degree) [11]. These types of loosely detached compartments of internally well-connected habitat patches can affect age and sex structures of populations in the landscape [83], and it could potentially contribute to genetic differentiations even at small temporal and spatial scales (cf. [84,85]). The inability to assess such structures is a consequence of the binary perspective of the component-based analysis; either two patches are being considered as connected, or else they are considered as being completely isolated. But what if a set of patches are very well connected internally, but rather weakly connected to the outside landscape? Using component-based analysis, such set of patches would not be identifiable unless they are completely isolated from the remaining landscape.



Ecological Topology and Networks, Figure 6

Decomposition of the largest component in the landscape (from Fig. 4b), using the Community Structure method [45], taken at the hierarchical level with the best "fit" ("Modularity" equal to 0.84, see [18]). Note that the largest component is decomposed into ten smaller compartments

Bodin and Norberg [11] addressed this possible limitation of the component-based analysis by using a divisive hierarchical decomposition algorithm called *Community Structure* recently proposed by Girvan and Newman [45]. This method fulfilled three potentially important topological criteria: (1) groups of coherent sets of patches could be assessed even though not necessarily being completely isolated from each other, (2) the algorithm is hierarchical and would therefore create non-overlapping groups of patches which means that it would help to divide the landscape into spatially separated compartments, and (3) it did not discriminate against peripheral patches (e.g., patches with just one link would be assigned the same group membership as their neighbor instead of being treated as isolated). Figure 6 illustrates how this method can be applied to decompose a single large component into a set of spatially distinguishable groups of patches.

Pascual-Hortal and Saura [86] have systematically studied how a set of ten different network-based measures, including the size of the largest component, respond to a number of different spatially explicit fragmentation scenarios. They suggest a new measure called *Integral index of connectivity* (IIC) which extends the previous size-of-largest-components analysis by also incorporating the area of the patches, as well as the actual topological distances between patches, into a single normalized measure of connectivity. They showed that the only measure that responded in a desirable and consistent way to the full set of fragmentation scenarios was the proposed IIS. However, the size of the largest component did also respond reasonably well.

Assessing Critical Habitat Patches

The concept of stepping stones is similar to the topological importance assigned to certain keystone species in food webs. A stepping stone patch can be seen as a bridge connecting otherwise separated groups of internally interconnected patches. A particular patch may also, although not being the very last remaining bridging patch in a otherwise largely fragmented landscape, exert influence since it could make it much more difficult for dispersing species to reach over the whole landscape if it is removed (i.e., its presence in the landscape provide shortcuts, thus reducing the overall topological distance between patches, see also [16]). Accordingly, assessing the impact individual patches exert on the overall landscape connectivity would be of uttermost importance in management targeted to preserve functional and well-connected landscapes (e.g., [11,75,86]).

A reasonable approach to assess individual patches' contribution to the overall connectivity of the landscape is to remove them, one by one, from the modeled network and then measure the effect on some relevant measures of landscape connectivity [75,78,86]. If, for example, the issue of interest is the topological traversability of the landscape, it is relevant to measure the size, or diameter, of the largest component following the removal of each and every patch, and then rank the patches according to the induced change their removal inflicted on the largest component [75,78]. Similarly, in the previously mentioned study by Pascual-Hortal and Saura [86], the effect on the chosen set of 10 different connectivity measures were assessed for a number of spatially explicit fragmentation scenarios wherein individual patches occupying different kinds of topological positions were removed.

Large changes in the size of the largest component following the removal of a certain patch would, however, only occur if the removal would split the largest component into two or more non-trivial components. From a management perspective, it would be desirable if such patches were identified before they become that critical (as a result of continuous fragmentation). If such early identification was possible, it would enable managers to more proactively target conservation efforts in order to protect patches that could become critical. The trial-and-error method of removing nodes, one-by-one, lacks this ability.

To enable such proactive topological analysis of individual patches' contribution to the landscapes connectivity, various measures of network centrality may be plausible. Following this line of thought, the betweenness centrality index [19] has been suggested as a potentially relevant measure in the context of landscape fragmentation [11,87]. Patches with a high score of betweenness centrality are e.g., expected to significantly contribute to: (1) Reducing the overall topological distances between all pair of patches, and (2) Linking otherwise separated groups of patches [11]. As opposed to the trial-and-error approach described earlier, it appears that by using the betweenness centrality as the criterion for ranking the topological importance of individual patches, potentially critical patches could be identified even in cases where their removal would not split the network into isolated components. Furthermore, even in cases where the overall connectivity of the landscape is fairly high, the betweenness centrality index could be of use in identifying patches that may be critical in the future as a consequence of further fragmentation ([11], but see also [87]).

Another approach to identifying critical habitat patches is to study the minimal spanning tree (MST) of the

network representation of the landscape. The MST would visually represent the “backbone” of the connected patches in the landscape, and from there important patches can be identified [75]. Furthermore, if the strength of the links are accounted for (using e. g., Eq. 1), a link-weighted MST can be constructed where the specific patches which contribute, more than others, to species dispersals stand out by their high number of links [78]. Furthermore, Fall and colleagues [81] have developed a modeling approach, which they call spatial graphs, where they explicitly incorporate geographical details in a network representation of a fragmented landscape. A spatial graph is e. g., spatially explicit in describing the geometrics of the least-cost paths a dispersing organism would follow when moving from patch to patch. Using this representation of the landscape, they suggest that critical passages, barriers and stepping stone patches can be visually identified and targeted for management and/or restoration efforts. This modeling approach has also been tested using empirical data on the movement of woodland caribou (*Rangifer tarandus caribou*), where a strong relationship between the distribution of caribou and larger clusters of high-quality habitat (identified using spatial graphs) was shown [88].

Summary

Three topological characteristics, degree distribution, modularity and centrality, all appear to be of relevance in studying different aspects of ecological systems. In food webs, the degree distribution affects the stability in terms of the risk for secondary extinctions following species loss. In addition, the density of links is believed to have effects on the stability of ecosystems, although how, and in what direction, is not entirely clear. Many food webs have a modular structure, i. e., the web is divided into several groups that are only weakly connected to each other. This may effect how far disturbances spread throughout the food web. In a web with many dense but separated groups of species, disturbances may very well be confined within groups. On the other hand, a highly modular structure implies there are fewer opportunities for species to compensate if some of their prey species would decline.

Furthermore, some species may be more influential than others, and that influence may be attributed to their structural position in the food web (structural keystone species). Influence may result from having many links to others, but it could also derive from the possession of structurally important links that for example connects otherwise disconnected groups of species.

In order to analyze the spatial structure of connectivity of fragmented landscapes, a network representation

where nodes are patches and links are dispersal possibilities paves the way for topological analyses at the level of landscapes. It appears that simple network characteristics, such as the number of components, can help in assessing how connected different species, with varying dispersal capabilities, actually experience the landscape. Furthermore, network analysis targeted at finding compartments of internally well-connected habitat patches can help in identifying population and metapopulation boundaries. Finally, by assessing the structural importance of individual habitat patches (similar to structural keystone species), land management could be made more efficient by letting different network centrality measures provide guidance in prioritizing conservation efforts.

Future Directions

Topological analyses alone can provide important insights into complex ecological systems. However, in many cases, it is desirable to also include the strengths of the links in the analyses. Many network analytical approaches support the inclusion of link strengths, but there are many others that do not. Furthermore, the network perspective has traditionally been rather static, and the analytical focus has been the topological structure currently at hand. But networks evolve and develop dynamically; thus the topology given at one moment in time may be outdated later on. Better understanding of processes that shape network evolution are needed, as are methods capable of longitudinal network analyses applicable in an ecological context. The last statement touches upon the fact that a topological perspective will only be of any significant value when we possess knowledge and/or well-grounded assumptions of important processes. Thus topological approaches need to develop alongside theoretical advances of understanding key ecological processes.

Bibliography

Primary Literature

1. Volterra V (1926) Fluctuations in the abundance of a species considered mathematically. *Nature* 118:558
2. Lotka AJ (1925) *Elements of physical biology*. Williams and Wilkins, Baltimore
3. Odum EP, Odum HT (1953) *Fundamentals of ecology*. Saunders, Philadelphia
4. Odum EP (1977) The emergence of ecology as a new integrative discipline. *Science* 195:1289
5. Paine RT (1980) Food Webs: Linkage, interaction strength and community infrastructure. *J Anim Ecol* 49:666
6. Keitt TH, Urban DL, Milne BT (1997) Detecting critical scales in fragmented landscapes. *Conserv Ecol* 1

7. Webb C, Bodin Ö (2008) A network perspective on modularity and control of flow in robust systems. In: Norberg J, Cumming G (eds) *Complexity theory for a sustainable future*. Columbia, New York: Chichester, pp 85–118
8. Levin SA (1998) Ecosystems and the biosphere as complex adaptive systems. *Ecosystems* 1:431
9. Freeman LC (2004) The development of social network analysis – A study in the sociology of science. Empirical Press, Vancouver
10. Luczkovich JJ, Borgatti SP, Johnson JC, Everett MG (2003) Defining and measuring trophic role similarity in food webs using regular equivalence. *J Theor Biol* 220:303
11. Bodin Ö, Norberg J (2007) A network approach for analyzing spatially structured populations in fragmented landscape. *Landsc Ecol* 22:31
12. Benedek Z, Jordán F, Báldi A (2007) Topological keystone species complexes in ecological interaction networks. *Community Ecol* 8:1
13. Barabási AL (2002) *Linked: The new science of networks*. Perseus, Cambridge
14. Albert R, Jeong H, Barabási AL (2000) Error and attack tolerance of complex networks. *Nature* 406:378
15. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. *Nature* 393:440
16. Borgatti SP, Everett MG, Shirey PR (1990) LS sets, lambda sets and other cohesive subsets. *Soc Netw* 12:337
17. Wasserman S, Faust K (1994) *Social network analysis – methods and applications*. Cambridge University Press, Cambridge
18. Newman MEJ, Girvan M (2004) Finding and evaluating community structure in networks. *Phys Rev E* 69:026113
19. Freeman L (1979) Centrality in social networks. *Conceptual clarifications*. *Soc Netw* 1:215
20. Sabidussi G (1966) The centrality index of a graph. *Psychometrika* 31:581
21. Woodward G et al (2005) Body size in ecological networks. *Trends Ecol Evol* 20:402
22. Berlow EL et al (2004) Interaction strengths in food webs: issues and opportunities. *J Anim Ecol* 73:585–598
23. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW (2003) Compartments revealed in food-web structure. *Nature* 426:282
24. Scotti M, Podani J, Jordán F (2007) Weighting, scale dependence and indirect effects in ecological networks: A comparative study. *Ecol Complex* 4:148–159
25. Dunne JA (2006) The network structure of food webs. In: Pascual M, Dunne JA (eds) *Ecological networks: Linking structure to dynamics in food webs*. Oxford University Press, Oxford, pp 27–86
26. Wootton JT (1997) Estimates and tests of per capita interaction strength: diet, abundance, and impact of intertidally foraging birds. *Ecol Monogr* 67:45
27. Paine RT (1992) Food-web analysis through field measurement of per capita interaction strength. *Nature* 355:73
28. Goldwasser L, Roughgarden J (1993) Construction and analysis of a large caribbean food web. *Ecology* 74:1216
29. Winemiller KO (1990) Spatial and temporal variation in tropical fish trophic networks. *Ecol Monogr* 60:331–367
30. Warren PH (1990) Variation in food-web structure—the determinants of connectance. *Am Nat* 136:689–700
31. de Ruiter PC, Neutel AM, Moore JC (1995) Energetics, patterns of interaction strengths, and stability in real ecosystems. *Science* 269:1257
32. Solé RV, Bascompte J (2006) *Self-organization in complex ecosystems*. Princeton University Press, New Jersey
33. Dunne JA, Williams RJ, Martinez ND (2002) Network structure and biodiversity loss in food webs: robustness increases with connectance. *Ecol Lett* 5:558–567
34. Seidman SB (1983) LS sets and cohesive subsets of graphs and hypergraphs. *Soc Netw* 5:92
35. Melián CJ, Bascompte J (2004) Food web cohesion. *Ecology* 85:352
36. Pastor-Satorras R, Vázquez A, Vespignani A (2001) Dynamical and correlation properties of the internet. *Phys Rev Lett* 87:258701
37. Melián CJ, Bascompte J (2002) Complex networks: Two ways to be robust? *Ecol Lett* 5:705
38. Garlaschelli D, Caldarelli G, Pietronero L (2003) Universal scaling relations in food webs. *Nature* 423:165–168
39. Dunne JA, Williams RJ, Martinez ND (2002) Food-web structure and network theory: The role of connectance and size. *Proc Natl Acad Sci USA* 99:12917–12922
40. Camacho J, Guimerá R, Amaral LAN (2002) Robust patterns in food web structure. *Phys Rev Lett* 88:228102
41. Williams RJ, Berlow EL, Dunne JA, Barabási AL, Martinez ND (2002) Two degrees of separation in complex food webs. *Proc Natl Acad Sci USA* 99:12913–12916
42. Pimm SL, Lawton JH (1980) Are food webs divided into compartments? *J Anim Ecol* 49:879
43. Webb C, Bodin Levin SA (2005) Cross-system perspectives on the ecology and evolution of resilience. In: Jen E (ed) *Robust design: A repertoire of biological, ecological, and engineering case studies*. SFI Lecture Note Series. Oxford University Press, Oxford, pp 151–172
44. Amaral LAN, Scala A, Barthélémy M, Stanley HE (2000) Classes of small-world networks. *Proc Natl Acad Sci USA* 97:11149–11152
45. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. *Proc Natl Acad Sci USA* 99:7821
46. Frank KA (1995) Identifying Cohesive Subgroups. *Soc Netw* 17:27
47. McCann K, Hastings A, Huxel GR (1998) Weak trophic interactions and the balance of nature. *Nature* 395:794
48. May RM (1974) Biological populations with nonoverlapping generations: stable points, stable cycles, and chaos. *Science* 186:645
49. Power ME et al (1996) Challenges in the quest for keystones. *Bioscience* 46:609
50. Solé RV, Montoya JM (2001) Complexity and fragility in ecological networks. *Proc Biol Sci* 268:2039
51. Steneck RS et al (2002) Kelp forest ecosystem: biodiversity, stability, resilience and their future. *Environ Conserv* 29:436
52. Estes JA, Palmisano JF (1974) Sea otters: their role in structuring nearshore communities. *Science* 185:1058
53. Levin SA (2000) *Fragile Dominion: Complexity and the Commons*. Perseus Publishing, Cambridge
54. Estes JA, Duggins DO (1995) Sea otters and kelp forests in Alaska—generality and variation in a community ecological paradigm. *Ecol Monogr* 65:75
55. Burt R (1992) *Structural holes: The social structure of competition*. Harvard University Press, Cambridge

56. Jordán F, Liu W-C, Davis AJ (2006) Topological keystone species: measures of positional importance in food webs. *Oikos* 112:535
57. Jordán F, Benedek Z, Podani J (2007) Quantifying positional importance in food webs: A comparison of centrality indices. *Ecol Model* 205:270–275
58. Estrada E (2007) Characterization of topological keystone species – Local, global and “meso-scale” centralities in food webs. *Ecol Complex* 4:48
59. Yodzis P (1988) The indeterminacy of ecological interactions, as perceived through perturbation experiments. *Ecology* 69:508
60. Eklöf A, Ebenman B (2006) Species loss and secondary extinctions in simple and complex model communities. *J Anim Ecol* 75:239
61. Tilman D, May RM, Lehman CL, Nowak MA (1994) Habitat destruction and the extinction debt. *Nature* 371:65
62. Borrvall C, Ebenman B (2006) Early onset of secondary extinctions in ecological communities following the loss of top predators. *Ecol Lett* 9:435–442
63. Vasas V, Jordán F (2006) Topological keystone species in ecological interaction networks: Considering link quality and non-trophic effects. *Ecol Model* 196:365–378
64. Jordano P, Bascompte J, Olesen JM (2003) Invariant properties in coevolutionary networks of plant-animal interactions. *Ecol Lett* 6:69
65. Bascompte J, Jordano P, Melián CJ, Olesen JM (2003) The nested assembly of plant–animal mutualistic networks. *Proc Natl Acad Sci USA* 100:9383
66. Borgatti SP, Everett MG (1999) Models of core/periphery structures. *Soc Netw* 21:375
67. Turing A (1952) On the chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 237:37
68. Levin R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological. *Control Bull Entomol Soc Am* 15:237
69. Lande R (1987) Extinction thresholds in demographic models of territorial populations. *Am Nat* 130:624
70. Bascompte J, Solé RV (1996) Habitat fragmentation and extinction thresholds in spatially explicit metapopulation models. *J Anim Ecol* 65:465
71. Hanski I (1994) A practical model of metapopulation dynamics. *J Anim Ecol* 63:151
72. Hanski I (2001) Spatially realistic theory of metapopulation ecology. *Naturwissenschaften* 88:372
73. Hanski I, Ovaskainen O (2000) The metapopulation capacity of a fragmented landscape. *Nature* 404:755–758
74. Taylor PD, Fahrig L, Henein K, Merriam G (1993) Connectivity is a vital element of landscape structure. *Oikos* 68:571–573
75. Urban D, Keitt T (2001) Landscape connectivity: A graph-theoretic perspective. *Ecology* 82:1205
76. Bodin Ö, Tengö M, Norman A, Lundberg J, Elmqvist T (2006) The value of small size: Loss of forest patches and ecological thresholds in southern Madagascar. *Ecol Appl* 16:440
77. Baum KA, Haynes KJ, Dilleuth FP, Cronin JT (2004) The matrix enhances the effectiveness of corridors and stepping stones. *Ecology* 85:2671
78. Bunn AG, Urban DL, Keitt TH (2000) Landscape connectivity: A conservation application of graph theory. *J Environ Manage* 59:265
79. Ricketts TH (2001) The matrix matters: Effective isolation in fragmented landscapes. *Am Nat* 158:87
80. Verbeylen G, Bruyn LD, Adriaensen F, Matthysen E (2003) Does matrix resistance influence Red squirrel (*Sciurus vulgaris* L. 1758) distribution in an urban landscape? *Landscape Ecol* 18:791
81. Fall A, Fortin MJ, Manseau M, O'Brien D (2006) Spatial graphs: Principles and applications for habitat connectivity. *Ecosystem* 10:448–461
82. Brooks CP (2003) A scalar analysis of landscape connectivity. *Oikos* 102:433
83. Sutherland GD, Harestad AS, Price K, Lertzman KP (2000) Scaling of natal dispersal distances in terrestrial birds and mammals. *Conserv Ecol* 4
84. Garant D, Kruuk LEB, Wilkin TA, McCleery RH, Sheldon BC (2005) Evolution driven by differential dispersal within a wild bird population. *Nature* 433:60
85. Postma E, Van Noordwijk AJ (2005) Gene flow maintains a large genetic difference in clutch size at a small spatial scale. *Nature* 433:65
86. Pascual-Hortal L, Saura S (2006) Comparison and development of new graph-based landscape connectivity indices: towards the prioritization of habitat patches and corridors for conservation. *Landscape Ecol* 21:959
87. Minor ES, Urban DL (2007) Graph theory as a proxy for spatially explicit population models in conservation planning. *Ecol Appl* 17:1771
88. O'Brien D, Manseau M, Fall A, Fortin MJ (2006) Testing the importance of spatial configuration of winter habitat for woodland caribou: An application of graph theory. *Biol Conserv* 130:70–83

Books and Reviews

- Buchanan M (2002) *Small world: Uncovering nature's hidden networks*. Weidenfeld & Nicolson, London
- Levin SA (2000) *Fragile dominion: Complexity and the commons*. Perseus Publishing, Cambridge
- Pascual M, Dunne JA (eds) (2006) *Ecological networks: Linking structure to dynamics in food webs*. Oxford University Press, Oxford
- Pimm SL (1982) *Food webs*. Chapman and Hall, London

Econometrics: Models of Regime Changes

JEREMY PIGER¹

Department of Economics,
University of Oregon, Eugene, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Threshold and Markov-Switching Models
of Regime Change](#)

[Estimation of a Basic Markov-Switching Model](#)
[Extensions of the Basic Markov-Switching Model](#)
[Specification Testing for Markov-Switching Models](#)
[Empirical Example: Identifying Business Cycle](#)
[Turning Points](#)
[Future Directions](#)
[Bibliography](#)

Glossary

- Filtered probability of a regime** The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period t , conditional on observing sample information up to period t .
- Gibbs sampler** An algorithm to generate a sequence of samples from the joint probability distribution of a group of random variables by repeatedly sampling from the full set of conditional distributions for the random variables.
- Markov chain** A process that consists of a finite number of states, or regimes, where the probability of moving to a future state conditional on the present state is independent of past states.
- Markov-switching model** A regime-switching model in which the shifts between regimes evolve according to an unobserved Markov chain.
- Regime-Switching Model** A parametric model of a time series in which parameters are allowed to take on different values in each of some fixed number of regimes.
- Smooth transition threshold model** A threshold model in which the effect of a regime shift on model parameters is phased in gradually, rather than occurring abruptly.
- Smoothed probability of a regime** The probability that the unobserved Markov chain for a Markov-switching model is in a particular regime in period t , conditional on observing all sample information.
- Threshold model** A regime-switching model in which the shifts between regimes are triggered by the level of an observed economic variable in relation to an unobserved threshold.
- Time-varying transition probability** A transition probability for a Markov chain that is allowed to vary depending on the outcome of observed information.
- Transition probability** The probability that a Markov chain will move from state j to state i .

Definition of the Subject

Regime-switching models are time-series models in which parameters are allowed to take on different values in each of some fixed number of “regimes.” A stochastic process assumed to have generated the regime shifts is included as part of the model, which allows for model-based forecasts that incorporate the possibility of future regime shifts. In certain special situations the regime in operation at any point in time is directly observable. More generally the regime is unobserved, and the researcher must conduct inference about which regime the process was in at past points in time. The primary use of these models in the applied econometrics literature has been to describe changes in the dynamic behavior of macroeconomic and financial time series.

Regime-switching models can be usefully divided into two categories: “threshold” models and “Markov-switching” models. The primary difference between these approaches is in how the evolution of the state process is modeled. Threshold models, introduced by Tong [91], assume that regime shifts are triggered by the level of observed variables in relation to an unobserved threshold. Markov-switching models, introduced to econometrics by [16,39,41], assume that the regime shifts evolve according to a Markov chain.

Regime-switching models have become an enormously popular modeling tool for applied work. Of particular note are regime-switching models of measures of economic output, such as real Gross Domestic Product (GDP), which have been used to model and identify the phases of the business cycle. Examples of such models include [3,7,41,57,60,61,73,75,77,90,93]. A sampling of other applications include modeling regime shifts in inflation and interest rates [2,25,34], high and low volatility regimes in equity returns [23,46,48,92], shifts in the Federal Reserve’s policy “rule” [55,83], and time variation in the response of economic output to monetary policy actions [35,53,69,81].

Introduction

There is substantial interest in modeling the dynamic behavior of macroeconomic and financial quantities observed over time. A challenge for this analysis is that these time series likely undergo changes in their behavior over reasonably long sample periods. This change may occur in the form of a “structural break”, in which there is a shift in the behavior of the time series due to some permanent change in the economy’s structure. Alternatively, the change in behavior might be temporary, as in the case of wars or “pathological” macroeconomic episodes such as

¹I am grateful to Jim Hamilton and Bruce Mizraich for comments on an earlier draft.

economic depressions, hyperinflations, or financial crises. Finally, such shifts might be both temporary and recurrent, in that the behavior of the time series might cycle between regimes. For example, early students of the business cycle argued that the behavior of economic variables changed dramatically in business cycle expansions vs. recessions.

The potential for shifts in the behavior of economic time series means that constant parameter time series models might be inadequate for describing their evolution. As a result, recent decades have seen extensive interest in econometric models designed to incorporate parameter variation. One approach to describing this variation, denoted a “regime-switching” model in the following, is to allow the parameters of the model to take on different values in each of some fixed number of regimes, where, in general, the regime in operation at any point in time is unobserved by the econometrician. However, the process that determines the arrival of new regimes is assumed known, and is incorporated into the stochastic structure of the model. This allows the econometrician to draw inference about the regime that is in operation at any point in time, as well as form forecasts of which regimes are most likely in the future.

Applications of regime-switching models are usually motivated by economic phenomena that appear to involve cycling between recurrent regimes. For example, regime-switching models have been used to investigate the cycling of the economy between business cycle phases (expansion and recession), “bull” and “bear” markets in equity returns, and high and low volatility regimes in asset prices. However, regime switching models need not be restricted to parameter movement across recurrent regimes. In particular, the regimes might be non-recurrent, in which case the models can capture permanent “structural breaks” in model parameters.

There are a number of formulations of regime-switching time-series models in the recent literature, which can be usefully divided into two broad approaches. The first models regime change as arising from the observed behavior of the level of an economic variable in relation to some threshold value. These “threshold” models were first introduced by Tong [91], and are surveyed by [78]. The second models regime change as arising from the outcome of an unobserved, discrete, random variable, which is assumed to follow a Markov process. These models, commonly referred to as “Markov-switching” models, were introduced in econometrics by [16,39], and became popular for applied work following the seminal contribution of Hamilton [41]. Hamilton and Raj [47] and Hamilton [44] provide surveys of Markov-switching models, while Hamil-

ton [43] and Kim and Nelson [62] provide textbook treatments.

There are by now a number of empirical applications of regime-switching models that establish their empirical relevance over constant parameter alternatives. In particular, a large amount of literature has evaluated the statistical significance of regime-switching autoregressive models of measures of US economic activity. While the early literature did not find strong evidence for simple regime-switching models over the alternative of a constant parameter autoregression for US real GDP (e.g. [33]), later researchers have found stronger evidence using more complicated models of real GDP [57], alternative measures of economic activity [45], and multivariate techniques [63]. Examples of other studies finding statistical evidence in favor of regime-switching models include Garcia and Perron [34], who document regime switching in the conditional mean of an autoregression for the US real interest rate, and Guidolin and Timmermann [40], who find evidence of regime-switching in the conditional mean and volatility of UK equity returns.

This article surveys the literature surrounding regime-switching models, focusing primarily on Markov-switching models. The organization of the article is as follows. Section “[Threshold and Markov-Switching Models of Regime Change](#)” describes both threshold and Markov-switching models using a simple example. The article then focuses on Markov-switching models, with Sect. “[Estimation of a Basic Markov-Switching Model](#)” discussing estimation techniques for a basic model, Sect. “[Extensions of the Basic Markov-Switching Model](#)” surveying a number of primary extensions of the basic model, and Sect. “[Specification Testing for Markov-Switching Models](#)” surveying issues related to specification analysis. Section “[Empirical Example: Identifying Business Cycle Turning Points](#)” gives an empirical example, discussing how Markov-switching models can be used to identify turning points in the US business cycle. The article concludes by highlighting some particular avenues for future research.

Threshold and Markov-Switching Models of Regime Change

This section describes the threshold and Markov-switching approaches to modeling regime-switching using a specific example. In particular, suppose we are interested in modeling the sample path of a time series, $\{y_t\}_{t=1}^T$, where y_t is a scalar, stationary, random variable. A popular choice is an autoregressive (AR) model of order k :

$$y_t = \alpha + \sum_{j=1}^k \phi_j y_{t-j} + \varepsilon_t, \quad (1)$$

where the disturbance term, ε_t , is assumed to be normally distributed, so that $\varepsilon_t \sim N(0, \sigma^2)$. The AR(k) model in (1) is a parsimonious description of the data, and has a long history as a tool for establishing stylized facts about the dynamic behavior of the time series, as well as an impressive record in forecasting.

In many cases however, we might be interested in whether the behavior of the time series changes across different periods of time, or regimes. In particular, we may be interested in the following regime-switching version of (1):

$$y_t = \alpha_{S_t} + \sum_{j=1}^k \phi_{j,S_t} y_{t-j} + \varepsilon_t, \quad (2)$$

where $\varepsilon_t \sim N(0, \sigma_{S_t}^2)$. In (2), the parameters of the AR(k) depend on the value of a discrete-valued state variable, $S_t = i, i = 1, \dots, N$, which denotes the regime in operation at time t . Put simply, the parameters of the AR(k) model are allowed to vary among one of N different values over the sample period.

There are several items worth emphasizing about the model in (2). First, conditional on being inside of any particular regime, (2) is simply a constant parameter linear regression. Such models, which are commonly referred to as “piecewise linear”, make up the vast majority of the applications of regime-switching models. Second, if the state variable were observed, the model in (2) is simply a linear regression model with dummy variables, a fact that will prove important in our discussion of how the parameters of (2) might be estimated. Third, although the specification in (2) allows for all parameters to switch across all regimes, more restrictive models are certainly possible, and indeed are common in applied work. For example, a popular model for time series of asset prices is one in which only the variance of the disturbance term is allowed to vary across regimes. Finally, the shifts in the parameters of (2) are modeled as occurring abruptly. An example of an alternative approach, in which parameter shifts are phased in gradually, can be found in the literature investigating “smooth transition” threshold models. Such models will not be described further here, but are discussed in detail in [93].

Threshold and Markov-switching models differ in the assumptions made about the state variable, S_t . Threshold models assume that S_t is a deterministic function of an observed variable. In most applications this variable is taken to be a particular lagged value of the process itself, in which case regime shifts are said to be “self-exciting”. In particular, define $N - 1$ “thresholds” as $\tau_1 < \tau_2 < \dots < \tau_{N-1}$. Then, for a self-exciting threshold model, S_t is defined as

follows:

$$\begin{aligned} S_t &= 1 & y_{t-d} < \tau_1, \\ S_t &= 2 & \tau_1 \leq y_{t-d} < \tau_2, \\ &\vdots & \vdots \\ S_t &= N & \tau_{N-1} \leq y_{t-d}. \end{aligned} \quad (3)$$

In (3), d is known as the “delay” parameter. In most cases S_t is unobserved by the econometrician, because the delay and thresholds, d and τ_i , are generally not observable. However, d and τ_i can be estimated along with other model parameters. [78] surveys classical and Bayesian approaches to estimation of the parameters of threshold models.

Markov-switching models also assume that S_t is unobserved. In contrast to threshold models however, S_t is assumed to follow a particular stochastic process, namely an N -state Markov chain. The evolution of Markov chains are described by their transition probabilities, given by:

$$\begin{aligned} P(S_t = i | S_{t-1} = j, S_{t-2} = q, \dots) \\ = P(S_t = i | S_{t-1} = j) = p_{ij}, \end{aligned} \quad (4)$$

where, conditional on a value of j , we assume $\sum_{i=1}^N p_{ij} = 1$. That is, the process in (4) specifies a complete probability distribution for S_t . In the general case, the Markov process allows regimes to be visited in any order and for regimes to be visited more than once. However, restrictions can be placed on the p_{ij} to restrict the order of regime shifts. For example, [12] notes that the transition probabilities can be restricted in such a way so that the model in (2) becomes a “changepoint” model in which there are $N - 1$ structural breaks in the model parameters. Finally, the vast majority of the applied literature has assumed that the transition probabilities in (4) evolve independently of lagged values of the series itself, so that

$$\begin{aligned} P(S_t = i | S_{t-1} = j, S_{t-2} = q, \dots, y_{t-1}, y_{t-2}, \dots) \\ = P(S_t = i | S_{t-1} = j) = p_{ij}, \end{aligned} \quad (5)$$

which is the polar opposite of the threshold process described in (3). For this reason, Markov-switching models are often described as having regimes that evolve “exogenously” of the series, while threshold models are said to have “endogenous” regimes. However, while popular in practice, the restriction in (5) is not necessary for estimation of the parameters of the Markov-switching model. Section “[Extensions of the Basic Markov-Switching Model](#)” of this article discusses models in which the transition probabilities of the Markov process are allowed to be partially determined by lagged values of the series.

The threshold and Markov-switching approaches are best viewed as complementary, with the “best” model likely to be application specific. Certain applications appear tailor-made for the threshold assumption. For example, we might have good reason to think that the behavior of time series such as an exchange rate or inflation will exhibit regime shifts when the series moves outside of certain thresholds, as this will trigger government intervention. The Markov-switching model might instead be the obvious choice when one does not wish to tie the regime shifts to the behavior of a particular observed variable, but instead wishes to let the data speak freely as to when regime shifts have occurred.

In the remainder of this article I will survey various aspects regarding the econometrics of Markov-switching models. For readers interested in learning more about threshold models, the survey article of Potter [78] is an excellent starting point.

Estimation of a Basic Markov-Switching Model

This section discusses estimation of the parameters of Markov-switching models. The existing literature has focused almost exclusively on likelihood-based methods for estimation. I retain this focus here, and discuss both maximum likelihood and Bayesian approaches to estimation. An alternative approach based on semi-parametric estimation is discussed in [4].

To aid understanding, we focus on a specific baseline case, which is the Markov-switching autoregression given in (2) and (5). We simplify further by allowing for $N = 2$ regimes, so that $S_t = 1$ or 2. It is worth noting that in many cases two regimes is a reasonable assumption. For example, in the literature using Markov-switching models to study business cycles phases, a two regime model, meant to capture an expansion and recession phase, is an obvious starting point that has been used extensively.

Estimation of Markov-switching models necessitates two additional restrictions over constant parameter models. First of all, the labeling of S_t is arbitrary, in that switching the vector of parameters associated with $S_t = 1$ and $S_t = 2$ will yield an identical model. A commonly used approach to normalize the model is to restrict the value of one of the parameters when $S_t = 1$ relative to its value when $S_t = 2$. For example, for the model in (2) we could restrict $\alpha_2 < \alpha_1$. For further details on the choice of normalization, see [49]. Second, the transition probabilities in (5) must be constrained to lie in $[0, 1]$. One approach to implement this constraint, which will be useful in later discussion, is to use a probit specification for S_t . In particular, the value of S_t is assumed to be determined by the realization of a random

variable, η_t , as follows:

$$S_t = \begin{cases} 1 & \text{if } \eta_t < \gamma_{S_{t-1}} \\ 2 & \text{if } \eta_t \geq \gamma_{S_{t-1}} \end{cases}, \quad (6)$$

where $\eta_t \sim i.i.d.N(0, 1)$. The specification in (6) depends on two parameters, γ_1 and γ_2 , which determine the transition probabilities of the Markov process as follows:

$$\begin{aligned} p_{1j} &= P(\eta_t < \gamma_j) = \Phi(\gamma_j) \\ p_{2j} &= 1 - p_{1j} \end{aligned}, \quad (7)$$

where $j = 1, 2$ and Φ is the standard normal cumulative distribution function.

There are two main items of interest on which to conduct statistical inference for Markov-switching models. The first are the parameters of the model, of which there are $2(k + 3)$ for the two-regime Markov-switching autoregression. In the following we collect these parameters in the vector

$$\theta = (\alpha_1, \phi_{1,1}, \phi_{2,1}, \dots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \dots, \phi_{k,2}, \sigma_2, \gamma_1, \gamma_2)', \quad (8)$$

The second item of interest is the regime indicator variable, S_t . In particular, as S_t is unobserved, we will be interested in constructing estimates of which regime was in operation at each point in time. These estimates will take the form of posterior probabilities that $S_t = i, i = 1, 2$. We assume that the econometrician has a sample of $T + k$ observations, $(y_T, y_{T-1}, y_{T-2}, \dots, y_{-(k-1)})$. The series of observations available up to time t is denoted as $\Omega_t = (y_t, y_{t-1}, y_{t-2}, \dots, y_{-(k-1)})$.

We begin with maximum likelihood estimation of θ . Maximum likelihood estimation techniques for various versions of Markov-switching regressions can be found in the existing literature of multiple disciplines, for example [52,76,79] in the speech recognition literature, and [16,41] in the econometrics literature. Here we focus on the presentation of the problem given in [41], who presents a simple iterative algorithm that can be used to construct the likelihood function of a Markov-switching autoregression, as well as compute posterior probabilities for S_t .

For a given value of θ , the conditional log likelihood function is given by:

$$L(\theta) = \sum_{t=1}^T \log f(y_t | \Omega_{t-1}; \theta). \quad (9)$$

Construction of the conditional log likelihood function then requires construction of the conditional density function, $f(y_t | \Omega_{t-1}; \theta)$, for $t = 1, \dots, T$. The “Hamilton

Filter” computes these conditional densities recursively as follows: Suppose for the moment that we are given $P(S_{t-1} = j|\Omega_{t-1}; \theta)$, which is the posterior probability that $S_{t-1} = j$ based on information observed through period $t - 1$. Equations (10) and (11) can then be used to construct $f(y_t|\Omega_{t-1}; \theta)$:

$$P(S_t = i|\Omega_{t-1}; \theta) = \sum_{j=1}^2 P(S_t = i|S_{t-1} = j, \Omega_{t-1}; \theta) \cdot P(S_{t-1} = j|\Omega_{t-1}; \theta), \quad (10)$$

$$f(y_t|\Omega_{t-1}; \theta) = \sum_{i=1}^2 f(y_t|S_t = i, \Omega_{t-1}; \theta) \cdot P(S_t = i|\Omega_{t-1}; \theta). \quad (11)$$

From (5), the first term in the summation in (10) is simply the transition probability, p_{ij} , which is known for any particular value of θ . The first term in (11) is the conditional density of y_t assuming that $S_t = i$, which, given the within-regime normality assumption for ε_t , is:

$$f(y_t|S_t = i, \Omega_0; \theta) = \frac{1}{\sigma_i \sqrt{2\pi}} \exp \left(- \frac{\left(y_t - \alpha_i - \sum_{j=1}^k \phi_{j,i} y_{t-j} \right)^2}{2\sigma_i^2} \right). \quad (12)$$

With $f(y_t|\Omega_{t-1}; \theta)$ in hand, the next step is then to update (10) and (11) to compute $f(y_{t+1}|\Omega_t; \theta)$. To do so requires $P(S_t = i|\Omega_t; \theta)$ as an input, meaning we must update $P(S_t = i|\Omega_{t-1}; \theta)$ to reflect the information contained in y_t . This updating is done using Bayes’ rule:

$$P(S_t = i|\Omega_t; \theta) = \frac{f(y_t|S_t = i, \Omega_{t-1}; \theta)P(S_t = i|\Omega_{t-1})}{f(y_t|\Omega_{t-1}; \theta)}, \quad (13)$$

where each of the three elements on the right-hand side of (13) are computable from the elements of (10) and (11). Given a value for $P(S_0 = i|\Omega_0; \theta)$ to initialize the filter, Eqs. (10) through (13) can then be iterated to construct $f(y_t|\Omega_{t-1}; \theta)$, $t = 1, \dots, T$, and therefore the log likelihood function, $L(\theta)$. The *maximum likelihood estimate* $\hat{\theta}_{MLE}$, is then the value of θ that maximizes $L(\theta)$, and can be obtained using standard numerical optimization techniques.

How do we set $P(S_0 = i|\Omega_0; \theta)$ to initialize the filter? As is discussed in [41], exact evaluation of this probability is rather involved. The usual practice, which is possible when S_t is an ergodic Markov chain, is to simply set

$P(S_0 = i|\Omega_0; \theta)$ equal to the unconditional probability, $P(S_0 = i)$. For the two-regime case considered here, these unconditional probabilities are given by:

$$P(S_0 = 1) = \frac{1 - p_{22}}{2 - p_{11} - p_{22}} \quad (14)$$

$$P(S_0 = 2) = 1 - P(S_0 = 1).$$

Alternatively, $P(S_0 = i|\Omega_0; \theta)$ could be treated as an additional parameter to be estimated. See Hamilton [43] and Kim and Nelson [62] for further details.

An appealing feature of the Hamilton filter is that, in addition to the likelihood function, the procedure also directly evaluates $P(S_t = i|\Omega_t; \theta)$, which is commonly referred to as a “filtered” probability. Inference regarding the value of S_t is then sometimes based on $P(S_t = i|\Omega_t; \hat{\theta}_{MLE})$, which is obtained by running the Hamilton filter with $\theta = \hat{\theta}_{MLE}$. In many circumstances, we might also be interested in the so-called “smoothed” probability of a regime computed using all available data, or $P(S_t = i|\Omega_T; \theta)$. [54] presents an efficient recursive algorithm that can be applied to compute these smoothed probabilities.

We now turn to Bayesian estimation of Markov-switching models. In the Bayesian approach, the parameters θ are themselves assumed to be random variables, and the goal is to construct the posterior density for these parameters given the observed data, denoted $f(\theta|\Omega_T)$. In all but the simplest of models, this posterior density does not take the form of any well known density whose properties can be analyzed analytically. In this case, modern Bayesian inference usually proceeds by sampling the posterior density repeatedly to form estimates of posterior moments and other objects of interest. These estimates can be made arbitrarily accurate by increasing the number of samples taken from the posterior. In the case of Markov-switching models, Albert and Chib [1] demonstrate that samples from $f(\theta|\Omega_T)$ can be obtained using a simulation-based approach known as the Gibbs Sampler. The Gibbs Sampler, introduced by [37,38,89], is an algorithm that produces random samples from the joint density of a group of random variables by repeatedly sampling from the full set of conditional densities for the random variables.

We will sketch out the main ideas of the Gibbs Sampler in the context of the two-regime Markov-switching autoregression. It will prove useful to divide the parameter space into $\theta = (\theta'_1, \theta'_2)'$, where $\theta_1 = (\alpha_1, \phi_{1,1}, \phi_{2,1}, \dots, \phi_{k,1}, \sigma_1, \alpha_2, \phi_{1,2}, \phi_{2,2}, \dots, \phi_{k,2}, \sigma_2)'$ and $\theta_2 = (\gamma_1, \gamma_2)'$. Suppose it is feasible to simulate draws from the three conditional distributions, $f(\theta_1|\theta_2, \hat{S}, \Omega_T)$, $f(\theta_2|\theta_1, \hat{S}, \Omega_T)$, and $P(\hat{S}|\theta_1, \theta_2, \Omega_T)$, where $\hat{S} = (S_1, S_2, \dots, S_T)'$. Then, conditional on arbitrary initial values, $\theta_2^{(0)}$ and

$\tilde{S}^{(0)}$, we can obtain a draw of θ_1 , denoted $\theta_1^{(1)}$, from $f(\theta_1|\theta_2^{(0)}, \tilde{S}^{(0)}, \Omega_T)$, a draw of θ_2 , denoted $\theta_2^{(1)}$, from $f(\theta_2|\theta_1^{(1)}, \tilde{S}^{(0)}, \Omega_T)$, and a draw of \tilde{S} , denoted $\tilde{S}^{(1)}$, from $P(\tilde{S}|\theta_1^{(1)}, \theta_2^{(1)}, \Omega_T)$. This procedure can be iterated to obtain $\theta_1^{(j)}$, $\theta_2^{(j)}$, and $\tilde{S}^{(j)}$, for $j = 1, \dots, J$. For large enough J , and assuming weak regularity conditions, these draws will converge to draws from $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$. Then, by taking a large number of such draws beyond J , one can estimate any feature of $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$, such as moments of interest, with an arbitrary degree of accuracy. For example, an estimate of $P(S_t = i|\Omega_T)$ can be obtained by computing the proportion of draws of \tilde{S} for which $S_t = i$.

Why is the Gibbs Sampler useful for a Markov-switching model? It turns out that although $f(\theta|\Omega_T)$ and $P(\tilde{S}|\Omega_T)$ cannot be sampled directly, it is straightforward, assuming natural conjugate prior distributions, to obtain samples from $f(\theta_1|\theta_2, \tilde{S}, \Omega_T)$, $f(\theta_2|\theta_1, \tilde{S}, \Omega_T)$, and $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$. This is most easily seen for the case of θ_1 , which, when \tilde{S} is conditioning information, represents the parameters of a linear regression with dummy variables, a case for which techniques to sample the parameter posterior distribution are well established (Zellner 96). An algorithm for obtaining draws of \tilde{S} from $P(\tilde{S}|\theta_1, \theta_2, \Omega_T)$ was first given in Albert and Chib [1], while Kim and Nelson [59] develop an alternative, efficient, algorithm based on the notion of “multi-move” Gibbs Sampling introduced in [6]. For further details regarding the implementation of the Gibbs Sampler in the context of Markov-switching models, see Kim and Nelson [62].

The Bayesian approach has a number of features that make it particularly attractive for estimation of Markov-switching models. First of all, the requirement of prior density functions for model parameters, considered by many to be a weakness of the Bayesian approach in general, is often an advantage for Bayesian analysis of Markov-switching models [42]. For example, priors can be used to push the model toward capturing one type of regime-switching vs. another. The value of this can be seen for Markov-switching models of the business cycle, for which the econometrician might wish to focus on portions of the likelihood surface related to business cycle switching, rather than those related to longer term regime shifts in productivity growth. Another advantage of the Bayesian approach is with regards to the inference drawn on S_t . In the maximum likelihood approach, the methods of [54] can be applied to obtain $P(S_t = i|\Omega_T; \hat{\theta}_{MLE})$. As these probabilities are conditioned on the maximum likelihood parameter estimates, uncertainty regarding the unknown values of the parameters has not been taken into account. By contrast, the Bayesian ap-

proach yields $P(S_t = i|\Omega_T)$, which is not conditional on a particular value of θ and thus incorporates uncertainty regarding the value of θ that generated the observed data.

Extensions of the Basic Markov-Switching Model

The basic, two-regime Markov-switching autoregression in (2) and (5) has been used extensively in the literature, and remains a popular specification in applied work. However, it has been extended in a number of directions in the substantial literature that follows [41]. This section surveys a number of these extensions.

The estimation techniques discussed in Sect. “Estimation of a Basic Markov-Switching Model” can be adapted in a straightforward manner to include several extensions to the basic Markov-switching model. For example, the filter used in (10) through (13) can be modified in obvious ways to incorporate the case of $N > 2$ regimes, as well as to allow y_t to be a vector of random variables, so that the model in (2) becomes a Markov-switching vector autoregression (MS-VAR). Hamilton [43] discusses both of these cases, while Krolzig [68] provides an extensive discussion of MS-VARs. [83] is a recent example of applied work using an MS-VAR with a large number of regimes. In addition, the (known) within-regime distribution of the disturbance term, ε_t , could be non-Gaussian, as in [23] or [45]. Further, the parameters of (2) could be extended to depend not just on S_t , but also on a finite number of lagged values of S_t , or even a second state variable possibly correlated with S_t . Indeed, such processes can generally be rewritten in terms of the current value of a single, suitably redefined, state variable. [58,66] provide examples of such a redefinition. For further discussion of all of these cases, see [43].

The specification for the transition probabilities in (5) restricted the probability $S_t = i$ to depend only on the value of S_{t-1} . However, in some applications we might think that these transition probabilities are driven in part by observed variables, such as the past evolution of the process. To this end, [21,28] develop Markov-switching models with time-varying transition probabilities (TVTP), in which the transition probabilities are allowed to vary depending on conditioning information. Suppose that z_t represents a vector of observed variables that are thought to influence the realization of the regime. The probit representation for the state process in (6) and (7) can then be extended as follows:

$$S_t = \begin{cases} 1 & \text{if } \eta_t < (\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}) \\ 2 & \text{if } \eta_t \geq (\gamma_{S_{t-1}} + z_t' \lambda_{S_{t-1}}) \end{cases}, \quad (15)$$

with associated transition probabilities:

$$\begin{aligned} p_{1j}(z_t) &= P(\eta_t < (\gamma_j + z'_t \lambda_j)) = \Phi(\gamma_j + z'_t \lambda_j) \\ p_{2j}(z_t) &= 1 - p_{1j}(z_t), \end{aligned} \quad (16)$$

where $j = 1, 2$ and Φ is again the standard normal cumulative distribution function. Estimation of the Markov-switching autoregression with TVTP is then straightforward. In particular, assuming that z_t contains lagged values of y_t or exogenous random variables, a maximum likelihood estimation proceeds by simply replacing p_{ij} with $p_{ij}(z_t)$ in the filter given in (10) through (13). Bayesian estimation of TVTP models via the Gibbs Sampler is also straightforward, and is discussed in [29]. Despite its intuitive appeal, the literature contains relatively few applications of the TVTP model. A notable example of the TVTP framework is found in Durland and McCurdy [24], Filardo and Gordon [29] and Kim and Nelson [59], who study business cycle “duration dependence”, or whether the probability of a business cycle phase shift depends on how long the economy has been in the current phase. Other applications include Ang and Bekaert [2], who model regime-switches in interest rates, and Lo and Piger [69], who investigate sources of time-variation in the response of output to monetary policy actions.

The TVTP model is capable of relaxing the restriction that the state variable, S_t , is independent of the lagged values of the series, y_t , and thus of lagged values of the disturbance term, ε_t . Kim, Piger and Startz [65] consider a Markov-switching model in which S_t is also correlated with the contemporaneous value of ε_t , and is thus “endogenous”. They model this endogenous switching by assuming that the shock to the probit process in (6), η_t , and ε_t are jointly normally distributed as follows:

$$\begin{bmatrix} \varepsilon_t \\ \eta_t \end{bmatrix} \sim N(0, \Sigma), \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \quad (17)$$

Kim, Piger and Startz [65] show that when $\rho \neq 0$, the conditional density in (12) is no longer Gaussian, but can be evaluated analytically. Thus, the likelihood function for the endogenous switching model can be evaluated with simple modifications to the recursive filter in (10) through (13). Tests of the null hypothesis that S_t is exogenous can also be implemented in a straightforward manner. Chib and Dueker [13] consider endogenous switching as in (17) from a Bayesian perspective.

The extensions listed above are primarily modifications to the stochastic process assumed to drive S_t . A more fundamental extension of (2) is to consider Markov-switching in time series models that are more complicated than simple autoregressions. An important example of this

is a state-space model with Markov-switching parameters. Allowing for Markov-switching in the state-space representation for a time series is particularly interesting because a large number of popular time-series models can be given a state-space representation. Thus, incorporating Markov-switching into a general state-space representation immediately extends the Markov-switching framework to these models.

To aid discussion, consider the following Markov-switching state-space representation for a vector of R random variables, $Y_t = (y_{1t}, y_{2t}, \dots, y_{Rt})'$, given as follows:

$$\begin{aligned} Y_t &= H'_{S_t} X_t + W_t \\ X_t &= A_{S_t} + F_{S_t} X_{t-1} + V_t \end{aligned} \quad (18)$$

where $X_t = (x_{1t}, x_{2t}, \dots, x_{Dt})'$, $W_t \sim N(0, B_{S_t})$ and $V_t \sim N(0, Q_{S_t})$. The parameters of the model undergo Markov switching, and are contained in the matrices $H_{S_t}, B_{S_t}, A_{S_t}, F_{S_t}, Q_{S_t}$. A case of primary interest is when some or all of the elements of X_t are unobserved. This is the case for a wide range of important models in practice, including models with moving average (MA) dynamics, unobserved components (UC) models, and dynamic factor models. However, in the presence of Markov-switching parameters, the fact that X_t is unobserved introduces substantial complications for construction of the likelihood function. In particular, as is discussed in detail in [54] and Kim and Nelson [62], exact construction of the conditional density $f(y_t | \Omega_{t-1}; \theta)$ requires that one consider all possible permutations of the entire history of the state variable, $S_t, S_{t-1}, S_{t-2}, \dots, S_1$. For even moderately sized values of t , this quickly becomes computationally infeasible.

To make inference via maximum likelihood estimation feasible, [54] develops a recursive filter that constructs an approximation to the likelihood function. This filter “collapses” the number of lagged regimes that are necessary to keep track of by approximating a nonlinear expectation with a linear projection. Kim and Nelson [62] provide a detailed description of the Kim [54] filter, as well as a number of examples of its practical use.

If one is willing to take a Bayesian approach to the problem, Kim and Nelson [59] show that inference can be conducted via the Gibbs Sampler without resorting to approximations. As before, the conditioning features of the Gibbs sampler greatly simplifies the analysis. For example, by conditioning on $\tilde{S} = (S_1, S_2, \dots, S_T)'$, the model in (18) is simply a linear, Gaussian, state-space model with dummy variables, for which techniques to sample the posterior distribution of model parameters and the unobserved elements of X_t are well established [6]. Kim and

Nelson [62] provide detailed descriptions of how the Gibbs Sampler can be implemented for a state-space model with Markov switching.

There are many applications of state space models with Markov switching. For example, a large literature uses UC models to decompose measures of economic output into trend and cyclical components, with the cyclical component often interpreted as a measure of the business cycle. Until recently, this literature focused on linear representations for the trend and cyclical components [14,51,72,94]. However, one might think that the processes used to describe the trend and cyclical components might display regime switching in a number of directions, such as that related to the phase of the business cycle or to longer-run structural breaks in productivity growth or volatility. A UC model with Markov switching in the trend and cyclical components can be cast as a Markov-switching state-space model as in (18). Applications of such regime-switching UC models can be found in [58,60,64,71,84]. Another primary example of a Markov-switching state-space model is a dynamic factor model with Markov-switching parameters, examples of which are given in [7,59]. Section “Empirical Example: Identifying Business Cycle Turning Points” presents a detailed empirical example of such a model.

Specification Testing for Markov-Switching Models

Our discussion so far has assumed that key elements in the specification of regime-switching models are known to the researcher. Chief among these is the number of regimes, N . However, in practice there is likely uncertainty about the appropriate number of regimes. This section discusses data-based techniques that can be used to select the value of N .

To fix ideas, consider a simple version of the Markov-switching model in (2):

$$y_t = \alpha_{S_t} + \varepsilon_t, \quad (19)$$

where $\varepsilon_t \sim N(0, \sigma^2)$. Consider the problem of trying to decide between a model with $N = 2$ regimes vs. the simpler model with $N = 1$ regimes. The model with one regime is a constant parameter model, and thus this problem can be interpreted as a decision between a model with regime-switching parameters vs. one without. An obvious choice for making this decision is to construct a test of the null hypothesis of $N = 1$ vs. the alternative of $N = 2$. For example, one might construct the likelihood ratio statistic:

$$LR = 2(L(\hat{\theta}_{MLE(2)}) - L(\hat{\theta}_{MLE(1)})), \quad (20)$$

where $\hat{\theta}_{MLE(1)}$ and $\hat{\theta}_{MLE(2)}$ are the maximum likelihood estimates under the assumptions of $N = 1$ and $N = 2$ respectively. Under the null hypothesis there are three fewer parameters to estimate, α_2 , γ_1 and γ_2 , than under the alternative hypothesis. Then, to test the null hypothesis, one might be tempted to proceed by constructing a p-value for LR using the standard $\chi^2(3)$ distribution.

However, this final step is not justified, and can lead to very misleading results in practice. In particular, the standard conditions for LR to have an asymptotic χ^2 distribution include that all parameters are identified under the null hypothesis [17]. In the case of the model in (19), the parameters γ_1 and γ_2 , which determine the transition probabilities p_{ij} , are not identified assuming the null hypothesis is true. In particular, if $\alpha_1 = \alpha_2$, then p_{ij} can take on any values without altering the likelihood function for the observed data. A similar problem exists when testing the general case of N vs. $N + 1$ regimes.

Fortunately, a number of contributions in recent years have produced asymptotically justified tests of the null hypothesis of N regimes vs. the alternative of $N + 1$ regimes. In particular, [33,50] provide techniques to compute asymptotically valid critical values for LR . Recently Carrasco, Hu and Ploberger [5] have developed an asymptotically optimal test for the null hypothesis of parameter constancy against the general alternative of Markov-switching parameters. Their test is particularly appealing because it does not require estimation of the model under the alternative hypothesis, as is the case with LR .

If one is willing to take a Bayesian approach, the comparison of models with N vs. $N + 1$ regimes creates no special considerations. In particular, one can proceed by computing standard Bayesian model comparison metrics, such as Bayes Factors or posterior odds ratios. Examples of such comparisons can be found in [11,63,78].

Empirical Example: Identifying Business Cycle Turning Points

This section presents an empirical example demonstrating how the Markov-switching framework can be used to model shifts between expansion and recession phases in the US business cycle. This example is of particular interest for two reasons. First, although Markov-switching models have been used to study a wide variety of topics, their most common application has been as formal statistical models of business cycle phase shifts. Second, the particular model we focus on here, a dynamic factor model with Markov-switching parameters, is of interest in its own right, with a number of potential applications.

The first presentation of a Markov-switching model of the business cycle is found in [41]. In particular, [41] showed that US real GDP growth could be characterized as an autoregressive model with a mean that switched between low and high growth regimes, where the estimated timing of the low growth regime corresponded closely to the dates of US recessions as established by the Business Cycle Dating Committee of the National Bureau of Economic Research (NBER). This suggested that Markov-switching models could be used as tools to identify the timing of shifts between business cycle phases, and a great amount of subsequent analysis has been devoted toward refining and using the Markov-switching model for this task.

The model used in [41] was univariate, considering only real GDP. However, as is discussed in [22], a long emphasized feature of the business cycle is comovement, or the tendency for business cycle fluctuations to be observed simultaneously in a large number of economic sectors and indicators. This suggests that, by using information from many economic indicators, the identification of business cycle phase shifts might be sharpened. One appealing way of capturing comovement in a number of economic indicators is through the use of dynamic factor models, as popularized by [85,86]. However, these models assumed constant parameters, and thus do not model business cycle phase shifts explicitly.

To simultaneously capture comovement and business cycle phase shifts, [7] introduces Markov-switching parameters into the dynamic factor model of [85,86]. Specifically, defining $y_{rt}^* = y_{rt} - \bar{y}_r$ as the demeaned growth rate of the r th economic indicator, the dynamic factor Markov-switching (DFMS) model has the form:

$$y_{rt}^* = \beta_r c_t + e_{rt}. \quad (21)$$

In (21), the demeaned first difference of each series is made up of a component common to each series, given by the dynamic factor c_t , and a component idiosyncratic to each series, given by e_{rt} . The common component is assumed to follow a stationary autoregressive process:

$$\phi(L)(c_t - \mu_{S_t}) = \varepsilon_t, \quad (22)$$

where $\varepsilon_t \sim i.i.d.N(0, 1)$. The unit variance for ε_t is imposed to identify the parameters of the model, as the factor loading coefficients, β_r , and the variance of ε_t are not separately identified. The lag polynomial $\phi(L)$ is assumed to have all roots outside of the unit circle. Regime switching is introduced by allowing the common component to have a Markov-switching mean, given by μ_{S_t} , where $S_t = \{1, 2\}$. The regime is normalized by setting $\mu_2 < \mu_1$.

Finally, each idiosyncratic component is assumed to follow a stationary autoregressive process:

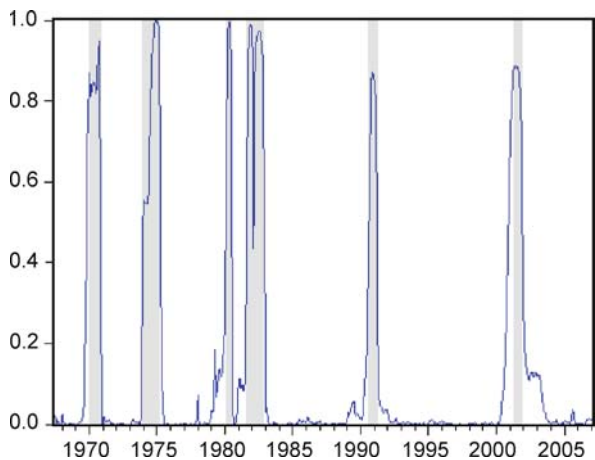
$$\theta_r(L)e_{rt} = \omega_{rt}. \quad (23)$$

where $\theta_r(L)$ is a lag polynomial with all roots outside the unit circle and $\omega_{rt} \sim N(0, \sigma_{\omega,r}^2)$.

[7] estimates the DFMS model for US monthly data on non-farm payroll employment, industrial production, real manufacturing and trade sales, and real personal income excluding transfer payments, which are the four monthly variables highlighted by the NBER in their analysis of business cycles. The DFMS model can be cast as a state-space model with Markov switching of the type discussed in Sect. “Extensions of the Basic Markov-Switching Model”. Chauvet estimates the parameters of the model via maximum likelihood, using the approximation to the likelihood function given in [54]. Kim and Nelson [59] instead use Bayesian estimation via the Gibbs Sampler to estimate the DFMS model.

Here I update the estimation of the DFMS model presented in [59] to a sample period extending from February 1967 through February 2007. For estimation, I use the Bayesian Gibbs Sampling approach, with prior distributions and specification details identical to those given in [59]. Figure 1 displays $P(S_t = 2 | \Psi_T)$ obtained from the Gibbs Sampler, which is the estimated probability that the low growth regime is active. For comparison, Fig. 1 also indicates NBER recession dates with shading.

There are two items of particular interest in Fig. 1. First of all, the estimated probability of the low growth regime is very clearly defined, with $P(S_t = 2 | \Psi_T)$ generally close to either zero or one. Indeed, of the 481 months



Econometrics: Models of Regime Changes, Figure 1
Probability of US Recession from Dynamic Factor Markov-Switching Model

in the sample, only 32 had $P(S_t = 2|\Psi_T)$ fall between 0.2 and 0.8. Second, $P(S_t = 2|\Psi_T)$ is very closely aligned with NBER expansion and recession dates. In particular, $P(S_t = 2|\Psi_T)$ tends to be very low during NBER expansion phases and very high during NBER recession phases.

Figure 1 demonstrates the added value of employing the DFMS model, which considers the comovement between multiple economic indicators, over models considering only a single measure of economic activity. In particular, results for the Markov-switching autoregressive model of real GDP presented in [41] were based on a data sample ending in 1984, and it is well documented that Hamilton’s original model does not perform well for capturing the two NBER recessions since 1984. Subsequent research has found that allowing for structural change in the residual variance parameter [61,70] or omitting all linear dynamics in the model [1,9] improves the Hamilton model’s performance. By contrast, the results presented here suggest that the DFMS model accurately identifies the NBER recession dates without a need for structural breaks or the omission of linear dynamics.

In some cases, we might be interested in converting $P(S_t = 2|\Psi_T)$ into a specific set of dates establishing the timing of shifts between business cycle phases. To do so requires a rule for establishing whether a particular month was an expansion month or a recession month. Here we consider a simple rule, which categorizes any particular month as an expansion month if $P(S_t = 2|\Psi_T) \leq 0.5$ and a recession month if $P(S_t = 2|\Psi_T) > 0.5$. Table 1 displays the dates of turning points between expansion and recession phases (business cycle peaks), and the dates of turning points between recession and expansion phases (business cycle troughs) that are established by this rule. For comparison, Table 1 also lists the NBER peak and trough dates.

Table 1 demonstrates that the simple rule applied to $P(S_t = 2|\Psi_T)$ does a very good job of matching the NBER peak and trough dates. Of the twelve turning points in the sample, the DFMS model establishes eleven within

two months of the NBER date. The exception is the peak of the 2001 recession, for which the peak date from the DFMS model is four months prior to that established by the NBER. In comparing peak and trough dates, the DFMS model appears to do especially well at matching NBER trough dates, for which the date established by the DFMS model matches the NBER date exactly in five of six cases.

Why has the ability of Markov-switching models to identify business cycle turning points generated so much attention? There are at least four reasons. First, it is sometimes argued that recession and expansion phases may not be of any intrinsic interest, as they need not reflect any real differences in the economy’s structure. In particular, as noted by [95], simulated data from simple, constant parameter, time-series models, for which the notion of separate regimes is meaningless, will contain episodes that look to the eye like “recession” and “expansion” phases. By capturing the notion of a business cycle phase formally inside of a statistical model, the Markov-switching model is then able to provide statistical evidence as to the extent to which business cycle phases are a meaningful concept. Second, although the dates of business cycle phases and their associated turning points are of interest to many economic researchers, they are not compiled in a systematic fashion for many economies. Markov-switching models could then be applied to obtain business cycle turning point dates for these economies. An example of this is given in [74], who use Markov-switching models to establish business cycle phase dates for US states. Third, if economic time-series do display different behavior over business cycle phases, then Markov-switching models designed to capture such differences might be exploited to obtain more accurate forecasts of economic activity. Finally, the current probability of a new economic turning point is likely of substantial interest to economic policymakers. To this end, Markov-switching models can be used for “real-time” monitoring of new business cycle phase shifts. Indeed, Chauvet and Piger [10] provide evidence that Markov-switching mod-

Econometrics: Models of Regime Changes, Table 1
Dates of Business Cycle Turning Points Produced by NBER and Dynamic Factor Markov-Switching Model

Peaks			Troughs		
DFMS	NBER	Discrepancy	DFMS	NBER	Discrepancy
Oct 1969	Dec 1969	2M	Nov 1970	Nov 1970	0M
Dec 1973	Nov 1973	−1M	Mar 1975	Mar 1975	0M
Jan 1980	Jan 1980	0M	Jun 1980	Jul 1980	1M
Jul 1981	Jul 1981	0M	Nov 1982	Nov 1982	0M
Aug 1990	Jul 1990	−1M	Mar 1991	Mar 1991	0M
Nov 2000	Mar 2001	4M	Nov 2001	Nov 2001	0M

els are often quicker to establish US business cycle turning points, particularly at business cycle troughs, than is the NBER. For additional analysis of the ability of regime-switching models to establish turning points in real time, see [8,9].

Future Directions

Research investigating applied and theoretical aspects of regime-switching models should be an important component of the future research agenda in macroeconomics and econometrics. In this section I highlight three directions for future research which are of particular interest.

To begin, additional research oriented toward improving the forecasting ability of regime-switching models is needed. In particular, given that regime-switching models of economic data contain important deviations from traditional, constant parameter, alternatives, we might expect that they could also provide improved out-of-sample forecasts. However, as surveyed in [15], the forecasting improvements generated by regime-switching models over simpler alternatives is spotty at best. That this is true is perhaps not completely surprising. For example, the ability of a Markov-switching model to identify regime shifts in past data does not guarantee that the model will do well at detecting regime shifts quickly enough in real time to generate improved forecasts. This is particularly problematic when regimes are short lived. Successful efforts to improve the forecasting ability of Markov-switching models are likely to come in the form of multivariate models, which can utilize additional information for quickly identifying regime shifts.

A second potentially important direction for future research is the extension of the Markov-switching dynamic factor model discussed in Sects. “[Extensions of the Basic Markov-Switching Model](#)” and “[Empirical Example: Identifying Business Cycle Turning Points](#)” to settings with a large cross-section of data series. Indeed, applications of the DFMS model have been largely restricted to a relatively small number of variables, such as in the model of the US business cycle considered in Sect. “[Empirical Example: Identifying Business Cycle Turning Points](#)”. However, in recent years there have been substantial developments in the analysis of dynamic factor models comprising a large number of variables, as in [31,32,87,88,92]. Research extending the regime-switching framework to such “big data” factor models will be of substantial interest.

Finally, much remains to be done incorporating regime-switching behavior into structural macroeconomic models. A number of recent studies have begun this synthesis by considering the implications of regime-

switches in the behavior of a fiscal or monetary policymaker for the dynamics and equilibrium behavior of model economies [18,19,20,26,27]. This literature has already yielded a number of new and interesting results, and is likely to continue to do so as it expands. Less attention has been paid to reconciling structural models with a list of new “stylized facts” generated by the application of regime-switching models in reduced-form settings. As one example, there is now a substantial list of studies, including [3,45,57,58,82], and Kim and Nelson [60] finding evidence that the persistence of shocks to key macroeconomic variables varies dramatically over business cycle phases. However, such an asymmetry is absent from most modern structural macroeconomic models, which generally possess a symmetric propagation structure for shocks. Research designed to incorporate and explain business cycle asymmetries and other types of regime-switching behavior inside of structural macroeconomic models will be particularly welcome.

Bibliography

1. Albert J, Chib S (1993) Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *J Bus Econ Stat* 11:1–15
2. Ang A, Bekaert G (2002) Regime switches in interest rates. *J Bus Econ Stat* 20:163–182
3. Beaudry P, Koop G (1993) Do recessions permanently change output? *J Monet Econ* 31:149–163
4. Campbell SD (2002) Specification testing and semiparametric estimation of regime switching models: An examination of the us short term interest rate. Brown University Department of Economics Working Paper #2002–26, Providence
5. Carrasco M, Hu L, Ploberger W (2004) Optimal test for Markov switching. Working paper, University of Rochester, Rochester
6. Carter CK, Kohn P (1994) On Gibbs sampling for state space models. *Biometrika* 81:541–553
7. Chauvet M (1998) An Econometric Characterization of Business Cycle Dynamics with Factor Structure and Regime Switching. *Int Econ Rev* 39:969–996
8. Chauvet M, Hamilton J (2006) Dating Business Cycle Turning Points. In: Milas C, Rothman P, van Dijk D (eds) *Nonlinear Time Series Analysis of Business Cycles*. Elsevier, North Holland
9. Chauvet M, Piger J (2003) Identifying Business Cycle Turning Points in Real Time. *Fed Res Bank St. Louis Rev* 85:47–61
10. Chauvet M, Piger J (2004) A Comparison of the Real-Time Performance of Business Cycle Dating Methods. *J Bus Econ Stat* 26:42–49
11. Chib S (1995) Marginal Likelihood from the Gibbs Output. *J Am Stat Assoc* 90:1313–1321
12. Chib S (1998) Estimation and Comparison of Multiple Change-Point Models. *J Econ* 86:221–241
13. Chib S, Dueker M (2004) Non-Markovian Regime Switching with Endogenous States and Time Varying State Strengths. Federal Reserve Bank of St. Louis Working Paper #2004–030A, St. Louis

14. Clark PK (1987) The Cyclical Component of US Economic Activity. *Quart J Econ* 102:797–814
15. Clements MP, Franses PH, Swanson NR (2004) Forecasting Economic and Financial Time-Series with Non-Linear Models. *Int J Forecast* 20:169–183
16. Cosslett SR, Lee LF (1985) Serial Correlation in Discrete Variable Models. *J Econ* 27:79–97
17. Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* 64:247–254
18. Davig T, Leeper E (2005) Generalizing the Taylor Principle. *Am Econ Rev* 97:607–635
19. Davig T, Leeper E (2006) Endogenous Monetary Policy Regime Change. In: Reichlin L, West KD (eds) *International Seminar on Macroeconomics*. MIT Press, Cambridge
20. Davig T, Leeper E, Chung H (2004) Monetary and Fiscal Policy Switching. *J Mon Credit Bank* 39:809–842
21. Diebold FX, Lee JH, Weinbach G (1994) Regime Switching with Time-Varying Transition Probabilities. In: Hargreaves C (ed) *Non-stationary Time Series Analysis and Cointegration*. Oxford University Press, Oxford UK
22. Diebold FX, Rudebusch GD (1996) Measuring business cycles: A modern perspective. *Rev Econ Stat* 78:67–77
23. Dueker M (1997) Markov Switching in GARCH Processes and Mean-Reverting Stock-Market Volatility. *J Bus Econ Stat* 15:26–34
24. Durland JM, McCurdy TH (1994) Duration Dependent Transitions in a Markov Model of USGNP Growth. *J Bus Econ Stat* 12:279–288
25. Evans M, Wachtel P (1993) Inflation Regimes and the Sources of Inflation Uncertainty. *J Mon Credit Bank* 25:475–511
26. Farmer REA, Waggoner DF, Zha T (2006) Indeterminacy in a Forward Looking Regime Switching Model. NBER working paper no. 12540, Cambridge
27. Farmer REA, Waggoner DF, Zha T (2007) Understanding the New-Keynesian Model when Monetary Policy Switches Regimes. NBER working paper no. 12965, Cambridge
28. Filardo AJ (1994) Business-Cycle Phases and Their Transitional Dynamics. *J Bus Econ Stat* 12:299–308
29. Filardo AJ, Gordon SF (1998) Business Cycle Durations. *J Econ* 85:99–123
30. Forni M, Hallin M, Lippi F, Reichlin L (2000) The Generalized Dynamic Factor Model: Identification and Estimation. *Rev Econ Stat* 82:540–554
31. Forni M, Hallin M, Lippi F, Reichlin L (2002) The generalized dynamic factor model: consistency and convergence rates. *J Econ* 82:540–554
32. Forni M, Hallin M, Lippi F, Reichlin L (2005) The generalized dynamic factor model: one-sided estimation and forecasting. *J Am Stat Assoc* 100:830–840
33. Garcia R (1998) Asymptotic null distribution of the likelihood ratio test in Markov switching models. *Int Econ Rev* 39:763–788
34. Garcia R, Perron P (1996) An Analysis of the Real Interest Rate under Regime Shifts. *Rev Econ Stat* 78:111–125
35. Garcia R, Schaller H (2002) Are the Effects of Monetary Policy Asymmetric? *Econ Inq* 40:102–119
36. Granger CWJ, Teräsvirta T (1993) *Modelling Nonlinear Economic Relationships*. Oxford University Press, Oxford
37. Gelfand AE, Smith AFM (1990) Sampling-Based Approaches to Calculating Marginal Densities. *J Am Stat Assoc* 85:398–409
38. Geman S, Geman D (1984) Stochastic Relaxation, Gibbs Distributions and the Bayesian Restoration of Images. *IEEE Trans Patt Anal Machine Int* 6:721–741
39. Goldfeld SM, Quandt RE (1973) A Markov Model for Switching Regressions. *J Econ* 1:3–16
40. Guidolin M, Timmermann A (2005) Economic Implications of Bull and Bear Regimes in UK Stock and Bond Returns. *Econ J* 115:111–143
41. Hamilton JD (1989) A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica* 57:357–384
42. Hamilton JD (1991) A Quasi-Bayesian Approach to Estimating Parameters for Mixtures of Normal Distributions. *J Bus Econ Statistics* 9:27–39
43. Hamilton JD (1994) *Time Series Analysis*. Princeton University Press, Princeton NJ
44. Hamilton JD (2005) Regime-Switching Models. In: Durlauf S, Blume L (eds) *New Palgrave Dictionary of Economics*, 2nd edn. Palgrave MacMillan Ltd, Hampshire
45. Hamilton JD (2005) What's Real About the Business Cycle? *Fed Res Bank St. Louis Rev* 87:435–452
46. Hamilton JD, Lin G (1996) Stock Market Volatility and the Business Cycle. *J Appl Econ* 11:573–593
47. Hamilton JD, Raj B (2002) New Directions in Business Cycle Research and Financial Analysis. *Empir Econ* 27:149–162
48. Hamilton JD, Susmel R (1994) Autoregressive Conditional Heteroskedasticity and Changes in Regime. *J Econ* 64:307–333
49. Hamilton, Waggoner JD DF, Zha T (2004) Normalization in Econometrics. *Econ Rev* 26:221–252
50. Hansen BE (1992) The likelihood ratio test under nonstandard conditions: Testing the Markov switching model of GNP. *J Appl Econ* 7:S61–S82
51. Harvey AC (1985) Trends and Cycles in Macroeconomic Time Series. *J Bus Econ Stat* 3:216–227
52. Juang BH, Rabiner LR (1985) Mixture Autoregressive Hidden Markov Models for Speech Signals. *IEEE Trans Acoust Speech Signal Proc ASSP-30*:1404–1413
53. Kaufmann S (2002) Is there an Asymmetric Effect of Monetary Policy Over Time? A Bayesian Analysis using Austrian Data. *Empir Econ* 27:277–297
54. Kim CJ (1994) Dynamic linear models with Markov-switching. *J Econ* 60:1–22
55. Kim CJ (2004) Markov-Switching Models with Endogenous Explanatory Variables. *J Econ* 122:127–136
56. Kim CJ, Morley J, Nelson C (2002) Is there a Positive Relationship between Stock Market Volatility and the Equity Premium? *J Money Cred Bank* 36:339–360
57. Kim CJ, Morley J, Piger J (2005) Nonlinearity and the Permanent Effects of Recessions. *J Appl Econ* 20:291–309
58. Kim CJ, Murray CJ (2002) Permanent and transitory components of recessions. *Empir Econ* 27:163–183
59. Kim CJ, Nelson CR (1998) Business Cycle Turning Points, a New Coincident Index, and Tests of Duration Dependence Based on a Dynamic Factor Model with Regime Switching. *Rev Econ Stat* 80:188–201
60. Kim CJ, Nelson CR (1999b) Friedman's Plucking Model of Business Fluctuations: Tests and Estimates of Permanent and Transitory Components. *J Money Cred Bank* 31:317–34
61. Kim CJ, Nelson CR (1999c) Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. *Rev Econ Stat* 81:608–616

62. Kim CJ, Nelson C (1999a) State-Space Models with Regime Switching: Classical and Gibbs-Sampling Approaches with Applications. MIT Press, Cambridge
63. Kim CJ, Nelson CR (2001) A Bayesian approach to testing for Markov-switching in univariate and dynamic factor models. *Int Econ Rev* 42:989–1013
64. Kim CJ, Piger J (2002) Common stochastic trends, common cycles, and asymmetry in economic fluctuations. *J Monet Econ* 49:1189–1211
65. Kim CJ, Piger J, Startz R (2003) Estimation of Markov Regime-Switching Regression Models with Endogenous Switching. *J Econom*, (in press)
66. Kim CJ, Piger J, Startz R (2007) The Dynamic Relationship Between Permanent and Transitory Components of US Business Cycles. *J Money Cred Bank* 39:187–204
67. Koop G, Potter SM (1999) Bayes Factors and Nonlinearity: Evidence from Economic Time Series. *J Econ* 88:251–281
68. Krolzig HM (1997) Markov-Switching Vector Autoregressions: Modelling, Statistical Inference, and Application to Business Cycle Analysis. Springer, Berlin
69. Lo M, Piger J (2005) Is the Response of Output to Monetary Policy Asymmetric? Evidence from a Regime-Switching Coefficients Model. *J Money Cred Bank* 37:865–887
70. McConnell MM, Quiros GP (2000) Output Fluctuations in the United States: What has Changed Since the Early (1980s)? *Am Econ Rev* 90:1464–1476
71. Mills TC, Wang P (2002) Plucking Models of Business Cycle Fluctuations: Evidence from the G-7 Countries. *Empir Econ* 27:255–276
72. Morley JC, Nelson CR, Zivot E (2003) Why Are the Beveridge-Nelson and Unobserved-Components Decompositions of GDP So Different? *Review Econ Stat* 85:235–243
73. Öcal N, Osborn DR (2000) Business cycle non-linearities in UK consumption and production. *J Appl Econ* 15:27–44
74. Owyang MT, Piger J, Wall HJ (2005) Business Cycle Phases in US States. *Rev Econ Stat* 87:604–616
75. Pesaran MH, Potter SM (1997) A floor and ceiling model of US output. *J Econ Dyn Control* 21:661–695
76. Poritz AB (1982) Linear Predictive Hidden Markov Models and the Speech Signal. *Acoustics, Speech and Signal Processing IEEE Conference on ICASSP '82*, vol 7:1291–1294
77. Potter SM (1995) A Nonlinear Approach to US GNP. *J Appl Econ* 10:109–125
78. Potter SM (1999) Nonlinear Time Series Modelling: An Introduction. *J Econ Surv* 13:505–528
79. Rabiner LR (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proc IEEE* 77:257–286
80. Rapach DE, Wohar ME (2002) Regime Changes in International Real Interest Rates: Are They a Monetary Phenomenon? *J Mon Cred Bank* 37:887–906
81. Ravn M, Sola M (2004) Asymmetric Effects of Monetary Policy in the United States. *Fed Res Bank St. Louis Rev* 86:41–60
82. Sichel DE (1994) Inventories and the three phases of the business cycle. *J Bus Econ Stat* 12:269–277
83. Sims C, Zha T (2006) Were there Regime Changes in US Monetary Policy? *Am Econ Rev* 96:54–81
84. Sinclair T (2007) Asymmetry in the Business Cycle: A New Unobserved Components Model. George Washington University working paper, Washington
85. Stock JH, Watson MW (1989) New Indexes of Coincident and Leading Economic Indicators. *NBER Macroeconomics Annual* 4:351–393
86. Stock JH, Watson MW (1991) A Probability Model of the Coincident Economic Indicators. In: Lahiri K, Moore GH (eds) *Leading Economic Indicators: New Approaches and Forecasting Records*. Cambridge University Press, Cambridge
87. Stock JH, Watson MW (2002a) Forecasting using principal components from a large number of predictors. *J Am Stat Assoc* 97:1167–1179
88. Stock JH, Watson MW (2002b) Macroeconomic Forecasting Using Diffusion Indexes (with James H Stock). *J Bus Econ Stat* 20(2):147–162
89. Tanner M, Wong W (1987) The Calculation of Posterior Distributions by Data Augmentation. *J Am Stat Assoc* 82:528–550
90. Tiao GC, Tsay RS (1994) Some advances in non-linear and adaptive modeling in time-series analysis. *J Forecast* 13:109–131
91. Tong H (1983) Threshold models in non-linear time series analysis, *Lecture Notes in Statistics*, No. 21. Springer, Heidelberg
92. Turner CM, Startz R, Nelson CR (1989) A Markov Model of Heteroskedasticity, Risk, and Learning in the Stock Market. *J Financ Econ* 25:3–22
93. van Dijk D, Franses PH (1999) Modeling multiple regimes in the business cycle. *Macroecon Dyn* 3:311–340
94. Watson MW (1986) Univariate Detrending Methods with Stochastic Trends. *J Monet Econ* 18:49–75
95. Watson MW (2005) Comment on James D Hamilton's 'What's Real about the Business Cycle?' *Fed Res Bank St. Louis Rev* 87:453–458
96. Zellner A (1971) *An Introduction to Bayesian Inference in Econometrics*. Wiley, New York

Econometrics: Non-linear Cointegration

JUAN-CARLOS ESCANCIANO¹, ALVARO ESCRIBANO²

¹ Department of Economics, Indiana University, Bloomington, USA

² Department of Economics, Universidad Carlos III de Madrid, Madrid, Spain

Article Outline

Glossary

Definition of the Subject

Introduction

Linear Measures of Memory

and Linear Error Correction Models

Nonlinear Error Correction (NEC) Models

Nonlinear Cointegration

Future Directions

Bibliography

Glossary

Cointegration Cointegration is an econometric property relating time series variables. If two or more series are themselves nonstationary, but a linear combination of them is stationary, then the series are said to be cointegrated.

Short memory A time series is said to be short memory if its information decays through time. In particular, we say that a variable is short memory in mean (in distribution), if the conditional mean (distribution) of the variable at time t given the information at time $t - h$ converges to a constant (to an unconditional distribution) as h diverges to infinity. Shocks in short memory time series have transitory effects.

Extended memory A time series is said to be extended memory in mean (in distribution), if it is not short memory in mean (distribution). Shocks in extended memory time series have permanent effects.

Nonlinear cointegration If two or more series are of extended memory, but a nonlinear transformation of them is short memory, then the series are said to be nonlinearly cointegrated.

Error correction model An Error Correction Model is a dynamic model in which the rate of growth of the variables in any period is related to the previous period's gap from long-run equilibrium.

Definition of the Subject

This paper is a selective review of the literature on nonlinear cointegration and nonlinear error correction models. The concept of cointegration plays a major role in macroeconomics, finance and econometrics. It was introduced by Granger in [42] and since then, it has achieved immense popularity among econometricians and applied economists. In fact in 2003 the Royal Swedish Academy of Science gave the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel to C. W. J. Granger for his contribution to the analysis of economic relationships based on cointegrated variables. In this paper we discuss the nonlinear extensions of the linear cointegration theory. Some authors consider nonlinear cointegration as a particular case of nonlinear error correction models. Although both concepts are related, we believe that it is useful to distinguish between them. After making this point clear, by relating linear and nonlinear error correction models, we discuss alternative measures of temporal dependence (memory) and co-dependence that are useful to characterize the usual notion of integration of order zero, $I(0)$, and cointegration in nonlinear contexts. We discuss parametric and nonparametric notions of nonlinear cointegration.

Finally, we conclude pointing out several lines of research that we think are promising in nonlinear and nonstationary contexts and therefore deserve further analysis.

Introduction

Granger in [42] introduced the concept of *cointegration* in a linear context; for further development see [20,64,65,85]. The alternative ways to deal with integrated and cointegrated series are now clear only in the linear context; see for example [43,52,57,59,67,77,105].

In macroeconomic and financial applications there are many cases where *nonlinearities* have been found in nonstationary contexts and therefore, there is a need for a theoretical justification of those empirical results. To reach this goal is not an easy target since the usual difficulties analyzing nonlinear time series models within a stationary and ergodic framework are enhanced in nonstationary contexts.

The purpose of this survey on *nonlinear cointegration* is to give a selected overview on the state of the art of econometrics that simultaneously analyzes nonstationarities and nonlinearities. The structure of this paper is the following: Sect. “Linear Measures of Memory and Linear Error Correction Models” discusses linear concepts of memory and dependence, cointegrated and error correction models. Section “Nonlinear Error Correction (NEC) Models” introduces nonlinear error correction models. Section “Nonlinear Cointegration” investigates nonlinear measures of memory and dependence and nonlinear cointegration. Finally, Sect. “Future Directions” concludes and mentions some open questions for future research.

Linear Measures of Memory and Linear Error Correction Models

The time series x_t is *integrated of order d* , denoted $x_t \sim I(d)$, if $\Delta^d x_t = (1 - L)^d x_t \sim I(0)$, where L is the lag operator such that $L^k x_t = x_{t-k}$ and d is an integer number. Here $I(0)$ denotes a covariance stationary short memory process with positive and bounded spectral density. We can extrapolate the concepts of integration to the *fractional case* where now d is not an integer but a real number. However, in this paper we will not cover fractional integration nor fractional cointegration; see Chapter 9.4.1 in [103] for a review.

Following the ideas of the seminal paper of [42], the most simple *definition of cointegration* could be the following; we say that two $I(1)$ series, y_t and x_t , are *cointegrated* if there is a linear combination $(1, -\beta)(y_t, x_t)'$ that is $I(0)$; $z_t = y_t - \beta x_t$ is $I(0)$, but any other linear combination, $z_t = y_t - \alpha x_t$, is $I(1)$ where $\alpha \neq \beta$. For sim-

plicity, through all this paper we will assume a bivariate system with a single cointegrating vector. Notice that the second condition of the above definition of cointegration ($z_t = y_t - \alpha x_t$, is $I(1)$ for any $\alpha \neq \beta$) is redundant in the linear context. However, it will be useful for the identification of the cointegrating vector in nonlinear cointegration, see Definition 3.

A nonparametric characterization of cointegration was introduced by [1] and [78]. Let x_t , y_t be the two $I(d)$ time series of interest, $d = 1$, and let $\gamma_{yx}(\tau, t)$ represent the *cross-covariance function* (CCF) of x_t , y_t , defined by $\gamma_{yx}(\tau, t) = \text{cov}(y_t, x_{t-\tau})$, where we make explicit the time dependence in $\gamma_{yx}(\tau, t)$ to allow for some degree of heterogeneity in the series. Similarly, define $\gamma_x(\tau, t) = \text{cov}(x_t, x_{t-\tau})$. Cointegration implies that the rates of convergence of $\gamma_{yx}(\tau, t)$ and $\gamma_x(\tau, t)$ should be the same as τ increases without bound and $\tau = o(t)$. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in terms of the long-run linear forecast of y_t . For example, suppose $x_t, y_t \sim I(1)$ and $z_t = y_t - \beta x_t$ a sequence of independent and identically distributed (i.i.d.) random variables independent of x_t . In this case, $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = \beta$ for all $\tau, t, \tau \leq t$. In more general cases, z_t might have serial correlation and might not be independent of x_t and therefore, the constancy of this ratio will only take place for τ 's beyond some value, see the Monte Carlo simulation results in [78]. On the other hand, in the spurious cointegration case where x_t, y_t are stochastically independent, $\lim_{\tau \rightarrow \infty} \gamma_{yx}(\tau, t)/\gamma_x(\tau, t) = 0$ for all $\tau, t, \tau \leq t$, therefore the ratio $\gamma_{yx}(\tau, t)/\gamma_x(\tau, t)$ is consistent against this type of spurious alternative hypothesis. As we will see later on, this notion of cointegration accepts nonlinear generalizations (nonlinear cointegration).

The most simple version of *Granger's Representation Theorem*, see [20], states that two series y_t and x_t are cointegrated if and only if they have an error correction representation, see Eqs (1a)–(1c) below. Therefore, either x_t Granger-causes y_t or y_t Granger-causes x_t or both.

Let y_t and x_t be two $I(1)$ series, where x_t is a pure random walk and y_t is generated by the following linear error correction (EC) model (1a) with linear cointegration (1b),

$$\Delta y_t = \psi_1 \Delta x_t + \gamma z_{t-1} + v_t \quad (1a)$$

$$y_t = \beta x_t + z_t \quad (1b)$$

$$\Delta x_t = \varepsilon_t \quad (1c)$$

where all the random error terms (v_t, z_t, ε_t) are $I(0)$. The errors z_t of (1b) form the error correction terms of (1a) and have usually more persistence (longer memory) than

the other two random error terms (v_t, ε_t). Therefore, in the system of Eqs. (1a)–(1c), x_t Granger-causes y_t but not the other way around.

Notice that we can write Eq. (1a), with $\phi_1 = \psi_1 - \beta$, in an equivalent way that will be very useful to introduce later on nonlinear error correction (NEC) models,

$$z_t = z_{t-1} + \phi_1 \Delta x_t + \gamma z_{t-1} + v_t. \quad (2)$$

Several alternative estimation procedures have been discussed in the literature to estimate the cointegrating parameter β :

- i) Maximum likelihood approach of [66] and [7]. Assumes that the conditional distribution of y given x and the lagged values of x and y is *Normal* and that the bivariate data generating process (DGP) of y and x is a *VAR of finite autoregressive order k* , $\text{VAR}(k)$ in error correction form. Furthermore, if the contemporaneous x -variable is *weakly exogenous*, then the partial maximum likelihood estimators is obtained by nonlinear least squares (NLS) on the error correction model obtained substituting (1b) in (1a). [98] and [38], derived the asymptotic properties of the NLS estimator of the error correction model (1a) and (1b), without the Normality assumption.
- ii) *Two-step approach* of Engle and Granger, see [20]. In the first step, Eq. (1b) is estimated by ordinary least squares (OLS) to get the residuals (z). In the second step, Eq. (1a) is estimated by OLS after substituting z_{t-1} by the corresponding lagged residuals from the first step. The OLS estimator of the first step is *super-consistent but biased*, and the limiting distribution depends on nuisance parameters. However, if z_t is serially uncorrelated and x_t is strictly exogenous, then the OLS estimator in (1b) coincides with the fully modified estimator and therefore it is asymptotically efficient, see [38].
- iii) *Fully modified OLS*, FM-OLS. This is a 2-step procedure developed by [79,80,86], and [81]. In the *first step*, Eq. (1b) is estimated by OLS. In the second step, semiparametric corrections are made for the *serial correlation* of the residuals z_t and for the *endogeneity* of the x -regressors. Under general conditions the fully modified estimator, is *asymptotically efficient*. The small sample behavior of these estimators was analyzed by Monte Carlo simulations by [53,56,57,58,71,86].
- iv) *Fully modified instrumental variable estimator*, FM-IV, of [71,86]. [78] showed that their nonparametric notion of cointegration has an instrumental variable (IV) interpretation if the instruments are the lagged

values of x . Furthermore, they showed that choosing those instruments has an extra advantage; we do not need to make the usual two corrections (endogeneity and serial correlation) to obtain a fully modified estimator. This particular IV-estimator has important advantages (bias reductions) over OLS in small samples.

- v) Recently [89] also studied the asymptotic properties of instrumental variables estimators (IV) in a fractional cointegration context, as in [78]. They propose to use IV estimates based on single equations estimation like (1b) employing exclusion and normalization restrictions, without correcting for the serial correlation of z_t .
- vi) [100] and [91], suggested a parametric correction for the endogeneity of the regressor (x_t) when estimating (1b) by OLS. The idea, based on the work of [97] about testing for causality, is to include additional future and past values of the Δx_t in Eq. (1b) when estimating it by OLS.
- vii) [87] proposed to add integral error correction terms (lagged values of the EC terms), to the procedure described in vi) in order to parsimoniously correct for serial correlation.

[63], using Monte Carlo simulations, compares some of these parametric and semi parametric estimators of the cointegrating vector. In the context of normally distributed errors, [63] recommends to model explicitly the dynamics instead of using nonparametric corrections based on fully modified estimators.

Nonlinear Error Correction (NEC) Models

There are interesting macroeconomic applications where nonlinearities have been found in nonstationary contexts. The first example of a nonlinear error correction (NEC) model is the UK money demand from 1878 to 1970 of [25,27]. Later on [61] used this nonlinear error correction strategy in their money demand estimation as an improvement over the usual linear money demands equations suggested by [37,60,76].

The variables of the usual money demand are: $m = \log$ money stock (millions), $y = \log$ real net national product Y , $p = \log$ of the price deflator of Y , $rs = \log$ of short term interest rate, $rl = \log$ of long-term interest rate, and $RS = \log$ of short term interest rate (without logs). Let V be the velocity of circulation of money, a version of the *quantity theory of money* says that $MV(RS) = PY$ or in logs $m + v(RS) = p + y$. Rearranging terms we can write $(m - p - y) = -v(RS)$ as a *long run money demand*.

[27] applied the *2-step approach* of [20] obtaining the following results:

1st Step:

$$(m - p - y)_t = -0.31 - 7RS_t + \hat{u}_t \quad (3a)$$

where \hat{u}_t are the residuals from 1878 to 2000 of the cointegrating relationship estimated by the super-consistent ordinary least squares (OLS) estimator. The inverse of the log of velocity of circulation of money, $(m - p - y) = \log\left(\frac{M}{PY}\right) = -v(RS)$, is $I(1)$ and the short run interest rate (RS) is also $I(1)$. Therefore, since the error term \hat{u}_t is *stable and significant* it is $I(0)$, see conditions (e) and (f) of Theorem 1 below. Equation (3a), or (3b), is the first example of *nonlinear cointegration* given by:

$$\frac{M}{PY} = \exp(-0.31 - 7RS + \hat{u}) \quad (3b)$$

Similar nonlinear cointegrating relationships based on long run money demand equations are recently estimated by [3].

[22] and [27] showed that, even if OLS might not be a consistent estimator (see [95]) when the errors of (3a), (3b) are nonlinear, the OLS estimates of (3a) and the NLS estimates of (4) in 1-step are very similar.

2nd Step:

$$\begin{aligned} (1-L)(m-p)_t &= 0.45(1-L)(m-p)_{t-1} \\ &\quad - (1-L)^2(m-p)_{t-2} - 0.60(1-L)p_t \\ &\quad + 0.39(1-L)p_{t-1} - 0.021(1-L)rs_t \\ &\quad - 0.062(1-L^2)rl_t - 2.55(\hat{u}_{t-1} - 0.2)\hat{u}_{t-1}^2 \\ &\quad + 0.005 + 3.7(D1 + D3) + \hat{\varepsilon}_t \end{aligned} \quad (4)$$

where D_1 and D_3 are dummy variables for the two world wars. The second nonlinear characteristic of model (4), apart from the nonlinear cointegration relationship, comes from the fact that the \hat{u}_{t-1} term enters in a *cubic polynomial form* as a particular nonlinear error correction (NEC) model, see also Sect. 3.3 in [17,18,95] discuss the inconsistencies derived from the 2-step approach OLS estimator in the context of nonlinear smooth transition error correction model. However, Monte Carlo simulations should be done to identify the type of nonlinearities that create series biases and inconsistencies using the 2-step estimation of NEC models.

A Nonlinear Version of Granger Representation Theorem

To justify this type of nonlinear models, we need to generalize the linear notions of temporal memory based on

the linear ARIMA concepts of integration, usually $I(1)$ and $I(0)$, to nonlinear measures of dependence. Several generalizations have been proposed in the literature, as we will see later on. Our first definition is motivated from asymptotic theory, more concretely from functional central limit theorems (FCLT). See Subsect. “A Nonlinear Version of Granger Representation Theorem” for alternative definitions.

FCLT-Based Definition of $I(0)$: A sequence $\{m_t\}$ is $I(0)$ if the “high level” condition that m_t verifies a FCLT is satisfied, i. e.

$$T^{-1/2} \sum_{t=1}^{[Tr]} m_t \xrightarrow{d} B(r)$$

where $B(r)$ is a Brownian motion, see [73,74,99].

Definition 1 (Strong mixing) Let $\{v_t\}$ be a sequence of random variables. Let $\mathfrak{F}_s^t \equiv \sigma(v_s, \dots, v_t)$ be the generated sigma-algebra. Define the α -mixing coefficients

$$\alpha_m \equiv \sup_t \sup_{\{F \in \mathfrak{F}_{-\infty}^t, G \in \mathfrak{F}_{t+m}^\infty\}} |P(G \cap F) - P(G)P(F)|$$

The process $\{v_t\}$ is said to be strong mixing (also α -mixing) if $\alpha_m \rightarrow 0$ as $m \rightarrow \infty$. If $\alpha_m \leq m^{-a}$ we say that $\{v_t\}$ is strong mixing of size $-a$.

Definition 2 (NED) Let $\{w_t\}$ be a sequence of random variables with $E\{w_t^2\} < \infty$ for all t . It is said that $\{w_t\}$ is NED on the underlying sequence $\{v_t\}$ of size $-a$ if $\phi(n)$ is of size $-a$, where $\phi(n)$ given by

$$\sup \|w_t - E_{t-n}^{t+n}(w_t)\|_2 \equiv \phi(n)$$

where $E_{t-n}^{t+n}(w_t) = E(w_t | v_{t-n}, \dots, v_{t+n})$ and $\|\cdot\|_2$ is the L_2 norm of a random vector, defined as $E^{1/2} |\cdot|^2$ where $|\cdot|$ denotes the Euclidean norm.

Weak-Dependence-Based Definition of $I(0)$: A sequence is $I(0)$ if it is NED on an underlying α -mixing sequence $\{v_t\}$ but the sequence $\{x_t\}$ given by $x_t = \sum_{s=1}^t w_s$ is not NED on $\{v_t\}$. In this case, we will say that x_t is $I(1)$.

Notice that if x_t is $I(1)$ then Δx_t is $I(0)$. This definition excludes $I(-1)$ series as $I(0)$, like $z_t = e_t - e_{t-1}$ for α -mixing sequences e_t , since in this case $\sum_{s=1}^t z_s$ is α -mixing.

Definition 3 Two $I(1)$ sequences $\{y_t\}$ and $\{x_t\}$ are (linearly) cointegrated with cointegrating vector $[1, -\beta]'$, if $y_t - \beta x_t$ is NED on a particular α -mixing sequence but $y_t - \delta_{12} x_t$ is not NED for $\delta_{12} \neq \beta$.

Theorem 1 (Granger’s Representation Theorem, see [31]) Consider the nonlinear correction model (NEC) for the (2×1) vector $X_t = (y_t, x_t)'$, given by

$$\Delta X_t = \Psi \Delta X_{t-1} + F(X_{t-1}) + v_t \quad (5)$$

Assume that:

- (a) $v_t = (v_{y_t}, v_{x_t})'$ is α -mixing of size $-s/(s-2)$ for $s > 2$
- (b) $\sum_t v_t$ is not NED on α -mixing sequence
- (c) $E \|v_t\|^2 \leq \Delta_v$
- (d) $F(X_{t-1}) = J(Z_{t-1})$, where $Z_t \equiv y_t - \beta x_t$ and $J(\cdot)$ is a continuously differentiable function, which satisfies a generalized Lipschitz condition of Lemma 2 of [31].
- (e) Let $SR(\Psi) < 1$, where $SR(M)$ is the spectral radius of the matrix M , and
- (f) for some fixed $\delta \in (0, 1)$

$$SR \begin{pmatrix} \Psi & \nabla_z J(z) \\ \beta' \Psi & I_r + \beta' \nabla_z J(z) \end{pmatrix} \leq 1 - \delta.$$

The above conditions ensure that;

- (i) ΔX_t and Z_t are simultaneously NED on the α -mixing sequence (v_t, u_t) , where $u_t = v_{y,t} - \beta' v_{x,t}$; and
- (ii) X_t is $I(1)$.

This theorem gives sufficient conditions for cointegrated variables to be generated by a nonlinear error correction model.

Single-Equation Parametric NEC Models with Linear Cointegration

Consider the following NEC with linear cointegration

$$\begin{aligned} \Delta y_t &= \psi_1 \Delta x_t + f(z_{t-1}; \gamma) + v_t \\ y_t &= \beta x_t + z_t. \end{aligned}$$

As we said in the previous section, it is not difficult to generalize this model to include other variables, lags and cointegrating relations. Consider two independent α -mixing sequences $\{a_t\}$ and $\{v_t\}$ with a zero mean. Then the following three equations represent the DGP,

$$x_t = x_{t-1} + a_t \quad (6a)$$

$$z_t = z_{t-1} + \phi_1 \Delta x_t + f(z_{t-1}, \gamma) + v_t \quad (6b)$$

$$y_t = \beta x_t + z_t \quad (6c)$$

where the nonlinear function $f(z_{t-1}, \gamma)$ form the nonlinear error correction term and $\beta \neq 0$. Notice that x_t is $I(1)$ by construction from (6a). Notice the similarity between Eqs. (6b) and the linear error correction of Eq. (2).

If we can ensure that z_t is NED then y_t is also $I(1)$ and linearly cointegrated with x_t , where the cointegration relationship, $y_t - \beta x_t$, is linear. If we apply the difference operator to (6c) and substitute in (6b) we obtain (7),

$$\Delta y_t = (\beta + \phi_1) \Delta x_t + f(z_{t-1}, \gamma) + v_t \quad (7)$$

which is a nonlinear error correction model with linear cointegration, with $\psi_1 = \beta + \phi_1$. For the sake of simplicity, and without loss of generality, we impose a *common factor restriction* so that $\phi_1 = 0$ on (7) obtaining,

$$z_t = z_{t-1} + f(z_{t-1}, \gamma) + v_t \quad (8)$$

and then ψ_1 equals β , the cointegration parameter, and therefore (8) is a nonlinear extension of the Dickey–Fuller equation used in unit root testing. The errors of the cointegration relation are given by $z_{t-1} = y_{t-1} - \beta x_{t-1}$, and the OLS residuals are given by $\hat{z}_{t-1} = y_{t-1} - \hat{\beta} x_{t-1}$, where $\hat{\beta}$ is the value of β estimated in the OLS regression (6c). Substituting z_t by \hat{z}_{t-1} in (8) we obtain a nonlinear version of Engle and Granger's cointegration test (cf. [20]).

Differentiating (8) with respect to z_{t-1} we obtain

$$\frac{d}{dz_{t-1}} z_t = 1 + \frac{d}{dz_{t-1}} f(z_{t-1}, \gamma)$$

and therefore, our boundedness condition (see assumptions (e) and (f) of Theorem 1) is $-1 < \frac{d}{dz_{t-1}} z_t < 1$, or $-2 < \frac{df(z_{t-1}, \gamma)}{dz_{t-1}} < 0$ (models (6b), (7) and (8) are error correcting), which is sufficient to ensure that the series z_t is near epoch dependent (NED) and therefore y_t and x_t are cointegrated, see [26,27] and [31].

We discuss now few alternative nonlinear error correction (or equilibrium correction) functions $f(\cdot)$ that could generate the series z_t from the system (6a) to (6c).

NEC Model 1: Arctan, [32]

$$f(z, \delta_1, \delta_2, \gamma_2) = -\gamma_2 \arctan(\delta_1 z + \delta_2) \quad \text{for } \gamma_2 > 0.$$

NEC Model 2: Rational Polynomial, [27]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = -\gamma_2 ((z + \delta_1)^3 + \delta_2) / ((z + \delta_3)^2 + \delta_4) \quad \text{for } \gamma_2 > 0.$$

In the first two models, the derivatives are in the desired region (satisfy assumptions (e) and (f)) for appropriate values of some of the parameters but not for all. However, within the class of rational polynomials the model considered can satisfy the condition on the absolute value of the derivative, see [27] and [30,32]. Other

empirical examples of nonlinear error correction models are [11,22,29,33,49,61,72].

An important body of the literature has focused on *threshold models*, see [5,6,39,48,54,75,94], among others.

NEC Model 3: Switching Exponential, [30,32]

$$f(z, \delta_1, \delta_2, \delta_3, \delta_4, \gamma_2) = \gamma_2 (\exp(-\delta_1 z) - \delta_2) I_{\{z \geq 0\}} + \gamma_2 (\delta_4 - \exp(\delta_3 z)) I_{\{z < 0\}},$$

where $I_{\{S\}}$ is the characteristic function of the set S , $\gamma_2 > 0$, $\delta_1 > 0$ and δ_3 .

NEC Model 4: Regime Switching Error Correction Models, [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^3 1(z \in R_s) \gamma_s z,$$

where $1(\cdot)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

NEC Model 5: Random Regime Switching Error Correction Models, [6] and [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^3 1(z + \eta \in R_s) \gamma_s z,$$

where $1(\cdot)$ is the indicator function selecting the three regimes, $R_1 = (-\infty, c_1]$, $R_2 = (c_1, c_2]$ and $R_3 = (c_2, \infty)$.

Another important literature is related to *smooth transition error correction models*, see [50,90,92].

NEC Model 6: Smooth Transition Error Correction Models, [102] and [92]

$$f(z, \delta_1, \delta_2, \gamma_2) = \sum_{s=1}^3 h(z) \gamma_s z$$

$$h(z) = \begin{cases} 1 - L_1(z), & s = 1 \\ L_1(z) - L_2(z), & s = 2 \\ L_3(z), & s = 3 \end{cases}$$

$$\text{where } L_s(z) = (1 + \exp\{-\gamma(z - c_s)\}) \quad \text{for } \gamma > 0 \quad \text{and } s = 1, 2.$$

Notice, that many smooth transition error correction models allow the nonlinear error correction function to affect all the parameters of the model, and not only the error correction term. However, in this paper we do not discuss them since they belong to a more general class of time varying models which is out of the context of this survey, see for example [50,101,102].

Nonparametric NEC Models with Linear Cointegration

[25,27] applied the semi-parametric *smoothing splines* estimation procedure of [21,96,104] to the estimation of the unknown nonlinear error correction function of Eq. (4). They found that in the long run money demand of the UK, see Eq. (3a), there are either multiple equilibria, or a threshold error correction with two attraction points (two equilibria); one negative and equal to -0.05 and one positive and equal to 0.2 . They suggest estimating those unknown thresholds using a cubic polynomial parametric functional form, see Eq. (4). Notice that the corresponding cubic polynomial error correction term, $-2.55(\hat{u}_{t-1} - 0.2)\hat{u}_{t-1}^2$, identifies perfectly one of the thresholds, the one that is equal to 0.2 . The second threshold could be obtained from the roots of the polynomial. Other empirical examples of threshold error correction models are [5,10,48,54]. In fact [10] used a similar nonparametric approach to estimate the nonlinear error correction function, but instead of using smoothing splines they used the Nadaraya–Watson kernel estimator discussed in [55].

Nonlinear Cointegration

In the recent years several proposals have been considered to extend linear cointegration and linear error correction of Granger [42] to a nonlinear framework. One possibility is to allow for a NEC model in the Granger's representation. We have discussed such an approach in the previous section. Alternatively, one may consider a nonlinear cointegration relation.

Despite the fact that many macroeconomic and financial time series dynamics are nonlinear, there are still today relatively few useful analytical tools capable of assessing the dependence and persistence behavior of nonlinear time series appropriately (cf. [50]). This problem is even more accentuated by the fact that traditional measures of dependence, such as autocorrelations and periodograms, may be inappropriate when the underlying time series is nonlinear and/or non-Gaussian. Then, it is generally accepted that new measures of dependence have to be defined in order to develop a new concept of nonlinear cointegration. We have already discussed measures based on FCLT and on NED concepts. We shall explore several alternative measures in this section. All the measures considered can be grouped in measures of conditional mean dependence or in distributional dependence. Higher order conditional moments, other than the mean, can of course be considered. In any case, we shall use the general terminology *extended memory* and *short memory* to indicate

a nonlinear persistence and non-persistence process, respectively (cf. [44]).

Once a concept of nonlinear persistence is introduced, a general definition of nonlinear cointegration is as follows. We say that two “extended memory” series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = f(y_t, x_t)$ is short memory. This definition is more appropriate when dealing with distributional persistence, and it is perhaps too general to be fully operative. Identification problems arise in this general context, as noted by many authors, so one should restrict the class of functions f to avoid such identification problems. [46] considered functions of the form $z_t = g(y_t) - h(x_t)$, and estimate g and h nonparametrically by means of the Alternating Conditional Expectations (ACE) algorithm. See also [40] for a related approach. It is still an open problem the theoretical justification of these nonparametric estimation procedures.

A less ambitious approach is to consider transformations of the form $z_t = y_t - f(x_t)$. This framework is especially convenient with conditional mean persistence measures. We review the existing measures in the next section.

Nonlinear Measures of Memory

As already discussed by [46], a generalization of linear cointegration to a nonlinear set-up goes through proper extensions of the linear concepts of $I(0)$ and $I(1)$. We introduce in this section alternative definitions of nonlinear $I(0)$ and $I(1)$ processes. We first focus on conditional mean persistence, we shall discuss distributional dependence at the end of this section. Define the conditional mean function $E(y_{t+h}|I_t)$, where $I_t = (x_t, x_{t-1}, \dots)$ is the conditioning set at time t . [44] defines the Short Memory in Mean (SMM) and Extended Memory in Mean (EMM) concepts as follows.

Definition 4 (SMM and EMM) $\{y_t\}$ is said to be SMM if for all t , $M(t, h) = E(y_{t+h}|I_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E|M(t, h) - \mu|^2 < c(h)$, where $c(h) \equiv c(h, t)$ is some positive sequence that tends to zero as h increases to infinity, for all t . If $\{y_t\}$ does not satisfy the previous condition is called EMM.

Note that to be mathematically precise in Definition 4, we should specify that $\{y_t\}$ is SMM or EMM with respect to $\{x_t\}$. Referring to this definition, [44] considered the case $x_t = y_t$. [46] replaced the name of EMM by long memory in mean, and [40] denoted EMM and SMM by nonlinear integrated (NLI) and nonlinear integrated of order zero (NLI(0)), respectively. As noted by [44] the concepts

of SMM and EMM are related to a kind of “mixing in mean” property, more precisely to the concept of mixing-ale, see [16].

[23] introduced the pairwise equivalent measures of the previous concepts, which, although weaker, are more operative because they only involve finite-dimensional random variables.

Definition 5 (PSMM and PEMM) $\{y_t\}$ is said to be Pairwise SMM (PSMM) if for all t , $m(t, h) = E(y_{t+h}|x_t)$, $h > 0$, tends to a constant μ as h becomes large. More precisely, $E|m(t, h) - \mu|^2 < c(h)$, where $c(h) \equiv c(h, t)$ is some positive sequence that tends to zero as h increases to infinity, for all t . If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM (PEMM).

From the previous definitions and the law of iterated expectations, we easily observe that a process SMM is PSMM. The reciprocal is false. There exist processes which are PSMM but not SMM, although they are rare in practice.

We now discuss generalizations of the usual autocovariances and crosscovariances to a nonlinear framework. These generalizations were introduced by [23]. It is well-known that in the presence of nonlinearity (or non-Gaussianity) the autocovariances do not characterize the dependence and the practitioner needs more reliable measures such as the pairwise regression functions $m(t, h)$. In general, inference on these functions involves nonparametric estimation with bandwidth choices, hampering their application to practical situations. By a measure-theoretic argument, the regression function $m(t, h)$ can be characterized by the integrated regression function $\gamma_{t,h}(x)$ given by

$$\begin{aligned}\gamma_{t,h}(x) &= E[(y_{t+h} - \mu_t)1(x_t \leq x)] \\ &= E[m(t, h)1(x_t \leq x)],\end{aligned}$$

where the second equality follows by the law of iterated expectations. The measures $\gamma_{t,h}(x)$ are called the Integrated Pairwise Regression Functions (IPRF), see [24]. Extensions to other weight functions different from the indicator weight $1(x_t \leq x)$ are possible. The integrated measures of dependence $\gamma_{t,h}(x)$ are useful for testing interesting hypotheses in a nonlinear time series framework and, unlike $m(t, h)$, they do not need of smoothing estimation and are easily estimated by the sample analogue. Moreover, they characterize the pairwise versions of the concepts introduced by [44], making these concepts more operative. First we need a definition, a norm $\|\cdot\|$ is nondecreasing if for all f and g with $|f(x)| \leq |g(x)|$ for all x , it holds that $\|f\| \leq \|g\|$. Associated to the norm $\|\cdot\|$, we de-

fine the distance $d(f, g) = \|f - g\|$. Usual nondecreasing norms are the L_2 norm and the supremum norm.

Definition 6 (PSMM_d and PEMM_d) $\{y_t\}$ is said to be Pairwise SMM relative to d (PSMM_d) if for all t , $\|\gamma_{t,h}(x)\|$, $h > 0$, tends to zero as h becomes large for any t . More precisely, $\|\gamma_{t,h}(x)\| < c(h)$, where $c(h) \equiv c(h, t)$ is some positive sequence that tends to zero as h increases to infinity for all t . If $\{y_t\}$ does not satisfy the previous condition is called Pairwise EMM relative to d (PEMM_d).

Theorem 2 (Relationship between PSMM and PSMM_d [23]) If the norm $\|\cdot\|$ is non-decreasing and $E\{y_t^2\} < \infty$ for all t , then $\{y_t\}$ is PSMM if and only if $\{y_t\}$ is PSMM_d.

Based on these concepts we define nonlinear cointegration as follows. We say that two PEMM_d series y_t and x_t are nonlinear cointegrated if there exist a function f such that $z_t = y_t - f(x_t)$ is PSMM_d.

In analogy with the linear world and based on the results in [1,78], a possible nonparametric characterization of nonlinear cointegration can be based on the rates of convergence of $\gamma_{t,h}(x)$ and $\gamma_{t,h}^x(x) = E[(x_{t+h} - \mu_t)1(x_t \leq x)]$ as h diverges to infinity. Intuitively, under cointegration, the remote past of x_t should be as useful as the remote past of y_t in long-run non-linearly forecasting y_t .

Similarly, we can define distributional measures of persistence and nonlinear cointegration. [45] defines a persistent process in distribution using the bivariate and marginal densities at different lags. [41] considered parametric and nonparametric methods for studying serial distributional dependence. In the nonparametric case [45] consider series expansions estimators for the nonlinear canonical analysis of the series. These authors apply their results to study the dynamics of the inter-trade durations of the Alcatel-stock on the Paris Bourse and find evidence of nonlinear strong persistence in distribution.

Regarding distributional dependence, we formalize a definition given in [45]. Let $f_{t,h}(y, x)$, $k_{t+h}(y)$ and $g_t(x)$ be, respectively, the bivariate and marginal densities of y_{t+h} and x_t . To define persistence in distribution one can use the Hellinger distance

$$H_{t,h} = \iint \left| f_{t,h}^{1/2}(y, x) - k_{t+h}^{1/2}(y) g_t^{1/2}(x) \right|^2 dy dx,$$

and define Pairwise Short Memory in Distribution (PSMD) according to the decay of $H_{t,h}$ to zero as h diverges to infinity. Alternative definitions can be given in terms of other divergence measures or distances, see [51] and references therein. This approach is explored in [1], who define nonlinear cointegration using mutual information measures.

Persistence in distribution is related to mixing concepts. In fact, uniformly in t , $H_{t,h} \leq 2\alpha(h)$, where $\alpha(h)$ is certain α -mixing coefficient, see [23] for details.

The aforementioned measures of nonlinear distributional dependence need of smoothing estimation, e. g. kernel estimators. Similarly to the case of conditional mean measures, we can avoid smoothing by means of the integrated measures of dependence

$$\eta_{t,h}(y, x) = \text{cov}(1(y_{t+h} \leq y), 1(x_t \leq x)) \\ = F_{t,h}(y, x) - K_{t+h}(y)G_t(x),$$

where $F_{t,h}(y, x)$, $K_{t+h}(y)$ and $G_t(x)$ are, respectively, the bivariate and marginal cumulative distribution functions (cdf) of y_{t+h} and x_t . The measures $\eta_{t,h}(y, x)$ can be estimated at different lags by using the sample analogue, i. e. the empirical distribution functions. Similar definitions to Definition 6 can be given for distributional persistence based on $\eta_{t,h}(y, x)$. See [23] for further generalizations and definitions. Definitions of nonlinear cointegration can be formulated along the lines in [1]. For instance, we can say that two persistent (in distribution) series y_t and x_t are nonlinear cointegrated (in distribution) if

$$\lim_{\tau \rightarrow \infty} \left\| \frac{\eta_{t,\tau}(\cdot)}{\eta_{t,\tau}^x(\cdot)} - \beta \right\| = 0$$

for all τ , $\tau \leq t$, where $\eta_{t,h}^x(y, x)$ is defined as $\eta_{t,h}(y, x)$ but replacing y_t by x_t there, β is a real number and $\|\cdot\|$ a suitable norm.

Integration and Cointegration Based on Order Statistics

[9,36,47] have considered rank based unit roots test to avoid the extreme sensitivity of usual test like the Dickey-Fuller type test to presence of nonlinearities and outliers, see [19] for an overview of the problems of unit root tests in general contexts. [2] suggested a range unit root test (RUR) based on the first differences of the ranges of a series. The range is the difference between the maximum and the minimum taken by a series at any given point in time. Therefore, the difference of the ranges is a measure of records. Counting the number of new records is an interesting way of distinguishing between stationary and non-stationary series since the frequency of new records vanishes faster for stationary series than for series containing unit roots. They have shown that this RUR test is robust to monotonic transformations, distributions of the errors and many structural breaks and additive outliers.

[8] suggests using the differences between the sequences of ranks. If there is no cointegration the sequence

of ranks tends to diverge, while under cointegration they evolve similarly. [34] consider a record counting cointegration test (RCC) based on the synchronicity of the jumps between cointegrated series. They suggest a test statistic based on counting the number of jumps (records) that simultaneously occur in both series. Certainly, those series that are cointegrated have a much larger number of simultaneous jumps in the ranges of the series. They show that the cointegration test based on RCC is robust to monotonic nonlinearities and many structural breaks and does not require a prior estimation of the nonlinear or linear cointegrating function. There is a large literature on the effects of structural breaks and outliers on unit root and on cointegration testing but it is out of the scope of this paper, see for example the references in [62].

Another cointegration test robust to nonlinearities and structural breaks is based on induced order statistics. In particular [35] consider that two series y_t and x_t are cointegrated (either linear or nonlinear) if the corresponding induced order series are plotted around the 45° line. Their test-statistic compares the two induced order series by comparing their corresponding empirical distributions, the empirical distribution of y_t and the empirical distribution of y_t induced by x_t , using the Kolmogorov-Smirnov type statistic.

Parametric Nonlinear Cointegration

As previously discussed, an important body of research has focused on nonlinear cointegration relations. The model used being a nonlinear cointegration regression or a nonlinear regression model with integrated regressors. References on this line of research include [12,13,82,84]. [93] study smooth transition regressions with integrated regressors. An example considered by these authors is the nonlinear extension of (1b)

$$y_t = \alpha x_t + \delta x_t g(x_t - c, \gamma) + v_t,$$

where $g(x - c, \gamma) = -1/(1 + e^{-\gamma(x-c)})$ is the logistic function. Under the restriction $\delta = 0$ the latter model reduces to the linear model in (1b), see [14] for linearity tests under this framework. Other examples of parametric nonlinear cointegration are given by Eqs. (3a), (3b) and by the exponential model of [68], see also the references given in [19]. Those models are a particular case of the more general *nonlinear parametric cointegration* model of the form,

$$y_t = f(x_t, \beta) + v_t,$$

where x_t is $p \times 1$ vector of $I(1)$ regressors, v_t is zero-mean stationary error term, and $f(x_t, \beta)$ a smooth function of

the process x_t , known up to the finite-dimensional parameter vector β .

At least, there are two different asymptotic justifications within these nonlinear cointegration models. In the classical asymptotic theory (e.g. [82,84]) all the existing literature has been confined to the case $p = 1$, although some extensions to single-index type regressions have been studied, see [83]. The main reason for the restriction to the univariate case is that commonly used asymptotic techniques are not appropriate for the case $p > 1$, e.g. asymptotics based on local times are not available for $p > 2$. Intrinsic to this problem is the non-recurrent property of the p -variate Brownian motion when $p > 2$. On the other hand, the triangular array asymptotics used in [93] allow for a general $p > 0$.

In the classical asymptotic theory, the properties of estimators of β , e.g. the nonlinear least squares estimator (NLSE), depend on the specific class of functions where $f(x_t, \beta)$ belongs. Commonly used classes are integrable functions, asymptotic homogeneous functions or exponential functions, see [82]. The rate of convergence of the NLSE is class-specific and, in some cases, involves random scaling. In the triangular array asymptotic theory of [93] the distribution theory of estimators of β , e.g. rates of convergence, does not depend on the specific class of functions.

Several authors exploit the previous asymptotic results on β to develop tests of nonlinear cointegration in this parametric framework. In [15] the so-called KPSS test is applied to the parametric residuals. Since the resulting limiting distribution depends on nuisance parameters, these authors implement the test with the assistance of a subsampling procedure as a smoothing device.

Nonparametric Nonlinear Cointegration

Nonparametric estimates of nonlinear cointegration relations were already computed by [46], but it has not been until the recent works by [69,70,88] that a nonparametric estimation theory for nonstationary processes has been developed. [88] considered a theory based on local time arguments, whereas [69,70] used the theory of null recurrent Markov processes. A comparison of both methodologies is discussed in [4], where near-integrated nonparametric asymptotics are studied.

More specifically, these authors estimate the transfer function $f(x_t)$ in the nonlinear regression model

$$y_t = f(x_t) + v_t,$$

where the series y_t and x_t are univariate observed nonstationary processes and v_t is a non-observed stationary

process. [70] study the nonparametric kernel estimation of $f(x_t)$ as

$$\hat{f}(x) = \frac{\sum_{t=0}^n y_t K_{x,h}(x_t)}{\sum_{t=0}^n K_{x,h}(x_t)},$$

where $K_{x,h}(x_t) = h^{-1}K((y - x)/h)$, K is a kernel function and h is a bandwidth parameter. These authors investigate the asymptotic theory for $\hat{f}(x)$ under some regularity conditions and different assumptions on the dependence relation between x_t and v_t . Especially convenient for the nonlinear cointegration framework are those assumptions that allow for dependence between x_t and v_t . The family of nonstationary processes considered by these authors is the class of the so-called β -null recurrent Markov processes satisfying a restriction on the tail distribution of the recurrence time. The class is large enough to contain the random walk, unit-root processes as well as other nonlinear nonstationary processes. It is shown that the nonparametric estimation theory is different to that in the stationary case, with slower rates of convergences, as expected. This new nonparametric asymptotic theory opens the door for future developments in inferences in nonlinear cointegration models.

Future Directions

This chapter has provided a selected overview of the available nonlinear extensions of this concept. While in the linear set-up there exists a complete theory and set of tools for studying the cointegration problem, it has been made clear that a nonlinear version of this theory possesses non-trivial challenges. A first natural nonlinear extension is to allow for a NEC model but still a linear cointegration regression. On the other hand, one can consider a nonlinear regression cointegration equation. It has been recognized that an extension of the concept of linear cointegration to a nonlinear set-up needs of appropriate extensions of the linear concepts of $I(0)$ and $I(1)$ to nonlinear time series (cf. [46]). Several extensions have been provided and discussed. We recommend operative integrated measures of dependence, since they are simple to estimate and avoid smoothing of the data, which can be a challenging problem when dealing with nonstationary variables (cf. [69]). An important line of future research is the development of inferential procedures for nonlinear cointegration based on these new integrated measures. This is currently investigated by the authors.

There is a large evidence of empirical applications in economics and finance where nonlinearities are found in nonstationary contexts. However, given the difficulty of

the theory involved, only few papers provide a sound justification of the empirical use of cointegration regressions (nonlinear cointegration) in nonlinear frameworks. The difficulties analyzing nonlinear time series models within a stationary and ergodic frameworks are substantially enhanced in nonstationary contexts. In particular, the classical asymptotic theory for nonlinear transformations of nonstationary variables becomes case-dependent (i. e. depends on the specific class of functions), and the available results are confined to the univariate case. A challenging and important line of research deals with the extension of this theory to multivariate frameworks.

Recently, an important step towards the development of a nonlinear cointegration theory has been accomplished by the nonparametric estimation theory of [69,70]. The application of this theory to inference in nonlinear cointegrated models is not fully explored yet. Residual-based tests for testing nonlinear cointegration, such as the so-called KPSS test (cf. [73]), can be constructed using nonparametric residuals. Moreover, model specification tests for nonlinear parametric cointegration can be based on the comparison between parametric and nonparametric fits. Finally, in a recent unpublished lecture, Clive Granger suggested to extend the concept of cointegration to quantiles. These are promising lines of future research which deserve serious attention in the economics literature.

Bibliography

Primary Literature

1. Aparicio F, Escribano A (1999) Information-theoretic analysis of serial correlation and cointegration. *Stud Nonlinear Dyn Econ* 3:119–140
2. Aparicio F, Escribano A, Sipols AE (2006) Range unit root (RUR) tests: Robust against nonlinearities, error distributions, structural breaks and outliers. *J Time Ser Anal* 27:545–576
3. Bae Y, de Jong RM (2005) Money demand function estimation by nonlinear cointegration. Working Paper, Department of Economics, Ohio State University
4. Bandi FM (2004) On persistente and nonparametric estimation (with an application to stock return predictability). Unpublished manuscript
5. Balke NS, Fomby TB (1997) Threshold cointegration. *Int Econ Rev* 38:627–645
6. Bec F, Rahbek A (2004) Vector equilibrium correction models with nonlinear discontinuous adjustments. *Econ J* 7:628–651
7. Boswijk HP (1995) Efficient inference on cointegration parameters in structural error correction models. *J Econ* 69:113–158
8. Breitung J (2001) Rank tests for nonlinear cointegration. *J Bus Econ Stat* 19:331–340
9. Breitung J, Goriéroux C (1997) Rank tests for unit roots. *J Econ* 81:7–27
10. Breitung J, Wulff C (2001) Nonlinear error correction and the efficient market hypothesis: The case of german dual-class shares. *Ger Econ Rev* 2:419–434
11. Burgess SM (1992) Nonlinear dynamics in a structural model of employment. *J Appl Econ* 7:101–118
12. Chang Y, Park JY (2003) Index models with integrated time series. *J Econ* 114:73–16
13. Chang Y, Park JY, Phillips PCB (2001) Nonlinear econometric models with cointegrated and deterministically trending regressors. *Econ J* 4:1–36
14. Choi I, Saikkonen P (2004) Testing linearity in cointegrating smooth transition regressions. *Econ J* 7:341–365
15. Choi I, Saikkonen P (2005) Tests for nonlinear cointegration. Unpublished manuscript
16. Davidson J (1994) *Stochastic limit theory*. Oxford U. P., New York
17. de Jong RM (2001) Nonlinear estimation using estimated cointegrating relations. *J Econ* 101:109–122
18. de Jong RM (2002) Nonlinear minimization estimators in the presence of cointegrating relations. *J Econ* 110:241–259
19. Dufrénot G, Mignon V (2002) Recent developments in nonlinear cointegration with applications to macroeconomics and finance. Kluwer, Boston
20. Engle RF, Granger CWJ (1987) Co-integration and error correction: Representation, estimation, and testing. *Econometrica* 55:251–276
21. Engle RF, Granger CWJ, Rice J, Weiss A (1986) Semiparametric estimates of the relationship between weather and electricity sales. *J Am Stat Assoc* 81:310–320
22. Ericsson NR, Hendry DF, Prestwich KM (1998) The demand for broad money in the United Kingdom, 1878–1993. *Scand J Econ* 100:289–324
23. Escanciano JC, Hualde J (2005) Persistence and long memory in nonlinear time series. Unpublished manuscript
24. Escanciano JC, Velasco C (2006) Testing the martingale difference hypothesis using integrated regression functions. *Comput Stat Data Anal* 51:2278–2294
25. Escribano A (1986) Non-Linear error-correction: The case of money demand in the UK (1878–1970), ch IV. Ph.D. Dissertation, University of California, San Diego
26. Escribano A (1987) Error-Correction systems: Nonlinear adjustment to linear long-run relationships. Core Discussion Paper 8730, C.O.R.E
27. Escribano A (2004) Nonlinear error correction: The case of money demand in the UK (1878–2000). *Macroecon Dyn* 8:76–116
28. Escribano A, Aparicio F (1999) Cointegration: Linearity, non-linearity, outliers and structural breaks. In: Dahiya SB (ed) *The current state of economic science*. Spellbound Publications, pp 383–408
29. Escribano A, Granger CWJ (1998) Investigating the relationship between gold and silver prices. *J Forecast* 17:81–107
30. Escribano A, Mira S (1997) Nonlinear error correction models. Working Paper 97-26. Universidad Carlos III de Madrid
31. Escribano A, Mira S (2002) Nonlinear error correction models. *J Time Ser Anal* 23:509–522
32. Escribano A, Mira S (2007) Specification of nonlinear error correction models: a simulation study. Mimeo, Universidad Carlos III de Madrid
33. Escribano A, Pfann GA (1998) Non-linear error correction, asymmetric adjustment and cointegration. *Econ Model* 15:197–216
34. Escribano A, Sipols AE, Aparicio F (2006) Nonlinear cointegration and nonlinear error correction: Applications of tests

- based on first differences of ranges. *Commun Stat Simul Comput* 35:939–956
35. Escribano A, Santos MT, Sipols AE (2008) Testing for cointegration using induced order statistics. *Comput Stat* 23:131–151
 36. Fotopoulos SB J, Ahn SK (2001) Rank based Dickey-Fuller tests statistics. *J Time Ser Anal* 24:647–662
 37. Friedman M, Schwartz A (1982) Monetary trends in the united states and the United Kingdom: Their relation to income, prices, and interest rates, 1867–1975. University of Chicago Press, Chicago
 38. Gonzalo J (1994) Five alternative methods of estimating long run equilibrium relationships. *J Econ* 60:1–31
 39. Gonzalo J, Pitarakis J-Y (2006) Threshold cointegrating relationships. *Oxford Bull Econ Stat* 68:813–833
 40. Gouriéroux C, Jasiak J (1999) Nonlinear persistence and copersistence. Unpublished manuscript
 41. Gouriéroux C, Jasiak J (2002) Nonlinear autocorrelograms: An application to inter-trade durations. *J Time Ser Anal* 23:127–154
 42. Granger CWJ (1981) Some properties of time series data and their use in econometric model specification. *J Econ* 16:121–130
 43. Granger CWJ (1986) Developments in the study of cointegrated variables. *Oxford Bull Econ Stat* 48:213–228
 44. Granger CWJ (1995) Modelling nonlinear relationships between extended-memory variables. *Econometrica* 63:265–279
 45. Granger CWJ (2002) Long memory, volatility, risk and distribution. Unpublished manuscript
 46. Granger CWJ, Hallman J (1991) Long memory series with attractors. *Oxford Bull Econ Stat* 53:11–26
 47. Granger CWJ, Hallman J (1991) Nonlinear transformations of integrated time series. *J Time Ser Anal* 12:207–224
 48. Granger CWJ, Lee TH (1989) Investigation of production, sales and inventory relationships using multicointegration and non-symmetric error correction models. *J Appl Econ* 4:145–159
 49. Granger CWJ, Swanson N (1996) Further developments in the study of cointegrated variables. *Oxford Bull Econ Stat* 58:537–553
 50. Granger CWJ, Teräsvirta T (1993) Modelling nonlinear economic relationships. Oxford University Press, New York
 51. Granger CWJ, Maasoumi E, Racine J (2004) A dependence metric for possibly nonlinear processes. *J Time Ser Anal* 25:649–669
 52. Hamilton JD (1994) Time series analysis. Princeton University Press, Princeton
 53. Hansen B, Phillips PCB (1990) Estimation and inference in models of cointegration: A simulation study. *Adv Econ* 8:225–248
 54. Hansen B, Seo B (2002) Testing for two-regime threshold cointegration in vector error correction models. *J Econ* 110:293–318
 55. Härdle W (1990) Applied nonparametric regression. *Econetric Society Monographs*, vol 19. Cambridge University Press, Cambridge
 56. Hargreaves CP (1994) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
 57. Hargreaves CP (1994) A review of methods of estimating cointegrating relationships. In: Hargreaves CP (ed) Nonstationary time series analysis and cointegration. Oxford University Press, Oxford
 58. Haug AA (2002) Testing linear restrictions on cointegrating vectors: Sizes and powers of wald and likelihood ratio tests in finite samples. *Econ Theory* 18:505–524
 59. Hendry DF (1995) Dynamic econometrics. Oxford University Press, Oxford
 60. Hendry DF, Ericsson N (1983) Assertion without empirical basis: An econometric appraisal of 'monetary trends in the ... the United Kingdom' by Friedman M, Schwartz AJ. In: Monetary trends in the United Kingdom. Bank of England Panel of Academic Consultants, Paper 22, 45–101
 61. Hendry DF, Ericsson NR (1991) An econometric analysis of the uk money demand in 'monetary trends in the united states and the United Kingdom' by Friedman M, Schwartz AJ. *Am Econ Rev* 81:8–38
 62. Hendry DF, Massmann M (2007) Co-breaking: Recent advances and a synopsis of the literature. *J Bus Econ Stat* 25:33–51
 63. Inder B (1993) Estimating long-run relationships in econometrics: A comparison of different approaches. *J Econ* 57:53–68
 64. Johansen S (1988) Statistical analysis of cointegration vectors. *J Econ Dyn Control* 12:231–54
 65. Johansen S (1991) Estimation and hypothesis testing of cointegration vectors, in gaussian vector autoregressive models. *Econometrica* 59:1551–1580
 66. Johansen S (1992) Cointegration in partial systems and the efficiency of single-equation analysis. *J Econ* 52:389–402
 67. Juselius K (2006) The cointegrated VAR model: Methodology and applications. Oxford University Press, Oxford
 68. Jusmah A, Kunst RM (2008) Modelling macroeconomic sub-aggregates: An application of non-linear cointegration. *Economics Series*, Institute for Advanced Studies. *Macroecon Dyn* 12:151–171
 69. Karlsen HA, Tjøstheim D (2001) Nonparametric estimation in null recurrent time series. *Ann Stat* 29:372–416
 70. Karlsen HA, Myklebust T, Tjøstheim D (2007) Nonparametric estimation in a nonlinear cointegration type model. *Ann Stat* 35(1):252–299
 71. Kitamura Y, Phillips PCB (1995) Efficient IV estimation in nonstationary regression: An overview and simulation study. *Econ Theory* 11:1095–1130
 72. Kunst RM (1992) Threshold cointegration in interest rates. Working Paper, Institute for Advanced Studies, Vienna
 73. Kwiatkowski D, Phillips PCB, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root. *J Econ* 54:159–178
 74. Lo AW (1991) Long-term memory in stock market prices. *Econometrica* 59:1279–1313
 75. Lo M, Zivot E (2001) Threshold cointegration and nonlinear adjustment to the law of one price. *Macroecon Dyn* 5:533–576
 76. Longbottom A, Holly S (1985) Econometric methodology and monetarism: Professor Friedman and Professor Hendry on the demand for money, Discussion Paper No. 131. London Business School
 77. Lütkepohl H (2005) New introduction to multiple time series analysis. Springer
 78. Marmol F, Escribano A, Aparicio FM (2002) Instrumental variable interpretation of cointegration with inference results for fractional cointegration. *Econ Theory* 18:646–672

79. Park JY (1992) Canonical cointegrating regressions. *Econometrica* 60:119–143
80. Park JY, Phillips PCB (1988) Statistical inference in regressions with integrated process: Part 1. *Econ Theory* 4:468–498
81. Park JY, Phillips PCB (1989) Statistical inference in regressions with integrated process: Part 2. *Econ Theory* 4:95–132
82. Park JY, Phillips PCB (1989) Asymptotics for nonlinear transformations of integrated time series. *Econ Theory* 15:269–298
83. Park JY, Phillips PCB (2000) Nonstationary binary choice. *Econometrica* 68:1249–1280
84. Park JY, Phillips PCB (2001) Nonlinear regressions with integrated time series. *Econometrica* 69:117–161
85. Phillips PCB (1991) Optimal inference in cointegrated systems. *Econometrica* 59:283–306
86. Phillips PCB, Hansen BE (1990) Statistical inference in instrumental variable regression with $I(1)$ processes. *Rev Econ Stud* 57:99–125
87. Phillips PCB, Loretan M (1991) Estimating long-run economic equilibria. *Rev Econ Stud* 59:407–436
88. Phillips PCB, Park JY (1998) Nonstationary density estimation and kernel autoregression. Unpublished manuscript
89. Robinson PM, Gerolimetto M (2006) Instrumental variables estimation of stationary and nonstationary cointegrating regressions. *Econ J* 9:291–306
90. Rothman P, van Dijk D, Franses PH (2001) A multivariate star analysis of the relationship between money and output. *Macroecon Dyn* 5:506–532
91. Saikkonen P (1991) Asymptotically efficient estimation of cointegration regressions. *Econ Theory* 7:1–21
92. Saikkonen P (2005) Stability results for nonlinear error correction models. *J Econ* 127:69–81
93. Saikkonen P, Choi I (2004) Cointegrating smooth transition regressions. *Econ Theory* 20:301–340
94. Seo MH (2006) Bootstrap testing for the null of no cointegration in a threshold vector error correction model. *J Econ* 134:129–150
95. Seo MH (2007) Estimation of nonlinear error correction models. Unpublished manuscript
96. Silverman BW (1985) Some aspects of the spline smoothing approach to nonparametric regression curve fitting. *J Roy Stat Soc Ser B* 47:1–52
97. Sims CA (1972) Money, income, and causality. *Am Econ Rev* 62:540–552
98. Stock JH (1987) Asymptotic properties of least squares estimation of cointegrating vectors. *Econometrica* 55:1035–1056
99. Stock JH (1994) Deciding between $I(1)$ and $I(0)$. *J Econ* 63:105–131
100. Stock JH, Watson MW (1993) A simple estimator of cointegration vectors in higher order integrated systems. *Econometrica* 61:783–820
101. Teräsvirta T (1998) Modelling economic relationships with smooth transition regressions. In: Ullah A, Giles DEA (eds) *Handbook of applied economic statistics*. Marcel Dekker, New York, pp 507–552
102. Teräsvirta T, Eliasson AC (2001) Non-linear error correction and the uk demand for broad money, 1878–1993. *J Appl Econ* 16:277–288
103. Velasco C (2006) Semiparametric estimation of long-memory models. In: Patterson K, Mills TC (eds) *Palgrave handbook of econometrics*, vol 1. *Econometric Theory*. MacMillan, Palgrave, pp 353–395
104. Wahba G (1975) Smoothing noisy data with spline functions. *Num Math* 24:309–63–75
105. Watson M (1994) Large sample estimation and hypothesis testing, vector autoregression and cointegration. In: Engle RF, McFadden DL (eds) *Handbook of econometrics*, vol IV. Elsevier Science, Amsterdam

Books and Reviews

- Franses PH, Terasvirta T (eds) (2001) Special issue on nonlinear modeling of multivariate macroeconomic relations. *Macroecon Dyn* 5(4)
- Banerjee A, Dolado JJ, Galbraith JW, Hendry DF (1993) Co-integration, error correction and the econometric analysis of non-stationary data. Oxford University Press, Oxford
- Bierens HJ (1981) Robust methods and asymptotic theory in nonlinear econometrics. *Lecture Notes in Economics and Mathematical Systems*, vol 192. Springer, Berlin
- Clements MP, Hendry DF (1999) Forecasting non-stationary economic time series. MIT Press, Cambridge
- Dhrymes P (1998) Time series, unit roots, and cointegration. Academic Press, San Diego
- Enders W (1995) Applied econometric time series. Wiley
- Engle RF, Granger CWJ (eds) (1991) Long-run economic relationships: Readings in cointegration. Oxford University Press, Oxford
- Franses PH, van Dijk D (2000) Nonlinear time series models in empirical finance. Cambridge University Press, Cambridge
- Gonzalo J, Dolado JJ, Marmol F (2001) A primer in cointegration. In: Baltagui BH (ed) *A companion to theoretical econometrics*. Blackwell, New York, ch 30
- Granger CWJ (2001) Overview of nonlinear macroeconomic empirical models. *Macroecon Dyn* 5:466–481
- Granger CWJ (2007) Some thoughts on the past and future of cointegration. Mimeo
- Hatanaka M (1996) Time series-based econometrics. Oxford University Press, Oxford
- Maddala GS, Kim I-M (1998) Unit roots, cointegration and structural change. Cambridge University Press, Cambridge
- Phillips PCB (1986) Understanding spurious regressions in econometrics. *J Econ* 33:311–340

Econometrics: Panel Data Methods

JEFFREY M. WOOLDRIDGE

Department of Economics, Michigan State University,
East Lansing, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Overview of Linear Panel Data Models](#)

[Sequentially Exogenous Regressors and Dynamic Models](#)

[Unbalanced Panel Data Sets](#)

[Nonlinear Models](#)

Future Directions

Bibliography

Glossary

Panel data Data on a set of cross-sectional units followed over time.

Unobserved effects Unobserved variables that affect the outcome which are constant over time.

Fixed effects estimation An estimation method that removes the unobserved effects, implying that the unobserved effects can be arbitrarily related to the observed covariates.

Correlated random effects An approach to modeling where the dependence between the unobserved effects and the history of the covariates is parametrically modeled. The traditional random effects approach is a special case under the assumption that the unobserved effects are independent of the covariates.

Average partial effect The partial effect of a covariate averaged across the distribution of the unobserved effects.

Definition of the Subject

Panel data consist of repeated observations over time on the same set of cross-sectional units. These units can be individuals, firms, schools, cities, or any collection of units one can follow over time. Special econometric methods have been developed to recognize and exploit the rich information available in panel data sets. Because the time dimension is a key feature of panel data sets, issues of serial correlation and dynamic effects need to be considered. Further, unlike the analysis of cross-sectional data, panel data sets allow the presence of systematic, unobserved differences across units that can be correlated with observed factors whose effects are to be measured. Distinguishing between persistence due to unobserved heterogeneity and that due to dynamics in the underlying process is a leading challenge for interpreting estimates from panel data models.

Panel data methods are the econometric tools used to estimate parameters, compute partial effects of interest in nonlinear models, quantify dynamic linkages, and perform valid inference when data are available on repeated cross sections. For linear models, the basis for many panel data methods is ordinary least squares applied to suitably transformed data. The challenge is to develop estimators, assumptions with good properties under reasonable assumptions, and to ensure that statistical inference is valid. Maximum likelihood estimation

plays a key role in the estimation of nonlinear panel data models.

Introduction

Many questions in economics, especially those with foundations in the behavior of relatively small units, can be empirically studied with the help of panel data. Even when detailed cross-sectional surveys are available, collecting enough information on units to account for systematic differences is often unrealistic. For example, in evaluating the effects of a job training program on labor market outcomes, unobserved factors might affect both participation in the program and outcomes such as labor earnings. Unless participation in the job training program is randomly assigned, or assigned on the basis of observed covariates, cross-sectional regression analysis is usually unconvincing. Nevertheless, one can control for this individual heterogeneity – including unobserved, time-constant human capital – by collecting a panel data set that includes data points both before and after the training program.

Some of the earliest econometric applications of panel data methods were to the estimation of agricultural production functions, where the worry was that unobserved inputs – such as soil quality, technical efficiency, or managerial skill of the farmer – would generally be correlated with observed inputs such as capital, labor, and amount of land. Classic examples are [31,45].

The nature of unobserved heterogeneity was discussed early in the development of panel data models. An important contribution is [46], which argued persuasively that in applications with many cross-sectional units and few time periods, it always makes sense to treat unit-specific heterogeneity as outcomes of random variables, rather than parameters to estimate. As Mundlak made clear, for economic applications the key issue is whether the unobserved heterogeneity can be assumed to be independent, or at least uncorrelated, with the observed covariates. [25] developed a testing framework that can be used, and often is, to test whether unobserved heterogeneity is correlated with observed covariates. Mundlak's perspective has had a lasting impact on panel data methods, and his insights have been applied to a variety of dynamic panel data models with unobserved heterogeneity.

The 1980s witnessed an explosion in both methodological developments and applications of panel data methods. Following the approach in [15,16,46], and [17] provided a unified approach to linear and nonlinear panel data models, and explicitly dealt with issues of inference in cases where full distributions were not specified. Dynamic linear models, and the problems they pose for estimation

and inference, were considered in [4]. Dynamic discrete response models were analyzed in [29,30]. The hope in estimating dynamic models that explicitly contain unobserved heterogeneity is that researchers can measure the importance of two causes for persistence in observed outcomes: unobserved, time-constant heterogeneity and so-called *state dependence*, which describes the idea that, conditional on observed and unobserved factors, the probability of being in a state in the current time period is affected by last period's state.

In the late 1980s and early 1990s, researchers began using panel data methods to test economic theories such as rational expectations models of consumption. Unlike macro-level data, data at the individual or family level allows one to control for different preferences, and perhaps different discount rates, in testing the implications of rational expectations. To avoid making distributional assumptions on unobserved shocks and heterogeneity, researchers often based estimation on conditions on expected values that are implied by rational expectations, as in [40].

Other developments in the 1990s include studying standard estimators under fewer assumptions – such as the analysis in [53] of the fixed effects Poisson estimator under distributional misspecification and unrestricted serial dependence – and the development of estimators in nonlinear models that are consistent for parameters under no distributional assumptions – such as the new estimator proposed in [33] for the panel data censored regression model.

The past 15 years has seen continued development of both linear and nonlinear models, with and without dynamics. For example, on the linear model front, methods have been proposed for estimating models where the effects of time-invariant heterogeneity can change over time – as in [2]. Semiparametric methods for estimating production functions, as in [48], and dynamic models, as in the dynamic censored regression model in [34], have been developed. Flexible parametric models, estimated by maximum likelihood, have also been proposed (see [57]).

Many researchers are paying closer attention to estimation of partial effects, and not just parameters, in nonlinear models – with or without dynamics. Results in [3] show how partial effects, with the unobserved heterogeneity appropriately averaged out, can be identified under weak assumptions.

The next several sections outline a modern approach to panel data methods. Section “[Future Directions](#)” provides an account of more recent advances, and discusses where those advances might head in the future.

Overview of Linear Panel Data Models

In panel data applications, linear models are still the most widely used. When drawing data from a large population, random sampling is often a realistic assumption; therefore, we can treat the observations as independent and identically distributed outcomes. For a random draw i from the population, the linear panel data model with additive heterogeneity can be written as

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, \dots, T, \quad (1)$$

where T is the number of time periods available for each unit and t indexes time periods. The time periods are often years, but the span between periods can be longer or shorter than a year. The distance between any two time periods need not be the same, although different spans can make it tricky to estimate certain dynamic models. As written, Eq. (1) assumes that we have the same time periods available for each cross-sectional unit. In other words, the panel data set is *balanced*.

As in any regression analysis, the left-hand-side variable is the dependent variable or the response variable. The terms η_t , which depend only on time, are treated here as parameters. In most microeconomic applications, the cross-sectional sample size, denoted N , is large – often very large – compared with T . Therefore, the η_t can be estimated precisely in most cases. Almost all applications should allow for aggregate time effects as captured by η_t . Including such time effects allows for secular changes in the economic environment that affect all units in the same way (such as inflation or aggregate productivity). For example, in studying the effects of school inputs on performance using school-level panel data for a particular state, including η_t allows for trends in statewide spending along with separate, unrelated trends in statewide test performance. It could be that, say, real spending rose at the same time that the statewide standardized tests were made easier; a failure to account for such aggregate trends could lead to a spurious association between performance and spending. Only occasionally are the η_t the focus of a panel data analysis, but it is sometimes interesting to study the pattern of aggregate changes once the covariates contained in the $1 \times K$ vector \mathbf{x}_{it} are netted out.

The parameters of primary interest are contained in the $K \times 1$ vector $\boldsymbol{\beta}$, which contains the coefficients on the set of explanatory variables. With the presence of η_t in (1), \mathbf{x}_{it} cannot include variables that change only over time. For example, if y_{it} is a measure of labor earnings for individual i in year t for a particular state in the US, \mathbf{x}_{it} cannot contain the state-level unemployment rate. Unless interest centers on how individual earnings depend on the

state-level unemployment rate, it is better to allow for different time intercepts in an unrestricted fashion: this way, any aggregate variables that affect each individual in the same way are accounted for without even collecting data on them. If the η_t are restricted to be functions of time – for example, a linear time trend – then aggregate variables can be included, but this is always more restrictive than allowing the η_t to be unrestricted.

The composite error term in (1), $c_i + u_{it}$, is an important feature of panel data models. With panel data, it makes sense to view the unobservables that affect y_{it} as consisting of two parts: the first is the time-constant variable, c_i , which is often called an *unobserved effect* or *unit-specific heterogeneity*. This term aggregates all factors that are important for unit i 's response that do not change over time. In panel data applications to individuals, c_i is often interpreted as containing cognitive ability, family background, and other factors that are essentially determined prior to the time periods under consideration. Or, if i indexes different schools across a state, and (1) is an equation to see if school inputs affect student performance, c_i includes historical factors that can affect student performance and also might be correlated with observed school inputs (such as class sizes, teacher competence, and so on). The word “heterogeneity” is often combined with a qualifier that indicates the unit of observation. For example, c_i might be “individual-specific heterogeneity” or “school-specific heterogeneity”. Often in the literature c_i is called a “random effect” or “fixed effect”, but these labels are not ideal. Traditionally, c_i was considered a random effect if it was treated as a random variable, and it was considered a fixed effect if it was treated as a parameter to estimate (for each i). The flaws with this way of thinking are revealed in [46]: the important issue is not whether c_i is random, but whether it is correlated with \mathbf{x}_{it} .

The sequence of errors $\{u_{it} : t = 1, \dots, T\}$ are specific to unit i , but they are allowed to change over time. Thus, these are the time-varying unobserved factors that affect y_{it} , and they are often called the *idiosyncratic errors*. Because u_{it} is in the error term at time t , it is important to know whether these unobserved, time-varying factors are uncorrelated with the covariates. It is also important to recognize that these idiosyncratic errors can be serially correlated, and often are.

Before treating the various assumptions more formally in the next subsection, it is important to recognize the asymmetry in the treatment of the time-specific effects, η_t , and the unit-specific effects, c_i . Language such as “both time and school fixed effects are included in the equation” is common in empirical work. While the language itself is harmless, with large N and small T it is best to view the

time effects, η_t , as parameters to estimate because they can be estimated precisely. As already mentioned earlier, viewing c_i as random draws is the most general, and natural, perspective.

Assumptions and Estimators for the Basic Model

The assumptions discussed in this subsection are best suited to cases where random sampling from a (large) population is realistic. In this setting, it is most natural to describe large-sample statistical properties as the cross-sectional sample size, N , grows, with the number of time periods, T , fixed.

In describing assumptions in the model (1), it probably makes more sense to drop the i subscript in (1) to emphasize that the equation holds for an entire population. Nevertheless, (1) is useful for emphasizing which factors change i , or t , or both. It is sometimes convenient to subsume the time dummies in \mathbf{x}_{it} , so that the separate intercepts η_t need not be displayed.

The traditional starting point for studying (1) is to rule out correlation between the idiosyncratic errors, u_{it} , and the covariates, \mathbf{x}_{it} . A useful assumption is that the sequence of explanatory variables $\{\mathbf{x}_{it} : t = 1, \dots, T\}$ is *contemporaneously exogenous conditional on c_i* :

$$E(u_{it} | \mathbf{x}_{it}, c_i) = 0, \quad t = 1, \dots, T. \quad (2)$$

This assumption essentially defines β in the sense that, under (1) and (2),

$$E(y_{it} | \mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\beta + c_i, \quad (3)$$

so the β_j are partial effects holding fixed the unobserved heterogeneity (and covariates other than \mathbf{x}_{it}). Strictly speaking, c_i need not be included in the conditioning set in (2), but including it leads to the useful Eq. (3). Plus, for purposes of stating assumptions for inference, it is convenient to express the contemporaneous exogeneity assumption as in (2).

Unfortunately, with a small number of time periods, β is not identified by (2), or by the weaker assumption $\text{Cov}(\mathbf{x}_{it}, u_{it}) = \mathbf{0}$. Of course, if c_i is assumed to be uncorrelated with the covariates, that is $\text{Cov}(\mathbf{x}_{it}, c_i) = \mathbf{0}$ for any t , then the composite error, $v_{it} = c_i + u_{it}$ is uncorrelated with \mathbf{x}_{it} , and then β is identified and can be consistently estimated by a cross section regression using a single time period t , or by using pooled regression across t . (See Chaps. 7 and 10 in [55] for further discussion.) But one of the main purposes in using panel data is to allow the unobserved effect to be correlated with time-varying \mathbf{x}_{it} .

Arbitrary correlation between c_i and $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT})$ is allowed if the sequence of explanatory variables

is strictly exogenous conditional on c_i ,

$$E(u_{it}|\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{iT}, c_i) = 0, \quad t = 1, \dots, T, \quad (4)$$

which can be expressed as

$$E(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + c_i. \quad (5)$$

Clearly, assumption (4) implies (2). Because the entire history of the covariates is in (4) for all t , (4) implies that \mathbf{x}_{ir} and u_{it} are uncorrelated for all r and t , including $r = t$. By contrast, (2) allows arbitrary correlation between \mathbf{x}_{ir} and u_{it} for any $r \neq t$. The strict exogeneity assumption (4) can place serious restrictions on the nature of the model and dynamic economic behavior. For example, (4) can never be true if \mathbf{x}_{it} contains lags of the dependent variable. Of course, (4) would be false under standard econometric problems, such as omitted time-varying variables, just as would (2). But there are important cases where (2) can hold but (4) might not. If, say, a change in u_{it} causes reactions in future values of the explanatory variables, then (4) is generally false. In applications to the social sciences, the potential for these kind of “feedback effects” is important. For example, in using panel data to estimate a firm-level production function, a shock to production today (captured by changes in u_{it}) might affect the amount of capital and labor inputs in the next time period. In other words, u_{it} and $\mathbf{x}_{i,t+1}$ would be correlated, violating (4).

How does assumption (4) (or (5)) identify the parameters? In fact, it only allows estimation of coefficients on time-varying elements of \mathbf{x}_{it} . Intuitively, because (4) puts no restrictions on the dependence between c_i and \mathbf{x}_i , it is not possible to distinguish between the effect of a time-constant observable covariate and that of the unobserved effect, c_i . For example, in an equation to describe the amount of pension savings invested in the stock market, c_i might include innate tolerance for risk, assumed to be fixed over time. Once c_i is allowed to be correlated with any observable covariate – including, say, gender – the effects of gender on stock market investing cannot be identified because gender, like c_i , is constant over time. Mechanically, common estimation methods eliminate c_i along with any time-constant explanatory variables. (What is meant by “time-varying” x_{itj} is that for at least some i , x_{itj} changes over time. For some units i , x_{itj} might be constant). When a full set of year intercepts – or even just a linear time trend – is included, the effects of variables that increase by the same amount in each period – such as a person’s age – cannot be included in \mathbf{x}_{it} . The reason is that the beginning age of each person is indistinguishable

from c_i , and then, once the initial age is known, each subsequent age is a deterministic – in fact, linear – function of time.

Perhaps the most common method of estimating $\boldsymbol{\beta}$ (and the η_t) is so-called *fixed effects* (FE) or *within* estimation. The FE estimator is obtained as a pooled OLS regression on variables that have had the unit-specific means removed. More precisely, let $\tilde{y}_{it} = y_{it} - T^{-1} \sum_{r=1}^T y_{ir} = y_{it} - \bar{y}_i$ be the deviation of y_{it} from the average over time for unit i , \tilde{y}_i and similarly for $\tilde{\mathbf{x}}_{it}$ (which is a vector). Then,

$$\tilde{y}_{it} = \tilde{\eta}_t + \tilde{\mathbf{x}}_{it}\boldsymbol{\beta} + \tilde{u}_{it}, \quad t = 1, \dots, T, \quad (6)$$

where the year intercepts and idiosyncratic errors are, of course, also demeaned. Consistency of pooled OLS (for fixed T and $N \rightarrow \infty$) applied to (6) essentially requires rests on $\sum_{t=1}^T E(\tilde{\mathbf{x}}'_{it}\tilde{u}_{it}) = \sum_{t=1}^T E(\tilde{\mathbf{x}}'_{it}u_{it}) = \mathbf{0}$, which means the error u_{it} should be uncorrelated with \mathbf{x}_{ir} for all r and t . This assumption is implied by (4). A rank condition on the demeaned explanatory variables is also needed. If $\tilde{\eta}_t$ is absorbed into $\tilde{\mathbf{x}}_{it}$, the condition is $\text{rank} \sum_{t=1}^T E(\tilde{\mathbf{x}}'_{it}\tilde{\mathbf{x}}_{it}) = K$, which rules out time constant variables and other variables that increase by the same value for all units in each time period (such as age).

A different estimation method is based on an equation in first differences. For $t > 1$, define $\Delta y_{it} = y_{it} - y_{i,t-1}$, and similarly for the other quantities. The first-differenced equation is

$$\Delta y_{it} = \delta_t + \Delta \mathbf{x}_{it}\boldsymbol{\beta} + \Delta u_{it}, \quad t = 2, \dots, T, \quad (7)$$

where $\delta_t = \eta_t - \eta_{t-1}$ is the change in the intercepts. The *first-difference* (FD) estimator is pooled OLS applied to (7). Any element x_{itj} of \mathbf{x}_{it} such that Δx_{itj} is constant for all i and t (most often zero) drops out, just as in FE estimation. Assuming suitable time variation in the covariates, $E(\Delta \mathbf{x}'_{it}\Delta u_{it}) = \mathbf{0}$ is sufficient for consistency. Naturally, this assumption is also implied by assumption (4).

Whether FE or FD estimation is used – and it is often prudent to try both approaches – inference about $\boldsymbol{\beta}$ can and generally should be made fully robust to heteroskedasticity and serial dependence. The robust asymptotic variance of both FE and FD estimators has the so-called “sandwich” form, which allows the vector of idiosyncratic errors, $\mathbf{u}_i = (u_{i1}, \dots, u_{iT})'$, to contain arbitrary serial correlation and heteroskedasticity, where the conditional covariances and variances can depend on \mathbf{x}_i in an unknown way. For notational simplicity, absorb dummy variables for the different time periods into \mathbf{x}_{it} . Let $\hat{\boldsymbol{\beta}}_{\text{FE}}$ denote the fixed effects estimator and $\hat{\mathbf{u}}_i = \tilde{\mathbf{y}}_i - \tilde{\mathbf{X}}_i\hat{\boldsymbol{\beta}}_{\text{FE}}$ the $T \times 1$ vector of fixed effects residuals for unit i . Here, $\tilde{\mathbf{X}}_i$ is the $T \times K$ matrix with t^{th} row $\tilde{\mathbf{x}}_{it}$.

Then a fully robust estimator of the asymptotic variance of $\hat{\beta}_{FE}$ is

$$\widehat{\text{Avar}}(\hat{\beta}_{FE}) = \left(\sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1} \left(\sum_{i=1}^N \ddot{\mathbf{X}}_i' \hat{\mathbf{u}}_i \hat{\mathbf{u}}_i' \ddot{\mathbf{X}}_i \right) \cdot \left(\sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i \right)^{-1}, \quad (8)$$

where it is easily seen that $\sum_{i=1}^N \ddot{\mathbf{X}}_i' \ddot{\mathbf{X}}_i = \sum_{i=1}^N \sum_{t=1}^T \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it}$ and the middle part of the sandwich consists of terms $\hat{\mathbf{u}}_{it} \hat{\mathbf{u}}_{it}' \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it}$ for all $r, t = 1, \dots, T$. See Chap. 10 in [55] for further discussion. A similar expression holds for $\hat{\beta}_{FD}$ but where the demeaned quantities are replaced by first differences.

When $T = 2$, it can be shown that the FE and FD estimation and inference about β are identical. If $T > 2$, the procedures generally differ. If (4) holds and $T > 2$, how does one choose between the FE and FD approaches? Because both are consistent and \sqrt{N} -asymptotically normal, the only way to choose is from efficiency considerations. Efficiency of the FE and FD estimators hinges on second moment assumptions concerning the idiosyncratic errors. Briefly, if $E(\mathbf{u}_i \mathbf{u}_i' | \mathbf{x}_i) = E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$, then the FE estimator is efficient. Practically, the most important implication of this assumption is that the idiosyncratic errors are serially uncorrelated. But they should also be homoskedastic, which means the variances can neither depend on the covariates nor change over time. The FD estimator is efficient if the errors in (7) are serially uncorrelated and homoskedasticity, which can be stated as $E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i' | \mathbf{x}_i) = E(\Delta \mathbf{u}_i \Delta \mathbf{u}_i') = \sigma_e^2 \mathbf{I}_{T-1}$, where $e_{it} = u_{it} - u_{i,t-1}$ and $\Delta \mathbf{u}_i$ is the $T-1$ vector of first-differenced errors. These two sets of conditions – that $\{u_{it} : t = 1, \dots, T\}$ is a serially uncorrelated sequence (for FE to be efficient) versus $\{u_{it} : t = 1, \dots, T\}$ is a random walk (for FD to be efficient) – represent extreme cases. Of course, there is much in between. In fact, probably neither condition should be assumed to be true, which is a good argument for robust inference. More efficient estimation can be based on generalized method of moments (GMM – see Chap. 8 in [55] – or minimum distance estimation, as in [16]).

It is good practice to compute both FE and FD estimates to see if they differ in substantive ways. It is also helpful to have a formal test of the strict exogeneity assumption that is easily computable and that maintains only strict exogeneity under the null – in particular, that takes no stand on whether the FE or FD estimator is asymptotically efficient. Because lags of covariates can al-

ways be included in a model, the primary violation of (4) that is of interest is due to feedback. Therefore, it makes sense to test that $\mathbf{x}_{i,t+1}$ is uncorrelated with u_{it} . Actually, let \mathbf{w}_{it} be a subset of \mathbf{x}_{it} that is suspected of failing the strict exogeneity assumption, and consider the augmented model

$$y_{it} = \eta_t + \mathbf{x}_{it} \beta + \mathbf{w}_{i,t+1} \delta + c_i + u_{it}, \quad t = 1, \dots, T-1. \quad (9)$$

Under the null hypothesis that $\{\mathbf{x}_{it} : t = 1, \dots, T\}$ is strictly exogenous, $H_0: \delta = \mathbf{0}$, and this is easily tested using fixed effects (using all but the last time period) or first differencing (where, again, the last time period is lost). It makes sense, as always, to make the test fully robust to serial correlation and heteroskedasticity. This test may probably has little power for detecting contemporaneous endogeneity, that is, correlation between \mathbf{w}_{it} and u_{it} .

A third common approach to estimation of unobserved effects models is so-called *random effects* estimation. RE estimation differs from FE and FD by leaving c_i in the error term and then accounting for its presence via generalized least squares (GLS). Therefore, the exogeneity requirements of the covariates must be strengthened. The most convenient way of stating the key random effects (RE) assumption is

$$E(c_i | \mathbf{x}_i) = E(c_i), \quad (10)$$

which ensures that every element of \mathbf{x}_i – that is, all explanatory variables in all time periods – is uncorrelated with c_i . Together with (4), (10) implies

$$E(v_{it} | \mathbf{x}_i) = 0, \quad t = 1, \dots, T, \quad (11)$$

where $v_{it} = c_i + u_{it}$ is the composite error. Condition (11) is the key condition for general least squares methods that exploit serial correlation in v_{it} to be consistent (although zero correlation would be sufficient). The random effects estimator uses a special structure for the variance-covariate matrix of \mathbf{v}_i , the $T \times 1$ vector of composite errors. If $E(\mathbf{u}_i \mathbf{u}_i') = \sigma_u^2 \mathbf{I}_T$ and c_i is uncorrelated with each u_{it} (as implied by assumption (4)), then

$$\text{Var}(\mathbf{v}_i) = \begin{pmatrix} \sigma_c^2 + \sigma_u^2 & \sigma_c^2 & \cdots & \sigma_c^2 \\ \sigma_c^2 & \sigma_c^2 + \sigma_u^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \sigma_c^2 \\ \sigma_c^2 & \cdots & \sigma_c^2 & \sigma_c^2 + \sigma_u^2 \end{pmatrix}. \quad (12)$$

Both σ_c^2 and σ_u^2 can be estimated after, say, preliminary estimation by pooled OLS (which is consistent under (11)) –

see, for example, Chap. 10 in [55] – and then a feasible GLS is possible. If (12) holds, along with the system homoskedasticity assumption $\text{Var}(\mathbf{v}_i|\mathbf{x}_i) = \text{Var}(\mathbf{v}_i)$, then feasible GLS is efficient, and the inference is standard. Even if $\text{Var}(\mathbf{v}_i|\mathbf{x}_i)$ is not constant, or $\text{Var}(\mathbf{v}_i)$ does not have the random effects structure in (12), the RE estimator is consistent provided (11) holds (Again, this is with N growing and T fixed). Therefore, although it is still not common, a good case can be made for using robust inference – that is, inference that allows an unknown form of $\text{Var}(\mathbf{v}_i|\mathbf{x}_i)$ – in the context of random effects. The idea is that the RE estimator can be more efficient than pooled OLS even if (12) fails, yet inference should not rest on (12). Chapter 10 in [55] contains the sandwich form of the estimator.

Under the key RE assumption (11), \mathbf{x}_{it} can contain time-constant variables. In fact, one way to ensure that the omitted factors are uncorrelated with the key covariates is to include a rich set of time-constant controls in \mathbf{x}_{it} . RE estimation is most convincing when many good time-constant controls are available. In some applications of RE, the key variable of interest does not change over time, which is why FE and FD cannot be used. (Methods proposed in [26] can be used when some covariates are correlated with c_i , but enough others are assumed to be uncorrelated with c_i).

Rather than eliminate c_i using the FE or FD transformation, or assuming (10) and using GLS, a different approach is to explicitly model the correlation between c_i and \mathbf{x}_i . A general approach is to write

$$c_i = \psi + \mathbf{x}_i\boldsymbol{\lambda} + a_i, \quad (13)$$

$$E(a_i) = 0 \quad \text{and} \quad E(\mathbf{x}_i' a_i) = 0, \quad (14)$$

where $\boldsymbol{\lambda}$ is a $TK \times 1$, vector of parameters. Equations (13) and (14) are definitional, and simply define the population linear regression of c_i on the entire set of covariates, \mathbf{x}_i . This representation is due to [16], and is an example of a *correlated random effects* (CRE) model. The uncorrelated random effects model occurs when $\boldsymbol{\lambda} = \mathbf{0}$.

A special case of (13) was used in [46], assuming that each \mathbf{x}_{it} has the same set of coefficients. Plus, [46] actually used conditional expectations (which is unnecessary but somewhat easier to work with):

$$c_i = \psi + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i \quad (15)$$

$$E(a_i|\mathbf{x}_i) = 0, \quad (16)$$

where recall that $\bar{\mathbf{x}}_i = T^{-1} \sum_{t=1}^T \mathbf{x}_{it}$. This formulation conserves on degrees of freedom, and extensions are useful for nonlinear models.

Plugging (15) into the original equation gives

$$y_{it} = \eta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\xi} + a_i + u_{it}, \quad (17)$$

where ψ is absorbed into the time intercepts. The composite error $a_i + u_{it}$ satisfies $E(a_i + u_{it}|\mathbf{x}_i) = 0$, and so pooled OLS or random effects applied to (17) produces consistent, \sqrt{N} -asymptotically normal estimators of all parameters, including $\boldsymbol{\xi}$. In fact, if the original model satisfies the second moments ideal for random effects, then so does (17). Interesting, both pooled OLS and RE applied to (17) produce the fixed effects estimate of $\boldsymbol{\beta}$ (and the η_t). Therefore, the FE estimator can be derived from a correlated random effects model. (Somewhat surprisingly, the same algebraic equivalence holds using Chamberlain's more flexible device. Of course, the pooled OLS estimator is not generally efficient, and [16] shows how to obtain the efficient minimum distance estimator. See also Chap. 11 in [55]).

One advantage of Eq. (17) is that it provides another interpretation of the FE estimate: it is obtained by holding fixed the time averages when obtaining the partial effects of each x_{itj} . This results in a more convincing analysis than not controlling for systematic differences in the levels of the covariates across i .

Equation (17) has other advantages over just using the time-demeaned data in pooled OLS: time-constant variables can be included in (17), and the resulting equation gives a simple, robust way of testing whether the time-varying covariates are uncorrelated with It is helpful to write the original equation as

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + c_i + u_{it}, \quad t = 1, \dots, T, \quad (18)$$

where \mathbf{g}_t is typically a vector of time period dummies but could instead include other variables that change only over time, including linear or quadratic trends, \mathbf{z}_i is a vector of time-constant variables, and \mathbf{w}_{it} contains elements that vary across i and t . It is clear that, in comparing FE to RE estimation, $\boldsymbol{\gamma}$ can play no role because it cannot be estimated by FE. What is less clear, but also true, is that the coefficients on the aggregate time variables, $\boldsymbol{\eta}$, cannot be included in any comparison, either. Only the $M \times 1$ estimates of $\boldsymbol{\delta}$, say $\hat{\boldsymbol{\delta}}_{\text{FE}}$ and $\hat{\boldsymbol{\delta}}_{\text{RE}}$, can be compared. If $\hat{\boldsymbol{\eta}}_{\text{FE}}$ and $\hat{\boldsymbol{\eta}}_{\text{RE}}$ are included, the asymptotic variance matrix of the difference in estimators has a nonsingularity in the asymptotic variance matrix. (In fact, RE and FE estimation only with aggregate time variables are identical.) The Mundlak equation is now

$$y_{it} = \mathbf{g}_t\boldsymbol{\eta} + \mathbf{z}_i\boldsymbol{\gamma} + \mathbf{w}_{it}\boldsymbol{\delta} + \bar{\mathbf{w}}_i\boldsymbol{\xi} + a_i + u_{it}, \quad t = 1, \dots, T, \quad (19)$$

where the intercept is absorbed into \mathbf{g}_t . A test of the key RE assumption is $H_0: \xi = \mathbf{0}$ is obtained by estimating (19) by RE, and this equation makes it clear there M restrictions to test. This test was described in [5,46] proposed the robust version. The original test based directly on comparing the RE and FE estimators, as proposed in [25], it more difficult to compute and not robust because it maintains that the RE estimator is efficient under the null.

The model in (19) gives estimates of the coefficients on the time-constant variables \mathbf{z}_i . Generally, these can be given a causal interpretation only if

$$E(c_i|\mathbf{w}_i, \mathbf{z}_i) = E(c_i|\mathbf{w}_i) = \psi + \tilde{\mathbf{w}}_i \xi, \quad (20)$$

where the first equality is the important one. In other words, \mathbf{z}_i is uncorrelated with c_i once the time-varying covariates are controlled for. This assumption is too strong in many applications, but one still might want to include time-constant covariates.

Before leaving this subsection, it is worth point out that generalized least squares methods with an unrestricted variance-covariance matrix can be applied to every estimating equation just presented. For example, after eliminating c_i by removing the time averages, the resulting vector of errors, $\tilde{\mathbf{u}}_i$, can have an unrestricted variance matrix. (Of course, there is no guarantee that this matrix is the same as the variance matrix conditional on the matrix of time-demeaned regressors, $\tilde{\mathbf{X}}_i$.) The only glitch in practice is that $\text{Var}(\tilde{\mathbf{u}}_i)$ has rank $T - 1$, not T . As it turns out, GLS with an unrestricted variance matrix for the original error vector, \mathbf{u}_i , can be implemented on the time-demeaned equation with any of the T time periods dropped. The so-called *fixed effects GLS* estimates are invariant to whichever equation is dropped. See [41] or [37] for further discussion. The initial estimator used to estimate the variance covariance matrix would probably be the usual FE estimator (applied to all time periods).

Feasible GLS can be applied directly the first differenced equation, too. It can also be applied to (19), allowing the composite errors $a_i + u_{it}$, $t = 1, \dots, T$, to have an unrestricted variance-covariance matrix. In all cases, the assumption that the conditional variance matrix equals the unconditional variance can fail, and so one should use fully robust inference even after using FGLS. Chapter 10 in [55] provides further discussion. Such options are widely available in software, sometimes under the rubric of *generalized estimating equations* (GEE). See, for example, [43].

Models with Heterogeneous Slopes

The basic model described in the previous subsection introduces a single source of heterogeneity in the additive effect, c_i . The form of the model implies that the partial effects of the covariates depend on a fixed set of population values (and possibly other unobserved covariates if interactions are included in \mathbf{x}_{it}). It seems natural to extend the model to allow interactions between the observed covariates and time-constant, unobserved heterogeneity:

$$y_{it} = c_i + \mathbf{x}_{it} \mathbf{b}_i + u_{it} \quad (21)$$

$$E(u_{it}|\mathbf{x}_i, c_i, \mathbf{b}_i) = 0, \quad t = 1, \dots, T, \quad (22)$$

where \mathbf{b}_i is $K \times 1$. With small T , one cannot precisely estimate \mathbf{b}_i . Instead, attention usually focuses on the *average partial effect* (APE) or *population averaged effect* (PAE). In (21), the vector of APEs is $\boldsymbol{\beta} \equiv E(\mathbf{b}_i)$, the $K \times 1$ vector of means. In this formulation, aggregate time effects are in \mathbf{x}_{it} . This model is sometimes called a *correlated random slopes* model – which means the slopes are allowed to be correlated with the covariates.

Generally, allowing (c_i, \mathbf{b}_i) and \mathbf{x}_i to be arbitrarily correlated requires $T > K + 1$ – see [56]. With a small number of time periods and even a modest number of regressors, this condition often fails in practice. Chapter 11 in [55] discusses how to allow only a subset of coefficients to be unit specific. Of interest here is the question: if the usual FE estimator is applied – that is, ignoring the unit-specific slopes \mathbf{b}_i – does this ever consistently estimate the APEs in $\boldsymbol{\beta}$? In addition to the usual rank condition and the strict exogeneity assumption (22), [56] shows that a simple sufficient condition is

$$E(\mathbf{b}_i|\tilde{\mathbf{x}}_{it}) = E(\mathbf{b}_i) = \boldsymbol{\beta}, \quad t = 1, \dots, T. \quad (23)$$

Importantly, condition (23) allows the slopes, \mathbf{b}_i , to be correlated with the regressors \mathbf{x}_{it} through permanent components. It rules out correlation between idiosyncratic movements in \mathbf{x}_{it} and \mathbf{b}_i . For example, suppose the covariates can be decomposed as $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{r}_{it}$, $t = 1, \dots, T$. Then (23) holds if $E(\mathbf{b}_i|\mathbf{r}_{i1}, \mathbf{r}_{i2}, \dots, \mathbf{r}_{iT}) = E(\mathbf{b}_i)$. In other words, \mathbf{b}_i is allowed to be arbitrarily correlated with the permanent component, \mathbf{f}_i . Condition (23) is similar in spirit to the key assumption in [46] for the intercept c_i : the correlation between the slopes b_{ij} and the entire history $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ is through the time averages, and not through deviations from the time averages. If \mathbf{b}_i changes across i , ignoring it by using the usual FE estimator effectively puts $\tilde{\mathbf{x}}_{it}(\mathbf{b}_i - \boldsymbol{\beta})$ in the error term, which induces heteroskedasticity and serial correlation in the composite error even if the $\{u_{it}\}$ are homoskedastic and serially

independent. The possible presence of this term provides another argument for making inference with FE fully robust to arbitrary conditional and unconditional second moments.

The (partial) robustness of FE to the presence of correlated random slopes extends to a more general class of estimators that includes the usual fixed effects estimator. Write an extension of the basic model as

$$y_{it} = \mathbf{g}_t \mathbf{a}_i + \mathbf{x}_{it} \mathbf{b}_i + u_{it}, \quad t = 1, \dots, T, \quad (24)$$

where \mathbf{g}_t is a set of deterministic functions of time. A leading case is $\mathbf{g}_t = (1, t)$, so that each unit has its own time trend along with a level effect. (The resulting model is sometimes called a *random trend model*). Now, assume that the random coefficients, \mathbf{a}_i , are swept away by regressing y_{it} and \mathbf{x}_{it} each on \mathbf{g}_t for each i . The residuals, \tilde{y}_{it} and $\tilde{\mathbf{x}}_{it}$, have had unit-specific trends removed, but the \mathbf{b}_i are treated as constant in the estimation. The key condition for consistently estimating $\boldsymbol{\beta}$ can still be written as in (23), but now $\tilde{\mathbf{x}}_{it}$ has had more features removed at unit-specific level. When $\mathbf{g}_t = (1, t)$, each covariate has been demeaned within each unit. Therefore, if $\mathbf{x}_{it} = \mathbf{f}_i + \mathbf{h}_i t + \mathbf{r}_{it}$, then \mathbf{b}_i can be arbitrarily correlated with $(\mathbf{f}_i, \mathbf{h}_i)$. Of course, individually detrending the \mathbf{x}_{it} requires at least three time periods, and it decreases the variation in $\tilde{\mathbf{x}}_{it}$ compared with the usual FE estimator. Not surprisingly, increasing the dimension of \mathbf{g}_t (subject to the restriction $\dim(\mathbf{g}_t) < T$), generally leads to less precision of the estimator. See [56] for further discussion.

Sequentially Exogenous Regressors and Dynamic Models

The summary of models and estimators from Sect. “Overview of Linear Panel Data Models” used the strict exogeneity assumption $E(u_{it} | \mathbf{x}_i, c_i) = 0$ for all t , and added an additional assumption for models with correlated random slopes. As discussed in Sect. “Overview of Linear Panel Data Models”, strict exogeneity is not an especially natural assumption. The contemporaneous exogeneity assumption $E(u_{it} | \mathbf{x}_{it}, c_i) = 0$ is attractive, but the parameters are not identified. In this section, a middle ground between these assumptions, which has been called a *sequential exogeneity assumption*, is used. But first, it is helpful to understand properties of the FE and FD estimators when strict exogeneity fails.

Behavior of Estimators Without Strict Exogeneity

Both the FE and FD estimators are inconsistent (with fixed T , $N \rightarrow \infty$) without the strict exogeneity assumption stated in Eq. (4). But it is also pretty well known that,

at least under certain assumptions, the FE estimator can be expected to have less “bias” for larger T . Under the contemporaneous exogeneity assumption (2) and the assumption that the data series $\{(\mathbf{x}_{it}, u_{it}) : t = 1, \dots, T\}$ is “weakly dependent” – in time series parlance, “integrated of order zero”, or $I(0)$ – then it can be shown that

$$\text{plim } \hat{\boldsymbol{\beta}}_{\text{FE}} = \boldsymbol{\beta} + O(T^{-1}) \quad (25)$$

$$\text{plim } \hat{\boldsymbol{\beta}}_{\text{FD}} = \boldsymbol{\beta} + O(1); \quad (26)$$

see Chap. 11 in [55]. In some very special cases, such as the simple AR(1) model discussed below, the “bias” terms can be calculated, but not generally.

Interestingly, the same results can be shown if $\{\mathbf{x}_{it} : t = 1, \dots, T\}$ has unit roots as long as $\{u_{it}\}$ is $I(0)$ and contemporaneous exogeneity holds. However, there is a catch: if $\{u_{it}\}$ is $I(1)$ – so that the time series version of the “model” would be a spurious regression (y_{it} and \mathbf{x}_{it} are not “cointegrated”), then (25) is no longer true. On the other hand, first differencing means any unit roots are eliminated and so there is little possibility of a spurious regression. The bottom line is that using “large T ” approximations such as those in (25) and (26) to choose between FE over FD obligates one to take the time series properties of the panel data seriously; one must recognize the possibility that the FE estimation is essentially a spurious regression.

Consistent Estimation Under Sequential Exogeneity

Because both the FE and FD estimators are inconsistent for fixed T , it makes sense to search for estimators that are consistent for fixed T . A natural specification for dynamic panel data models, and one that allows consistent estimation under certain assumptions, is

$$E(y_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{it}, c_i) = E(y_{it} | \mathbf{x}_{it}, c_i) = \eta_t + \mathbf{x}_{it} \boldsymbol{\beta} + c_i, \quad (27)$$

which says that \mathbf{x}_{it} contains enough lags so that further lags of variables are not needed. When the model is written in error form, (27) is the same as

$$E(u_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{it}, c_i) = 0, \quad t = 1, 2, \dots, T. \quad (28)$$

Under (28), the covariates $\{\mathbf{x}_{it} : t = 1, \dots, T\}$ are said to be *sequentially exogenous conditional on c_i* . Some estimation methods are motivated by a weaker version of (28), namely,

$$E(\mathbf{x}'_{is} u_{it}) = 0, \quad s = 1, \dots, t, \quad t = 1, \dots, T, \quad (29)$$

but (28) is natural in most applications.

Assumption (28) is appealing in that it allows for finite distributed lag models as well as models with lagged dependent variables. For example, the finite distributed lag model

$$y_{it} = \eta_t + \mathbf{z}_{it}\delta_0 + \mathbf{z}_{i,t-1}\delta_1 + \cdots + \mathbf{z}_{i,t-L}\delta_L + c_i + u_{it} \quad (30)$$

allows the elements of \mathbf{z}_{it} to have effects up to L time periods after a change. With $\mathbf{x}_{it} = (\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \dots, \mathbf{z}_{i,t-L})$, Assumption (28) implies

$$\begin{aligned} E(y_{it} | \mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, \dots, c_i) \\ = E(y_{it} | \mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \mathbf{z}_{i,t-2}, c_i) \\ = \eta_t + \mathbf{z}_{it}\delta_0 + \mathbf{z}_{i,t-1}\delta_1 + \cdots + \mathbf{z}_{i,t-L}\delta_L + c_i, \end{aligned} \quad (31)$$

which means that the distributed lag dynamics are captured by L lags. The important difference with the strict exogeneity assumption is that sequential exogeneity allows feedback from u_{it} to \mathbf{z}_{ir} for $r > t$.

How can (28) be used for estimation? The FD transformation is natural because of the sequential nature of the restrictions. In particular, write the FD equation as

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\beta + \Delta u_{it}, \quad t = 2, \dots, T. \quad (32)$$

Then, under (29),

$$\begin{aligned} E(\mathbf{x}'_{is} \Delta u_{it}) = 0, \quad s = 1, \dots, t-1; \\ t = 2, \dots, T, \end{aligned} \quad (33)$$

which means any \mathbf{x}_{is} with $s < t$ can be used as an instrument for the time t FD equation. An efficient estimator that uses (33) is obtained by stacking the FD equations as

$$\Delta \mathbf{y}_i = \Delta \mathbf{X}_i \beta + \Delta \mathbf{u}_i, \quad (34)$$

where $\Delta \mathbf{y}_i = (\Delta y_{i2}, \Delta y_{i3}, \dots, \Delta y_{iT})'$ is the $(T-1) \times 1$ vector of first differences and $\Delta \mathbf{X}_i$ is the $(T-1) \times K$ matrix of differences on the regressors. (Time period dummies are absorbed into \mathbf{x}_{it} for notational simplicity.) To apply a system estimation method to (34), define

$$\mathbf{x}_{it}^o \equiv (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \dots, \mathbf{x}_{it}), \quad (35)$$

which means the valid instruments at time t are in $\mathbf{x}_{i,t-1}^o$ (minus redundancies, of course). The matrix of instruments to apply to (34) is

$$\mathbf{W}_i = \text{diag}(\mathbf{x}_{i1}^o, \mathbf{x}_{i2}^o, \dots, \mathbf{x}_{i,T-1}^o), \quad (36)$$

which has $T-1$ rows and a large number of columns. Because of sequential exogeneity, the number of valid instruments increases with t .

Given \mathbf{W}_i , it is routine to apply generalized method of moments estimation, as summarized in [27,55]. A simpler strategy is available that can be used for comparison or as the first-stage estimator in computing the optimal weighting matrix. First, estimate a reduced form for $\Delta \mathbf{x}_{it}$ separately for each t . In other words, at time t , run the regression $\Delta \mathbf{x}_{it}$ on $\mathbf{x}_{i,t-1}^o$, $i = 1, \dots, N$, and obtain the fitted values, $\widehat{\Delta \mathbf{x}_{it}}$. Of course, the fitted values are all $1 \times K$ vectors for each t , even though the number of available instruments grows with t . Then, estimate the FD Eq. (32) by pooled IV using $\widehat{\Delta \mathbf{x}_{it}}$ as instruments (not regressors). It is simple to obtain robust standard errors and test statistics from such a procedure because the first stage estimation to obtain the instruments can be ignored (asymptotically, of course).

One potential problem with estimating the FD equation using IVs that are simply lags of \mathbf{x}_{it} is that changes in variables over time are often difficult to predict. In other words, $\Delta \mathbf{x}_{it}$ might have little correlation with $\mathbf{x}_{i,t-1}^o$. This is an example of the so-called “weak instruments” problem, which can cause the statistical properties of the IV estimators to be poor and the usual asymptotic inference misleading. Identification is lost entirely if $\mathbf{x}_{it} = \boldsymbol{\lambda}_t + \mathbf{x}_{i,t-1} + \mathbf{q}_{it}$, where $E(\mathbf{q}_{it} | \mathbf{x}_{i,t-1}, \dots, \mathbf{x}_{i1}) = \mathbf{0}$ – that is, the elements of \mathbf{x}_{it} are random walks with drift. Then, then $E(\Delta \mathbf{x}_{it} | \mathbf{x}_{i,t-1}, \dots, \mathbf{x}_{i1}) = \mathbf{0}$, and the rank condition for IV estimation fails. Of course, if some elements of \mathbf{x}_{it} are strictly exogenous, then their changes act as their own instruments. Nevertheless, typically at least one element of \mathbf{x}_{it} is suspected of failing strict exogeneity, otherwise standard FE or FD would be used.

In situations where simple estimators that impose few assumptions are too imprecise to be useful, sometimes one is willing to improve estimation of β by adding more assumptions. How can this be done in the panel data case under sequential exogeneity? There are two common approaches. First, the sequential exogeneity condition can be strengthened to the assumption that the conditional mean model is *dynamically complete*, which can be written in terms of the errors as

$$\begin{aligned} E(u_{it} | \mathbf{x}_{it}, y_{i,t-1}, \mathbf{x}_{i,t-1}, \dots, y_{i1}, \mathbf{x}_{i1}, c_i) = 0, \\ t = 1, \dots, T. \end{aligned} \quad (37)$$

Clearly, (37) implies (28). Dynamic completeness is neither stronger nor weaker than strict exogeneity, because the latter includes the entire history of the covariates while (37) conditions only on current and past \mathbf{x}_{it} . Dynamic completeness is natural when \mathbf{x}_{it} contains lagged dependent variables, because it basically means enough lags have been included to capture all of the dynamics. It

is often too restrictive in finite distributed lag models such as (30), where (37) would imply

$$\begin{aligned} E(y_{it} | \mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}, \dots, y_{i1}, \mathbf{z}_{i1}, c_i) \\ = E(y_{it} | \mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \dots, \mathbf{z}_{i-L}, c_i), \quad t = 1, \dots, T, \end{aligned} \quad (38)$$

which puts strong restrictions on the fully dynamic conditional mean: values y_{it} , $r \leq t-1$, do not help to predict y_{it} once $(\mathbf{z}_{it}, \mathbf{z}_{i,t-1}, \dots)$ are controlled for. FDLs are of interest even if (38) does not hold. Imposing (37) in FDLs implies that the idiosyncratic errors must be serially uncorrelated, something that is often violated in FDLs.

Dynamic completeness is natural in a model such as

$$y_{it} = \rho y_{i,t-1} + \mathbf{z}_{it} \delta_0 + \mathbf{z}_{i,t-1} \delta_1 + c_i + u_{it}. \quad (39)$$

Usually – although there are exceptions – (39) is supposed to represent the conditional mean $E(y_{it} | \mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}, \dots, y_{i1}, \mathbf{z}_{i1}, c_i)$, and then the issue is whether one lag of y_{it} and \mathbf{z}_{it} suffice to capture the dynamics.

Regardless of what is contained in \mathbf{x}_{it} , assumption (37) implies some additional moment conditions that can be used to estimate β . The extra moment conditions, first proposed in [1] in the context of the AR(1) unobserved effects model, can be written as

$$\begin{aligned} E[(\Delta y_{i,t-1} - \Delta \mathbf{x}_{i,t-1} \beta)'(y_{it} - \mathbf{x}_{it} \beta)] = 0, \\ t = 3, \dots, T; \end{aligned} \quad (40)$$

see also [9]. The conditions can be used in conjunction with those in Eq. (33) in a method of moments estimation method. In addition to imposing dynamic completeness, the moment conditions in (40) are nonlinear in parameters, which makes them more difficult to implement than just using (33). Nevertheless, the simulation evidence in [1] for the AR(1) model shows that (40) can help considerably when the coefficient ρ is large.

[7] suggested a different set of restrictions,

$$\text{Cov}(\Delta \mathbf{x}'_{it}, c_i) = 0, \quad t = 2, \dots, T. \quad (41)$$

Interestingly, this assumption is very similar in spirit to assumption (23), except that it is in terms of the first difference of the covariates, not the time-demeaned covariates. Condition (41) generates moment conditions in the levels of equation,

$$E[\Delta \mathbf{x}'_{it}(y_{it} - \alpha - \mathbf{x}_{it} \beta)] = 0, \quad t = 2, \dots, T, \quad (42)$$

where α allows for a nonzero mean for c_i . [10] applies these moment conditions, along with the usual conditions

in (33), to estimate firm-level production functions. Because of persistence in the data, they find the moments in (33) are not especially informative for estimating the parameters, whereas (42) along with (33) are. Of course, (42) is an extra set of assumptions.

The previous discussion can be applied to the AR(1) model, which has received much attention. In its simplest form the model is

$$y_{it} = \rho y_{i,t-1} + c_i + u_{it}, \quad t = 1, \dots, T, \quad (43)$$

so that, by convention, the first observation on y is at $t = 0$. The minimal assumptions imposed are

$$E(y_{is} u_{it}) = 0, \quad s = 0, \dots, t-1, \quad t = 1, \dots, T, \quad (44)$$

in which case the available instruments at time t are $\mathbf{w}_{it} = (y_{i0}, \dots, y_{i,t-2})$ in the FD equation

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta u_{it}, \quad t = 2, \dots, T. \quad (45)$$

Written in terms of the parameters and observed data, the moment conditions are

$$\begin{aligned} E[y_{is}(\Delta y_{it} - \rho \Delta y_{i,t-1})] = 0, \\ s = 0, \dots, t-2, \quad t = 2, \dots, T. \end{aligned} \quad (46)$$

[4] proposed pooled IV estimation of the FD equation with the single instrument $y_{i,t-2}$ (in which case all $T-1$ periods can be used) or $\Delta y_{i,t-2}$ (in which case only $T-2$ periods can be used). A better approach is pooled IV where $T-1$ separate reduced forms are estimated for $\Delta y_{i,t-1}$ as a linear function of $(y_{i0}, \dots, y_{i,t-2})$. The fitted values $\hat{\Delta y}_{i,t-1}$, can be used as the instruments in (45) in a pooled IV estimation. Of course, standard errors and inference should be made robust to the MA(1) serial correlation in Δu_{it} . [6] suggested full GMM estimation using all of the available instruments $(y_{i0}, \dots, y_{i,t-2})$, and this estimator uses the conditions in (44) efficiently.

Under the dynamic completeness assumption

$$E(u_{it} | y_{i,t-1}, y_{i,t-2}, \dots, y_{i0}, c_i) = 0, \quad (47)$$

the extra moment conditions in [1] become

$$\begin{aligned} E[(\Delta y_{i,t-1} - \rho \Delta y_{i,t-2})(y_{it} - \rho y_{i,t-1})] = 0, \\ t = 3, \dots, T. \end{aligned} \quad (48)$$

[10] noted that if the condition

$$\text{Cov}(\Delta y_{i1}, c_i) = \text{Cov}(y_{i1} - y_{i0}, c_i) = 0 \quad (49)$$

is added to (47) then the combined set of moment conditions becomes

$$E[\Delta y_{i,t-1}(y_{it} - \alpha - \rho y_{i,t-1})] = 0, \quad t = 2, \dots, T, \quad (50)$$

which can be added to the usual moment conditions (46). Conditions (46) and (50) combined are attractive because they are linear in the parameters, and they can produce much more precise estimates than just using (46).

As discussed in [10], condition (49) can be interpreted as a restriction on the initial condition, y_{i0} , and the steady state. When $|\rho| < 1$, the steady state of the process is $c_i/(1 - \rho)$. Then, it can be shown that (49) holds if the deviation of y_{i0} from its steady state is uncorrelated with c_i . Statistically, this condition becomes more useful as ρ approaches one, but this is when the existence of a steady state is most in doubt. [22] shows theoretically that such restrictions can greatly increase the information about ρ .

Other approaches to dynamic models are based on maximum likelihood estimation. Approaches that condition on the initial condition y_{i0} , suggested by [10,13,15], seem especially attractive. Under normality assumptions, maximum likelihood conditional on y_{i0} is tractable.

If some strictly exogenous variables are added to the AR(1) model, then it is easiest to use IV methods on the FD equation, namely,

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta \mathbf{z}_{it} \boldsymbol{\gamma} + \Delta u_{it}, \quad t = 1, \dots, T. \quad (51)$$

The available instruments (in addition to time period dummies) are $(\mathbf{z}_i, y_{i,t-2}, \dots, y_{i0})$, and the extra conditions (42) can be used, too. If sequentially exogenous variables, say \mathbf{h}_{it} , are added, then $(\mathbf{h}_{i,t-1}, \dots, \mathbf{h}_{i1})$ would be added to the list of instruments (and $\Delta \mathbf{h}_{it}$ would appear in the equation).

Unbalanced Panel Data Sets

The previous sections considered estimation of models using balanced panel data sets, where each unit is observed in each time period. Often, especially with data at the individual, family, or firm level, data are missing in some time periods – that is, the panel data set is *unbalanced*. Standard methods, such as fixed effects, can often be applied to produce consistent estimators, and most software packages that have built-in panel data routines typically allow unbalanced panels. However, determining whether applying standard methods to the unbalanced panel produces consistent estimators requires knowing something about the mechanism generating the missing data.

Methods based on removing the unobserved effect warrant special attention, as they allow some nonrandomness in the sample selection. Let $t = 1, \dots, T$ denote the time periods for which data can exist for each unit from the population, and again consider the model

$$y_{it} = \eta_t + \mathbf{x}_{it} \boldsymbol{\beta} + c_i + u_{it}, \quad t = 1, \dots, T. \quad (52)$$

It is helpful to have, for each i and t , a binary selection variable, s_{it} , equal to one of the data for unit i in time t can be used, and zero otherwise. For concreteness, consider the case where time averages are removed to eliminate c_i , but where the averages necessarily only include the $s_{it} = 1$ observations. Let $\tilde{y}_{it} = y_{it} - T_i^{-1} \sum_{r=1}^T s_{ir} y_{ir}$ and $\tilde{\mathbf{x}}_{it} = \mathbf{x}_{it} - T_i^{-1} \sum_{r=1}^T s_{ir} \mathbf{x}_{ir}$ be the time-demeaned quantities using the observed time periods for unit i , where $T_i = \sum_{t=1}^T s_{it}$ is the number of time periods observed for unit i – properly viewed as a random variable. The fixed effects estimator on the unbalanced panel can be expressed as

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{FE}} &= \left(N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right)^{-1} \\ &\quad \cdot \left(N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} \tilde{\mathbf{x}}_{it}' \tilde{y}_{it} \right) \\ &= \boldsymbol{\beta} + \left(N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} \tilde{\mathbf{x}}_{it}' \tilde{\mathbf{x}}_{it} \right)^{-1} \\ &\quad \cdot \left(N^{-1} \sum_{i=1}^N \sum_{t=1}^T s_{it} \tilde{\mathbf{x}}_{it}' u_{it} \right). \end{aligned} \quad (53)$$

With fixed T and $N \rightarrow \infty$ asymptotics, the key condition for consistency is

$$\sum_{t=1}^T E(s_{it} \tilde{\mathbf{x}}_{it}' u_{it}) = \mathbf{0}. \quad (54)$$

In evaluating (54), it is important to remember that $\tilde{\mathbf{x}}_{it}$ depends on $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, s_{i1}, \dots, s_{iT})$, and in a nonlinear way. Therefore, it is not sufficient to assume $(\mathbf{x}_{ir}, s_{ir})$ are uncorrelated with u_{it} for all r and t . A condition that is sufficient for (54) is

$$E(u_{it} | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, s_{i1}, \dots, s_{iT}, c_i) = 0, \quad t = 1, \dots, T. \quad (55)$$

Importantly, (55) allows arbitrary correlation between the heterogeneity, c_i , and selection, s_{it} , in any time period t . In other words, some units are allowed to be more likely to

be in or out of the sample in any time period, and these probabilities can change across t . But (55) rules out some important kinds of sample selection. For example, selection at time t , s_{it} , cannot be correlated with the idiosyncratic error at time t , u_{it} . Further, feedback is not allowed: in affect, like the covariates, selection must be strictly exogenous conditional on c_i .

Testing for no feedback into selection is easy in the context of FE estimation. Under (55), $s_{i,t+1}$ and u_{it} should be uncorrelated. Therefore, $s_{i,t+1}$ can be added to the FE estimation on the unbalanced panel – where the last time period is lost for all observations – and a t test can be used to determine significance. A rejection means (55) is false. Because serial correlation and heteroskedasticity are always a possibility, the t test should be made fully robust.

Contemporaneous selection bias – that is, correlation between s_{it} and u_{it} – is more difficult to test. Chapter 17 in [55] summarizes how to derive tests and corrections by extending the corrections in [28] (so-called “Heckman corrections”) to panel data.

First differencing can be used on unbalanced panels, too, although straight first differencing can result in many lost observations: a time period is used only if it is observed along with the previous or next time period. FD is more useful in the case of attrition in panel data, where a unit is observed until it drops out of the sample and never reappears. Then, if a data point is observed at time t , it is also observed at time $t - 1$. Differencing can be combined with the approach in [28] to solve bias due to attrition – at least under certain assumptions. See Chap. 17 in [55].

Random effects methods can also be applied with unbalanced panels, but the assumptions under which the RE estimator is consistent are stronger than for FE. In addition to (55), one must assume selection is unrelated to c_i . A natural assumption, that also imposes exogeneity on the covariates with respect to c_i , is

$$E(c_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, s_{i1}, \dots, s_{iT}) = E(c_i). \quad (56)$$

The only case beside randomly determined sample selection where (56) holds is when s_{it} is essentially a function of the observed covariates. Even in this case, (56) requires that the unobserved heterogeneity is mean independent of the observed covariates – as in the typical RE analysis on balanced panel.

Nonlinear Models

Nonlinear panel data models are considerably more difficult to interpret and estimate than linear models. Key issues concern how the unobserved heterogeneity appears

in the model and how one accounts for that heterogeneity in summarizing the effects of the explanatory variables on the response. Also, in some cases, conditional independence of the response is used to identify certain parameters and quantities.

Basic Issues and Quantities of Interest

As in the linear case, the setup here is best suited for situations with small T and large N . In particular, the asymptotic analysis underlying the discussion of estimation is with fixed T and $N \rightarrow \infty$. Sampling is assumed to be random from the population. Unbalanced panels are generally difficult to deal with because, except in special cases, the unobserved heterogeneity cannot be completely eliminated in obtaining estimating equations. Consequently, methods that model the conditional distribution of the heterogeneity conditional on the entire history of the covariates – as we saw with the Chamberlain–Mundlak approach – are relied on heavily, and such approaches are difficult when data are missing on the covariates for some time periods. Therefore, this section considers only balanced panels. The discussion here takes the response variable, y_{it} , as a scalar for simplicity.

The starting point for nonlinear panel data models with unobserved heterogeneity is the conditional distribution

$$D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i), \quad (57)$$

where \mathbf{c}_i is the unobserved heterogeneity for observation i drawn along with the observables. Often there is a particular feature of this distribution, such as $E(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i)$, or a conditional median, that is of primary interest. Even focusing on the conditional mean raises some tricky issues in models where \mathbf{c}_i does not appear in an additive or linear form. To be precise, let $E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t, \mathbf{c})$ be the mean function. If x_{ij} is continuous, then the partial effect can be defined as

$$\theta_j(\mathbf{x}_t, \mathbf{c}) \equiv \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}}. \quad (58)$$

For discrete (or continuous) variables, (58) can be replaced with discrete changes. Either way, a key question is: How can one account for the unobserved \mathbf{c} in (58)? In order to estimate magnitudes of effects, sensible values of \mathbf{c} need to be plugged into (58), which means knowledge of at least some distributional features of \mathbf{c}_i is needed. For example, suppose $\boldsymbol{\mu}_c = E(\mathbf{c}_i)$ is identified. Then the *partial effect at the average* (PEA),

$$\theta_j(\mathbf{x}_t, \boldsymbol{\mu}_c), \quad (59)$$

can be identified if the regression function m_t is identified. Given more information about the distribution of \mathbf{c}_i , different quantiles can be inserted into (59), or a certain number of standard deviations from the mean.

An alternative to plugging in specific values for \mathbf{c} is to average the partial effects across the distribution of \mathbf{c}_i :

$$\text{APE}(\mathbf{x}_t) = E_{\mathbf{c}_i}[\theta_j(\mathbf{x}_t, \mathbf{c}_i)], \quad (60)$$

the so-called *average partial effect* (APE). The difference between (59) and (60) can be nontrivial for nonlinear mean functions. The definition in (60) dates back at least to [17], and is closely related to the notion of the *average structural function* (ASF), as introduced in [12]. The ASF is defined as

$$\text{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i)]. \quad (61)$$

Assuming the derivative passes through the expectation results in (60); computing a discrete change in the ASF always gives the corresponding APE. A useful feature of APEs is that they can be compared across models, where the functional form of the mean or the distribution of the heterogeneity can be different. In particular, APEs in general nonlinear models are comparable to the estimated coefficients in a standard linear model.

Average partial effects are not always identified, even when parameters are. Semi-parametric panel data methods that are silent about the distribution of \mathbf{c}_i , unconditionally or conditional on $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$, cannot generally deliver estimates of APEs, essentially by design. Instead, an index structure is usually imposed so that parameters can be consistently estimated. A common setup with scalar heterogeneity is

$$m_t(\mathbf{x}_t, c) = G(\mathbf{x}_t\boldsymbol{\beta} + c), \quad (62)$$

where, say, $G(\cdot)$ is strictly increasing and continuously differentiable. The partial effects are proportional to the parameters:

$$\theta_j(\mathbf{x}_t, c) = \beta_j g(\mathbf{x}_t\boldsymbol{\beta} + c), \quad (63)$$

where $g(\cdot)$ is the derivative of $G(\cdot)$. Therefore, if β_j is identified, then so is the sign of the partial effect, and even the relative effects of any two continuous variables: the ratio of partial effects for x_{tj} and x_{th} is β_j/β_h . However, even if $G(\cdot)$ is specified (the common case), the magnitude of the effect evidently cannot be estimated without making assumptions about the distribution of c_i ; otherwise, the term $E[g(\mathbf{x}_t\boldsymbol{\beta} + c_i)]$ cannot generally be estimated. The probit example below shows how the APEs can be estimated in index models under distributional assumptions for c_i .

The previous discussion holds regardless of the exogeneity assumptions on the covariates. For example, the definition of the APE for a continuous variable holds whether \mathbf{x}_t contains lagged dependent variables or only contemporaneous variables. However, approaches for estimating the parameters and the APEs depend critically on exogeneity assumptions.

Exogeneity Assumptions on the Covariates

As in the case of linear models, it is not nearly enough to simply specify a model for the conditional distribution of interest, $D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, or some feature of it, in order to estimate parameters and partial effects. This section offers two exogeneity assumptions on the covariates that are more restrictive versions of the linear model assumptions.

It is easiest to deal with estimation under a strict exogeneity assumption. The most useful definition of strict exogeneity for nonlinear panel data models is

$$D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \quad (64)$$

which means that \mathbf{x}_{ir} , $r \neq t$, does not appear in the conditional distribution of y_{it} once \mathbf{x}_{it} and \mathbf{c}_i have been counted for. [17] labeled (64) *strict exogeneity conditional on the unobserved effects* \mathbf{c}_i . Sometimes, a conditional mean version is sufficient:

$$E(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), \quad (65)$$

which already played a role in linear models. Assumption (64), or its conditional mean version, are less restrictive than if \mathbf{c}_i is not in the conditioning set, as discussed in [17]. Indeed, it is easy to see that, if (64) holds and $D(\mathbf{c}_i|\mathbf{x}_i)$ depends on \mathbf{x}_i , then strict exogeneity without conditioning on \mathbf{c}_i , $D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = D(y_{it}|\mathbf{x}_{it})$, cannot hold. Unfortunately, both (64) and (65) rule out lagged dependent variables, as well as other situations where there may be feedback from idiosyncratic changes in y_{it} to future movements in \mathbf{x}_{ir} , $r > t$. Nevertheless, the conditional strict exogeneity assumption underlies the most common estimation methods for nonlinear models.

More natural is *sequential exogeneity conditional on the unobserved effects*, which, in terms of conditional distributions, is

$$D(y_{it}|\mathbf{x}_{i1}, \dots, \mathbf{x}_{it}, \mathbf{c}_i) = D(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i). \quad (66)$$

Assumption (66) allows for lagged dependent variables and does not restrict feedback. Unfortunately, (66) is substantially more difficult to work with than (64) for general nonlinear models.

Because \mathbf{x}_{it} is conditioned on, neither (64) nor (66) allows for contemporaneous endogeneity of \mathbf{x}_{it} as would arise with measurement error, time-varying omitted variables, or simultaneous equations. This chapter does not treat such cases. See [38] for a recent summary.

Conditional Independence Assumption

The exogeneity conditions stated in Subsect. “Exogeneity Assumptions on the Covariates” generally do not restrict the dependence in the responses, $\{y_{it}: t = 1, \dots, T\}$. Often, a *conditional independence* assumption is explicitly imposed, which can be written generally as

$$D(y_{i1}, \dots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^T D(y_{it} | \mathbf{x}_i, \mathbf{c}_i). \quad (67)$$

Equation (67) simply means that, conditional on the entire history $\{\mathbf{x}_{it}: t = 1, \dots, T\}$ and the unobserved heterogeneity \mathbf{c}_i , the responses are independent across time. One way to think about (67) is that time-varying unobservables are independent over time. Because (67) conditions on \mathbf{x}_i , it is useful only in the context of the strict exogeneity assumption (64). Then, conditional independence can be written as

$$D(y_{i1}, \dots, y_{iT} | \mathbf{x}_i, \mathbf{c}_i) = \prod_{t=1}^T D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i). \quad (68)$$

Therefore, under strict exogeneity and conditional independence, the panel data modeling exercise reduces to specifying a model for $D(y_{it} | \mathbf{x}_{it}, \mathbf{c}_i)$, and then determining how to treat the unobserved heterogeneity, \mathbf{c}_i . In random effects and correlated RE frameworks, conditional independence can play a critical role in being able to estimate the parameters and the distribution of \mathbf{c}_i . As it turns out, conditional independence plays no role in estimating APEs for a broad class of models. Before explaining how that works, the key issue of dependence between the heterogeneity and covariates needs to be addressed.

Assumptions About the Unobserved Heterogeneity

For general nonlinear models, the *random effects assumption* is independence between \mathbf{c}_i and $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$:

$$D(\mathbf{c}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT}) = D(\mathbf{c}_i). \quad (69)$$

Assumption (69) is very strong. To illustrate how strong it is, suppose that (69) is combined with a model for the conditional mean, $E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t, \mathbf{c}_i = \mathbf{c}) = m_t(\mathbf{x}_t, \mathbf{c})$. Without any additional assumptions, the average partial effects

are nonparametrically identified. In particular, the APEs can be obtained directly from the conditional mean

$$r_t(\mathbf{x}_t) \equiv E(y_{it} | \mathbf{x}_{it} = \mathbf{x}_t). \quad (70)$$

(The argument is a simple application of the law of iterated expectations; it is discussed in [56]). Nevertheless, (69) is still common in many applications, especially when the explanatory variables of interest do not change over time.

As in the linear case, a *correlated random effects* (CRE) framework allows dependence between \mathbf{c}_i and \mathbf{x}_i , but the dependence is restricted in some way. In a parametric setting, a CRE approach involves specifying a distribution for $D(\mathbf{c}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$, as in [15,17,46], and many subsequent authors; see, for example, [55] and [14]. For many models – see, for example, Subsect. “Binary Response Models” – one can allow $D(\mathbf{c}_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ to depend on $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ in a “nonexchangeable” manner, that is, the distribution need not be symmetric on its conditioning arguments. However, allowing nonexchangeability usually comes at the expense of potentially restrictive distributional assumptions, such as homoskedastic normal with a linear conditional mean. For estimating APEs, it is sufficient to assume, along with strict exogeneity,

$$D(\mathbf{c}_i | \mathbf{x}_i) = D(\mathbf{c}_i | \bar{\mathbf{x}}_i), \quad (71)$$

without specifying $D(\mathbf{c}_i | \bar{\mathbf{x}}_i)$ or restricting any feature of this distribution. (See, for example, [3,56].) As a practical matter, it makes sense to adopt (71) – or perhaps allow other features of $\{\mathbf{x}_{it}: t = 1, \dots, T\}$ – in a flexible parametric analysis.

Condition (71) still imposes restrictions on $D(\mathbf{c}_i | \mathbf{x}_i)$. Ideally, as in the linear model, one could estimate at least some features of interest without making any assumption about $D(\mathbf{c}_i | \mathbf{x}_i)$. Unfortunately, the scope for allowing unrestricted $D(\mathbf{c}_i | \mathbf{x}_i)$ is limited to special nonlinear models, at least with small T . Allowing $D(\mathbf{c}_i | \mathbf{x}_i)$ to be unspecified is the hallmark of a “fixed effects” analysis, but the label has not been used consistently. Often, fixed effects has been used to describe a situation where the \mathbf{c}_i are treated as parameters to be estimated, along with parameters that do not vary across i . Except in special cases or with large T , estimating the unobserved heterogeneity is prone to an *incidental parameters problem*. Namely, using a fixed T , $N \rightarrow \infty$ framework, one cannot get consistent estimators of the \mathbf{c}_i , and the inconsistency in, say, $\hat{\mathbf{c}}_i$, generally transmits itself to the parameters that do not vary with i . The incidental parameters problem does not arise in estimating the coefficients β in a linear model because the estimator obtained by treating the \mathbf{c}_i as parameters to estimate is equivalent to pooled OLS on the time-demeaned data –

that is, the fixed effects estimator can be obtained by eliminating the c_i using the within transformation or estimating the c_i along with β . This occurrence is rare in nonlinear models. Section “Future Directions” further discusses this issue, as there is much ongoing research that attempts to reduce the asymptotic bias in nonlinear models.

The “fixed effects” label has also been applied to settings where the c_i are not treated as parameters to estimate; rather, the c_i can be eliminated by conditioning on a sufficient statistic. Let w_i be a function of the observed data, (x_i, y_i) , such that

$$D(y_{i1}, \dots, y_{iT} | x_i, c_i, w_i) = D(y_{i1}, \dots, y_{iT} | x_i, w_i). \quad (72)$$

Then, provided the latter conditional distribution depends on the parameters of interest, and can be derived or approximated from the original specification of $D(y_{i1}, \dots, y_{iT} | x_i, c_i)$, maximum likelihood methods can be used. Such an approach is also called *conditional maximum likelihood estimation* (CMLE), where “conditional” refers to conditioning on a function of y_i . (In traditional treatments of MLE, conditioning on so-called “exogenous” variables is usually implicit.) In most cases where the CMLE approach applies, the conditional independence assumption (67) is maintained, although one conditional MLE is known to have robustness properties: the so-called “fixed effects” Poisson estimator (see [53]).

Maximum Likelihood Estimation and Partial MLE

There are two common approaches to estimating the parameters in nonlinear, unobserved effects panel data models when the explanatory variables are strictly exogenous. (A third approach, generalized method of moments, is available in special cases but is not treated here. See, for example, Chap. 19 in [55].) The first approach is full maximum likelihood (conditional on the entire history of covariates). Most commonly, full MLE is applied under the conditional independence assumption, although sometimes models are used that explicitly allow dependence in $D(y_{i1}, \dots, y_{iT} | x_i, c_i)$. Assuming strict exogeneity, conditional independence, a model for the density of y_{it} given (x_{it}, c_i) (say, $f_t(y_{it} | x_{it}, c; \theta)$), and a model for the density of c_i given x_i (say, $h(c | x; \delta)$), the log likelihood for random draw i from the cross section is

$$\log \left\{ \left[\prod_{t=1}^T f_t(y_{it} | x_{it}, c; \theta) \right] h(c | x_i; \delta) \right\}. \quad (73)$$

This log-likelihood function “integrates out” the unobserved heterogeneity to obtain the joint density of

(y_{i1}, \dots, y_{iT}) conditional on x_i . In the most commonly applied models, including logit, probit, Tobit, and various count models (such as the Poisson model), the log likelihood in (73) identifies all of the parameters. Computation can be expensive but is typically tractable. The main methodological drawback to the full MLE approach is that it is not robust to violations of the conditional independence assumption, except for the linear model where normal conditional distributions are used for y_{it} and c_i .

The *partial* MLE ignores temporal dependence in the responses when estimating the parameters – at least when the parameters are identified. In particular, obtain the density of y_{it} given x_i by integrating the marginal density for y_{it} against the density for the heterogeneity:

$$g_t(y_t | x; \theta, \delta) = \int f_t(y_t | x_t, c; \theta) h(c | x; \delta) dc. \quad (74)$$

The *partial* MLE (PMLE) (or pooled MLE) uses, for each i , the partial log likelihood

$$\sum_{t=1}^T \log [g_t(y_{it} | x_i; \theta, \delta)]. \quad (75)$$

Because the partial MLE ignores the serial dependence caused by the presence of c_i , it is essentially never efficient. But in leading cases, such as probit, Tobit, and Poisson models, $g_t(y_t | x; \theta, \delta)$ has a simple form when $h(c | x; \delta)$ is chosen judiciously. Further, the PMLE is fully robust to violations of (67). Inference is complicated by the neglected serial dependence, but an appropriate adjustment to the asymptotic variance is easily obtained; see Chap. 13 in [55].

One complication with PMLE is that in the cases where it leads to a simple analysis (probit, ordered probit, and Tobit, to name a few), not all of the parameters in θ and δ are separately identified. The conditional independence assumption and the use of full MLE serves to identify all parameters. Fortunately, the PMLE does identify the parameters that index the average partial effects, a claim that will be verified for the probit model in Subsect. “Binary Response Models”.

Dynamic Models

General models with only sequentially exogenous variables are difficult to estimate. [8] considered binary response models and [54] suggested a general strategy that requires modeling the dynamic distribution of the variables that are not strictly exogenous.

Much more is known about the specific case where the model contains lagged dependent variables along with strictly exogenous variables. The starting point is a model for the dynamic distribution,

$$D(y_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}, \dots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i), \\ t = 1, \dots, T, \quad (76)$$

where \mathbf{z}_{it} are variables strictly exogenous (conditional on \mathbf{c}_i) in the sense that

$$D(y_{it}|\mathbf{z}_i, y_{i,t-1}, \mathbf{z}_{i,t-1}, \dots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i) \\ = D(\mathbf{y}_{it}|\mathbf{z}_{it}, y_{i,t-1}, \mathbf{z}_{i,t-1}, \dots, y_{i1}, \mathbf{z}_{i1}, y_{i0}, \mathbf{c}_i), \quad (77)$$

where \mathbf{z}_i is the entire history $\{\mathbf{z}_{it}: t = 1, \dots, T\}$.

In the leading case, (76) depends only on $(\mathbf{z}_{it}, y_{i,t-1}, \mathbf{c}_i)$ (although putting lags of strictly exogenous variables only slightly changes the notation). Let $f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta)$ denote a model for the conditional density, which depends on parameters θ . The joint density of (y_{i1}, \dots, y_{iT}) given $(y_{i0}, \mathbf{z}_i, \mathbf{c}_i)$ is

$$\prod_{t=1}^T f_t(y_t|\mathbf{z}_t, y_{t-1}, \mathbf{c}; \theta). \quad (78)$$

The problem with using (78) for estimation is that, when it is turned into a log likelihood by plugging in the “data”, \mathbf{c}_i must be inserted. Plus, the log likelihood depends on the initial condition, y_{i0} . Several approaches have been suggested to address these problems: (i) Treat the \mathbf{c}_i as parameters to estimate (which results in an incidental parameters problem). (ii) Try to estimate the parameters without specifying conditional or unconditional distributions for \mathbf{c}_i . (This approach is available for very limited situations, and other restrictions are needed. And, generally, one cannot estimate average partial effects.) (iii) Find, or, more practically, approximate $D(y_{i0}|\mathbf{c}_i, \mathbf{z}_i)$ and then model $D(\mathbf{c}_i|\mathbf{z}_i)$. Integrating out \mathbf{c}_i gives the density for $D(y_{i0}, y_{i1}, \dots, y_{iT}|\mathbf{z}_i)$, which can be used in an MLE analysis (conditional on \mathbf{z}_i), (iv) Model $D(\mathbf{c}_i|y_{i0}, \mathbf{z}_i)$. Then, integrate out \mathbf{c}_i conditional on (y_{i0}, \mathbf{z}_i) to obtain the density for $D(y_{i1}, \dots, y_{iT}|y_{i0}, \mathbf{z}_i)$. Now, MLE is conditional on (y_{i0}, \mathbf{z}_i) . As shown by [57], in some leading cases – probit, ordered probit, Tobit, Poisson regression – there is a density $h(\mathbf{c}|y_0, \mathbf{z})$ that mixes with the density $f(y_1, \dots, y_T|y_0, \mathbf{z}, \mathbf{c})$ to produce a log-likelihood that is in a common family and programmed in standard software packages.

If $m_t(\mathbf{x}_t, \mathbf{c}, \theta)$ is the mean function $E(y_t|\mathbf{x}_t, \mathbf{c})$, with $\mathbf{x}_t = (\mathbf{z}_t, y_{t-1})$, then APEs are easy to obtain. The average

structural function is

$$\text{ASF}(\mathbf{x}_t) = E_{\mathbf{c}_i}[m_t(\mathbf{x}_t, \mathbf{c}_i, \theta)] \\ = E\left\{\left[\int m_t(\mathbf{x}_t, \mathbf{c}, \theta)h(\mathbf{c}|y_{i0}, \mathbf{z}_i, \boldsymbol{\gamma})d\mathbf{c}\right] | y_{i0}, \mathbf{z}_i\right\}. \quad (79)$$

The term inside the brackets, say $r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \theta, \boldsymbol{\gamma})$ is available, at least in principle, because $m_t()$ and $h()$ have been specified. Often, they have simple forms, or they can be simulated. A consistent estimator of the ASF is obtained by averaging out (y_{i0}, \mathbf{z}_i) :

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^T r_t(\mathbf{x}_t, y_{i0}, \mathbf{z}_i, \hat{\theta}, \hat{\boldsymbol{\gamma}}). \quad (80)$$

Partial derivatives and differences with respect to elements of \mathbf{x}_t (which, remember, includes functions of y_{t-1}) can be computed. With large N and small T , the panel data bootstrap – where resampling is carried out in the cross section so that every time period is kept when a unit i is resampled – can be used for standard errors and inference. The properties of the nonparametric bootstrap are standard in this setting because the resampling is carried out in the cross section.

Binary Response Models

Unobserved effects models – static and dynamic – have been estimated for various kinds of response variables, including binary responses, ordered responses, count data, and corner solutions. Most of the issues outlined above can be illustrated by binary response models, which is the topic of this subsection.

The standard specification for the unobserved effects (UE) probit model is

$$P(y_{it} = 1|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \quad t = 1, \dots, T, \quad (81)$$

where \mathbf{x}_{it} does not contain an overall intercept but would usually include time dummies, and c_i is the scalar heterogeneity. Without further assumptions, neither $\boldsymbol{\beta}$ nor the APEs are identified. The traditional RE probit model imposes a strong set of assumptions: strict exogeneity, conditional independence, and independence between c_i and \mathbf{x}_i with $c_i \sim \text{Normal}(\mu_c, \sigma_c^2)$. Under these assumptions, $\boldsymbol{\beta}$ and the parameters in the distribution of c_i are identified and are consistently estimated by full MLE (conditional on \mathbf{x}_i).

Under the strict exogeneity assumption (64), a correlated random effects version of the model is obtained from

the Chamberlain–Mundlak device under conditional normality:

$$c_i = \psi + \bar{\mathbf{x}}_i \boldsymbol{\xi} + a_i, a_i | \mathbf{x}_i \sim \text{Normal}(0, \sigma_a^2). \quad (82)$$

The less restrictive version $c_i = \psi + \mathbf{x}_i \boldsymbol{\xi} + a_i = \psi + \mathbf{x}_{i1} \xi_1 + \dots + \mathbf{x}_{iT} \xi_T + a_i$ can be used, but the time average conserves on degrees of freedom.

As an example, suppose that y_{it} is a binary variable indicating whether firm i in year t was awarded at least one patent, and the key explanatory variable in \mathbf{x}_{it} is current and past spending on research and development (R&D). It makes sense that R&D spending is correlated, at least on average, with unobserved firm heterogeneity, and so a correlated random effects model seems natural. Unfortunately, the strict exogeneity assumption might be problematic: it could be that being awarded a patent in year t might affect future values of spending on R&D. Most studies assume this is not the case, but one should be aware that, as in the linear case, the strict exogeneity assumption imposes restrictions on economic behavior.

When the conditional independence assumption (67) is added to (81), strict exogeneity, and (82), all parameters in (81) and (82) are identified (assuming that all elements of \mathbf{x}_{it} are time-varying) and the parameters can be efficiently estimated by maximum likelihood (conditional on \mathbf{x}_i). Afterwards, the mean of c_i can be consistently estimated as $\hat{\mu}_c = \hat{\psi} + (N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i) \hat{\boldsymbol{\xi}}$ and the variance as $\hat{\sigma}_c^2 = \hat{\boldsymbol{\xi}}' (N^{-1} \sum_{i=1}^N \bar{\mathbf{x}}_i \bar{\mathbf{x}}_i') \hat{\boldsymbol{\xi}} + \hat{\sigma}_a^2$. Because a_i is normally distributed, c_i is not normally distributed unless $\bar{\mathbf{x}}_i \boldsymbol{\xi}$ is. A normal approximation for $D(c_i)$ gets better as T gets large. In any case, the estimated mean and standard deviation can be used to plug in values of c that are a certain number of estimated standard deviations from $\hat{\mu}_c$, say $\hat{\mu}_c \pm \hat{\sigma}_c$ or $\hat{\mu}_c \pm 2\hat{\sigma}_c$.

The APEs are identified from the ASF, which is consistently estimated by

$$\widehat{\text{ASF}}(\mathbf{x}_t) = N^{-1} \sum_{i=1}^N \Phi(\mathbf{x}_t \hat{\boldsymbol{\beta}}_a + \hat{\psi}_a + \bar{\mathbf{x}}_i \hat{\boldsymbol{\xi}}_a) \quad (83)$$

where the “ a ” subscript means that a coefficient has been divided by $(1 + \hat{\sigma}_a^2)^{1/2}$, for example, $\hat{\boldsymbol{\beta}}_a = \hat{\boldsymbol{\beta}} / (1 + \hat{\sigma}_a^2)^{1/2}$. The derivatives or changes of $\widehat{\text{ASF}}(\mathbf{x}_t)$ with respect to elements of \mathbf{x}_t can be compared with fixed effects estimates from a linear model. Often, to obtain a single scale factor, a further averaging across \mathbf{x}_{it} is done. The APEs computed from such averaging can be compared to linear FE estimates.

The CRE probit model is an example of a model where the APEs are identified without the conditional indepen-

dence assumption. Without (67) – or any restriction on the joint distribution – it can still be shown that

$$P(y_{it} = 1 | \mathbf{x}_i) = \Phi(\mathbf{x}_{it} \boldsymbol{\beta}_a + \psi_a + \bar{\mathbf{x}}_i \boldsymbol{\xi}_a), \quad (84)$$

which means a number of estimation approaches identify the scaled coefficients $\boldsymbol{\beta}_a$, ψ_a , and $\boldsymbol{\xi}_a$. The estimates of these scaled coefficients can be inserted directly into (83). The unscaled parameters and σ_a^2 are not separately identified, but in most cases this is a small price to pay for relaxing the conditional independence assumption. Note that for determining directions of effects and relative effects, $\boldsymbol{\beta}_a$ is just as useful as $\boldsymbol{\beta}$. Plus, it is $\boldsymbol{\beta}_a$ that appears in the APEs. The partial effects at the mean value of c_i are not identified.

Using pooled probit can be inefficient for estimating the scaled parameters. Full MLE, with a specified correlation matrix for the $T \times 1$ vector \mathbf{u}_i , is possible in principle but difficult in practice. An alternative approach, the *generalized estimating equations* (GEE) approach, can be more efficient than pooled probit but just as robust in that only (84) is needed for consistency. See [38] for a summary of how GEE – which is essentially the same as multivariate weighted nonlinear least squares – applies to the CRE probit model.

A simple test of the strict exogeneity assumption is to add selected elements of $\mathbf{x}_{i,t+1}$, say $\mathbf{w}_{i,t+1}$, to the model and computing a test of joint significance. Unless the full MLE is used, the test should be made robust to serial dependence of unknown form. For example, as a test of strict exogeneity of R&D spending when y_{it} is a patent indicator, one can just include next year’s value of R&D spending and compute a t test. In carrying out the test, the last time period is lost for all firms.

Because there is nothing sacred about the standard model (81) under (82) – indeed, these assumptions are potentially quite restrictive – it is natural to pursue other models and assumptions. Even with (81) as the starting point, and under strict exogeneity, there are no known ways of identifying parameters or partial effects without restricting $D(c_i | \mathbf{x}_i)$. Nevertheless, as mentioned in Subsect. “[Assumptions About the Unobserved Heterogeneity](#)”, there are nonparametric restrictions on $D(c_i | \mathbf{x}_i)$ that do identify the APEs under strict exogeneity – even if (81) is dropped entirely. As shown in [3], the restriction $D(c_i | \mathbf{x}_i) = D(c_i | \bar{\mathbf{x}}_i)$ identifies the APEs. While fully nonparametric methods can be used, some simple strategies are evident. For example, because the APEs can be obtained from $D(y_{it} | \mathbf{x}_{it}, \bar{\mathbf{x}}_i)$, it makes sense to apply flexible parametric models directly to this distribution – without worrying about the original models for $D(y_{it} | \mathbf{x}_{it}, c_i)$ and $D(c_i | \mathbf{x}_i)$.

As an example of this approach, a flexible parametric model, such as

$$\begin{aligned} P(y_{it} = 1 | \mathbf{x}_{it}, \bar{\mathbf{x}}_i) \\ = \Phi[\theta_t + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i\boldsymbol{\gamma} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\boldsymbol{\delta} + (\mathbf{x}_{it} \otimes \bar{\mathbf{x}}_i)\boldsymbol{\eta}], \end{aligned} \quad (85)$$

might provide a reasonable approximation. The average structural function is estimated as

$$\begin{aligned} \widehat{\text{ASF}}(\mathbf{x}_i) = \\ N^{-1} \sum_{i=1}^N \Phi[\hat{\theta}_t + \mathbf{x}_i\hat{\boldsymbol{\beta}} + \bar{\mathbf{x}}_i\hat{\boldsymbol{\gamma}} + (\bar{\mathbf{x}}_i \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\delta}} + (\mathbf{x}_i \otimes \bar{\mathbf{x}}_i)\hat{\boldsymbol{\eta}}], \end{aligned} \quad (86)$$

where the estimates can come from pooled MLE, GEE, or a method of moments procedure. The point is that extensions of the basic probit model such as (85) can provide considerable flexibility and deliver good estimators of the APEs. The drawback is that one has to be willing to abandon standard underlying models for $P(y_{it} = 1 | \mathbf{x}_{it}, c_i)$ and $D(c_i | \mathbf{x}_i)$; in fact, it seems very difficult to characterize models for these two features that would lead to an expression such as (85).

An alternative model for the response probability is the logit model

$$P(y_{it} = 1 | \mathbf{x}_{it}, c_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + c_i), \quad (87)$$

where $\Lambda(z) = \exp(z)/[1 + \exp(z)]$. In cross section applications, researchers often find few practical differences between (81) and (87). But when unobserved heterogeneity is added in a panel data context, the logit formulation has an advantage: under the conditional independence assumption (and strict exogeneity), the parameters $\boldsymbol{\beta}$ can be consistently estimated, with a \sqrt{N} -asymptotic normal distribution, without restricting $D(c_i | \mathbf{x}_i)$. The method works by conditioning on the number of “successes” for each unit, that is, $n_i = \sum_{t=1}^T y_{it}$. [17] shows that $D(y_{i1}, \dots, y_{iT} | \mathbf{x}_i, c_i, n_i) = D(y_{i1}, \dots, y_{iT} | \mathbf{x}_i, n_i)$, and the latter depends on $\boldsymbol{\beta}$ (at least when all elements of \mathbf{x}_{it} are time varying). The conditional MLE – sometimes called the “fixed effects logit” estimator – is asymptotically efficient in the class of estimators putting no assumptions on $D(c_i | \mathbf{x}_i)$. While this feature of the logit CMLE is attractive, the method has two drawbacks. First, it does not appear to be robust to violations of the conditional independence assumption, and little is known about the practical effects of serial dependence in $D(y_{i1}, \dots, y_{iT} | \mathbf{x}_i, c_i)$. Second, and perhaps more importantly, because $D(c_i | \mathbf{x}_i)$ and $D(c_i)$ are not restricted, it is not clear how one estimates

magnitudes of the effects of the covariates on the response probability. The logit CMLE is intended to estimate the parameters, which means the effects of the covariates on the log-odds ratio, $\log\{[P(y_{it} = 1 | \mathbf{x}_{it}, c_i)]/[1 - P(y_{it} = 1 | \mathbf{x}_{it}, c_i)]\} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i$, can be estimated. But the magnitudes of the effects of covariates on the response probability are not available. Therefore, there are tradeoffs when choosing between CRE probit and “fixed effects” logit: the CRE probit identifies average partial effects with or without the conditional independence assumptions, at the cost of specifying $D(c_i | \mathbf{x}_i)$, while the FE logit estimates parameters without specifying $D(c_i | \mathbf{x}_i)$, but requires conditional independence and still does not deliver estimates of partial effects. As often is the case in econometrics, there are tradeoffs between assumptions between the logit and probit approaches, and also tradeoffs. See [38] for further discussion.

Estimation of parameters and APEs is more difficult in simple dynamic probit models. Consider

$$P(y_{it} = 1 | \mathbf{z}_{it}, y_{i,t-1}, c_i) = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + c_i), \quad (88)$$

which assumes first-order dynamics and strict exogeneity of $\{\mathbf{z}_{it} : t = 1, \dots, T\}$. Treating the c_i as parameters to estimate causes inconsistency in $\boldsymbol{\delta}$ and ρ because of the incidental parameters problem. A simple analysis is available under the assumption

$$c_i | y_{i0}, \mathbf{z}_i \sim \text{Normal}(\psi + \xi_0 y_{i0} + \mathbf{z}_i \boldsymbol{\xi}, \sigma_a^2). \quad (89)$$

Then,

$$\begin{aligned} P(y_{it} = 1 | \mathbf{z}_i, y_{i,t-1}, \dots, y_{i0}, a_i) \\ = \Phi(\mathbf{z}_{it}\boldsymbol{\delta} + \rho y_{i,t-1} + \psi + \xi_0 y_{i0} + \mathbf{z}_i \boldsymbol{\xi} + a_i), \end{aligned} \quad (90)$$

where $a_i \equiv c_i - \psi - \xi_0 y_{i0} - \mathbf{z}_i \boldsymbol{\xi}$. Because a_i is independent of (y_{i0}, \mathbf{z}_i) , it turns out that standard random effects probit software can be used, with explanatory variables $(1, \mathbf{z}_{it}, y_{i,t-1}, y_{i0}, \mathbf{z}_i)$ in time period t . All parameters, including σ_a^2 , are consistently estimated, and the ASF is estimated by averaging out (y_{i0}, \mathbf{z}_i) :

$$\begin{aligned} \widehat{\text{ASF}}(\mathbf{z}_t, y_{t-1}) = \\ N^{-1} \sum_{i=1}^N \Phi(\mathbf{z}_i \hat{\boldsymbol{\delta}}_a + \hat{\rho}_a y_{t-1} + \hat{\psi}_a + \hat{\xi}_{a0} y_{i0} + \mathbf{z}_i \hat{\boldsymbol{\xi}}_a), \end{aligned} \quad (91)$$

where the coefficients are multiplied by $(1 + \delta_a^2)^{-1/2}$. APEs are gotten, as usual, by taking differences or derivatives with respect to elements of (\mathbf{z}_t, y_{t-1}) . Both (88) and

the model for $D(c_i|y_{i0}, \mathbf{z}_i)$ can be made more flexible (such as including interactions, or letting $\text{Var}(c_i|\mathbf{z}_i, y_{i0})$ be heteroskedastic). See [57] for further discussion.

Similar analyses hold for other nonlinear models, although the particulars differ. For count data, maximum likelihood methods are available – based on correlated random effects or conditioning on a sufficient statistic. In this case, the CMLE based on the Poisson distribution has very satisfying robustness properties, requiring only the conditional mean in the unobserved effects model to be correctly specified along with strict exogeneity (Conditional independence is not needed). These and dynamic count models are discussed in Chap. 19 in [55,57].

Correlated random effects Tobit models are specified and estimated in a manner very similar to CRE probit models; see Chap. 16 in [55]. Unfortunately, there are no known conditional MLEs that eliminate the unobserved heterogeneity in Tobit models. Nevertheless, [33,34] show how the parameters in models for corner solutions can be estimated without distributional assumptions on $D(c_i|\mathbf{x}_i)$. Such methods do place exchangeability restrictions on $D(y_{i1}, \dots, y_{iT}|\mathbf{x}_i, c_i)$, but they are not as strong as conditional independence with identical distributions.

Future Directions

Research in panel data methods continues unabated. Dynamic linear models are a subject of ongoing interest. The problem of feedback in linear models when the covariates are persistent – and the weak instrument problem that it entails – is important for panels with small T . For example, with firm-level panel data, the number of time periods is typically small and inputs into a production function would often be well-approximated as random walks with perhaps additional short-term dependence. The estimators described in Sect. “[Sequentially Exogenous Regressors and Dynamic Models](#)” that impose additional assumptions should be studied when those assumptions fail. Perhaps the lower variance of the estimators from the misspecified model is worth the additional bias.

Models with random coefficients, especially when those random coefficients are on non-strictly exogenous variables (such as lagged dependent variables), have received some attention, but many of the proposed solutions require large T . (See, for example, [49,50]). An alternative approach is flexible MLE, as in [57], where one models the distribution of heterogeneity conditional on the initial condition and the history of covariates. See [19] for any application to dynamic product choice.

When T is large enough so that it makes sense to use large-sample approximations with large T , as well as

large N , one must make explicit assumptions about the time series dependence in the data. Such frameworks are sensible for modeling large geographical units, such as states, provinces, or countries, where long stretches of time are observed. The same estimators that are attractive for the fixed T case, particularly fixed effects, can have good properties when T grows with N , but the properties depend on whether unit-specific effects, time-specific effects, or both are included. The rates at which T and N are assumed to grow also affect the large-sample approximations. See [52] for a survey of linear model methods with T and N are both assumed to grow in the asymptotic analysis. A recent study that considers estimation when the data have unit roots is [44]. Unlike the fixed T case, a unified theory for linear models, let alone nonlinear models, remains elusive when T grows with N and is an important area for future research.

In the models surveyed here, a single coefficient is assumed for the unobserved heterogeneity, whereas the effect might change over time. In the linear model, the additive c_i can be replaced with $\psi_t c_i$ (with $\psi_1 = 1$ as a normalization). For example, the return to unobserved managerial talent in a firm production function can change over time. Conditions under which ignoring the time-varying loads, ψ_t , and using the usual fixed effects estimator, consistently estimates the coefficients on \mathbf{x}_{it} are given in [47]. But one can also estimate the ψ_t along with β using method of moments frameworks. Examples are [2,32]. An area for future research is to allow heterogeneous slopes on observed covariates along with time-varying loads on the unobserved heterogeneity. Allowing for time-varying loads and heterogeneous slopes in nonlinear models can allow for significant flexibility, but only parametric approaches to estimation have been studied.

There is considerable interest in estimating production functions using proxy variables, such as investment, for time-varying, unobserved productivity. The pioneering work is [48]; see also [42]. Estimation in this case does not rely on differencing or time-demeaning to remove unobserved heterogeneity, and so the estimates can be considerably more precise than the FE or FD estimators. But the assumption that a deterministic function of investment can proxy for unobserved productivity is strong. [11] provides an analysis that explicitly allows for unobserved heterogeneity and non-strictly exogenous inputs using the methods described in Sect. “[Sequentially Exogenous Regressors and Dynamic Models](#)”. An interesting challenge for future researchers is to unify the two approaches to exploit the attractive features of each.

The parametric correlated random effects approach for both static and dynamic nonlinear models is now fairly

well understood in the balanced case. Much less attention has been paid to the unbalanced case, and missing data, especially for fully dynamic models, is a serious challenge. [57] discusses the assumptions under which using a balanced subset produces consistent estimates.

Identification of average partial effects (equivalently, the average structural function) has recently received the attention that it deserves, although little is known about how robust are the estimated APEs under various misspecifications of parametric models. One might hope that using flexible models for nonlinear responses might provide good approximations, but evidence on this issue is lacking.

As mentioned earlier, recent research in [3] has shown how to identify and estimate partial effects without making parametric assumptions about $E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$ or $D(\mathbf{c}_i|\mathbf{x}_i)$. The setup in [3] allows for $D(\mathbf{c}_i|\mathbf{x}_i)$ to depend on $\mathbf{x}_i = (\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$ in an exchangeable way. The simplest case is the one given in (71), $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\bar{\mathbf{x}}_i)$. Under (71) and the strict exogeneity assumption $E(y_{it}|\mathbf{x}_i, \mathbf{c}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i)$, the average structural function is identified as

$$\text{ASF}_t(\mathbf{x}_t) = E_{\bar{\mathbf{x}}_i}[r_t(\mathbf{x}_t, \bar{\mathbf{x}}_i)], \quad (92)$$

where $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i) = E(y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$. Because $r_t(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ can be estimated very generally – even using nonparametric regression of y_{it} on $(\mathbf{x}_{it}, \bar{\mathbf{x}}_i)$ for each t – the average partial effects can be estimated without any parametric assumptions. Research in [3] shows how $D(\mathbf{c}_i|\mathbf{x}_i)$ can depend on other exchangeable functions of $(\mathbf{x}_{i1}, \dots, \mathbf{x}_{iT})$, such as sample variances and covariances. As discussed in [38], nonexchangeable functions, such as trends and growth rates, can be accommodated, provided these functions are known. For example, for each i , let $(\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$ be the vectors of intercepts and slopes from the regression \mathbf{x}_{it} on $1, t, t = 1, \dots, T$. Then, an extension of (71) is $D(\mathbf{c}_i|\mathbf{x}_i) = D(\mathbf{c}_i|\hat{\mathbf{f}}_i, \hat{\mathbf{g}}_i)$. It appears these kinds of assumptions have not yet been applied, but they are a fertile area for future research because they extend the typical CRE setup.

Future research on nonlinear models will likely consider the issue of the kinds of partial effects that are of most interest. [3] studies identification and estimation of the *local average response* (LAR). The LAR at \mathbf{x}_t for a continuous variable x_{tj} is

$$\int \frac{\partial m_t(\mathbf{x}_t, \mathbf{c})}{\partial x_{tj}} dH_t(\mathbf{c}|\mathbf{x}_t), \quad (93)$$

where $m_t(\mathbf{x}_t, \mathbf{c})$ is the conditional mean of the response and $H_t(\mathbf{c}|\mathbf{x}_t)$ denotes the cdf of $D(\mathbf{c}_i|\mathbf{x}_{it} = \mathbf{x}_t)$. This is a “local” partial effect because it averages out the heterogeneity for the slice of the population described by the

vector of observed covariates, \mathbf{x}_t . The APE averages out over the entire distribution of \mathbf{c}_i , and therefore can be called a “global average response”. See also [21]. The results in [3] include general identification results for the LAR, and future empirical researchers using nonlinear panel data models may find the local nature of the LAR more appealing (although more difficult to estimate) than APEs.

A different branch of the panel data literature has studied identification of coefficients or, more often, scaled coefficients, in nonlinear models. For example, [35] shows how to estimate β in the model

$$y_{it} = 1[w_{it} + \mathbf{x}_{it}\beta + c_i + u_{it} \geq 0] \quad (94)$$

without distributional assumptions on the composite error, $c_i + u_{it}$. In this model, w_{it} is a special continuous explanatory variable (which need not be time varying). Because its coefficient is normalized to unity, w_{it} necessarily affects the response, y_{it} . More importantly, w_{it} is assumed to satisfy the distributional restriction $D(c_i + u_{it}|\mathbf{x}_{it}, \mathbf{z}_i) = D(c_i + u_{it}|\mathbf{x}_{it}, \mathbf{z}_i)$, which is a conditional independence assumption. The vector \mathbf{z}_i is assumed to be independent of u_{it} in all time periods. (So, if two time periods are used, \mathbf{z}_i could be functions of variables determined prior to the earliest time period). The most likely scenario where the framework in [35] applies is when w_{it} is randomized and therefore independent of the entire vector $(\mathbf{x}_{it}, \mathbf{z}_i, c_i + u_{it})$. The key condition seems unlikely to hold if w_{it} is related to past outcomes on y_{it} . The estimator of β derived in [35] is \sqrt{N} -asymptotically normal, and fairly easy to compute; it requires estimation of the density of w_{it} given $(\mathbf{x}_{it}, \mathbf{z}_i)$ and then a simple IV estimation. Essentially by construction, estimation of partial effects on the response probability is not possible.

Recently, [36] shows how to obtain bounds on parameters and APEs in dynamic models, including the dynamic probit model in Eq. (85) under the strict exogeneity assumption on $\{\mathbf{z}_{it}: t = 1, \dots, T\}$. A further assumption is that c_i and \mathbf{z}_i are independent. By putting restrictions on $D(c_i)$ – which nevertheless allow flexibility – [36] explains how to estimate bounds for the unknown ρ . The bounds allow one to determine how much information are in the data when few assumptions are made. Similar calculations can be made for average partial effects, so that the size of so-called state dependence – the difference between $E_{c_i}[\Phi(\mathbf{z}_i\delta + \rho + c_i) - \Phi(\mathbf{z}_i\delta + c_i)]$ – can be bounded.

Because CRE methods require some restriction on the distribution of heterogeneity, and estimation of scaled coefficients leaves partial effects unidentified, the theoretical literature has returned to the properties of parameter estimates and partial effects when the heterogeneity

is treated as unit-specific parameters to estimate. Recent work has focused on adjusting the “fixed effects” estimates (of the common population parameters) so that they have reduced bias.

An emerging question is whether the average partial effects might be estimated well even though the parameters themselves are biased. In other words, suppose that for a nonlinear model one obtains $\{\hat{\theta}, \hat{c}_1, \hat{c}_2, \dots, \hat{c}_N\}$, typically by maximizing a pooled log-likelihood function across all i and t . If $m_t(\mathbf{x}_t, \mathbf{c}, \theta) = E(y_{it} | \mathbf{x}_t, \mathbf{c})$ is the conditional mean function, the average partial effects can be estimated as

$$N^{-1} \sum_{i=1}^N \frac{\partial m_t(\mathbf{x}_t, \hat{c}_i, \hat{\theta})}{\partial x_{tj}}. \quad (95)$$

In the unobserved effects probit model, (95) becomes

$$N^{-1} \sum_{i=1}^N \hat{\beta}_j \phi(\mathbf{x}_t \hat{\boldsymbol{\beta}} + \hat{c}_i). \quad (96)$$

[20] studied the properties of (96) with strictly exogenous regressors under conditional independence, assuming that the covariates are weakly dependent over time. Interestingly, the bias in (96) is of order T^{-2} when there is no heterogeneity, which suggests that estimating the unobserved effects might not be especially harmful when the amount of heterogeneity is small. Unfortunately, these findings do not carry over to models with time heterogeneity or lagged dependent variables. While bias corrections are available, they are difficult to implement.

[24] proposes both jackknife and analytical bias corrections and show that they work well for the probit case. Generally, the jackknife procedure to remove the bias in $\hat{\theta}$ is simple but can be computationally intensive. The idea is this. The estimator based on T time periods has probability limit (as $N \rightarrow \infty$) that can be written as

$$\theta_T = \theta + \mathbf{b}_1/T + \mathbf{b}_2/T^2 + O(T^{-3}) \quad (97)$$

for vectors \mathbf{b}_1 and \mathbf{b}_2 . Now, let $\hat{\theta}_{(t)}$ denote the estimator that drops time period t . Then, assuming stability across t , it can be shown that the jackknife estimator,

$$\tilde{\theta} = T\hat{\theta} - (T-1)T^{-1} \sum_{t=1}^T \hat{\theta}_{(t)} \quad (98)$$

has asymptotic bias of $\tilde{\theta}$ on the order of T^{-2} .

Unfortunately, there are currently some practical limitations to the jackknife procedure, as well as to the analytical corrections derived in [24]. First, aggregate time effects

are not allowed, and they would be very difficult to include because the analysis is with $T \rightarrow \infty$. (In other words, time effects would introduce an incidental parameters problem in the time dimension, in addition to the incidental parameters problem in the cross section). Plus, heterogeneity in the distribution of the response y_{it} across t changes the bias terms \mathbf{b}_1 and \mathbf{b}_2 when a time period is dropped, and so the adjustment in (98) does not remove the bias terms. Second, [24] assumes independence across t conditional on c_i . It is a traditional assumption, but in static models it is often violated, and it must be violated in dynamic models. Plus, even without time heterogeneity, the jackknife does not apply to dynamic models; see [23].

Another area that has seen a resurgence is so-called pseudo panel data, as initially explicated in [18]. A pseudo-panel data set is constructed from repeated cross sections across time, where the units appearing in each cross section are not repeated (or, if they are, it is a coincidence and is ignored). If there is a natural grouping of the cross-sectional units – for example, for individuals, birth year cohorts – one can create a pseudo-panel data set by constructing group or cohort averages in each time period. With relatively few cohorts and large cross sections, one can identify pseudo panels in the context of minimum distance estimation. With a large number of groups, a different large-sample analysis might be warranted. A recent contribution is [39] and [38] includes a recent survey. Open questions include the most efficient way to use the full set of restrictions in the underlying individual-level model.

As mentioned earlier, this chapter did not consider panel data model with explanatory variables that are endogenous in the sense that they are correlated with time-varying unobservables. For linear models, the usual fixed effects and first differencing transformations can be combined with instrumental variables methods. In nonlinear models, the Chamberlain–Mundlak approach can be combined with so-called “control function” methods, provided the endogenous explanatory variables are continuous. [38] includes a discussion of some recent advances for complicated models such as multinomial response models; see also [51]. Generally, structural estimation in discrete response models with unobserved heterogeneity and endogenous explanatory variables is an area of great interest.

Bibliography

Primary Literature

1. Ahn SC, Schmidt P (1995) Efficient estimation of models for dynamic panel data. *J Econom* 68:5–27

2. Ahn SC, Lee YH, Schmidt P (2001) GMM estimation of linear panel data models with time-varying individual effects. *J Econom* 101:219–255
3. Altonji JG, Matzkin RL (2005) Cross section and panel data estimators for nonseparable models with endogenous regressors. *Econometrica* 73:1053–1102
4. Anderson TW, Hsiao C (1982) Formulation and estimation of dynamic models using panel data. *J Econom* 18:47–82
5. Arellano M (1993) On the testing of correlated effects with panel data. *J Econom* 59:87–97
6. Arellano M, Bond SR (1991) Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Rev Econ Stud* 58:277–297
7. Arellano M, Bover O (1995) Another look at the instrumental variable estimation of error components models. *J Econom* 68:29–51
8. Arellano M, Carrasco R (2003) Binary choice panel data models with predetermined variables. *J Econom* 115:125–157
9. Arellano M, Honoré B (2001) Panel data models: Some recent developments. In: Heckman JJ, Leamer E (eds) *Handbook of econometrics*, vol 5. North Holland, Amsterdam, pp 3229–3296
10. Blundell R, Bond SR (1998) Initial conditions and moment restrictions in dynamic panel data models. *J Econom* 87:115–143
11. Blundell R, Bond SR (2000) GMM estimation with persistent panel data: An application to production functions. *Econom Rev* 19:321–340
12. Blundell R, Powell JL (2003) Endogeneity in nonparametric and semiparametric regression models. In: Dewatripont M, Hansen LP, Turnovsky SJ (eds) *Advances in economics and econometrics: Theory and applications*, 8th World Congress, vol 2. Cambridge University Press, Cambridge, pp 312–357
13. Blundell R, Smith RJ (1991) Initial conditions and efficient estimation in dynamic panel data models – an application to company investment behaviours. *Ann d'écon stat* 20–21:109–124
14. Cameron AC, Trivedi PK (2005) *Microeconometrics: Methods and applications*. Cambridge University Press, Cambridge
15. Chamberlain G (1980) Analysis of covariance with qualitative data. *Rev Econ Stud* 47:225–238
16. Chamberlain G (1982) Multivariate regression models for panel data. *J Econom* 1:5–46
17. Chamberlain G (1984) Panel data. In: Griliches Z, Intriligator MD (eds) *Handbook of econometrics*, vol 2. North Holland, Amsterdam, pp 1248–1318
18. Deaton A (1985) Panel Data from time series of cross-sections. *J Econom* 30:109–126
19. Erdem T, Sun B (2001) Testing for choice dynamics in panel data. *J Bus Econ Stat* 19:142–152
20. Fernández-Val I (2007) Fixed effects estimation of structural parameters and marginal effects in panel probit models. Mimeo. Boston University, Department of Economics
21. Florens JP, Heckman JJ, Meghir C, Vytlacil E (2007) Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects. Mimeo. Columbia University, Department of Economics
22. Hahn J (1999) How informative is the initial condition in the dynamic panel model with fixed effects? *J Econom* 93:309–326
23. Hahn J, Kuersteiner G (2002) Asymptotically unbiased inference for a dynamic panel model with fixed effects when both n and T are large. *Econometrica* 70:1639–1657
24. Hahn J, Newey WK (2004) Jackknife and analytical bias reduction for nonlinear panel models. *Econometrica* 72:1295–1319
25. Hausman JA (1978) Specification tests in econometrics. *Econometrica* 46:1251–1271
26. Hausman JA, Taylor WE (1981) Panel data and unobservable individual effects. *Econometrica* 49:1377–1398
27. Hayashi F (2000) *Econometrics*. Princeton University Press, Princeton
28. Heckman JJ (1976) The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann Econ Soc Meas* 5:475–492
29. Heckman JJ (1981) Statistical models for discrete panel data. In: Manski CF, McFadden DL (eds) *Structural analysis of discrete data and econometric applications*. MIT Press, Cambridge, pp 114–178
30. Heckman JJ (1981) The incidental parameters problem and the problem of initial condition in estimating a discrete time-discrete data stochastic process. In: Manski CF, McFadden DL (eds) *Structural analysis of discrete data and econometric applications*. MIT Press, Cambridge, pp 179–195
31. Hoch I (1962) Estimation of production function parameters combining time-series and cross-section data. *Econometrica* 30:34–53
32. Holtz-Eakin D, Newey W, Rosen HS (1988) Estimating vector autoregressions with panel data. *Econometrica* 56:1371–1395
33. Honoré BE (1992) Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica* 60:533–565
34. Honoré BE, Hu L (2004) Estimation of cross sectional and panel data censored regression models with endogeneity. *J Econom* 122:293–316
35. Honoré BE, Lewbel A (2002) Semiparametric binary choice panel data models without strictly exogenous regressors. *Econometrica* 70:2053–2063
36. Honoré BE, Tamer E (2006) Bounds on parameters in panel dynamic discrete choice models. *Econometrica* 74:611–629
37. Im KS, Ahn SC, Schmidt P, Wooldridge JM (1999) Efficient estimation of panel data models with strictly exogenous explanatory variables. *J Econom* 93:177–201
38. Imbens GW, Wooldridge JM (2007) What's new in econometrics? Lecture Notes. National Bureau of Economic Research, Summer Institute
39. Inoue A (2008) Efficient estimation and inference in linear pseudo-panel data models. *J Econom* 142:449–466
40. Keane MP, Runkle DE (1992) On the estimation of panel-data models with serial correlation when instruments are not strictly exogenous. *J Bus Econ Stat* 10:1–9
41. Kiefer NM (1980) Estimation of fixed effect models for time series of cross-sections with arbitrary intertemporal covariance. *J Econom* 14:195–202
42. Levinshohn J, Petrin A (2003) Estimating production functions using inputs to control for unobservables. *Rev Econ Stud* 70:317–341
43. Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. *Biometrika* 73:13–22
44. Moon HR, Phillips PCB (2004) GMM estimation of autoregressive roots near unity with panel data. *Econometrica* 72:467–522
45. Mundlak Y (1961) Empirical production function free of management bias. *Farm J Econ* 43:44–56

46. Mundlak Y (1978) On the pooling of time series and cross section data. *Econometrica* 46:69–85
47. Murtazashvili I, Wooldridge JM (2007) Fixed effects instrumental variables estimation in correlated random coefficient panel data models. *J Econom* 142:539–552
48. Olley S, Pakes A (1996) The dynamics of productivity in the telecommunications equipment industry. *Econometrica* 64:1263–1298
49. Pesaran MH, Takashi Y (2008) Testing slope homogeneity in large panels. *J Econom* 142:50–93
50. Pesaran MH, Smith RP, Im KS (1996) Dynamic linear models for heterogeneous panels. In: Mátyás L, Sevestre P (eds) *The econometrics of panel data*. Kluwer, Dordrecht, pp 145–195
51. Petrin A, Train KE (2005) Tests for omitted attributes in differentiated product models. Mimeo. University of Minnesota, Department of Economics
52. Phillips PCB, Moon HR (2000) Nonstationary panel data analysis: An overview of some recent developments. *Econom Rev* 19:263–286
53. Wooldridge JM (1999) Distribution-free estimation of some nonlinear panel data models. *J Econom* 90:77–97
54. Wooldridge JM (2000) A framework for estimating dynamic, unobserved effects panel data models with possible feedback to future explanatory variables. *Econ Lett* 68:245–250
55. Wooldridge JM (2002) *Econometric analysis of cross section and panel data*. MIT Press, Cambridge
56. Wooldridge JM (2005) Fixed-effects and related estimators for correlated random-coefficient and treatment-effect panel data models. *Rev Econ Stat* 87:385–390
57. Wooldridge JM (2005) Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *J Appl Econom* 20:39–54

Books and Reviews

- Arellano M (2003) *Panel data econometrics*. Oxford University Press, Oxford
- Baltagi BH (2001) *Econometric analysis of panel data*, vol 2e. Wiley, New York
- Hsiao C (2003) *Analysis of panel data*, vol 2e. Cambridge University Press, Cambridge

Econophysics, Observational

BERTRAND M. ROEHNER
Institute for Theoretical and High Energy Physics,
University of Paris 7, Paris, France

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[The Primacy of Observation](#)
[Investigating One Effect at a Time](#)
[What Guidance Can Physics Provide?](#)

[How Cross-National Observations Can Be Used to Test the Role of Different Factors](#)

[How Vested Interests May Affect the Accessibility and Reliability of Social Data](#)

[How Can Exogenous Factors be Taken into Account?](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Cross national comparisons Comparing cross-national data for a specific phenomenon, e.g. a surge in housing prices, is the key to distinguishing between essential factors which are common to all episodes and those which are accessory and context dependent.

Economathematicians Mathematicians or theoretical physicists who develop mathematical tools, models or simulations for social phenomena but do not try to confront these models to actual observations.

Econophysics A field of physics which originated in the mid-1990s. Throughout this article, we use the term in a broad sense which includes econophysics, sociophysics and historiophysics. As a matter of fact, these fields can hardly be studied separately in the sense that economic effects depend upon social reactions (e.g. reactions of consumers to advertising campaigns); furthermore, economic investigations crucially rely on statistics which typically must combine present-day data with data from former historical episodes.

Econophysicists Physicists who study social, economic or political issues.

Endogenous mechanisms Models usually describe endogenous mechanisms. For instance a population model would describe how people get married and have children.

Exogenous factors Exogenous factors are more or less unexpected external forces which act on the system. Thus, for a population wars or epidemics may bring about sudden population changes. It is only when exogenous factors are recurrent and fairly repetitive that they can be taken into account in models.

Experiment Apart from its standard meaning in physics or biology we also use this term to designate the process of (i) defining the phenomenon that one wants to study (ii) locating and collecting the data which are best suited for the investigation (iii) analyzing the data and deriving *regularity rules* or testing a model.

Model testing Before confronting the predictions of a model to statistical evidence it is necessary to ensure that the system was not subject to unexpected exogenous shocks. The impact of exogenous factors which

are not accounted for in the model must in some way be removed, that is to say the data must be corrected in a way which takes these shocks out of the picture. Usually, such corrections are very tricky to implement.

Definition of the Subject

“No science thrives in the atmosphere of direct practical aim. We should still be without most of the conveniences of modern life if physicists had been as eager for immediate applications as most economists are and always have been.” (J. Schumpeter p.6 in [11])

“The free fall is a very trivial physical phenomenon, but it was the study of this exceedingly simple fact and its comparison with the astronomical material which brought forth mechanics. The sound procedure [in every science] is to obtain first utmost precision and mastery in a limited field, and then to proceed to another, somewhat wider one and so on.” (J. von Neumann and O. Morgenstern [5])

These two quotes define fairly well the path that econophysics tries to follow. They both insist on the fact that one should begin by focusing on simple phenomena even if at first sight they have little practical implications. In what follows we will develop this point but first of all we must address a question which comes to the mind of all persons who hear about econophysics for the first time, namely:

“Why should physicists have something to say about economic and social phenomena. Admittedly, biology can benefit from physics because of the means of observation [e.g. exploration of protein molecules by X-ray scattering] that it provides, but there are no similar needs in economics.”

I have heard this question asked repeatedly by many of my colleagues. In my answer I usually emphasize that what matters is more the method of investigation than the phenomena by themselves. I stress that applying to the social sciences the experimental methodology invented by physicists and chemists would mark a great progress. However, with the benefit of insight, I realize that these answers may have appeared far fetched and unconvincing to many of my listeners. A better and more factual claim is to observe that over the past century several of the most renowned economists and sociologists were in fact econophysicists in the sense defined in the glossary. Indeed, back in the nineteenth century, the only way to get a decent mathematical training was to study astronomy, engineering,

mathematics or physics. When such people entered the social sciences this led to two kinds of approaches which we may designate as econophysics and economathematics (see Sect. “Glossary”). In the first category one may mention the astronomer Adolphe Quételet (1796–1874), Clément Juglar (1819–1905) educated as a medical doctor, Vilfredo Pareto (1848–1923) educated as an engineer, the mathematician Louis Bachelier (1870–1946), the physicist Elliott Montroll (1916–1983), the mathematician Benoît Mandelbrot (1924–). In the second category one may mention Léon Walras (1834–1910) who was educated as an engineer, the astronomer Simon Newcomb (1835–1905), the physicist Maurice Allais (1911–).

Of course, if the economic discipline had been highly successful there would be little need for an alternative approach. However, great doubts have been expressed by some of the most renowned economists about the attainments of their discipline. We have already cited Schumpeter’s opinion on this matter. In addition one may mention the judgments of Vassily Leontief, Anna Schwartz, Lawrence Summers or the thesis developed in a recent book by Masanao Aoki and Hiroshi Yoshikawa.

- Leontief and Schwartz emphasized that the present organization of economic research discourages observational research. In Schwartz’s words [12]¹

“The main disincentive to improve the handling and use of data is that the profession withholds recognition to those who devote their energies to measurement. Someone who introduces an innovation in econometrics, by contrast, will win plaudits.”

- The assessment made by Summers in a paper published in 1991 is well summarized by its title: “The scientific illusion of empirical macroeconomics”.
- In their book, Aoki and Yoshikawa (p. 25 in [1]) point out that the representative agent assumption which is supposed to provide a connection between micro- and macroeconomics is fundamentally flawed because it neglects both social variability and stochastic fluctuations. It may be true that in recent years a greater emphasis has been put on the issue of heterogeneity. Yet, is this the right way to tackle the problem? A model is a simplification of reality anyway, so if it is not tenable to use loosely defined representative agents, an alternative solution may be to focus on sharply defined agent’s

¹Leontief (p. xi in [3]) has even stronger words: “The methods used to maintain intellectual discipline in this country’s most influential economics departments can occasionally remind one of those employed by the Marines to maintain discipline on Parris Island [a training camp of US Marines].”

attitudes. For instance, whereas without further specification home buyers may not be well defined as a useful category, the behavior of investors during the final phases of speculative price peaks may be sufficiently recurrent to make up for a well defined category.

Introduction

What are the main characteristics of econophysics? In what follows we will try to summarize some basic principles. Each of them will be illustrated by one or several studies performed by econophysicists over the past decade. Although the wording that we use is fairly personal, we believe that fundamentally these principles are shared by many econophysicists. In the course of more than a decade, econophysics has become a big tree with many branches. Obviously it is impossible to describe all of them if only because the knowledge and understanding of the present author is limited. He apologizes in advance for his limitations and for the fact that the present selection is by necessity fairly subjective.

The Primacy of Observation

Econophysics started around 1995 in sync with the creation of huge computerized databases giving minute by minute transactions on financial markets such as the New York stock market, the dollar-yen exchange rate, the forward interest rates or providing individual income data for millions of people. It may be estimated that between 1995 and 2005 about two thirds of the papers published by econophysicists aimed at deriving *regularity rules* from such databases. Let us illustrate this point by the case of income data. Since Pareto's work we know that the distribution of high incomes can be described by a power law with an exponent α comprised between 1 and 1.5. With databases comprising millions of income data one can get high accuracy estimates for α and observe how α changes as the result of economic booms or stock market crashes. It turns out that α decreases during booms and increases in the wake of stock market collapses [6].

Other empirical investigations were carried out in the past decades. We list some of them below. The list is arranged by topic and by research teams.

- Stock transactions, (i) Boston University: see publications involving G. Stanley. (ii) CEA (i. e. Commissariat à l'Energie Atomique which means Institute for Atomic Research) and "Science-Finance": see publications involving J.P. Bouchaud. (iii) Nice University and UCLA: see publications involving D. Sornette. (iv) University of Warsaw: see publications involving J. Kertesz.
- Forward interest rates, Singapore University: see publications involving B. Baaquie.
- Exchange rates, Zurich: see publications involving M. Dacorogna.

To many physicists the statement that observation is supreme could seem self evident. In economics, however, such a statement represents a revolution. We already mentioned the fact that observation is a neglected topic in economics. As a matter of fact, before econophysics started it was impossible to publish a paper which would identify *regularity rules* without at the same time providing a model².

Investigating One Effect at a Time

In most natural phenomena different effects occur simultaneously. For instance, if one leaves a glass of cold water in the sun, the water will of course get warmer but if one looks at the mechanisms which are implied this involves many different effects: interaction of light and water, interaction of light and glass, conduction of heat, creation of convection currents between layers of water which are at different temperatures, and so on. One of the main challenges of physics was to identify these effects and to study them separately. Similarly, most social phenomena involve different effects; thus, one of the main tasks of the social sciences should be to disentangle and decompose complex phenomena into simple effects. In principle this is easier to do in physics than in the social sciences because one can change experimental conditions fairly easily. However, history shows that the main obstacle are conceptual. The previous phenomenon involves the transformation of one form of energy (light) into other forms of energy and it is well known that it took centuries for a clear understanding of these processes to emerge. In order to convince the reader that the same approach can be used in the social sciences we briefly describe a specific case.

Suicide is commonly considered as a phenomenon which is due to many factors. One of them is the strength of the marital bond. How can we isolate that factor? Of course, it is impossible to isolate it completely but one can at least make it so predominant that other factors become negligible. To achieve that objective, we consider a population in which the number of males is much larger than

²In what economists call "empirical econometrics" the researcher necessarily must provide a multivariate econometric model which means that even before he analyses the data he already knows the theory which rules the phenomenon. Moreover, all factors whether they have a weak or a strong impact are treated on the same footing. As we will see in the next point this has important implications.

the number of females. Such a population will necessarily have a large proportion of bachelors and therefore will be an ideal testing ground to study the role of the marital bond. Where can we find populations with a large excess of men? Almost all populations of immigrants are characterized by an excess of males. It turns out that due to specific circumstances, this imbalance was particularly large in the population of Chinese people living in the United States. By the end of the 19th century there were about 27 Chinese men for one Chinese woman³.

What makes the present principle important? Unless one is able to estimate the impact of each factor separately, one will never gain a *lasting* understanding. It is important to understand why. Let us for a moment return to the previous experiment. In the econometric approach one would conduct multivariate regressions of the temperature as a function of various (pre-conceived) parameters such as the volume of the liquid, the thickness of the glass and so on. Now suppose we wish to predict what happens when water is replaced by black ink. As a result of greater light absorption temperature differentials will be larger and convection currents will be stronger. The fact that many effects change at the same time will make the multivariate estimates irrelevant. Unless one has an understanding of the various individual effects it will be impossible to make any sound prediction. To sum up, any major change in business and social conditions will invalidate the previously accepted econometric models. This explains why the econometric approach fails to ensure that knowledge grows in a cumulative way.

What Guidance Can Physics Provide?

One can recall that the experimental methodology pioneered by researchers such as Tycho Brahe (1546–1601), Johannes Kepler (1571–1630) or Galileo (1564–1642) marked the beginning of modern physics. Two centuries later, that methodology was adapted to the exploration of the living world by people such as Claude Bernard (1813–1878), Louis Pasteur (1822–1895) and Gregor Mendel (1822–1884). In a sense it is a paradox that this method has been used successfully for the understanding of living organisms but has not yet gained broad acceptance in the social sciences for it can be argued with good reason that living organisms are more complex systems than are states or societies⁴. In short, applying the experimental method-

ology to the social sciences is a move which seems both natural and long overdue. Actually, serious efforts were made in this direction by social scientists such as Emile Durkheim (1858–1917) or Vilfredo Pareto (1848–1923) but this route seems to have been sidetracked in the second half of the 20th century.

Can we use the mathematical framework of physics in the investigation of social phenomena? This approach has been tried with some success by renowned econophysicists such as Belal Baaquie and coworkers (2004, 2007) and Jean-Philippe Bouchaud and coworkers [2,4]. In those cases the success must probably be attributed to the fact that the methods of theoretical physics which were used could be formulated in a purely mathematical way which did not rely on any physical concepts such as energy, momentum or temperature. As we do not yet know how these notions should be transposed to social systems, it seems impossible to apply the formalism of statistical mechanics to social phenomena⁵.

Our claim that the experimental methodology of physics can be used to explore social phenomena must be substantiated by explaining how it is possible to carry out “experiments” in social phenomena. This is the purpose of the next section.

How Cross-National Observations Can Be Used to Test the Role of Different Factors

Nowadays when a solid state physicist wants to measure, say, the interaction between ultraviolet light and a crystal of germanium, the experiment involves little uncertainties. That is so because this field of physics is already well understood. On the contrary, in the case of new and not well understood phenomena there is considerable uncertainty about the specific conditions of the experimental set up. In the two years after Léon Foucault demonstrated the Foucault pendulum experiment, at least twenty physicists tried to repeat it. Some succeeded while others did not. Indeed the experimental conditions, e. g. the length of the pendulum or the nature of the suspension wire, ensuring that the Foucault effect will be observed were not well understood.

societies appeared less than 100,000 years ago and states less than 10,000 years ago.

⁵It could be argued that one is free to define “social energy” in the way which one wishes. However, one should remember that the notion of energy is pivotal in physics only because it is ruled by (experimentally proved) conservation laws, such as the equivalence between heat and mechanical energy demonstrated by James Joule. Naturally, prior to defining a “social temperature”, it would seem natural to define a herd- or swarm-temperature describing aggregated populations of bacteria, insects or animals. As far as we know, no operational definition of this kind has yet been proposed.

³For more details about this case, see [9].

⁴We will not develop this point here but it can be observed that a bacteria or a cell contains thousands of different proteins which interact in various ways. In the same line of thought one may recall that living organisms have been around for several billions years whereas

It is only through various attempts with different settings that a better understanding progressively emerged. For instance it was realized that by using a pendulum of great length one would be able to reduce two undesirable effects (i) the sensitivity of the pendulum to exogenous noise⁶ (ii) the Puisseux effect which generates a rotation of the oscillation plane which interferes with the Foucault effect.

Few (if any) sociological phenomena are well understood which means that social researchers are basically in the same situation as those physicists in the years 1851–1852 who tried to observe the Foucault effect⁷. As an illustration suppose we wish to know if the publication of a specific type of news has an effect on the number of suicides⁸. Such an observation depends upon many parameters: the nature of the news and the amount of attention that it receives, the time interval (days, weeks or months?) between the publication of the news and the occurrence of the suicides. In addition one does not know if there will be an increase or a decrease in the number of suicides, if men will be more or less affected than women, and so on. All these questions can in principle be answered by conducting many observations in different countries and in different periods of time. In other words, if we are sufficiently determined, patient and tenacious and if we can get access to the statistical data that are needed, we should be able to disentangle and unravel the phenomenon under consideration in the same way as experimenters have been able to determine how the Foucault effect can be observed.

How Vested Interests May Affect the Accessibility and Reliability of Social Data

So far we have emphasized the similarities between natural and social phenomena but there are also some stumbling blocks which are specific to the social sciences. One of them is the fact that some data may have been altered or swept under the carpet by some sort of ideological, political or social bias, pressure or interference. Needless to say, extreme care must be exercised in such cases before making use of the data.

As an illustration, suppose that an econophysicist or a sociologist wants to study episodes of military occupation of one country by another. Such episodes are of

particular interest from a sociological perspective because they bring about strong interactions and can serve to probe the characteristics of a society. Moreover, because armies display many similarities no matter their country of origin, such episodes offer a set of *controlled experiments*. Naturally, in order to be meaningful the comparison must rely on trustworthy accounts for each of the episodes. Unfortunately, it turns out that in many cases only scant and fairly unreliable information is available. Consider for instance the occupation of Iceland by British and American forces during World War II. Among all occupation episodes this one was particularly massive with troops representing 50% of the population of Iceland prior to the occupation. The same proportion in a country such as Japan would have meant 30 million occupation troops that is 60 times more than the peak number of 500,000 reached at the end of 1945. Quite understandably for such a high density of troops, there were many incidents with the population of Iceland⁹; yet, it is difficult to find detailed evidence. Due to the paucity of data a superficial investigation would easily lead to the conclusion that there were in fact only few incidents. It does not require much imagination to understand why this information has not been released. The fact that in a general way all countries whatsoever are reluctant to recognize possible misconduct of their military personnel explains why the information is still classified in British and American archives. Because Iceland and the United States became close allies after 1945, one can also understand that the Icelandic National Archive is reluctant to release information about these incidents. The same observation also applies (and for the same reasons) to the occupation of Japan, 1946–1951; for more details see Roehner pp. 90–98 in [9] and [10]. Naturally, similar cases abound. Due to a variety of reasons well-meaning governments, archivists and statistical offices keep sensitive files closed to social scientists. Most often it is in fact sufficient to catalog sensitive file units in a fairly obscure way. The plain effect is that the information will not be found except perhaps by pure luck, a fairly unlikely prospect in big archives.

How Can Exogenous Factors be Taken into Account?

This question is not specific to social phenomena, it is also of importance in physics. As a matter of fact, in astronomy

⁶Indeed, it is when the speed of the pendulum goes through zero that it is particularly sensitive to external perturbations; increasing the length of the pendulum reduces the number of oscillations in a given time interval and therefore the drift due to noise.

⁷As a more recent and even less understood case, one can mention the physicists who keep on trying to observe the cold fusion effect.

⁸This question is connected to what is known in sociology as the Werther effect; for more details see the papers written by Phillips (in particular [7]) and Chap. 3 in [9].

⁹According to a report that Prime Minister Hermann Jonasson sent to the American Headquarters, there were 136 incidents between troops and Icelanders during the period between July 1941 (arrival of the American troops) and April 1942 (Hunt 1966) in Reykjavik alone. Unfortunately, no copy of this report seems to be available at the National Archives of Iceland.

it provides a powerful method for observing objects that cannot be observed directly. Thus, we know the existence of exoplanets only from the perturbing effect which they have on the position of the star around which they move. However, for social phenomena the problem of exogenous factors is much more serious because (i) they may not be known to observers (ii) even once they are identified it is very difficult to correct the data in a reliable way. One of the main pitfalls in the modeling of socio-economic phenomena is to explain them through endogenous mechanisms while they are in fact due to exogenous factors. The following examples make clear that this difficulty exists for many phenomena, whether they belong to the financial, economic or social sphere.

- In their paper of 2005 about consensus formation and shifts in opinion Michard and Bouchaud confront their theory to two classes of social phenomena: (i) the diffusion of cell phones (ii) the diffusion of birth rate patterns. In the first case it is clear that advertising campaigns may have played an important role. Of course, one could argue that these campaigns were part of the endogenous diffusion process. However, this argument does not hold for big telecom companies (e.g. Vodafone) which operate in many countries. In such cases the decision about the magnitude of the advertising campaigns are taken by the board of the company which means that such campaigns can hardly be considered as endogenous effects. Similarly, birth rates depend upon exogenous factors. For instance the length of time spent in higher education has an effect on the average age of marriage and the later has an effect on birth rates.
- On 21 July 2004 the share price of Converium, a Swiss reinsurance company listed on the New York Stock Exchange dropped 50%. Was this fall the result of an avalanche effect due to a movement of panic among investors? In fact, the most likely explanation is that it was the consequence of a decision taken by the board of Fidelity International, a major investment fund and one of the main shareholders of Converium. Indeed in a statement issued by Converium on August 3, 2004 it was announced that Fidelity had reduced its holdings from 9.87% to 3.81%. In other words, it would be completely irrelevant to explain such a fall through a herd effect model or through any other endogenous mechanism (more details can be found in [8]). Similar conclusions apply to corporate stock buybacks, as well as to mergers, acquisitions, buyouts and takeovers; in all these cases decisions taken by a few persons (the average board of directors has nine members) may trigger substantial changes in share prices. How should such effects be taken into account by stock market models?
- At the end of 2004 and in the first months of 2005 British housing prices began to decline after having risen rapidly during several years. Yet after May 2005, they suddenly began to pick up again at an annual rate of about 10%. This resurgence was particularly intriguing because at the same time US housing prices began to decline. To what factor should this unexpected rise be attributed? Most certainly this was the market response to a plan introduced by the Chancellor of the Exchequer Gordon Brown in late May (The Economist May 28, 2005). Under this plan which aimed at propping up house prices new buyers would benefit from a zero-interest loan for 12% of the price. In addition, the government would cover all losses incurred by banks as a result of possible bankruptcies of borrowers (at least so long as prices did not fall by more than 12%). It appears that the plan indeed propped up the market. Consequently, in order to confront the predictions of any model (e.g. see Richmond's paper which was published in 2007) with observation the impact of this plan effect must first be taken out of the picture.
- The same difficulty is also encountered in socio-political phenomena. Here is an illustration. On 5 October 2000, in protest against the publication of the results of the presidential election there was a huge mass demonstration in Belgrade which involved thousands of people from the provinces who were transported to the federal capital by hundreds of buses. It clearly showed that president Milosevic was no longer in control of the police and army and lead to his retirement from the political scene. Thus, what NATO air strikes (24 March-11 June 1999¹⁰) had not been able to achieve was accomplished by one night of street demonstrations. What was the part of exogenous factors in this event? Although in many similar cases it is very difficult to know what really happened, in this specific case a partial understanding is provided by a long article published in the New York Times¹¹. In this article

¹⁰It can be noted that similarly to what would happen in 2003 for the invasion of Iraq, these air strikes were carried out without the authorization of the United Nations Security Council.

¹¹New York Times, Sunday 26 November 2000, Magazine Section, p. 43, 7705 words; the article by Roger Cohen is entitled: "Who really brought down Milosevic". What makes this account particularly convincing is the fact that it was preceded by another article entitled: "US anti-Milosevic plan faces major test at polls" which appeared on September 23, 2000 (p. 6, 1150 words); this article described the way Milosevic would be removed from power two weeks *before* the events. The article makes clear that the course of events would be the same no matter what the results of the election would be.

we learn that several American organizations belonging to the intelligence network supported, financed and trained Serbian opposition groups. For instance the article mentions the Albert Einstein Foundation, the International Republican Institute, the National Endowment for Democracy, the US Agency for International Development. Although the amount of the total financial support is not known, the New York Times article says that it exceeded \$ 28 million. The plan comprised two facets: the organization of demonstrations on the one hand and the infiltration of the army and police on the other hand in order to undermine their loyalty and convince them to remain passive during the demonstrations. According to the article this second facet remains classified. With an exogenous interference of such a magnitude, it would clearly be meaningless to describe this upheaval as a purely endogenous process. Moreover, the fact that we have only partial knowledge about the exogenous forces makes it very difficult (if not altogether impossible) to come up with a satisfactory description. It should also be noted that the influence of these groups did not disappear overnight after October 4, which means that the subsequent history of Serbia must also take them into account at least to some extent.

Future Directions

In this article we have described the challenges and obstacles to which one is confronted in trying to understand socio-economic phenomena. In parallel we have shown that the econophysics approach has many assets. One of them which has not yet been mentioned is the fact that econophysicists are not subject to the rigid barriers which exist between various fields and subfields of the human sciences. Thus, if it turns out that in order to explain an economic phenomena one needs to understand a social effect, econophysicists would have no problem in shifting from one field to another. There is another historical chance that we have not mentioned so far, namely the development of the Internet. In the past decade 1997–2007 the amount of information to which one has access has increased tremendously. Electronic catalogs of major libraries or of national archives, indexes of newspaper, search engines on the Internet, searchable databases of books, all these innovations contributed to give the researcher easy access to information sources that have never been available before. In particular it has become fairly easy to find cross-national data. Thus, social scientists and econophysicists are in a better position than ever for carrying out the kind of comparative studies that we called for in this article.

Bibliography

Primary Literature

1. Aoki M, Yoshikawa H (2007) *Reconstructing macroeconomics*. Cambridge University Press, Cambridge
2. Bouchaud J-P, Potters M (2003) *Theory of financial risk and derivative pricing*. Cambridge University Press, Cambridge
3. Leontief W (1983) Foreword. In: Eichner AS (ed) *Why economics is not yet a science*. M.E. Sharpe, Armonk(New York)
4. Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. *Eur. Phys. J. B* 47:151–159
5. Neumann J von, Morgenstern O (1953) *Theory of games and economic behavior*. Princeton University Press, Princeton
6. Nirei M, Souma W (2007) Two factor model of income distribution dynamics. *Review of Income and Wealth* 53(3):440–459
7. Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. *Am. Sociol. Rev.* 39:340–354
8. Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) *Proceedings of the 3rd Nikkei Economics Symposium*, Tokyo. Springer, Tokyo, pp 262–271
9. Roehner BM (2007) *Driving forces in physical, biological and socio-economic phenomena*. Cambridge University Press, Cambridge
10. Roehner BM (2008) *Relations between Allied forces and the population of Japan*, Working Report UPMC, Paris
11. Schumpeter J (1933) The common sense of econometrics. *Econometrica* 1:5–12
12. Schwartz AJ (1995) An interview with Anna J. Schwartz. *NewsL. Cliometric Soc.* 10(2):3–7

Books and Reviews

Two observations are in order about this reference section:

Many of these references are not mentioned in the text; the objective is to give readers a starting point for further readings on various aspects of econophysics.

There is a fairly complete list of publications of the present author; it is given for the purpose of illustrating through one specific case the “trajectory” of an econophysicist in the course of time (1995–2007).

- Amaral LAN, Buldyrev SV, Havlin S, Leschhorn H, Maass P, Salinger A, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: I. Empirical results for company growth. *J Phys. I Fr.* 7:621–633
- Amaral LAN, Buldyrev SV, Havlin S, Salinger MA, Stanley HE (1998) Power law scaling for a system of interacting units with complex internal structure. *Phys. Rev. Lett.* 80(7):1385–1388
- Aoki M, Yoshikawa H (2007) *Reconstructing macroeconomics*. Cambridge University Press, Cambridge
- Baaquie BE (2004) *Quantum finance*. Cambridge University Press, Cambridge
- Baaquie BE (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. I. *Theory Phys. Rev. E* 75, 016703
- Baaquie BE, Liang C (2007) Feynman perturbation expansion for the price of coupon bond options and swaptions in quantum finance. II. *Empirical Phys. Rev. E* 75, 016704

- Baaquie BE, Srikant M (2004) Comparison of field theory models of interest rates with market data. *Phys. Rev. E* 69, 036129
- Borghesi C, Bouchaud J-P (2007) On songs and men. *Quality and Quantity* 41(4):557–568
- Bouchaud J-P, Marsili M, Roehner BM, Slanina F (eds) (2001) Application of physics in economic modelling. Proceedings of the NATO Advanced Research Workshop held in Prague, Czech Republic, 8–10 February 2001. *Physica A* 299(1–2):1–355
- Bouchaud J-P, Potters M (1997) *Théorie des risques financiers*. Aléa, Saclay
- Bouchaud J-P, Potters M (2003) *Theory of financial risk and derivative pricing*. Cambridge University Press, Cambridge
- Buldyrev SV, Amaral LAN, Havlin S, Leschhorn H, Maass P, Salinger MA, Stanley HE, Stanley MHR (1997) Scaling behavior in economics: II. Modeling of company growth. *J Phys. I Fr.* 7:635–650
- Chakraborti A, Chakraborti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. *Eur. Phys. J. B* 17:167–170
- de Oliveira SM, de Oliveira PMC, Stauffer D (1999) *Evolution, money, war and computer*. Teubner, Leipzig
- Deschâtres F, Sornette D (2005) The dynamics of book sales: endogenous versus exogenous shocks in complex networks. *Phys. Rev. E* 72, 016112
- Dragulescu A, Yakovenko VM (2000) Statistical mechanics of money. *Eur. Phys. J. B* 17(4):723–729
- Farmer JD (1999) Physicists attempt to scale the ivory towers of finance. *Comput. Sci. Eng.* Nov-Dec 1999, 26–39
- Farmer JD, Lillo F (2004) On the origin of power law tails in price fluctuations. *Quant. Finance* 4(1):7–11
- Feigenbaum JA, Freund PGO (1998) Discrete scale invariance and the second Black Monday. *Mod. Phys. Lett. B* 12(2–3):57–60
- Fu Y-Q, Zhang H, Cao Z, Zheng B, Hu G (2005) Removal of pinned spiral by generating target waves with a localized stimulus. *Phys. Rev. E* 72, 046206
- Galam S (2006) Opinion dynamics, minority spreading and heterogeneous beliefs. In: Chakraborti BK, Chakraborti A, Chatterjee A (eds) *Econophysics and Sociophysics*. Wiley-VCH, Weinheim
- Ghashghaie S, Breymann W, Peinke J, Talkner P, Dodge Y (1996) Turbulent cascades in foreign exchange markets. *Nature* 381:767–770
- Guillaume DM, Dacorogna MM, Davé R, Müller UA, Olsen RB, Pictet OV (1997) From the bird's eye to the microscope: a survey of new stylized facts of the intra-daily foreign exchange markets. *Finance Stoch.* 1:95–129
- Hunt JJ (1966) *The United States occupation of Iceland, 1941–1946*. Thesis. Georgetown University, Washington DC
- Johansen A, Sornette D (1999) Financial anti-bubbles: Log-periodicity in gold and Nikkei collapses. *Int. J. Mod Phys C* 10(4):563–575
- Johansen A, Sornette D (2001) Bubbles and anti-bubbles in Latin-American, Asian and Western stock markets: An empirical study. *Int. J. Theor. Appl. Finance* 4(6):853–920
- Juglar C (1862): *Des crises commerciales et de leur retour périodique en France, en Angleterre et aux Etats-Unis*. English translation (1893, 1966) *A brief history of panics and their periodical occurrence in the United States*. A.M. Kelley, New York
- Lai KK, Leung FKN, Tao B, Wang S (2000) Practices of preventive maintenance and replacement for engines: a case study. *Eur J Oper Res* 124:2
- Leontief W (1983) Foreword. In: Eichner AS (ed) *Why economics is not yet a science*. M.E. Sharpe, Armonk New York
- Li M, Wu J, Wang D, Zhou T, Di Z, Fan Y (2006) Evolving model of weighted networks inspired by scientific collaboration networks. *Physica A* 375(1):355–364
- Lillo F, Mike S, Farmer JD (2005) Theory for Long Memory in supply and demand. *Physical Review E* 7106 (6 pt 2) 287–297
- Lux T (1996) The stable Paretian hypothesis and the frequency of large returns: an examination of major German stocks. *Appl Financial Econ* 6:463–475
- Mandelbrot B (1997) Les fractales et la Bourse. *Pour Sci* 242:16–17
- Mantegna RN (1999) Hierarchical structure in financial markets. *Eur Phys J B* 11:193–197
- Mantegna RN, Stanley HE (1995) Scaling behavior in the dynamics of an economic index. *Nature* 376:46–49
- Mantegna RN, Stanley HE (1999) *Introduction to econophysics*. Cambridge University Press, Cambridge
- McCauley JL (2004) *Dynamics of markets*. Cambridge University Press, Cambridge
- Michard Q, Bouchaud J-P (2005) Theory of collective opinion shifts: from smooth trends to abrupt swings. *Eur Phys J B* 47:151–159
- Mimkes J (2006) A thermodynamic formulation of social science. In: Chakraborti BK, Chatterjee A (eds) *Econophysics and sociophysics: trends and perspectives*. Wiley-VCH, Weinheim, pp 279–310
- Müller UA, Dacorogna MM, Davé R, Olsen RB, Pictet OV, Weizsäcker J von (1997) Volatilities of different time resolutions. Analysing the dynamics of market components. *J Empir Finance* 4(2–3):213–240
- Müller UA, Dacorogna MM, Olsen RB, Pictet OV, Schwarz M (1990) Statistical study of foreign exchange rates, empirical evidence of a price scaling law, and intraday analysis. *J Bank Finance* 14:1189–1208
- Neumann J von, Morgenstern O (1953) *Theory of games and economic behavior*. Princeton University Press, Princeton
- Phillips DP (1974) The influence of suggestion on suicide: substantive and theoretical implications of the Werther effect. *Am Sociol Rev* 39:340–354
- Plerou V, Amaral LAN, Gopikrishnan P (1999) Similarities between the growth dynamics of university research and of competitive economic activities. *Nature* 400(6743):433–437
- Plerou V, Gopikrishnan P, Rosenow B, Amaral LA, Stanley H (1999) Universal and nonuniversal properties of cross-correlation in financial time series. *Phys Rev Lett* 83(7):1471–1474
- Richmond P (2007) A roof over your head; house price peaks in the UK and Ireland. *Physica A* 375(1,15):281–287
- Roehner BM (1995) *Theory of markets. Trade and space-time patterns of price fluctuations: a study in analytical economics*. Springer, Berlin
- Roehner BM (1997) Jesuits and the state. A comparative study of their expulsions (1500–1990). *Religion* 27:165–182
- Roehner BM (1997) The comparative way in economics: a reappraisal. *Econom Appl* 50(4):7–32
- Roehner BM (1999) Spatial analysis of real estate price bubbles: Paris 1984–1993. *Reg Sci Urban Econ* 29:73–88
- Roehner BM (1999) The space-time pattern of price waves. *Eur Phys J B* 8:151–159
- Roehner BM (2000) Determining bottom price-levels after a speculative peak. *Eur Phys J B* 17:341–345
- Roehner BM (2000) Identifying the bottom line after a stock market crash. *Int J Mod Phys C* 11(1):91–100

- Roehner BM (2000) Speculative trading: the price multiplier effect. *Eur Phys J B* 14:395–399
- Roehner BM (2000) The correlation length of commodity markets: 1. Empirical evidence. *Eur Phys J B* 13:175–187
- Roehner BM (2000) The correlation length of commodity markets: 2. Theoretical framework. *Eur Phys J B* 13:189–200
- Roehner BM (2001) Hidden collective factors in speculative trading: a study in analytical economics. Springer, Berlin
- Roehner BM (2001) To sell or not to sell? Behavior of shareholders during price collapses. *Int J Mod Phys C* 12(1):43–53
- Roehner BM (2001) Two classes of speculative peaks. *Physica A* 299:71–83
- Roehner BM (2002) Patterns of speculation: a study in observational econophysics. Cambridge University Press, Cambridge
- Roehner BM (2002) Patterns and repertoire. Harvard University Press, Cambridge Massachussets
- Roehner BM (2002) Separatism and integration. Rowman and Littlefield, Lanham Maryland
- Roehner BM (2004) Patterns of speculation in real estate and stocks. In: Takayasu H (ed) Proceedings of the 2nd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 103–116
- Roehner BM (2005) A bridge between liquids and socio-economic systems: the key-role of interaction strengths. *Physica A* 348:659–682
- Roehner BM (2005) Cohésion sociale. Odile Jacob, Paris
- Roehner BM (2005) Stock markets are not what we think they are: the key roles of cross-ownership and corporate treasury stock. *Physica A* 347:613–626
- Roehner BM (2006) Macroplayers in stock markets. In: Takayasu H (ed) Proceedings of the 3rd Nikkei Economics Symposium, Tokyo. Springer, Tokyo, pp 262–271
- Roehner BM (2006) Real estate price peaks: a comparative perspective. *Evol Institutional Econ Rev* 2(2):167–182
- Roehner BM (2007) Driving forces in physical, biological and socio-economic phenomena. Cambridge University Press, Cambridge
- Roehner BM (2008) Econophysics: Challenges and promises. *Evolutionary and Institutional Economics Review* 4(2):251–266
- Roehner BM, Jegu C (2006) White flight or flight from poverty? *J Econ Interact Coord* 1:75–87
- Roehner BM, Maslov S (2003) Does the price multiplier effect also hold for stocks? *Int J Mod Phys C* 14(10):1439–1451
- Roehner BM, Maslov S (2003) The conundrum of stock versus bond prices. *Physica A* 335:164–182 (2004)
- Roehner BM, Rahilly LJ (2002) Separatism and integration: a study in analytical history. Rowman and Littlefield, Lanham Maryland
- Roehner BM, Shiue C (2001) Comparing the correlation length of grain markets in China and France. *Int J Mod Phys C* 11(7):1383–1410
- Roehner BM, Sornette D (1998) The sharp peak – flat trough pattern and critical speculation. *Eur Phys J B* 4:387–399
- Roehner BM, Sornette D (1999) Analysis of the phenomenon of speculative trading in one of its basic manifestations: postage stamp bubbles. *Int J Mod Phys C* 10(6):1099–1116
- Roehner BM, Sornette D (2000) Thermometers of speculative frenzy. *Eur Phys J B* 16:729–739
- Roehner BM, Sornette D, Andersen J (2004) Response functions to critical shocks in social sciences: an empirical and numerical study. *Int J Mod Phys C* 15(6):809–834
- Roehner BM, Syme T (2002) Pattern and repertoire in history: an introduction to analytical history. Harvard University Press, Cambridge Massachussets
- Schumpeter J (1933) The common sense of econometrics. *Econometrica* 1:5–12
- Schwartz AJ (1995) An interview with Anna J Schwartz. *News Econometric Soc* 10(2):3–7
- Sornette D (2003) Why stock markets crash. Critical events in complex financial systems. Princeton University Press, Princeton
- Stauffer D, Sornette D (1999) Self-organized percolation model for stock market fluctuations. *Physica A* 271(3–4):496–506
- Summers LH (1991) The scientific illusion in empirical macroeconomics. *Scand J Econ* 93(2):129–148
- Takayasu H (ed) (2004) The application of econophysics. In: Proceedings of the 2nd Nikkei Econophysics Symposium. Springer, Tokyo
- Takayasu H (ed) (2006) Practical fruits of econophysics. In: Proceedings of the 3rd Nikkei Econophysics Symposium. Springer, Tokyo
- Turchin P (2003) Historical dynamics. Why states rise and fall. Princeton University Press, Princeton
- Wyatt M, Bouchaud J-P (2003) Self referential behaviour, overreaction and conventions in financial markets. *Conduct* 03033584
- Zhou W-X, Sornette D (2003) 2000–2003 real estate bubble in the UK and not in the USA. *Physica A* 329(1–2):249–263
- Zhou W-X, Sornette D (2003) Evidence of a worldwide stock market log-periodic anti-bubble since mid-2000. *Physica A* 330:543–583
- Zhou W-X, Sornette D (2004) Antibubble and Prediction of China's stock market and Real-Estate. *Physica A* 337(1–2):243–268

Econophysics, Statistical Mechanics Approach to

VICTOR M. YAKOVENKO

Department of Physics, University of Maryland,
College Park, USA

“Money, it’s a gas.” Pink Floyd

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Historical Introduction](#)

[Statistical Mechanics of Money Distribution](#)

[Statistical Mechanics of Wealth Distribution](#)

[Data and Models for Income Distribution](#)

[Other Applications of Statistical Physics](#)

[Future Directions, Criticism, and Conclusions](#)

[Bibliography](#)

Glossary

Probability density $P(x)$ is defined so that the probability of finding a random variable x in the interval from x to $x + dx$ is equal to $P(x) dx$.

Cumulative probability $C(x)$ is defined as the integral $C(x) = \int_x^\infty P(x) dx$. It gives the probability that the random variable exceeds a given value x .

The Boltzmann–Gibbs distribution gives the probability of finding a physical system in a state with the energy ε . Its probability density is given by the exponential function (1).

The Gamma distribution has the probability density given by a product of an exponential function and a power-law function, as in (9).

The Pareto distribution has the probability density $P(x) \propto 1/x^{1+\alpha}$ and the cumulative probability $C(x) \propto 1/x^\alpha$ given by a power law. These expressions apply only for high enough values of x and do not apply for $x \rightarrow 0$.

The Lorenz curve was introduced by American economist Max Lorenz to describe income and wealth inequality. It is defined in terms of two coordinates $x(r)$ and $y(r)$ given by (19). The horizontal coordinate $x(r)$ is the fraction of the population with income below r , and the vertical coordinate $y(r)$ is the fraction of income this population accounts for. As r changes from 0 to ∞ , x and y change from 0 to 1, parametrically defining a curve in the (x, y) -plane.

The Gini coefficient G was introduced by the Italian statistician Corrado Gini as a measure of inequality in a society. It is defined as the area between the Lorenz curve and the straight diagonal line, divided by the area of the triangle beneath the diagonal line. For perfect equality (everybody has the same income or wealth) $G = 0$, and for total inequality (one person has all income or wealth, and the rest have nothing) $G = 1$.

The Fokker–Planck equation is the partial differential equation (22) that describes evolution in time t of the probability density $P(r, t)$ of a random variable r experiencing small random changes Δr during short time intervals Δt . It is also known in mathematical literature as the Kolmogorov forward equation. The diffusion equation is an example of the Fokker–Planck equation.

Definition of the Subject

Econophysics is an interdisciplinary research field applying methods of statistical physics to problems in economics and finance. The term “econophysics” was first introduced by the prominent theoretical physicist Eu-

gene Stanley in 1995 at the conference *Dynamics of Complex Systems*, which was held in Calcutta (now known as Kolkata) as a satellite meeting to the STATPHYS-19 conference in China [1,2]. The term appeared in print for the first time in the paper by Stanley et al. [3] in the proceedings of the Calcutta conference. The paper presented a manifesto of the new field, arguing that “behavior of large numbers of humans (as measured, e.g., by economic indices) might conform to analogs of the scaling laws that have proved useful in describing systems composed of large numbers of inanimate objects” [3]. Soon the first econophysics conferences were organized: *International Workshop on Econophysics*, Budapest, 1997 and *International Workshop on Econophysics and Statistical Finance*, Palermo, 1998 [2], and the book *An Introduction to Econophysics* [4] was published.

The term “econophysics” was introduced by analogy with similar terms, such as “astrophysics”, “geophysics”, and “biophysics”, which describe applications of physics to different fields. Particularly important is the parallel with biophysics, which studies living creatures, which still obey the laws of physics. It should be emphasized that econophysics does not literally apply the laws of physics, such as Newton’s laws or quantum mechanics, to humans, but rather uses mathematical methods developed in statistical physics to study statistical properties of complex economic systems consisting of a large number of humans. So, it may be considered as a branch of applied theory of probabilities. However, statistical physics is distinctly different from mathematical statistics in its focus, methods, and results.

Originating from physics as a quantitative science, econophysics emphasizes quantitative analysis of large amounts of economic and financial data, which became increasingly available with the massive introduction of computers and the Internet. Econophysics distances itself from the verbose, narrative, and ideological style of political economy and is closer to **econometrics** in its focus. Studying mathematical models of a large number of interacting economic agents, econophysics has much common ground with the **agent-based modeling and simulation**. Correspondingly, it distances itself from the representative-agent approach of traditional economics, which, by definition, ignores statistical and heterogeneous aspects of the economy.

Two major directions in econophysics are applications to finance and economics. Observational aspects are covered in the article ► [Econophysics, Observational](#). The present article, ► [Econophysics, Statistical Mechanics Approach to](#), concentrates primarily on statistical distributions of money, wealth, and income among interacting economic agents.

Another direction related to econophysics has been advocated by the theoretical physicist Serge Galam since the early 1980s under the name “**sociophysics**” [5], with the first appearance of the term in print in [6]. It echoes the term *physique sociale* proposed in the nineteenth century by Auguste Comte, the founder of sociology. Unlike econophysics, the term “sociophysics” did not catch on when first introduced, but it is coming back with the popularity of econophysics and active promotion by some physicists [7,8,9]. While the principles of both fields have a lot in common, econophysics focuses on the narrower subject of economic behavior of humans, where more quantitative data are available, whereas sociophysics studies a broader range of social issues. The boundary between econophysics and sociophysics is not sharp, and the two fields enjoy a good rapport [10]. A more detailed description of the historical development is presented in Sect. “[Historical Introduction](#)”.

Historical Introduction

Statistical mechanics was developed in the second half of the nineteenth century by James Clerk Maxwell, Ludwig Boltzmann, and Josiah Willard Gibbs. These physicists believed in the existence of atoms and developed mathematical methods for describing their statistical properties, such as the probability distribution of velocities of molecules in a gas (the Maxwell–Boltzmann distribution) and the general probability distribution of states with different energies (the Boltzmann–Gibbs distribution). There are interesting connections between the development of statistical physics and statistics of social phenomena, which were recently brought up by the science journalist Philip Ball [11,12].

Collection and study of “social numbers”, such as the rates of death, birth, and marriage, has been growing progressively since the seventeenth century (see Chap. 3 in [12]). The term “statistics” was introduced in the eighteenth century to denote these studies dealing with the civil “states”, and its practitioners were called “statists”. Popularization of social statistics in the nineteenth century is particularly accredited to the Belgian astronomer Adolphe Quetelet. Before the 1850s, statistics was considered an empirical arm of political economy, but then it started to transform into a general method of quantitative analysis suitable for all disciplines. It stimulated physicists to develop statistical mechanics in the second half of the nineteenth century.

Rudolf Clausius started development of the kinetic theory of gases, but it was James Clerk Maxwell who made a decisive step of deriving the probability distribu-

tion of velocities of molecules in a gas. Historical studies show (see Chap. 3 in [12]) that, in developing statistical mechanics, Maxwell was strongly influenced and encouraged by the widespread popularity of social statistics at the time. This approach was further developed by Ludwig Boltzmann, who was very explicit about its origins (see p. 69 in [12]):

“The molecules are like individuals, ... and the properties of gases only remain unaltered, because the number of these molecules, which on the average have a given state, is constant.”

In his book *Populäre Schriften* from 1905 [13], Boltzmann praises Josiah Willard Gibbs for systematic development of statistical mechanics. Then, Boltzmann says (cited from [14]):

“This opens a broad perspective if we do not only think of mechanical objects. Let’s consider to apply this method to the statistics of living beings, society, sociology and so forth.”

(The author is grateful to Michael E. Fisher for bringing this quote to his attention.)

It is worth noting that many now-famous economists were originally educated in physics and engineering. Vilfredo Pareto earned a degree in mathematical sciences and a doctorate in engineering. Working as a civil engineer, he collected statistics demonstrating that distributions of income and wealth in a society follow a power law [15]. He later became a professor of economics at Lausanne, where he replaced Léon Walras, also an engineer by education. The influential American economist Irving Fisher was a student of Gibbs. However, most of the mathematical apparatus transferred to economics from physics was that of Newtonian mechanics and classical thermodynamics [16]. It culminated in the neoclassical concept of mechanistic equilibrium where the “forces” of supply and demand balance each other. The more general concept of statistical equilibrium largely eluded mainstream economics.

With time, both physics and economics became more formal and rigid in their specializations, and the social origin of statistical physics was forgotten. The situation is well summarized by Philip Ball (see p. 69 in [12]):

“Today physicists regard the application of statistical mechanics to social phenomena as a new and risky venture. Few, it seems, recall how the process originated the other way around, in the days when physical science and social science were the twin siblings of a mechanistic philosophy and when it was

not in the least disreputable to invoke the habits of people to explain the habits of inanimate particles”.

Some physicists and economists attempted to connect the two disciplines during the twentieth century. The theoretical physicist Ettore Majorana argued in favor of applying the laws of statistical physics to social phenomena in a paper published after his mysterious disappearance [17]. The statistical physicist Elliott Montroll co-authored the book *Introduction to Quantitative Aspects of Social Phenomena* [18]. Several economists applied statistical physics to economic problems [19,20,21,22]. An early attempt to bring together the leading theoretical physicists and economists at the Santa Fe Institute was not entirely successful [23]. However, by the late 1990s, the attempts to apply statistical physics to social phenomena finally coalesced into the robust movements of econophysics and sociophysics, as described in Sect. “Definition of the Subject”.

The current standing of econophysics within the physics and economics communities is mixed. Although an entry on econophysics has appeared in the *New Palgrave Dictionary of Economics* [24], it is fair to say that econophysics is not accepted yet by mainstream economics. Nevertheless, a number of open-minded, nontraditional economists have joined this movement, and the number is growing. Under these circumstances, econophysicists have most of their papers published in physics journals. The journal *Physica A: Statistical Mechanics and Its Applications* emerged as the leader in econophysics publications and has even attracted submissions from some bona fide economists. The mainstream physics community is generally sympathetic to econophysics, although it is not uncommon for econophysics papers to be rejected by *Physical Review Letters* on the grounds that “it is not physics”. There are regular conferences on econophysics, such as *Applications of Physics in Financial Analysis* (sponsored by the European Physical Society), *Nikkei Econophysics Symposium*, and *Econophysics Colloquium*. Econophysics sessions are included in the annual meetings of physical societies and statistical physics conferences. The overlap with economics is the strongest in the field of agent-based simulation. Not surprisingly, the conference series WEHIA/ESHIA, which deals with heterogeneous interacting agents, regularly includes sessions on econophysics.

Statistical Mechanics of Money Distribution

When modern econophysics started in the middle of the 1990s, its attention was primarily focused on analysis of financial markets. However, three influential pa-

pers [25,26,27], dealing with the subject of money and wealth distributions, were published in 2000. They started a new direction that is closer to economics than finance and created an expanding wave of follow-up publications. We start reviewing this subject with [25], whose results are the most closely related to the traditional statistical mechanics in physics.

The Boltzmann–Gibbs Distribution of Energy

The fundamental law of equilibrium statistical mechanics is the Boltzmann–Gibbs distribution. It states that the probability $P(\varepsilon)$ of finding a physical system or subsystem in a state with the energy ε is given by the exponential function

$$P(\varepsilon) = ce^{\frac{-\varepsilon}{T}}, \quad (1)$$

where T is the temperature, and c is a normalizing constant [28]. Here we set the Boltzmann constant k_B to unity by choosing the energy units for measuring the physical temperature T . Then, the expectation value of any physical variable x can be obtained as

$$\langle x \rangle = \frac{\sum_k x_k e^{\frac{-\varepsilon_k}{T}}}{\sum_k e^{\frac{-\varepsilon_k}{T}}}, \quad (2)$$

where the sum is taken over all states of the system. Temperature is equal to the average energy per particle: $T \sim \langle \varepsilon \rangle$, up to a numerical coefficient of the order of 1.

Equation (1) can be derived in different ways [28]. All derivations involve the two main ingredients: statistical character of the system and conservation of energy ε . One of the shortest derivations can be summarized as follows. Let us divide the system into two (generally unequal) parts. Then, the total energy is the sum of the parts, $\varepsilon = \varepsilon_1 + \varepsilon_2$, whereas the probability is the product of probabilities, $P(\varepsilon) = P(\varepsilon_1)P(\varepsilon_2)$. The only solution of these two equations is the exponential function (1).

A more sophisticated derivation, proposed by Boltzmann himself, uses the concept of entropy. Let us consider N particles with total energy E . Let us divide the energy axis into small intervals (bins) of width $\Delta\varepsilon$ and count the number of particles N_k having energies from ε_k to $\varepsilon_k + \Delta\varepsilon$. The ratio $N_k/N = P_k$ gives the probability for a particle having the energy ε_k . Let us now calculate the multiplicity W , which is the number of permutations of the particles between different energy bins such that the occupation numbers of the bins do not change. This quantity is given by the combinatorial formula in terms of the factorials

$$W = \frac{N!}{N_1!N_2!N_3!\dots}. \quad (3)$$

The logarithm of multiplicity of called the entropy $S = \ln W$. In the limit of large numbers, the entropy per particle can be written in the following form using the Stirling approximation for the factorials:

$$\frac{S}{N} = - \sum_k \frac{N_k}{N} \ln \left(\frac{N_k}{N} \right) = - \sum_k P_k \ln P_k. \quad (4)$$

Now we would like to find what distribution of particles between different energy states has the highest entropy, i. e. the highest multiplicity, provided that the total energy of the system, $E = \sum_k N_k \varepsilon_k$, has a fixed value. Solution of this problem can be easily obtained using the method of Lagrange multipliers [28], and the answer gives the exponential distribution (1).

The same result can be derived from the **ergodic theory**, which says that the many-body system occupies all possible states of a given total energy with equal probabilities. Then it is straightforward to show [29,30] that the probability distribution of the energy of an individual particle is given by (1).

Conservation of Money

The derivations outlined in Sect. “[The Boltzmann–Gibbs Distribution of Energy](#)” are very general and use only the statistical character of the system and the conservation of energy. So, one may expect that the exponential Boltzmann–Gibbs distribution (1) may apply to other statistical systems with a conserved quantity.

The economy is a big statistical system with millions of participating agents, so it is a promising target for applications of statistical mechanics. Is there a conserved quantity in economy? Drăgulescu and Yakovenko [25] argue that such a conserved quantity is money m . Indeed, the ordinary economic agents can only receive money from and give money to other agents. They are not permitted to “manufacture” money, e. g., to print dollar bills. When one agent i pays money Δm to another agent j for some goods or services, the money balances of the agents change as follows:

$$\begin{aligned} m_i &\rightarrow m'_i = m_i - \Delta m, \\ m_j &\rightarrow m'_j = m_j + \Delta m. \end{aligned} \quad (5)$$

The total amount of money of the two agents before and after the transaction remains the same,

$$m_i + m_j = m'_i + m'_j, \quad (6)$$

i. e., there is a local conservation law for money. The rule (5) for the transfer of money is analogous to the transfer of energy from one molecule to another in molecular

collisions in a gas, and (6) is analogous to conservation of energy in such collisions.

Addressing some misunderstandings developed in economic literature [31,32,33,34], we should emphasize that, in the model of [25], the transfer of money from one agent to another happens voluntarily, as a payment for goods and services in a market economy. However, the model only keeps track of money flow, and does not keep track of what kinds of goods and service are delivered. One reason for this is that many goods, e. g., food and other supplies, and most services, e. g., getting a haircut or going to a movie, are not tangible and disappear after consumption. Because they are not conserved and also because they are measured in different physical units, it is not very practical to keep track of them. In contrast, money is measured in the same unit (within a given country with a single currency) and is conserved in transactions, so it is straightforward to keep track of money flow.

Unlike ordinary economic agents, a central bank or a central government can inject money into the economy. This process is analogous to an influx of energy into a system from external sources, e. g., the Earth receives energy from the Sun. Dealing with these situations, physicists start with an idealization of a closed system in thermal equilibrium and then generalize to an open system subject to an energy flux. As long as the rate of money influx from central sources is slow compared with relaxation processes in the economy and does not cause hyperinflation, the system is in quasi-stationary statistical equilibrium with slowly changing parameters. This situation is analogous to heating a kettle on a gas stove slowly, where the kettle has a well-defined, but slowly increasing temperature at any moment of time.

Another potential problem with conservation of money is debt. This issue is discussed in more detail in Sect. “[Models with Debt](#)”. As a starting point, Drăgulescu and Yakovenko [25] first considered simple models, where debt is not permitted. This means that money balances of agents cannot go below zero: $m_i \geq 0$ for all i . Transaction (5) takes place only when an agent has enough money to pay the price, $m_i \geq \Delta m$, otherwise the transaction does not take place. If an agent spends all the money, the balance drops to zero $m_i = 0$, so the agent cannot buy any goods from other agents. However, this agent can still produce goods or services, sell them to other agents, and receive money for them. In real life, cash balance dropping to zero is not at all unusual for people who live from paycheck to paycheck.

The conservation law is the key feature for the successful functioning of money. If the agents were permitted to “manufacture” money, they would be printing money

and buying all goods for nothing, which would be a disaster. The physical medium of money is not essential, as long as the conservation law is enforced. Money may be in the form of paper cash, but today it is more often represented by digits in computerized bank accounts. The conservation law is the fundamental principle of accounting, whether in the single-entry or in the double-entry form. More discussion of banks and debt is given in Sect. “Models with Debt”.

The Boltzmann–Gibbs Distribution of Money

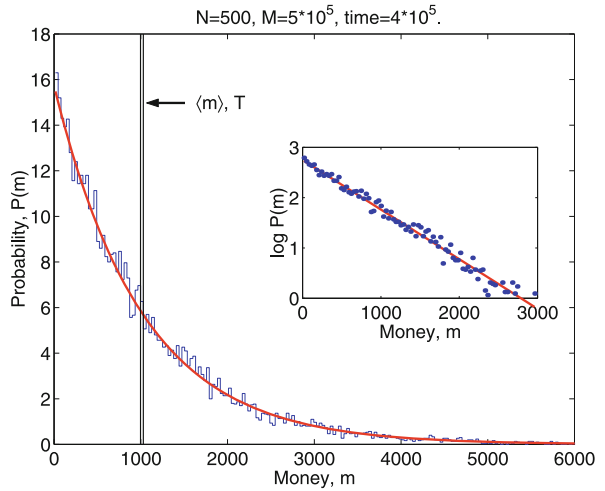
Having recognized the principle of money conservation, Drăgulescu and Yakovenko [25] argued that the stationary distribution of money should be given by the exponential Boltzmann–Gibbs function analogous to (1):

$$P(m) = ce^{\frac{-m}{T_m}}. \quad (7)$$

Here c is a normalizing constant, and T_m is the “money temperature”, which is equal to the average amount of money per agent: $T = \langle m \rangle = M/N$, where M is the total money, and N is the number of agents.

To verify this conjecture, Drăgulescu and Yakovenko [25] performed agent-based computer simulations of money transfers between agents. Initially all agents were given the same amount of money, say, \$ 1000. Then, a pair of agents (i, j) were randomly selected, the amount Δm was transferred from one agent to another, and the process was repeated many times. Time evolution of the probability distribution of money $P(m)$ can be seen in computer animation videos at the Web pages [35,36]. After a transitory period, money distribution converges to the stationary form shown in Fig. 1. As expected, the distribution is very well fitted by the exponential function (7).

Several different rules for Δm were considered in [25]. In one model, the amount transferred was fixed to a constant $\Delta m = 1\$$. Economically, it means that all agents were selling their products for the same price $\Delta m = 1\$$. Computer animation [35] shows that the initial distribution of money first broadens to a symmetric, Gaussian curve, characteristic for a diffusion process. Then, the distribution starts to pile up around the $m = 0$ state, which acts as the impenetrable boundary, because of the imposed condition $m \geq 0$. As a result, $P(m)$ becomes skewed (asymmetric) and eventually reaches the stationary exponential shape, as shown in Fig. 1. The boundary at $m = 0$ is analogous to the ground-state energy in statistical physics. Without this boundary condition, the probability distribution of money would not reach a stationary state. Computer animation [35,36] also shows how the entropy of money distribution, defined as $S/N =$



Econophysics, Statistical Mechanics Approach to, Figure 1

Stationary probability distribution of money $P(m)$ obtained in agent-based computer simulations. Solid curves: fits to the Boltzmann–Gibbs law (7). Vertical lines: the initial distribution of money. (Reproduced from [25])

$-\sum_k P(m_k) \ln P(m_k)$, grows from the initial value $S = 0$, when all agents have the same money, to the maximal value at the statistical equilibrium.

While the model with $\Delta m = 1$ is very simple and instructive, it is not very realistic, because all prices are taken to be the same. In another model considered in [25], Δm in each transaction is taken to be a random fraction of the average amount of money per agent, i. e., $\Delta m = \nu(M/N)$, where ν is a uniformly distributed random number between 0 and 1. The random distribution of Δm is supposed to represent the wide variety of prices for different products in the real economy. It reflects the fact that agents buy and consume many different types of products, some of them simple and cheap, some sophisticated and expensive. Moreover, different agents like to consume these products in different quantities, so there is variation in the amounts Δm paid, even though the unit price of the same product is constant. Computer simulation of this model produces exactly the same stationary distribution (7) as in the first model. Computer animation for this model is also available on the Web page [35].

The final distribution is universal despite different rules for Δm . To amplify this point further, Drăgulescu and Yakovenko [25] also considered a toy model, where Δm was taken to be a random fraction of the average amount of money of the two agents: $\Delta m = \nu(m_i + m_j)/2$. This model produced the same stationary distribution (7) as the other two models.

The pairwise models of money transfer are attractive in their simplicity, but they represent a rather prim-

itive market. The modern economy is dominated by big firms, which consist of many agents, so Drăgulescu and Yakovenko [25] also studied a model with firms. One agent at a time is appointed to become a “firm”. The firm borrows capital K from another agent and returns it with interest hK , hires L agents and pays them wages W , manufactures Q items of a product, sells them to Q agents at price R , and receives profit $F = RQ - LW - hK$. All of these agents are randomly selected. The parameters of the model are optimized following a procedure from economics textbooks [37]. The aggregate demand–supply curve for the product is used in the form $R(Q) = v/Q^\eta$, where Q is the quantity consumers would buy at price R , and η and v are some parameters. The production function of the firm has the traditional Cobb–Douglas form: $Q(L, K) = L^\chi K^{1-\chi}$, where χ is a parameter. Then the profit of the firm F is maximized with respect to K and L . The net result of the firm activity is a many-body transfer of money, which still satisfies the conservation law. Computer simulation of this model generates the same exponential distribution (7), independently of the model parameters. The reasons for the universality of the Boltzmann–Gibbs distribution and its limitations are discussed from a different perspective in Sect. “Additive Versus Multiplicative Models”.

Well after paper [25] appeared, Italian econophysicists [38] found that similar ideas had been published earlier in obscure journals in Italian by Eleonora Beninati [39,40]. They proposed calling these models the Beninati–Drăgulescu–Yakovenko game [41]. The Boltzmann distribution was independently applied to social sciences by Jürgen Mimkes using the Lagrange principle of maximization with constraints [42,43]. The exponential distribution of money was also found in [44] using a Markov chain approach to strategic market games. A long time ago, Benoit Mandelbrot observed (see p. 83 in [45]):

“There is a great temptation to consider the exchanges of money which occur in economic interaction as analogous to the exchanges of energy which occur in physical shocks between gas molecules”.

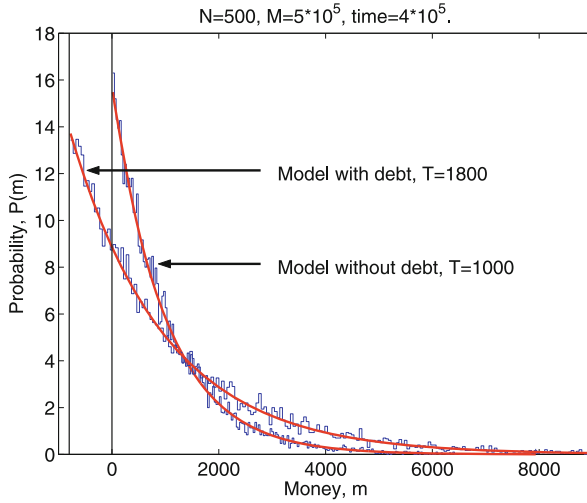
He realized that this process should result in the exponential distribution, by analogy with the barometric distribution of density in the atmosphere. However, he discarded this idea, because it does not produce the Pareto power law, and proceeded to study the stable Lévy distributions. Ironically, the actual economic data, discussed in Sect. “Empirical Data on Money and Wealth Distributions” and “Empirical Data on Income Distribution”, do show the exponential distribution for the majority of the population. Moreover, the data have finite variance, so the

stable Lévy distributions are not applicable because of their infinite variance.

Models with Debt

Now let us discuss how the results change when debt is permitted. Debt may be considered as negative money. When an agent borrows money from a bank (considered here as a big reservoir of money), the cash balance of the agent (positive money) increases, but the agent also acquires a debt obligation (negative money), so the total balance (net worth) of the agent remains the same, and the conservation law of total money (positive and negative) is satisfied. After spending some cash, the agent still has the debt obligation, so the money balance of the agent becomes negative. Any stable economic system must have a mechanism preventing unlimited borrowing and unlimited debt. Otherwise, agents can buy any products without producing anything in exchange by simply going into unlimited debt. The exact mechanisms of limiting debt in the real economy are complicated and obscured. Drăgulescu and Yakovenko [25] considered a simple model where the maximal debt of any agent is limited by a certain amount m_d . This means that the boundary condition $m_i \geq 0$ is now replaced by the condition $m_i \geq -m_d$ for all agents i . Setting interest rates on borrowed money to be zero for simplicity, Drăgulescu and Yakovenko [25] performed computer simulations of the models described in Sect. “The Boltzmann–Gibbs Distribution of Money” with the new boundary condition. The results are shown in Fig. 2. Not surprisingly, the stationary money distribution again has an exponential shape, but now with the new boundary condition at $m = -m_d$ and the higher money temperature $T_d = m_d + M/N$. By allowing agents to go into debt up to m_d , we effectively increase the amount of money available to each agent by m_d . So, the money temperature, which is equal to the average amount of effectively available money per agent, increases. A model with nonzero interest rates was also studied in [25].

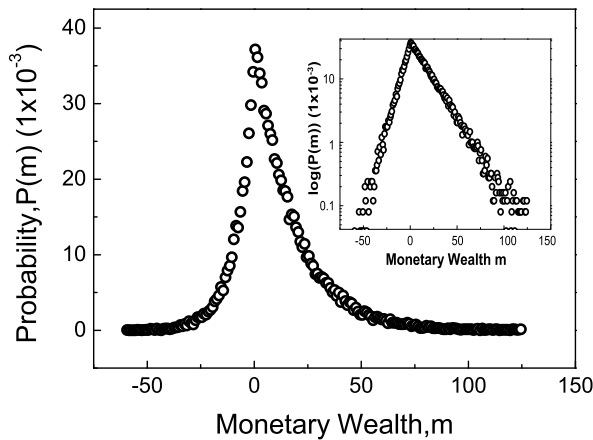
We see that debt does not violate the conservation law of money, but rather modifies boundary conditions for $P(m)$. When economics textbooks describe how “banks create money” or “debt creates money” [37], they count only positive money (cash) as money, but do not count liabilities (debt obligations) as negative money. With such a definition, money is not conserved. However, if we include debt obligations in the definition of money, then the conservation law is restored. This approach is in agreement with the principles of double-entry accounting, which records both assets and debts. Debt obligations are as real as positive cash, as many borrowers painfully real-



Econophysics, Statistical Mechanics Approach to, Figure 2
Stationary distributions of money with and without debt. The debt is limited to $m_d = 800$. Solid curves: fits to the Boltzmann-Gibbs laws with the “temperatures” $T = 1800$ and $T = 1000$. (Reproduced from [25])

ized in their experience. A more detailed study of positive and negative money and bookkeeping from the point of view of econophysics is presented in a series of papers by the physicist Dieter Braun [46,47,48].

One way of limiting the total debt in the economy is the so-called required reserve ratio r [37]. Every bank is required by law to set aside a fraction r of the money deposited with the bank, and this reserved money cannot be loaned further. If the initial amount of money in the sys-



Econophysics, Statistical Mechanics Approach to, Figure 3
The stationary distribution of money for the required reserve ratio $r = 0.8$. The distribution is exponential for positive and negative money with different “temperatures” T_+ and T_- , as illustrated by the inset on log-linear scale. (Reproduced from [49])

tem (the money base) is M_0 , then with loans and borrowing the total amount of positive money available to the agents increases to $M = M_0/r$, where the factor $1/r$ is called the money multiplier [37]. This is how “banks create money”. Where does this extra money come from? It comes from the increase of the total debt in the system. The maximal total debt is equal to $D = M_0/r - M_0$ and is limited by the factor r . When the debt is maximal, the total amounts of positive, M_0/r , and negative, $M_0(1-r)/r$, money circulate between the agents in the system, so there are effectively two conservation laws for each of them [49]. Thus, we expect to see the exponential distributions of positive and negative money characterized by two different temperatures: $T_+ = M_0/rN$ and $T_- = M_0(1-r)/rN$. This is exactly what was found in computer simulations in [49], shown in Fig. 3. Similar two-sided distributions were also found in [47].

Proportional Money Transfers and Saving Propensity

In the models of money transfer considered thus far, the transferred amount Δm is typically independent of the money balances of agents. A different model was introduced in the physics literature earlier [50] under the name multiplicative asset exchange model. This model also satisfies the conservation law, but the amount of money transferred is a fixed fraction γ of the payer’s money in (5):

$$\Delta m = \gamma m_i. \quad (8)$$

The stationary distribution of money in this model, shown in Fig. 4 with an exponential function, is close, but not exactly equal, to the Gamma distribution:

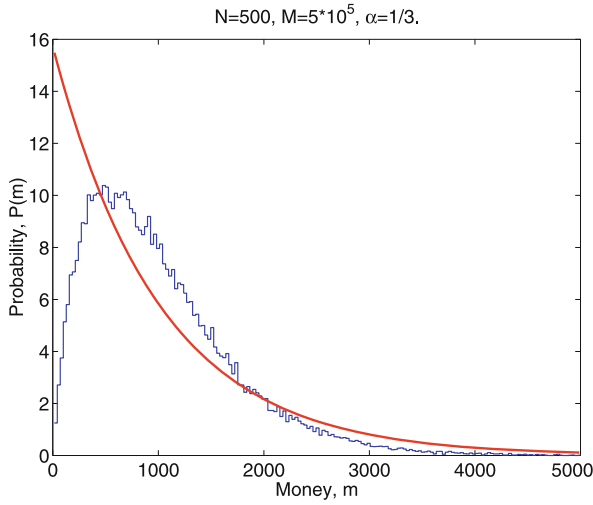
$$P(m) = cm^\beta e^{-\frac{m}{T}}. \quad (9)$$

Equation (9) differs from (7) by the power-law prefactor m^β . From the Boltzmann kinetic equation (discussed in more detail in Sect. “Additive Versus Multiplicative Models”), Ispolatov et al. [50] derived a formula relating the parameters γ and β in (8) and (9):

$$\beta = \frac{-1 - \ln 2}{\ln(1 - \gamma)}. \quad (10)$$

When payers spend a relatively small fraction of their money $\gamma < 1/2$, (10) gives $\beta > 0$, so the low-money population is reduced and $P(m \rightarrow 0) = 0$, as shown in Fig. 4.

Later, Thomas Lux brought to the attention of physicists [32] that essentially the same model, called the inequality process, had been introduced and studied much earlier by the sociologist John Angle [51,52,53,54,55], see also the review [56] for additional references. While Ispo-



Econophysics, Statistical Mechanics Approach to, Figure 4
Stationary probability distribution of money in the multiplicative random exchange model (8) for $\gamma = 1/3$. Solid curve: the exponential Boltzmann-Gibbs law. (Reproduced from [25])

latov et al. [50] did not give much justification for the proportionality law (8), Angle [51] connected this rule with the surplus theory of social stratification [57], which argues that inequality in human society develops when people can produce more than necessary for minimal subsistence. This additional wealth (surplus) can be transferred from original producers to other people, thus generating inequality. In the first paper by Angle [51], the parameter γ was randomly distributed, and another parameter, δ , gave a higher probability of winning to the agent with a higher money balance in (5). However, in the following papers, he simplified the model to a fixed γ (denoted as ω by Angle) and equal probabilities of winning for higher- and lower-balance agents, which makes it completely equivalent to the model of [50]. Angle also considered a model [55,56] where groups of agents have different values of γ , simulating the effect of education and other “human capital”. All of these models generate a Gamma-like distribution, well approximated by (9).

Another model with an element of proportionality was proposed in [26]. (This paper originally appeared as a follow-up preprint cond-mat/0004256 to the preprint cond-mat/0001432 of [25].) In this model, the agents set aside (save) some fraction of their money λm_i , whereas the rest of their money balance $(1 - \lambda)m_i$ becomes available for random exchanges. Thus, the rule of exchange (5) becomes

$$\begin{aligned} m'_i &= \lambda m_i + \xi(1 - \lambda)(m_i + m_j), \\ m'_j &= \lambda m_j + (1 - \xi)(1 - \lambda)(m_i + m_j). \end{aligned} \quad (11)$$

Here the coefficient λ is called the saving propensity, and the random variable ξ is uniformly distributed between 0 and 1. It was pointed out in [56] that, by the change of notation $\lambda \rightarrow (1 - \gamma)$, (11) can be transformed to the same form as (8), if the random variable ξ takes only discrete values 0 and 1. Computer simulations [26] of the model (11) found a stationary distribution close to the Gamma distribution (9). It was shown that the parameter β is related to the saving propensity λ by the formula $\beta = 3\lambda/(1 - \lambda)$ [38,58,59,60]. For $\lambda \neq 0$, agents always keep some money, so their balances never go to zero and $P(m \rightarrow 0) = 0$, whereas for $\lambda = 0$ the distribution becomes exponential.

In the subsequent papers by the Kolkata school [1] and related papers, the case of random saving propensity was studied. In these models, the agents are assigned random parameters λ drawn from a uniform distribution between 0 and 1 [61]. It was found that this model produces a power-law tail $P(m) \propto 1/m^2$ at high m . The reasons for stability of this law were understood using the Boltzmann kinetic equation [60,62,63], but most elegantly in the mean-field theory [64]. The fat tail originates from the agents whose saving propensity is close to 1, who hoard money and do not give it back [38,65]. An interesting matrix formulation of the problem was presented in [66]. Patriarca et al. [67] studied the relaxation rate in the money transfer models. Drăgulescu and Yakovenko [25] studied a model with taxation, which also has an element of proportionality. The Gamma distribution was also studied for conservative models within a simple Boltzmann approach in [68] and using much more complicated rules of exchange in [69,70].

Additive Versus Multiplicative Models

The stationary distribution of money (9) for the models of Sect. “Proportional Money Transfers and Saving Propensity” is different from the simple exponential formula (7) found for the models of Sect. “The Boltzmann-Gibbs Distribution of Money”. The origin of this difference can be understood from the Boltzmann kinetic equation [28,71]. This equation describes time evolution of the distribution function $P(m)$ due to pairwise interactions:

$$\begin{aligned} \frac{dP(m)}{dt} &= \iint \{ -f_{[m,m'] \rightarrow [m-\Delta, m'+\Delta]} P(m) P(m') \\ &\quad + f_{[m-\Delta, m'+\Delta] \rightarrow [m,m']} P(m - \Delta) \cdot P(m' + \Delta) \} dm' d\Delta. \end{aligned} \quad (12)$$

Here $f_{[m,m'] \rightarrow [m-\Delta, m'+\Delta]}$ is the probability of transferring money Δ from an agent with money m to an agent

with money m' per unit time. This probability, multiplied by the occupation numbers $P(m)$ and $P(m')$, gives the rate of transitions from the state $[m, m']$ to the state $[m - \Delta, m' + \Delta]$. The first term in (12) gives the depopulation rate of the state m . The second term in (12) describes the reverse process, where the occupation number $P(m)$ increases. When the two terms are equal, the direct and reverse transitions cancel each other statistically, and the probability distribution is stationary: $dP(m)/dt = 0$. This is the principle of detailed balance.

In physics, the fundamental microscopic equations of motion of particles obey time-reversal symmetry. This means that the probabilities of the direct and reverse processes are exactly equal:

$$f_{[m, m'] \rightarrow [m - \Delta, m' + \Delta]} = f_{[m - \Delta, m' + \Delta] \rightarrow [m, m']} \cdot \quad (13)$$

When (13) is satisfied, the detailed balance condition for (12) reduces to the equation $P(m)P(m') = P(m - \Delta)P(m' + \Delta)$, because the factors f cancel out. The only solution of this equation is the exponential function $P(m) = c \exp(-m/T_m)$, so the Boltzmann–Gibbs distribution is the stationary solution of the Boltzmann kinetic equation (12). Notice that the transition probabilities (13) are determined by the dynamical rules of the model, but the equilibrium Boltzmann–Gibbs distribution does not depend on the dynamical rules at all. This is the origin of the universality of the Boltzmann–Gibbs distribution. It shows that it may be possible to find out the stationary distribution without knowing details of the dynamical rules (which are rarely known very well), as long as the symmetry condition (13) is satisfied.

The models considered in Sect. “The Boltzmann–Gibbs Distribution of Money” have the time-reversal symmetry. The model with the fixed money transfer Δ has equal probabilities (13) of transferring money from an agent with balance m to an agent with balance m' and vice versa. This is also true when Δ is random, as long as the probability distribution of Δ is independent of m and m' . Thus, the stationary distribution $P(m)$ is always exponential in these models.

However, there is no fundamental reason to expect time-reversal symmetry in economics, so (13) may be not valid. In this case, the system may have a nonexponential stationary distribution or no stationary distribution at all. In model (8), the time-reversal symmetry is broken. Indeed, when an agent i gives a fixed fraction γ of his money m_i to an agent with balance m_j , their balances become $(1 - \gamma)m_i$ and $m_j + \gamma m_i$. If we try to reverse this process and appoint an agent j to be the payer and

to give the fraction γ of her money, $\gamma(m_j + \gamma m_i)$, to agent i , the system does not return to the original configuration $[m_i, m_j]$. As emphasized by Angle [56], the payer pays a deterministic fraction of his money, but the receiver receives a random amount from a random agent, so their roles are not interchangeable. Because the proportional rule typically violates the time-reversal symmetry, the stationary distribution $P(m)$ in multiplicative models is typically not exactly exponential.¹ Making the transfer dependent on the money balance of the payer effectively introduces a Maxwell’s demon into the model. That is why the stationary distribution is not exponential, and, thus, does not maximize entropy (4). Another view on the time-reversal symmetry in economic dynamics is presented in [72].

These examples show that the Boltzmann–Gibbs distribution does not hold for any conservative model. However, it is universal in a limited sense. For a broad class of models that have time-reversal symmetry, the stationary distribution is exponential and does not depend on the details of the model. Conversely, when the time-reversal symmetry is broken, the distribution may depend on the details of the model. The difference between these two classes of models may be rather subtle. Deviations from the Boltzmann–Gibbs law may occur only if the transition rates f in (13) explicitly depend on the agent’s money m or m' in an asymmetric manner. Drăgulescu and Yakovenko [25] performed a computer simulation where the direction of payment was randomly selected in advance for every pair of agents (i, j) . In this case, money flows along directed links between the agents: $i \rightarrow j \rightarrow k$, and the time-reversal symmetry is strongly violated. This model is closer to the real economy, where one typically receives money from an employer and pays it to a grocery store. Nevertheless, the Boltzmann–Gibbs distribution was found in this model, because the transition rates f do not explicitly depend on m and m' and do not violate (13).

In the absence of detailed knowledge of real microscopic dynamics of economic exchanges, the semiuniversal Boltzmann–Gibbs distribution (7) is a natural starting point. Moreover, the assumption of [25] that agents pay the same prices Δm for the same products, independent of their money balances m , seems very appropriate for the modern anonymous economy, especially for purchases over the Internet. There is no particular empirical evidence for the proportional rules (8) or (11). However, the differ-

¹However, when Δm is a fraction of the total money $m_i + m_j$ of the two agents, the model is time-reversible and has an exponential distribution, as discussed in Sect. “The Boltzmann–Gibbs Distribution of Money”.

ence between the additive (7) and multiplicative (9) distributions may be not so crucial after all. From the mathematical point of view, the difference is in the implementation of the boundary condition at $m = 0$. In the additive models of Sect. “[The Boltzmann–Gibbs Distribution of Money](#)”, there is a sharp cutoff of $P(m)$ at $m = 0$. In the multiplicative models of Sect. “[Proportional Money Transfers and Saving Propensity](#)”, the balance of an agent never reaches $m = 0$, so $P(m)$ vanishes at $m \rightarrow 0$ in a power-law manner. At the same time, $P(m)$ decreases exponentially for large m for both models.

By further modifying the rules of money transfer and introducing more parameters in the models, one can obtain even more complicated distributions [73]. However, one can argue that parsimony is the virtue of a good mathematical model, not the abundance of additional assumptions and parameters, whose correspondence to reality is hard to verify.

Statistical Mechanics of Wealth Distribution

In the econophysics literature on exchange models, the terms “money” and “wealth” are often used interchangeably; however, economists emphasize the difference between these two concepts. In this section, we review the models of wealth distribution, as opposed to money distribution.

Models with a Conserved Commodity

What is the difference between money and wealth? One can argue [25] that wealth w_i is equal to money m_i plus the other property that an agent i has. The latter may include durable material property, such as houses and cars, and financial instruments, such as stocks, bonds, and options. Money (paper cash, bank accounts) is generally liquid and countable. However, the other property is not immediately liquid and has to be sold first (converted into money) to be used for other purchases. In order to estimate the monetary value of property, one needs to know the price p . In the simplest model, let us consider just one type of property, say, stocks s . Then the wealth of an agent i is given by the formula

$$w_i = m_i + ps_i. \quad (14)$$

It is assumed that the price p is common for all agents and is established by some kind of market process, such as an auction, and may change in time.

It is reasonable to start with a model where both the total money $M = \sum_i m_i$ and the total stock $S = \sum_i s_i$ are

conserved [74,75,76]. The agents pay money to buy stock and sell stock to get money, and so on. Although M and S are conserved, the total wealth $W = \sum_i w_i$ is generally not conserved, because of the price fluctuation [75] in (14). This is an important difference from the money transfer models of Sect. “[Statistical Mechanics of Money Distribution](#)”. Here the wealth w_i of an agent i , not participating in any transactions, may change when transactions between other agents establish a new price p . Moreover, the wealth w_i of an agent i does not change after a transaction with an agent j . Indeed, in exchange for paying money Δm , agent i receives the stock $\Delta s = \Delta m/p$, so her total wealth (14) remains the same. In principle, the agent can instantaneously sell the stock back at the same price and recover the money paid. If the price p never changes, then the wealth w_i of each agent remains constant, despite transfers of money and stock between agents.

We see that redistribution of wealth in this model is directly related to price fluctuations. One mathematical model of this process was studied in [77]. In this model, the agents randomly change preferences for the fraction of their wealth invested in stocks. As a result, some agents offer stock for sale and some want to buy it. The price p is determined from the market-clearing auction matching supply and demand. Silver et al. [77] demonstrated in computer simulations and proved analytically using the theory of Markov processes that the stationary distribution $P(w)$ of wealth w in this model is given by the Gamma distribution, as in (9). Various modifications of this model [32], such as introducing monopolistic coalitions, do not change this result significantly, which shows the robustness of the Gamma distribution. For models with a conserved commodity, Chatterjee and Chakrabarti [75] found the Gamma distribution for a fixed saving propensity and a power law tail for a distributed saving propensity.

Another model with conserved money and stock was studied in [78] for an artificial stock market where traders follow different investment strategies: random, momentum, contrarian, and fundamentalist. Wealth distribution in the model with random traders was found have a power-law tail $P(w) \sim 1/w^2$ for large w . However, unlike in most other simulation, where all agents initially have equal balances, here the initial money and stock balances of the agents were randomly populated according to a power law with the same exponent. This raises the question whether the observed power-law distribution of wealth is an artifact of the initial conditions, because equilibration of the upper tail may take a very long simulation time.

Models with Stochastic Growth of Wealth

Although the total wealth W is not exactly conserved in the models considered in Sect. “[Models with a Conserved Commodity](#)”, W nevertheless remains constant on average, because the total money M and stock S are conserved. A different model for wealth distribution was proposed in [27]. In this model, time evolution of the wealth w_i of an agent i is given by the stochastic differential equation

$$\frac{dw_i}{dt} = \eta_i(t)w_i + \sum_{j(\neq i)} J_{ij}w_j - \sum_{j(\neq i)} J_{ji}w_i, \quad (15)$$

where $\eta_i(t)$ is a Gaussian random variable with mean $\langle \eta \rangle$ and variance $2\sigma^2$. This variable represents growth or loss of wealth of an agent due to investment in stock market. The last two terms describe transfer of wealth between different agents, which is taken to be proportional to the wealth of the payers with the coefficients J_{ij} . So, the model (15) is multiplicative and invariant under the scale transformation $w_i \rightarrow Zw_i$. For simplicity, the exchange fractions are taken to be the same for all agents: $J_{ij} = J/N$ for all $i \neq j$, where N is the total number of agents. In this case, the last two terms in (15) can be written as $J(\langle w \rangle - w_i)$, where $\langle w \rangle = \sum_i w_i/N$ is the average wealth per agent. This case represents a “mean-field” model, where all agents feel the same environment. It can be easily shown that the average wealth increases in time as $\langle w \rangle_t = \langle w \rangle_0 e^{(\langle \eta \rangle + \sigma^2)t}$. Then, it makes more sense to consider the relative wealth $\tilde{w}_i = w_i/\langle w \rangle_t$. Equation (15) for this variable becomes

$$\frac{d\tilde{w}_i}{dt} = (\eta_i(t) - \langle \eta \rangle - \sigma^2)\tilde{w}_i + J(1 - \tilde{w}_i). \quad (16)$$

The probability distribution $P(\tilde{w}, t)$ for the stochastic differential equation (16) is governed by the Fokker-Planck equation:

$$\frac{\partial P}{\partial t} = \frac{\partial [J(\tilde{w} - 1) + \sigma^2 \tilde{w}]P}{\partial \tilde{w}} + \sigma^2 \frac{\partial}{\partial \tilde{w}} \left(\tilde{w} \frac{\partial (\tilde{w}P)}{\partial \tilde{w}} \right). \quad (17)$$

The stationary solution ($\partial P/\partial t = 0$) of this equation is given by the following formula:

$$P(\tilde{w}) = c \frac{e^{-\frac{J}{\sigma^2 \tilde{w}}}}{\tilde{w}^{\frac{2+J}{\sigma^2}}}. \quad (18)$$

The distribution (18) is quite different from the Boltzmann-Gibbs (7) and Gamma (9) distributions. Equation (18) has a power-law tail at large \tilde{w} and a sharp cut-off at small \tilde{w} . Equation (15) is a version of the generalized Lotka-Volterra model, and the stationary distribution (18)

was also obtained in [79,80]. The model was generalized to include negative wealth in [81].

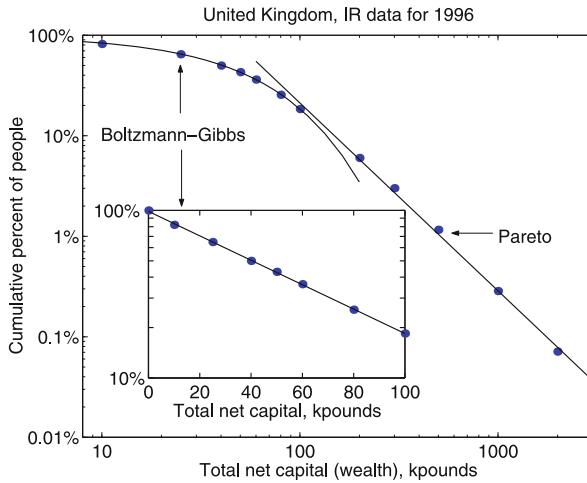
Bouchaud and Mézard [27] used the mean-field approach. A similar result was found for a model with pairwise interaction between agents in [82]. In this model, wealth is transferred between the agents using the proportional rule (8). In addition, the wealth of the agents increases by the factor $1 + \zeta$ in each transaction. This factor is supposed to reflect creation of wealth in economic interactions. Because the total wealth in the system increases, it makes sense to consider the distribution of relative wealth $P(\tilde{w})$. In the limit of continuous trading, Slanina [82] found the same stationary distribution (18). This result was reproduced using a mathematically more involved treatment of this model in [83]. Numerical simulations of the models with stochastic noise η in [69,70] also found a power-law tail for large w .

Let us contrast the models discussed in Sect. “[Models with a Conserved Commodity](#)” and “[Models with Stochastic Growth of Wealth](#)”. In the former case, where money and commodity are conserved and wealth does not grow, the distribution of wealth is given by the Gamma distribution with an exponential tail for large w . In the latter models, wealth grows in time exponentially, and the distribution of relative wealth has a power-law tail for large \tilde{w} . These results suggest that the presence of a power-law tail is a nonequilibrium effect that requires constant growth or inflation of the economy, but disappears for a closed system with conservation laws.

Reviews of the models discussed were also given in [84,85]. Because of lack of space, we omit discussion of models with wealth condensation [27,50,86,87,88], where a few agents accumulate a finite fraction of the total wealth, and studies of wealth distribution on networks [89,90,91,92]. Here we discuss the models with long-range interaction, where any agent can exchange money and wealth with any other agent. A local model, where agents trade only with the nearest neighbors, was studied in [93].

Empirical Data on Money and Wealth Distributions

It would be very interesting to compare theoretical results for money and wealth distributions in various models with empirical data. Unfortunately, such empirical data are difficult to find. Unlike income, which is discussed in Sect. “[Data and Models for Income Distribution](#)”, wealth is not routinely reported by the majority of individuals to the government. However, in many countries, when a person dies, all assets must be reported for the purpose of inheritance tax. So, in principle, there exist good statis-



Econophysics, Statistical Mechanics Approach to, Figure 5

Cumulative probability distribution of net wealth in the UK shown on log-log and log-linear (inset) scales. Points represent the data from the Inland Revenue, and solid lines are fits to the exponential (Boltzmann-Gibbs) and power (Pareto) laws. (Reproduced from [95])

tics of wealth distribution among dead people, which, of course, is different from the wealth distribution among living people. Using an adjustment procedure based on the age, gender, and other characteristics of the deceased, the UK tax agency, the Inland Revenue, reconstructed the wealth distribution of the whole population of the UK [94]. Figure 5 shows the UK data for 1996 reproduced from [95]. The figure shows the cumulative probability $C(w) = \int_w^\infty P(w')dw'$ as a function of the personal net wealth w , which is composed of assets (cash, stocks, property, household goods, etc.) and liabilities (mortgages and other debts). Because statistical data are usually reported at nonuniform intervals of w , it is more practical to plot the cumulative probability distribution $C(w)$ rather than its derivative, the probability density $P(w)$. Fortunately, when $P(w)$ is an exponential or a power-law function, then $C(w)$ is also an exponential or a power-law function.

The cumulative probability distribution in Fig. 5 is plotted on a log-log scale, where a straight line represents a power-law dependence. The figure shows that the distribution follows a power law $C(w) \propto 1/w^\alpha$ with exponent $\alpha = 1.9$ for wealth greater than about £100,000. The inset in Fig. 5 shows the data on log-linear scale, where the straight line represents an exponential dependence. We observe that below £100,000 the data are well fitted by the exponential distribution $C(w) \propto \exp(-w/T_w)$ with the effective “wealth temperature” $T_w = £60,000$, (which corresponds to the median wealth of £41,000). So, the dis-

tribution of wealth is characterized by the Pareto power law in the upper tail of the distribution and the exponential Boltzmann-Gibbs law in the lower part of the distribution for the great majority (about 90%) of the population. Similar results are found for the distribution of income, as discussed in Sect. “Data and Models for Income Distribution”. One may speculate that the wealth distribution in the lower part is dominated by distribution of money, because the corresponding people do not have other significant assets, so the results of Sect. “Statistical Mechanics of Money Distribution” give the Boltzmann-Gibbs law. On the other hand, the upper tail of the wealth distribution is dominated by investment assets, where the results of Sect. “Models with Stochastic Growth of Wealth” give the Pareto law. The power law was studied by many researchers for the upper-tail data, such as the Forbes list of the 400 richest people [96,97], but much less attention was paid to the lower part of the wealth distribution. Curiously, Abdul-Maghd [98] found that the wealth distribution in ancient Egyptian society was consistent with (18).

For direct comparison with the results of Sect. “Statistical Mechanics of Money Distribution”, it would be very interesting to find data on the distribution of money, as opposed to the distribution of wealth. Making a reasonable assumption that most people keep most of their money in banks, one can approximate the distribution of money by the distribution of balances on bank accounts. (Balances on all types of bank accounts, such as checking, saving, and money manager, associated with the same person should be added up.) Despite imperfections (people may have accounts with different banks or not keep all their money in banks), the distribution of balances on bank accounts would give valuable information about the distribution of money. The data for a big enough bank would be representative of the distribution in the whole economy. Unfortunately, it has not been possible to obtain such data thus far, even though it would be completely anonymous and not compromise the privacy of bank clients.

Measuring the probability distribution of money would be very useful, because it determines how much people can, in principle, spend on purchases without going into debt. This is different from the distribution of wealth, where the property component, such as house, car, or retirement investment, is effectively locked up and, in most cases, is not easily available for consumer spending. So, although wealth distribution may reflect the distribution of economic power, the distribution of money is more relevant for consumption. Money distribution can be useful for determining prices that maximize revenue of a manufacturer [25]. If a price p is set too high, few people can afford it, and, if a price is too low, the sales revenue is

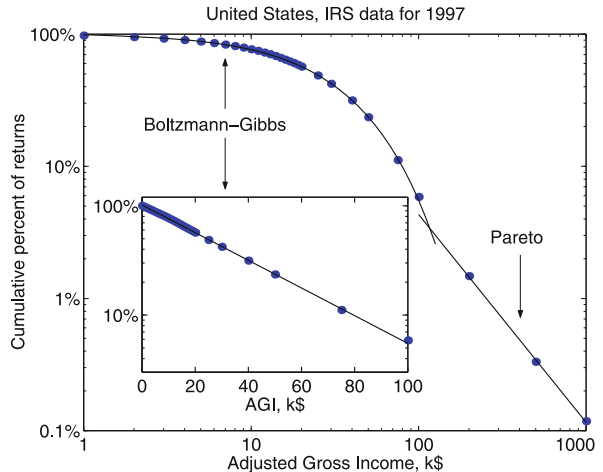
small, so the optimal price must be in-between. The fraction of the population who can afford to pay the price p is given by the cumulative probability $C(p)$, so the total sales revenue is proportional to $pC(p)$. For the exponential distribution $C(p) = \exp(-p/T_m)$, the maximal revenue is achieved at $p = T_m$, i.e., at the optimal price equal to the average amount of money per person [25]. Indeed, the prices of mass-market consumer products, such as computers, electronics goods, and appliances, remain stable for many years at a level determined by their affordability to the population, whereas the technical parameters of these products at the same price level improve dramatically owing to technological progress.

Data and Models for Income Distribution

In contrast to money and wealth distributions, a lot more empirical data are available for the distribution of income r from tax agencies and population surveys. In this section, we first present empirical data on income distribution and then discuss theoretical models.

Empirical Data on Income Distribution

Empirical studies of income distribution have a long history in the economics literature [99,100,101]. Following the work by Pareto [15], much attention was focused on the power-law upper tail of the income distribution and less on the lower part. In contrast to more complicated functions discussed in the literature, Drăgulescu and Yakovenko [102] introduced a new idea by demonstrating that the lower part of income distribution can be well fitted with a simple exponential function $P(r) = c \exp(-r/T_r)$ characterized by just one parameter, the “income temperature” T_r . Then it was recognized that the whole income distribution can be fitted by an exponential function in the lower part and a power-law function in the upper part [95,103], as shown in Fig. 6. The straight line on the log-linear scale in the inset of Fig. 6 demonstrates the exponential Boltzmann-Gibbs law, and the straight line on the log-log scale in the main panel illustrates the Pareto power law. The fact that income distribution consists of two distinct parts reveals the two-class structure of American society [104,105]. Coexistence of the exponential and power-law distributions is known in plasma physics and astrophysics, where they are called the “thermal” and “superthermal” parts [106,107,108]. The boundary between the lower and upper classes can be defined as the intersection point of the exponential and power-law fits in Fig. 6. For 1997, the annual income separating the two classes was about \$120,000. About 3% of the popula-



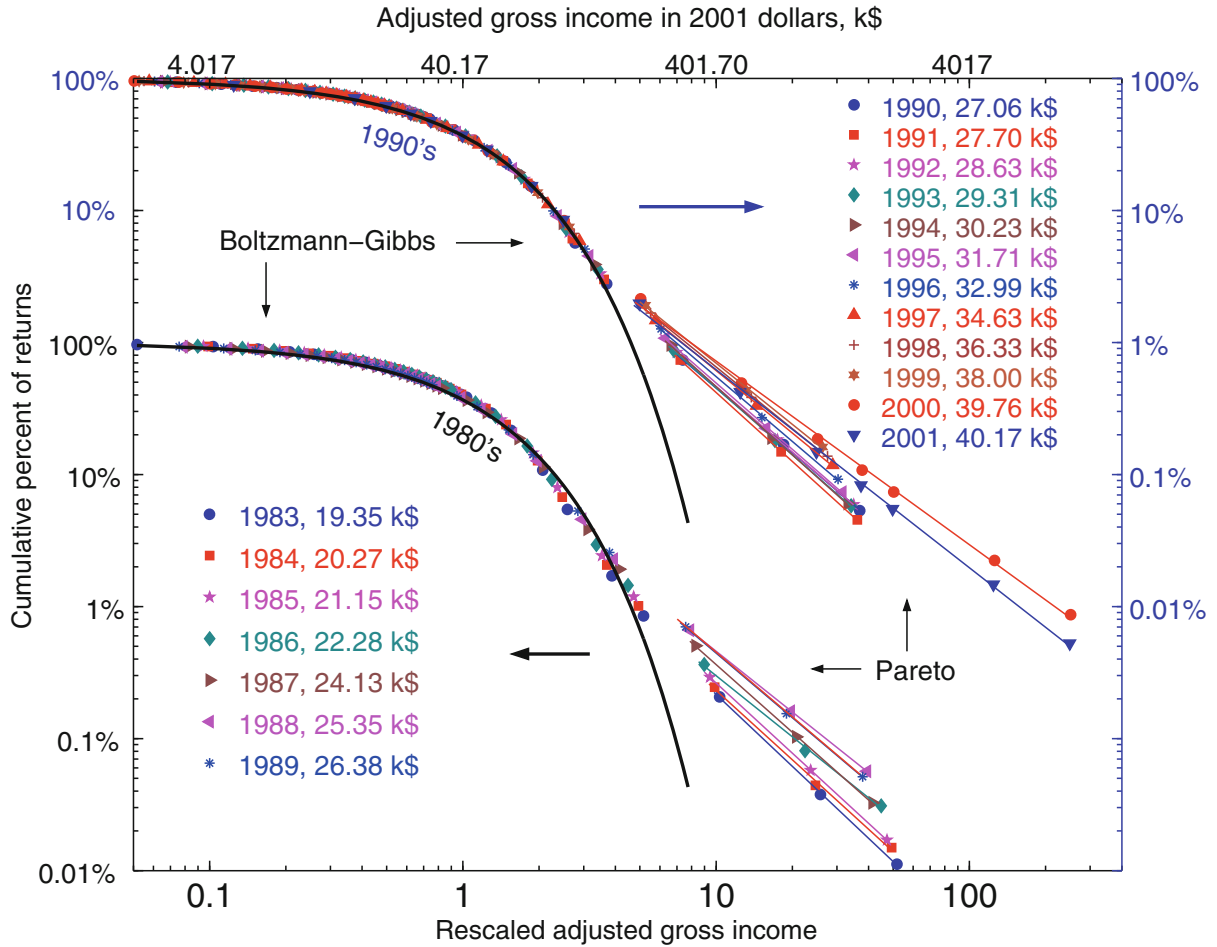
Econophysics, Statistical Mechanics Approach to, Figure 6

Cumulative probability distribution of tax returns for USA in 1997 shown on log-log and log-linear (inset) scales. Points represent the Internal Revenue Service (IRS) data, and solid lines are fits to the exponential and power-law functions. (Reproduced from [103])

tion belonged to the upper class, and 97% belonged to the lower class.

Silva and Yakovenko [105] studied time evolution of income distribution in the USA during 1983–2001 on the basis of data from the Internal Revenue Service (IRS), the government tax agency. The structure of the income distribution was found to be qualitatively the same for all years, as shown in Fig. 7. The average income in nominal dollars approximately doubled during this time interval. So, the horizontal axis in Fig. 7 shows the normalized income r/T_r , where the “income temperature” T_r was obtained by fitting of the exponential part of the distribution for each year. The values of T_r are shown in Fig. 7. The plots for the 1980s and 1990s are shifted vertically for clarity. We observe that the data points in the lower-income part of the distribution collapse on the same exponential curve for all years. This demonstrates that the shape of the income distribution for the lower class is extremely stable and does not change in time, despite the gradual increase of the average income in nominal dollars. This observation suggests that the lower-class distribution is in statistical, “thermal” equilibrium.

On the other hand, Fig. 7 shows that the income distribution in the upper class does not rescale and significantly changes in time. Silva and Yakovenko [105] found that the exponent α of the power law $C(r) \propto 1/r^\alpha$ decreased from 1.8 in 1983 to 1.4 in 2000. This means that the upper tail became “fatter”. Another useful parameter is the total in-



Econophysics, Statistical Mechanics Approach to, Figure 7

Cumulative probability distribution of tax returns plotted on log-log scale versus r/T_r (the annual income r normalized by the average income T_r in the exponential part of the distribution). The IRS data points are for 1983–2001, and the *columns of numbers* give the values of T_r for the corresponding years. (Reproduced from [105])

come of the upper class as the fraction f of the total income in the system. The fraction f increased from 4% in 1983 to 20% in 2000 [105]. However, in 2001, α increased and f decreased, indicating that the upper tail was reduced after the stock market crash at that time. These results indicate that the upper tail is highly dynamical and not stationary. It tends to swell during the stock market boom and shrink during the bust. Similar results were found for Japan [109,110,111,112].

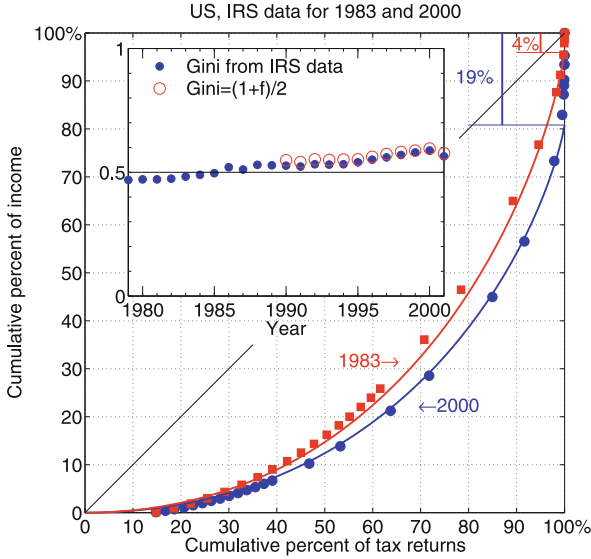
Although relative income inequality within the lower class remains stable, the overall income inequality in the USA has increased significantly as a result of the tremendous growth of the income of the upper class. This is illustrated by the Lorenz curve and the Gini coefficient shown in Fig. 8. The Lorenz curve [99] is a standard way of representing income distribution in the economics literature.

It is defined in terms of two coordinates $x(r)$ and $y(r)$ depending on a parameter r :

$$\begin{aligned} x(r) &= \int_0^r P(r') dr', \\ y(r) &= \frac{\int_0^r r' P(r') dr'}{\int_0^\infty r' P(r') dr'}. \end{aligned} \quad (19)$$

The horizontal coordinate $x(r)$ is the fraction of the population with income below r , and the vertical coordinate $y(r)$ is the fraction of the income this population accounts for. As r changes from 0 to ∞ , x and y change from 0 to 1 and parametrically define a curve in the (x, y) -plane.

Figure 8 shows the data points for the Lorenz curves in 1983 and 2000, as computed by the IRS [113]. Drăgulescu and Yakovenko [102] analytically derived the Lorenz curve



Econophysics, Statistical Mechanics Approach to, Figure 8

Lorenz plots for income distribution in 1983 and 2000. The data points are from the IRS [113], and the theoretical curves represent (20) with f from Fig. 7. Inset: The closed circles are the IRS data 113 for the Gini coefficient G , and the open circles show the theoretical formula $G = (1 + f)/2$. (Reproduced from [105])

formula $y = x + (1 - x) \ln(1 - x)$ for a purely exponential distribution $P(r) = c \exp(-r/T_r)$. This formula is shown by the red line in Fig. 8 and describes the 1983 data reasonably well. However, for 2000, it is essential to take into account the fraction f of income in the upper tail, which modifies the Lorenz formula as follows [103,104,105]:

$$y = (1 - f)[x + (1 - x) \ln(1 - x)] + f\Theta(x - 1). \quad (20)$$

The last term in (20) represents the vertical jump of the Lorenz curve at $x = 1$, where a very small percentage of the population in the upper class accounts for a substantial fraction f of the total income. The blue curve representing (20) fits the 2000 data in Fig. 8 very well.

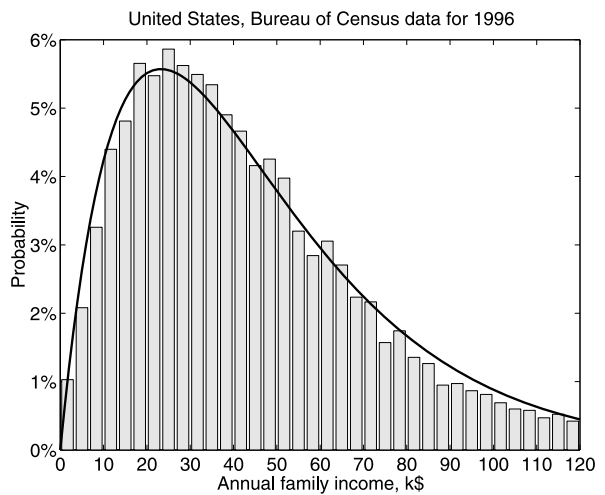
The deviation of the Lorenz curve from the straight diagonal line in Fig. 8 is a certain measure of income inequality. Indeed, if everybody had the same income, the Lorenz curve would be a diagonal line, because the fraction of income would be proportional to the fraction of the population. The standard measure of income inequality is the so-called Gini coefficient $0 \leq G \leq 1$, which is defined as the area between the Lorenz curve and the diagonal line, divided by the area of the triangle beneath the diagonal line [99]. Time evolution of the Gini coefficient, as computed by the IRS [113], is shown in the inset of Fig. 8. Drăgulescu and Yakovenko [102] derived analytically the result that $G = 1/2$ for a purely exponential distribution.

In the first approximation, the values of G shown in the inset of Fig. 8 are indeed close to the theoretical value $1/2$. If we take into account the upper tail using (20), the formula for the Gini coefficient becomes $G = (1 + f)/2$ [105]. The inset in Fig. 8 shows that this formula is a very good fit to the IRS data for the 1990s using the values of f deduced from Fig. 7. The values $G < 1/2$ for the 1980s cannot be captured by this formula, because the Lorenz data points are slightly above the theoretical curve for 1983 in Fig. 8. Overall, we observe that income inequality has been increasing for the last 20 years, because of swelling of the Pareto tail, but decreased in 2001 after the stock market crash.

Thus far we have discussed the distribution of individual income. An interesting related question is the distribution $P_2(r)$ of family income $r = r_1 + r_2$, where r_1 and r_2 are the incomes of spouses. If individual incomes are distributed exponentially $P(r) \propto \exp(-r/T_r)$, then

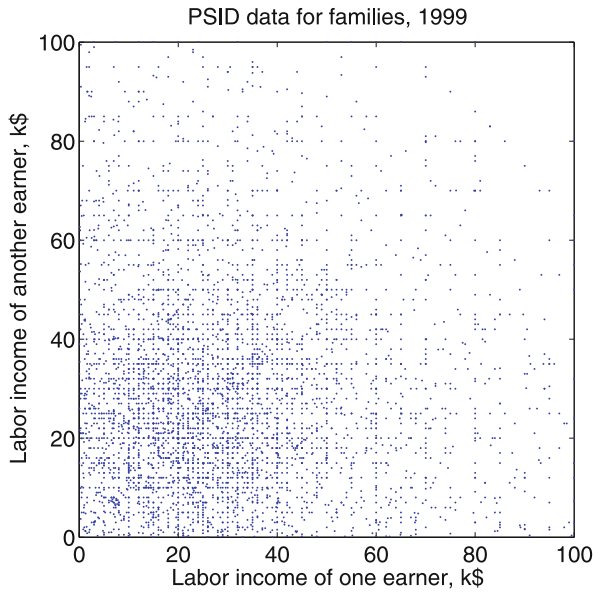
$$P_2(r) = \int_0^r dr' P(r') P(r - r') = c r \exp(-r/T_r), \quad (21)$$

where c is a normalization constant. Figure 9 shows that (21) is in good agreement with the family income distribution data from the US Census Bureau [102]. In (21), we assumed that incomes of spouses are uncorrelated. This simple approximation is indeed supported by the scatter plot of incomes of spouses shown in Fig. 10. Each family is represented in this plot by two points (r_1, r_2) and (r_2, r_1) for symmetry. We observe that the density of points is approximately constant along the lines of constant family in-



Econophysics, Statistical Mechanics Approach to, Figure 9

Probability distribution of family income for families with two adults (US Census Bureau data). Solid line: fit to (21). (Reproduced from [102])

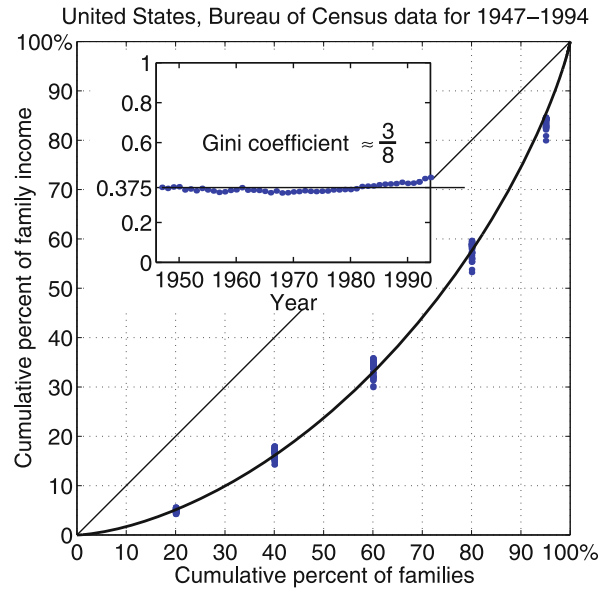


Econophysics, Statistical Mechanics Approach to, Figure 10 Scatter plot of the spouses' incomes (r_1, r_2) and (r_2, r_1) based on the data from the Panel Study of Income Dynamics (PSID). (Reproduced from [103])

come $r_1 + r_2 = \text{const}$, which indicates that incomes of spouses are approximately uncorrelated. There is no significant clustering of points along the diagonal $r_1 = r_2$, i. e., no strong positive correlation of spouses' incomes.

The Gini coefficient for the family income distribution (21) was calculated in [102] as $G = 3/8 = 37.5\%$. Figure 11 shows the Lorenz quintiles and the Gini coefficient for 1947–1994 plotted from the US Census Bureau data. The solid line, representing the Lorenz curve calculated from (21), is in good agreement with the data. The systematic deviation for the top 5% of earners results from the upper tail, which has a less pronounced effect on family income than on individual income, because of income averaging in the family. The Gini coefficient, shown in the inset of Fig. 11, is close to the calculated value of 37.5%. Moreover, the average G for the developed capitalist countries of North America and western Europe, as determined by the World Bank [103], is also close to the calculated value of 37.5%.

Income distribution has been examined in econophysics papers for different countries: Japan [68,109,110,111,112,114,115,116], Germany [117,118], the UK [68,85,116,117,118], Italy [118,119,120], the USA [117,121], India [97], Australia [91,120,122], and New Zealand [68,116]. The distributions are qualitatively similar to the results presented in this section. The upper tail follows a power law and comprises a small fraction of the pop-



Econophysics, Statistical Mechanics Approach to, Figure 11 Lorenz plot for family income calculated from (21), compared with the US Census data points. Inset: The US Census data points for the Gini coefficient for families, compared with the theoretically calculated value $3/8=37.5\%$. (Reproduced from [102])

ulation. To fit the lower part of the distribution, the use of exponential, Gamma, and log-normal distributions was reported in different papers. Unfortunately, income distribution is often reported by statistical agencies for households, so it is difficult to differentiate between one-earner and two-earner income distributions. Some papers reported the use of interpolating functions with different asymptotic behavior for low and high incomes, such as the Tsallis function [116] and the Kaniadakis function [118]. However, the transition between the lower and upper classes is not smooth for the US data shown in Figs. 6 and 7, so such functions would not be useful in this case. The special case is income distribution in Argentina during the economic crisis, which shows a time-dependent bimodal shape with two peaks [116].

Theoretical Models of Income Distribution

Having examined the empirical data on income distribution, let us now discuss theoretical models. Income r_i is the influx of money per unit time to an agent i . If the money balance m_i is analogous to energy, then the income r_i would be analogous to power, which is the energy flux per unit time. So, one should conceptually distinguish between the distributions of money and income. While money is regularly transferred from one agent to another in pairwise

transactions, it is not typical for agents to trade portions of their income. Nevertheless, indirect transfer of income may occur when one employee is promoted and another demoted, while the total annual budget is fixed, or when one company gets a contract, whereas another one loses it, etc. A reasonable approach, which has a long tradition in the economics literature [123,124,125], is to treat individual income r as a stochastic process and study its probability distribution. In general, one can study a Markov process generated by a matrix of transitions from one income to another. In the case where income r changes by a small amount Δr over a time period Δt , the Markov process can be treated as income diffusion. Then one can apply the general Fokker–Planck equation [71] to describe evolution in time t of the income distribution function $P(r, t)$ [105]:

$$\begin{aligned} \frac{\partial P}{\partial t} &= \frac{\partial}{\partial r} \left[AP + \frac{\partial(BP)}{\partial r} \right], \\ A &= -\frac{\langle \Delta r \rangle}{\Delta t}, \quad B = \frac{\langle (\Delta r)^2 \rangle}{2\Delta t}. \end{aligned} \quad (22)$$

The coefficients A and B in (22) are determined by the first and second moments of income changes per unit time. The stationary solution $\partial_t P = 0$ of (22) obeys the following equation with the general solution:

$$\begin{aligned} \frac{\partial(BP)}{\partial r} &= -AP, \\ P(r) &= \frac{c}{B(r)} \exp \left(-\int^r \frac{A(r')}{B(r')} dr' \right). \end{aligned} \quad (23)$$

For the lower part of the distribution, it is reasonable to assume that Δr is independent of r , i.e., the changes of income are independent of income itself. This process is called additive diffusion [105]. In this case, the coefficients in (22) are constants A_0 and B_0 . Then (23) gives the exponential distribution $P(r) \propto \exp(-r/T_r)$, with the effective income temperature $T_r = B_0/A_0$. Notice that a meaningful stationary solution (23) requires that $A > 0$, i.e., $\langle \Delta r \rangle < 0$. The coincidence of this result with the Boltzmann–Gibbs exponential law (1) and (7) is not accidental. Indeed, instead of considering pairwise interaction between particles, one can derive (1) by considering energy transfers between a particle and a big reservoir, as long as the transfer process is “additive” and does not involve a Maxwell–demon-like discrimination. Stochastic income fluctuations are described by a similar process. So, although money and income are different concepts, they may have similar distributions, because they are governed by similar mathematical principles. It was shown explicitly in [25,82,83] that the models of pairwise money transfer can be described in a certain limit by the Fokker–Planck equation.

On the other hand, for the upper tail of the income distribution, it is reasonable to expect that $\Delta r \propto r$, i.e., income changes are proportional to income itself. This is known as the proportionality principle of Gibrat [123], and the process is called multiplicative diffusion [105]. In this case, $A = ar$ and $B = br^2$, and (23) gives the power-law distribution $P(r) \propto 1/r^{\alpha+1}$, with $\alpha = 1 + a/b$.

Generally, lower-class income comes from wages and salaries, where the additive process is appropriate, whereas upper-class income comes from bonuses, investments, and capital gains, calculated in percentages, where the multiplicative process applies [126]. However, the additive and multiplicative processes may coexist. An employee may receive a cost-of-living rise calculated in percentages (the multiplicative process) and a merit rise calculated in dollars (the additive process). In this case, we have $A = A_0 + ar$ and $B = B_0 + br^2 = b(r_0^2 + r^2)$, where $r_0^2 = B_0/b$. Substituting these expressions into (23), we find

$$P(r) = c \frac{e^{-(\frac{r_0}{T_r}) \arctan(\frac{r}{r_0})}}{[1 + (\frac{r}{r_0})^2]^{\frac{1+a}{2b}}}. \quad (24)$$

The distribution (24) interpolates between the exponential law for low r and the power law for high r , because either the additive or the multiplicative process dominates in the corresponding limit. The crossover between the two regimes takes place at $r \sim r_0$, where the additive and multiplicative contributions to B are equal. The distribution (24) has three parameters: the “income temperature” $T_r = A_0/B_0$, the Pareto exponent $\alpha = 1 + a/b$, and the crossover income r_0 . It is a minimal model that captures the salient features of the empirical income distribution shown in Fig. 6. A mathematically similar, but more economically oriented model was proposed in [114,115], where labor income and asset accumulation are described by the additive and multiplicative processes correspondingly. A general stochastic process with additive and multiplicative noise was studied numerically in [127], but the stationary distribution was not derived analytically. A similar process with discrete time increments was studied by Kesten [128]. Recently, a formula similar to (24) was obtained in [129].

To verify the multiplicative and additive hypotheses empirically, it is necessary to have data on income mobility, i.e., the income changes Δr of the same people from one year to another. The distribution of income changes $P(\Delta r|r)$ conditional on income r is generally not available publicly, although it can be reconstructed by researchers at the tax agencies. Nevertheless, the multiplicative hypothesis for the upper class was quantitatively verified

in [111,112] for Japan, where tax identification data is published for the top taxpayers.

The power-law distribution is meaningful only when it is limited to high enough incomes $r > r_0$. If all incomes r from 0 to ∞ follow a purely multiplicative process, then one can change to a logarithmic variable $x = \ln(r/r_*)$ in (22) and show that it gives a Gaussian time-dependent distribution $P_t(x) \propto \exp(-x^2/2\sigma^2 t)$ for x , i.e., the log-normal distribution for r , also known as the Gibrat distribution [123]. However, the width of this distribution increases linearly in time, so the distribution is not stationary. This was pointed out by Kalecki [124] a long time ago, but the log-normal distribution is still widely used for fitting income distribution, despite this fundamental logical flaw in its justification. In a classic paper, Champenowne [125] showed that a multiplicative process gives a stationary power-law distribution when a boundary condition is imposed at $r_0 \neq 0$. Later, this result was rediscovered by econophysicists [130,131]. In our (24), the exponential distribution of the lower class effectively provides such a boundary condition for the power law of the upper class. Notice also that (24) reduces to (18) in the limit $r_0 \rightarrow 0$, which corresponds to purely multiplicative noise $B = br^2$.

There are alternative approaches to income distribution in the economics literature. One of them, proposed by Lydall [132], involves social hierarchy. Groups of people have leaders, who have leaders of a higher order, and so on. The number of people decreases geometrically (exponentially) with the increase of the hierarchical level. If individual income increases by a certain factor (i.e., multiplicatively) when moving to the next hierarchical level, then income distribution follows a power law [132]. However, the original argument of Lydall can be easily modified to produce an exponential distribution. If individual income increases by a certain amount, i.e., income increases linearly with the hierarchical level, then income distribution is exponential. The latter process seems to be more realistic for moderate incomes below \$100,000. A similar scenario is the Bernoulli trials [133], where individuals have a constant probability of increasing their income by a fixed amount. We see that the deterministic hierarchical models and the stochastic models of additive and multiplicative income mobility represent essentially the same ideas.

Other Applications of Statistical Physics

Statistical physics was applied to a number of other subjects in economics. Because of lack of space, only two such topics are briefly discussed in this section.

Economic Temperatures in Different Countries

As discussed in Sect. “Empirical Data on Money and Wealth Distributions” and “Empirical Data on Income Distribution”, the distributions of money, wealth, and income are often described by exponential functions for the majority of the population. These exponential distributions are characterized by the parameters T_m , T_w , and T_r , which are mathematically analogous to temperature in the Boltzmann–Gibbs distribution (1). The values of these parameters, extracted from the fits of the empirical data, are generally different for different countries, i.e., different countries have different economic “temperatures”. For example, Drăgulescu and Yakovenko [95] found that the US income temperature was 1.9 times higher than the UK income temperature in 1998 (using the exchange rate of dollars to pounds at that time). Also, there was $\pm 25\%$ variation between income temperatures of different states within the USA. [95].

In physics, a difference of temperatures allows one to set up a thermal machine. It was argued in [25] that the difference of money or income temperatures between different countries allows one to extract profit in international trade. Indeed, as discussed at the end of Sect. “Empirical Data on Money and Wealth Distributions”, the prices of goods should be commensurate with money or income temperature, because otherwise people cannot afford to buy those goods. So, an international trading company can buy goods at a low price T_1 in a “low-temperature” country and sell them at a high price T_2 in a “high-temperature” country. The difference of prices $T_2 - T_1$ would be the profit of the trading company. In this process, money (the analog of energy) flows from the “high-temperature” to the “low-temperature” country, in agreement with the second law of thermodynamics, whereas products flow in the opposite direction. This process very much resembles what is going on in the global economy now. In this framework, the perpetual trade deficit of the USA is the consequence of the second law of thermodynamics and the difference of temperatures between the USA and “low-temperature” countries, such as China. Similar ideas were developed in more detail in [134,135], including a formal Carnot cycle for international trade.

The statistical physics approach demonstrates that profit originates from statistical nonequilibrium (the difference of temperatures), which exists in the global economy. However, it does not answer the question what is the origin of this difference. By analogy with physics, one would expect that the money flow should reduce the temperature difference and, eventually, lead to equilibration.

of temperatures. In physics, this situation is known as the “thermal death of the universe”. In a completely equilibrated global economy, it would be impossible to make profit by exploiting differences of economic temperatures between different countries. Although globalization of the modern economy does show a tendency toward equilibration of living standards in different countries, this process is far from straightforward, and there are many examples contrary to equilibration. This interesting and timely subject certainly requires further study.

Society as a Binary Alloy

In 1971, Thomas Schelling [136] proposed the now-famous mathematical model of segregation. He considered a lattice, where the sites can be occupied by agents of two types, e.g., blacks and whites in the problem of racial segregation. He showed that if the agents have some probabilistic preference for the neighbors of the same type, the system spontaneously segregates into black and white neighborhoods. This mathematical model is similar to the so-called Ising model, which is a popular model for studying phase transitions in physics. In this model, each lattice site is occupied by a magnetic atom, whose magnetic moment has only two possible orientations, up or down. The interaction energy between two neighboring atoms depends on whether their magnetic moments point in the same or in the opposite directions. In physics language, the segregation found by Schelling represents a phase transition in this system.

Another similar model is the binary alloy, a mixture of two elements which attract or repel each other. It was noticed in [137] that the behavior of actual binary alloys is strikingly similar to social segregation. In the following papers [42,138], this mathematical analogy was developed further and compared with social data. Interesting concepts, such as the coexistence curve between two phases and the solubility limit, were discussed in this work. The latter concept means that a small amount of one substance dissolves in another up to some limit, but phase separation (segregation) develops for higher concentrations. Recently, similar ideas were rediscovered in [139,140,141]. The vast experience of physicists in dealing with phase transitions and alloys may be helpful for practical applications of such models [142].

Future Directions, Criticism, and Conclusions

The statistical models described in this review are quite simple. It is commonly accepted in physics that theoretical models are not intended to be photographic copies of

reality, but rather to be caricatures, capturing the most essential features of a phenomenon with a minimal number of details. With only few rules and parameters, the models discussed in Sect. “[Statistical Mechanics of Money Distribution](#)”, “[Statistical Mechanics of Wealth Distribution](#)”, and “[Data and Models for Income Distribution](#)” reproduce spontaneous development of stable inequality, which is present in virtually all societies. It is amazing that the calculated Gini coefficients, $G = 1/2$ for individuals and $G = 3/8$ for families, are actually very close to the US income data, as shown in Figs. 8 and 11. These simple models establish a baseline and a reference point for development of more sophisticated and more realistic models. Some of these future directions are outlined below.

Future Directions

Agents with a Finite Lifespan The models discussed in this review consider immortal agents who live forever, like atoms. However, humans have a finite lifespan. They enter the economy as young people and exit at an old age. Evolution of income and wealth as functions of age is studied in economics using the so-called overlapping-generations model. The absence of the age variable was one of the criticisms of econophysics by the economist Paul Anglin [31]. However, the drawback of the standard overlapping-generations model is that there is no variation of income and wealth between agents of the same age, because it is a representative-agent model. It would be best to combine stochastic models with the age variable. Also, to take into account inflation of average income, (22) should be rewritten for relative income, in the spirit of (17). These modifications would allow one to study the effects of demographic waves, such as baby boomers, on the distributions of income and wealth.

Agent-Based Simulations of the Two-Class Society

The empirical data presented in Sect. “[Empirical Data on Income Distribution](#)” show quite convincingly that the US population consists of two very distinct classes characterized by different distribution functions. However, the theoretical models discussed in Sect. “[Statistical Mechanics of Money Distribution](#)” and “[Statistical Mechanics of Wealth Distribution](#)” do not produce two classes, although they do produce broad distributions. Generally, not much attention has been paid in the agent-based literature to simulation of two classes. One exception is [143], in which spontaneous development of employers and employees classes from initially equal agents was simulated [36]. More work in this direction would be certainly desirable.

Access to Detailed Empirical Data A great amount of statistical information is publicly available on the Internet, but not for all types of data. As discussed in Sect. “[Empirical Data on Money and Wealth Distributions](#)”, it would be very interesting to obtain data on the distribution of balances on bank accounts, which would give information about the distribution of money (as opposed to wealth). As discussed in Sect. “[Theoretical Models of Income Distribution](#)”, it would be useful to obtain detailed data on income mobility, to verify the additive and multiplicative hypotheses for income dynamics. Income distribution is often reported as a mix of data on individual income and family income, when the counting unit is a tax return (joint or single) or a household. To have a meaningful comparison with theoretical models, it is desirable to obtain clean data where the counting unit is an individual. Direct collaboration with statistical agencies would be very useful.

Economies in Transition Inequality in developed capitalist countries is generally quite stable. The situation is very different for former socialist countries making a transition to a market economy. According to the World Bank data [103], the average Gini coefficient for family income in eastern Europe and the former Soviet Union jumped from 25% in 1988 to 47% in 1993. The Gini coefficient in the socialist countries before the transition was well below the equilibrium value of 37.5% for market economies. However, the fast collapse of socialism left these countries out of market equilibrium and generated a much higher inequality. One may expect that, with time, their inequality will decrease to the equilibrium value of 37.5%. It would be very interesting to trace how fast this relaxation takes place. Such a study would also verify whether the equilibrium level of inequality is universal for all market economies.

Relation to Physical Energy The analogy between energy and money discussed in Sect. “[Conservation of Money](#)” is a formal mathematical analogy. However, actual physical energy with low entropy (typically in the form of fossil fuel) also plays a very important role in the modern economy, as the basis of current human technology. In view of the looming energy and climate crisis, it is imperative to find realistic ways for making a transition from the current “disposable” economy based on “cheap” and “unlimited” energy and natural resources to a sustainable one. Heterogeneity of human society is one of the important factors affecting such a transition. Econophysics, at the intersection of energy, entropy, economy, and statistical physics, may play a useful role in this quest [144].

Criticism from Economists

As econophysics is gaining popularity, some criticism has appeared from economists [31], including those who are closely involved with the econophysics movement [32,33,34]. This reflects a long-standing tradition in economic and social sciences of writing critiques on different schools of thought. Much of the criticism is useful and constructive and is already being accommodated in econophysics work. However, some criticism results from misunderstanding or miscommunication between the two fields and some from significant differences in scientific philosophy. Several insightful responses to the criticism have been published [145,146,147]; see also [7,148]. In this section, we briefly address the issues that are directly related to the material discussed in this review.

Awareness of Previous Economics Literature One complaint of [31,32,33,34] is that physicists are not well aware of the previous economics literature and either rediscover known results or ignore well-established approaches. To address this issue, it is useful to keep in mind that science itself is a complex system, and scientific progress is an evolutionary process with natural selection. The sea of scientific literature is enormous, and nobody knows it all. Recurrent rediscovery of regularities in the natural and social world only confirms their validity. Independent rediscovery usually brings a different perspective, broader applicability range, higher accuracy, and better mathematical treatment, so there is progress even when some overlap with previous results exists. Physicists are grateful to economists for bringing relevant and specific references to their attention. Since the beginning of modern econophysics, many old references have been uncovered and are now routinely cited.

However, not all old references are relevant to the new development. For example, Gallegati et al. [33] complained that the econophysics literature on income distribution ignores the so-called Kuznets hypothesis [149]. The Kuznets hypothesis postulates that income inequality first rises during an industrial revolution and then decreases, producing an inverted-U-shaped curve. Gallegati et al. [33] admitted that, to date, the large amount of literature on the Kuznets hypothesis is inconclusive. Kuznets [149] mentioned that this hypothesis applies to the period from colonial times to the 1970s; however, the empirical data for this period are sparse and not very reliable. The econophysics literature deals with reliable volumes of data for the second half of the twentieth century, collected with the introduction of computers. It is not clear what is the definition of industrial revolution

and when exactly it starts and ends. The chain of technological progress seems to be continuous (steam engine, internal combustion engine, cars, plastics, computers, Internet), so it is not clear where the purported U-curve is supposed to be placed in time. Thus, the Kuznets hypothesis appears to be, in principle, unverifiable and unfalsifiable. The original paper by Kuznets [149] actually does not contain any curves, but it has one table filled with made-up, imaginary data! Kuznets admits that he has “neither the necessary data nor a reasonably complete theoretical model” (p. 12 in [149]). So, this paper is understandably ignored by the econophysics community. In fact, the data analysis for 1947–1984 shows amazing stability of income distribution [150], consistent with Fig. 11. The increase of inequality in the 1990s resulted from growth of the upper class relative to the lower class, but the relative inequality within the lower class remains very stable, as shown in Fig. 7.

Reliance on Visual Data Analysis Another complaint of [33] is that econophysicists favor graphic analysis of data over the formal and “rigorous” testing prescribed by mathematical statistics, as favored by economists. This complaint goes against the trend of all sciences to use increasingly sophisticated data visualization for uncovering regularities in complex systems. The thick IRS publication 1304 [151] is filled with data tables, but has virtually no graphs. Despite the abundance of data, it gives a reader no idea about income distribution, whereas plotting the data immediately gives insight. However, intelligent plotting is the art with many tools, which not many researchers have mastered. The author completely agrees with Gallegati et al. [33] that too many papers mindlessly plot any kind of data on a log–log scale, pick a finite interval, where any smooth curved line can be approximated by a straight line, and claim that there is a power law. In many cases, replotting the same data on a log–linear scale converts a curved line into a straight line, which means that the law is actually exponential.

Good visualization is extremely helpful in identifying trends in complex data, which can then be fitted to a mathematical function; however, for a complex system, such a fit should not be expected with infinite precision. The fundamental laws of physics, such as Newton’s law of gravity or Maxwell’s equations, are valid with enormous precision. However, the laws in condensed matter physics, uncovered by experimentalists with a combination of visual analysis and fitting, usually have much lower precision, at best 10% or so. Most of these laws would fail the formal criteria of mathematical statistics. Nevertheless these approximate laws are enormously useful in practice, and ev-

eryday devices engineered on the basis of these laws work very well for all of us.

Because of the finite accuracy, different functions may produce equally good fits. Discrimination between the exponential, Gamma, and log-normal functions may not be always possible [122]. However, the exponential function has fewer fitting parameters, so it is preferable on the basis of simplicity. The other two functions can simply mimic the exponential function with a particular choice of the additional parameters [122]. Unfortunately, many papers in mathematical statistics introduce too many fitting parameters into complicated functions, such as the generalized beta distribution mentioned in [33]. Such overparameterization is more misleading than insightful for data fitting.

Quest for Universality Gallegati et al. [33] criticized physicists for trying to find universality in economics data. They also seemed to equate the concepts of power law, scaling, and universality. These are three different, albeit overlapping, concepts. Power laws usually apply only to a small fraction of data at the high ends of various distributions. Moreover, the exponents of these power laws are usually nonuniversal and vary from case to case. Scaling means that the shape of a function remains the same when its scale changes. However, the scaling function does not have to be a power-law function. A good example of scaling is shown in Fig. 7, where income distributions for the lower class collapse on the same exponential line for about 20 years of data. We observe amazing universality of income distribution, unrelated to a power law. In a general sense, the diffusion equation is universal, because it describes a wide range of systems, from dissolution of sugar in water to a random walk in the stock market.

Universalities are not easy to uncover, but they form the backbone of regularities in the world around us. This is why physicists are so interested in them. Universalities establish the first-order effect, and deviations represent the second-order effect. Different countries may have somewhat different distributions, and economists often tend to focus on these differences. However, this focus on details misses the big picture that, in the first approximation, the distributions are quite similar and universal.

Theoretical Modeling of Money, Wealth, and Income

It was pointed out in [31,33,34] that many econophysics papers confuse or misuse the terms for money, wealth, and income. It is true that terminology is sloppy in many papers and should be refined. However, the terms in [25,26] are quite precise, and this review clearly distinguishes between these concepts in Sect. “Statistical Mechanics of Money Distribution”, “Statistical Mechanics of Wealth

Distribution”, and “Data and Models for Income Distribution”.

One contentious issue is about conservation of money. Gallegati et al. [33] agree that “transactions are a key economic process, and they are necessarily conservative”, i. e., money is indeed conserved in transactions between agents. However, Anglin [31], Gallegati et al. [33], and Lux [34] complain that the models of conservative exchange do not consider production of goods, which is the core economic process and the source of economic growth. Material production is indeed the ultimate goal of an economy, but it does not violate conservation of money by itself. One can grow coffee beans, but nobody can grow money on a money tree. Money is an artificial economic device that is designed to be conserved. As explained in Sect. “Statistical Mechanics of Money Distribution”, the money transfer models implicitly assume that money in transactions is voluntarily paid for goods and services generated by production for the mutual benefit of the parties. In principle, one can introduce a billion variables to keep track of every coffee bean and other product of the economy. What difference would it make for the distribution of money? Despite the claims in [31,33], there is no contradiction between models of conservative exchange and the classic work of Adam Smith and David Ricardo. The difference is only in the focus: We keep track of money, whereas they keep track of coffee beans, from production to consumption. These approaches address different questions, but do not contradict each other. Because money constantly circulates in the system as payment for production and consumption, the resulting statistical distribution of money may very well not depend on what exactly is produced and in what quantities.

In principle, the models with random transfers of money should be considered as a reference point for developing more sophisticated models. Despite the totally random rules and “zero intelligence” of the agents, these models develop well-characterized, stable, and stationary distributions of money. One can modify the rules to make the agents more intelligent and realistic and see how much the resulting distribution changes relative to the reference one. Such an attempt was made in [32] by modifying the model of [77] with various more realistic economic ingredients. However, despite the modifications, the resulting distributions were essentially the same as in the original model. This example illustrates the typical robustness and universality of statistical models: Modifying details of microscopic rules does not necessarily change the statistical outcome.

Another misconception, elaborated in [32,34], is that the money transfer models discussed in Sect. “Statistical

Mechanics of Money Distribution” imply that money is transferred by fraud, theft, and violence, rather than voluntarily. One should keep in mind that the catchy labels “theft-and-fraud”, “marriage-and-divorce”, and “yard-sale” were given to the money transfer models by the journalist Brian Hayes [152] in a popular article. Econophysicists who originally introduced and studied these models do not subscribe to this terminology, although the early work of Angle [51] did mention violence as one source of redistribution. In the opinion of the author, it is indeed difficult to justify the proportionality rule (8), which implies that agents with high balances pay proportionally greater amounts in transactions than agents with low balances. However, the additive model of [25], where money transfers Δm are independent of money balances m_i of the agents, does not have this problem. As explained in Sect. “The Boltzmann–Gibbs Distribution of Money”, this model simply means that all agents pay the same prices for the same product, although prices may be different for different products. So, this model is consistent with voluntary transactions in a free market.

McCauley [145] argued that conservation of money is violated by credit. As explained in Sect. “Models with Debt”, credit does not violate conservation law, but creates positive and negative money without changing net worth. Negative money (debt) is as real as positive money. McCauley [145] claimed that money can be easily created with the tap of a computer key via credit. Then why would an employer not tap the key and double salaries, or a funding agency double research grants? Because budget constraints are real. Credit may provide a temporary relief, but sooner or later it has to be paid back. Allowing debt may produce a double-exponential distribution as shown in Fig. 3, but it does not change the distribution fundamentally.

As discussed in Sect. “Conservation of Money”, a central bank or a central government can inject new money into the economy. As discussed in Sect. “Statistical Mechanics of Wealth Distribution”, wealth is generally not conserved. As discussed in Sect. “Data and Models for Income Distribution”, income is different from money and is described by a different model (22). However, the empirical distribution of income shown in Fig. 6 is qualitatively similar to the distribution of wealth shown in Fig. 5, and we do not have data on money distribution.

Conclusions

The “invasion” of physicists into economics and finance at the turn of the millennium is a fascinating phenomenon. The physicist Joseph McCauley proclaims that “Econo-

physics will displace economics in both the universities and boardrooms, simply because what is taught in economics classes doesn't work" [153]. Although there is some truth in his arguments [145], one may consider a less radical scenario. Econophysics may become a branch of economics, in the same way as games theory, psychological economics, and now agent-based modeling became branches of economics. These branches have their own interests, methods, philosophy, and journals. The main contribution from the infusion of new ideas from a different field is not in answering old questions, but in raising new questions. Much of the misunderstanding between economists and physicists happens not because they are getting different answers, but because they are answering different questions.

The subject of income and wealth distributions and social inequality was very popular at the turn of another century and is associated with the names of Pareto, Lorenz, Gini, Gibrat, and Champernowne, among others. Following the work by Pareto, attention of researchers was primarily focused on the power laws. However, when physicists took a fresh, unbiased look at the empirical data, they found a different, exponential law for the lower part of the distribution. The motivation for looking at the exponential law, of course, came from the Boltzmann–Gibbs distribution in physics. Further studies provided a more detailed picture of the two-class distribution in a society. Although social classes have been known in political economy since Karl Marx, the realization that they are described by simple mathematical distributions is quite new. Demonstration of the ubiquitous nature of the exponential distribution for money, wealth, and income is one of the new contributions produced by econophysics.

Bibliography

Primary Literature

- Chakrabarti BK (2005) Econophys-Kolkata: a short story. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of Wealth Distributions*. Springer, Milan, pp 225–228
- Carbone A, Kaniadakis G, Scarfone AM (2007) Where do we stand on econophysics? *Phys A* 382:xi–xiv
- Stanley HE et al (1996) Anomalous fluctuations in the dynamics of complex systems: from DNA and physiology to econophysics. *Phys A* 224:302–321
- Mantegna RN, Stanley HE (1999) *An introduction to econophysics: correlations and complexity in finance*. Cambridge University Press, Cambridge
- Galam S (2004) Sociophysics: a personal testimony. *Phys A* 336:49–55
- Galan S, Gefen Y, Shapir Y (1982) Sociophysics: a new approach of sociological collective behaviour. I. Mean-behaviour description of a strike. *J Math Soc* 9:1–13
- Stauffer D (2004) Introduction to statistical physics outside physics. *Phys A* 336:1–5
- Schweitzer F (2003) *Brownian agents and active particles: collective dynamics in the natural and social sciences*. Springer, Berlin
- Weidlich W (2000) *Sociodynamics: a systematic approach to mathematical modeling in the social sciences*. Harwood Academic Publishers, Amsterdam
- Chakrabarti BK, Chakraborti A, Chatterjee A (eds) (2006) *Econophysics and sociophysics: trends and perspectives*. Wiley-VCH, Berlin
- Ball P (2002) The physical modelling of society: a historical perspective. *Phys A* 314:1–14
- Ball P (2004) *Critical mass: how one thing leads to another*. Farrar, Straus and Giroux, New York
- Boltzmann L (1905) *Populäre Schriften*. Barth, Leipzig, p 360
- Austrian Central Library for Physics (2006) *Ludwig Boltzmann 1844–1906*. ISBN 3-900490-11-2. Vienna
- Pareto V (1897) *Cours d'Économie Politique*. L'Université de Lausanne
- Mirowski P (1989) *More heat than light: economics as social physics, physics as nature's economics*. Cambridge University Press, Cambridge
- Majorana E (1942) Il valore delle leggi statistiche nella fisica e nelle scienze sociali. *Scientia* 36:58–66 (English translation by Mantegna RN in: Bassani GF (ed) (2006) *Ettore Majorana Scientific Papers*. Springer, Berlin, pp 250–260)
- Montroll EW, Badger WW (1974) *Introduction to quantitative aspects of social phenomena*. Gordon and Breach, New York
- Föllmer H (1974) Random economies with many interacting agents. *J Math Econ* 1:51–62
- Blume LE (1993) The statistical mechanics of strategic interaction. *Games Econ Behav* 5:387–424
- Foley DK (1994) A statistical equilibrium theory of markets. *J Econ Theory* 62:321–345
- Durlauf SN (1997) Statistical mechanics approaches to socioeconomic behavior. In: Arthur WB, Durlauf SN, Lane DA (eds) *The Economy as a Complex Evolving System II*. Addison-Wesley, Redwood City, pp 81–104
- Anderson PW, Arrow KJ, Pines D (eds) (1988) *The economy as an evolving complex system*. Addison-Wesley, Reading
- Rosser JB (2008) Econophysics. In: Blume LE, Durlauf SN (eds) *New Palgrave Dictionary of Economics*, 2nd edn. Macmillan, London (in press)
- Drăgulescu AA, Yakovenko VM (2000) Statistical mechanics of money. *Europ Phys J B* 17:723–729
- Chakraborti A, Chakrabarti BK (2000) Statistical mechanics of money: how saving propensity affects its distribution. *Europ Phys J B* 17:167–170
- Bouchaud JP, Mézard M (2000) Wealth condensation in a simple model of economy. *Phys A* 282:536–545
- Wannier GH (1987) *Statistical physics*. Dover, New York
- Lopez-Ruiz R, Sanudo J, Calbet X (2007) Geometrical derivation of the Boltzmann factor. Available via DIALOG. <http://arxiv.org/abs/0707.4081>. Accessed 1 Jul 2008
- Lopez-Ruiz R, Sanudo J, Calbet X (2007) On the equivalence of the microcanonical and the canonical ensembles: a geometrical approach. Available via DIALOG. <http://arxiv.org/abs/0708.1866>. Accessed 1 Jul 2008
- Anglin P (2005) Econophysics of wealth distributions: a comment. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds)

- Econophysics of wealth distributions. Springer, New York, pp 229–238
32. Lux T (2005) Emergent statistical wealth distributions in simple monetary exchange models: a critical review. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 51–60
 33. Gallegati M, Keen S, Lux T, Ormerod P (2006) Worrying trends in econophysics. *Phys A* 370:1–6
 34. Lux T (2008) Applications of statistical physics in finance and economics. In: Rosser JB (ed) *Handbook of complexity research*. Edward Elgar, Cheltenham, UK and Northampton, MA (in press)
 35. Computer animation videos of money-transfer models. <http://www2.physics.umd.edu/~yakovenk/econophysics/animation.html>. Accessed 1 Jul 2008
 36. Wright I (2007) Computer simulations of statistical mechanics of money in mathematica. Available via DIALOG. <http://demonstrations.wolfram.com/StatisticalMechanicsOfMoney>. Accessed 1 Jul 2008
 37. McConnell CR, Brue SL (1996) *Economics: principles, problems, and policies*. McGraw-Hill, New York
 38. Patriarca M, Chakraborti A, Kaski K, Germano G (2005) Kinetic theory models for the distribution of wealth: Power law from overlap of exponentials. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 93–110
 39. Bennati E (1988) Un metodo di simulazione statistica per l'analisi della distribuzione del reddito. *Rivista Internazionale di Scienze Economiche e Commerciali* 35:735–756
 40. Bennati E (1993) Il metodo di Montecarlo nell'analisi economica. *Rassegna di Lavori dell'ISCO (Istituto Nazionale per lo Studio della Congiuntura)*, Anno X 4:31–79
 41. Scalas E, Garibaldi U, Donadio S (2006) Statistical equilibrium in simple exchange games I: methods of solution and application to the Bennati–Drăgulescu–Yakovenko (BDY) game. *Eur Phys J B* 53:267–272
 42. Mimkes J (2000) Society as a many-particle system. *J Therm Anal Calorim* 60:1055–1069
 43. Mimkes J (2005) Lagrange principle of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 61–69
 44. Shubik M (1999) *The theory of money and financial institutions*, vol 2. MIT Press, Cambridge, p 192
 45. Mandelbrot B (1960) The Pareto–Lévy law and the distribution of income. *Int Econ Rev* 1:79–106
 46. Braun D (2001) Assets and liabilities are the momentum of particles and antiparticles displayed in Feynman-graphs. *Phys A* 290:491–500
 47. Fischer R, Braun D (2003) Transfer potentials shape and equilibrate monetary systems. *Phys A* 321:605–618
 48. Fischer R, Braun D (2003) Nontrivial bookkeeping: a mechanical perspective. *Phys A* 324:266–271
 49. Xi N, Ding N, Wang Y (2005) How required reserve ratio affects distribution and velocity of money. *Phys A* 357:543–555
 50. Ispolatov S, Krapivsky PL, Redner S (1998) Wealth distributions in asset exchange models. *Eur Phys J B* 2:267–276
 51. Angle J (1986) The surplus theory of social stratification and the size distribution of personal wealth. *Soc Forces* 65:293–326
 52. Angle J (1992) The inequality process and the distribution of income to blacks and whites. *J Math Soc* 17:77–98
 53. Angle J (1992) Deriving the size distribution of personal wealth from 'the rich get richer, the poor get poorer'. *J Math Soc* 18:27–46
 54. Angle J (1996) How the Gamma Law of income distribution appears invariant under aggregation. *J Math Soc* 21:325–358
 55. Angle J (2002) The statistical signature of pervasive competition on wage and salary incomes. *J Math Soc* 26:217–270
 56. Angle J (2006) The Inequality Process as a wealth maximizing process. *Phys A* 367:388–414
 57. Engels F (1972) *The origin of the family, private property and the state*, in the light of the researches of Lewis H. Morgan. International Publishers, New York
 58. Patriarca M, Chakraborti A, Kaski K (2004) Gibbs versus non-Gibbs distributions in money dynamics. *Phys A* 340:334–339
 59. Patriarca M, Chakraborti A, Kaski K (2004) Statistical model with a standard Gamma distribution. *Phys Rev E* 70:016104
 60. Repetowicz P, Hutzler S, Richmond P (2005) Dynamics of money and income distributions. *Phys A* 356:641–654
 61. Chatterjee A, Chakrabarti BK, Manna SS (2004) Pareto law in a kinetic model of market with random saving propensity. *Phys A* 335:155–163
 62. Das A, Yarlagadda S (2005) An analytic treatment of the Gibbs-Pareto behavior in wealth distribution. *Phys A* 353:529–538
 63. Chatterjee S, Chakrabarti BK, Stinchcombe RB (2005) Master equation for a kinetic model of a trading market and its analytic solution. *Phys Rev E* 72:026126
 64. Mohanty PK (2006) Generic features of the wealth distribution in ideal-gas-like markets. *Phys Rev E* 74:011117
 65. Patriarca M, Chakraborti A, Germano G (2006) Influence of saving propensity on the power-law tail of the wealth distribution. *Phys A* 369:723–736
 66. Gupta AK (2006) Money exchange model and a general outlook. *Phys A* 359:634–640
 67. Patriarca M, Chakraborti A, Heinsalu E, Germano G (2007) Relaxation in statistical many-agent economy models. *Eur Phys J B* 57:219–224
 68. Ferrero JC (2004) The statistical distribution of money and the rate of money transference. *Phys A* 341:575–585
 69. Scafetta N, Picozzi S, West BJ (2004) An out-of-equilibrium model of the distributions of wealth. *Quant Financ* 4:353–364
 70. Scafetta N, Picozzi S, West BJ (2004) A trade-investment model for distribution of wealth. *Physica D* 193:338–352
 71. Lifshitz EM, Pitaevskii LP (1981) *Physical kinetics*. Pergamon Press, Oxford
 72. Ao P (2007) Boltzmann–Gibbs distribution of fortune and broken time reversible symmetry in econodynamics. *Commun Nonlinear Sci Numer Simul* 12:619–626
 73. Scafetta N, West BJ (2007) Probability distributions in conservative energy exchange models of multiple interacting agents. *J Phys Condens Matter* 19:065138
 74. Chakraborti A, Pradhan S, Chakrabarti BK (2001) A self-organising model of market with single commodity. *Phys A* 297:253–259
 75. Chatterjee A, Chakrabarti BK (2006) Kinetic market models with single commodity having price fluctuations. *Eur Phys J B* 54:399–404
 76. Ausloos M, Pekalski A (2007) Model of wealth and goods dynamics in a closed market. *Phys A* 373:560–568

77. Silver J, Slud E, Takamoto K (2002) Statistical equilibrium wealth distributions in an exchange economy with stochastic preferences. *J Econ Theory* 106:417–435
78. Raberto M, Cincotti S, Focardi SM, Marchesi M (2003) Traders' long-run wealth in an artificial financial market. *Comput Econ* 22:255–272
79. Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. *Phys A* 299:188–197
80. Solomon S, Richmond P (2002) Stable power laws in variable economies; Lotka-Volterra implies Pareto-Zipf. *Europ Phys J B* 27:257–261
81. Huang DW (2004) Wealth accumulation with random redistribution. *Phys Rev E* 69:057103
82. Slanina F (2004) Inelastically scattering particles and wealth distribution in an open economy. *Phys Rev E* 69:046102
83. Cordier S, Pareschi L, Toscani G (2005) On a kinetic model for a simple market economy. *J Statist Phys* 120:253–277
84. Richmond P, Repetowicz P, Hutzler S, Coelho R (2006) Comments on recent studies of the dynamics and distribution of money. *Phys A* 370:43–48
85. Richmond P, Hutzler S, Coelho R, Repetowicz P (2006) A review of empirical studies and models of income distributions in society. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) *Econophysics and sociophysics: trends and perspectives*. Wiley-VCH, Berlin
86. Burda Z, Johnston D, Jurkiewicz J, Kaminski M, Nowak MA, Papp G, Zahed I (2002) Wealth condensation in Pareto macroeconomies. *Phys Rev E* 65:026102
87. Pianegonda S, Iglesias JR, Abramson G, Vega JL (2003) Wealth redistribution with conservative exchanges. *Phys A* 322:667–675
88. Braun D (2006) Nonequilibrium thermodynamics of wealth condensation. *Phys A* 369:714–722
89. Coelho R, Nédá Z, Ramasco JJ, Santos MA (2005) A family-network model for wealth distribution in societies. *Phys A* 353:515–528
90. Iglesias JR, Gonçalves S, Pianegonda S, Vega JL, Abramson G (2003) Wealth redistribution in our small world. *Phys A* 327:12–17
91. Di Matteo T, Aste T, Hyde ST (2004) Exchanges in complex networks: income and wealth distributions. In: Mallamace F, Stanley HE (eds) *The physics of complex systems (New advances and perspectives)*. IOS Press, Amsterdam, p 435
92. Hu MB, Jiang R, Wu QS, Wu YH (2007) Simulating the wealth distribution with a Richest-Following strategy on scale-free network. *Phys A* 381:467–472
93. Bak P, Nørrelykke SF, Shubik M (1999) Dynamics of money. *Phys Rev E* 60:2528–2532
94. Her Majesty Revenue and Customs (2003) Distribution of personal wealth. Available via DIALOG. http://www.hmrc.gov.uk/stats/personal_wealth/wealth_oct03.pdf. Accessed 1 Jul 2008
95. Drăgulescu AA, Yakovenko VM (2001) Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States. *Phys A* 299:213–221
96. Klass OS, Biham O, Levy M, Malcai O, Solomon S (2007) The Forbes 400, the Pareto power-law and efficient markets. *Europ Phys J B* 55:143–147
97. Sinha S (2006) Evidence for power-law tail of the wealth distribution in India. *Phys A* 359:555–562
98. Abul-Magd AY (2002) Wealth distribution in an ancient Egyptian society. *Phys Rev E* 66:057104
99. Kakwani N (1980) *Income Inequality and Poverty*. Oxford University Press, Oxford
100. Champernowne DG, Cowell FA (1998) *Economic inequality and income distribution*. Cambridge University Press, Cambridge
101. Atkinson AB, Bourguignon F (eds) (2000) *Handbook of income distribution*. Elsevier, Amsterdam
102. Drăgulescu AA, Yakovenko VM (2001) Evidence for the exponential distribution of income in the USA. *Europ Phys J B* 20:585–589
103. Drăgulescu AA, Yakovenko VM (2003) Statistical mechanics of money, income, and wealth: a short survey. In: Garrido PL, Marro J (eds) *Modeling of complex systems: seventh granada lectures, Conference Proceedings 661*. American Institute of Physics, New York, pp 180–183
104. Yakovenko VM, Silva AC (2005) Two-class structure of income distribution in the USA: Exponential bulk and power-law tail. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 15–23
105. Silva AC, Yakovenko VM (2005) Temporal evolution of the 'thermal' and 'superthermal' income classes in the USA during 1983–2001. *Europhys Lett* 69:304–310
106. Hasegawa A, Mima K, Duong-van M (1985) Plasma distribution function in a superthermal radiation field. *Phys Rev Lett* 54:2608–2610
107. Desai MI, Mason GM, Dwyer JR, Mazur JE, Gold RE, Krimigis SM, Smith CW, Skoug RM (2003) Evidence for a suprathermal seed population of heavy ions accelerated by interplanetary shocks near 1 AU. *Astrophys J* 588:1149–1162
108. Collier MR (2004) Are magnetospheric suprathermal particle distributions (κ functions) inconsistent with maximum entropy considerations? *Adv Space Res* 33:2108–2112
109. Souma W (2001) Universal structure of the personal income distribution. *Fractals* 9:463–470
110. Souma W (2002) Physics of personal income. In: Takayasu H (ed) *Empirical science of financial fluctuations: the advent of econophysics*. Springer, Tokyo, pp 343–352
111. Fujiwara Y, Souma W, Aoyama H, Kaizoji T, Aoki M (2003) Growth and fluctuations of personal income. *Phys A* 321:598–604
112. Aoyama H, Souma W, Fujiwara Y (2003) Growth and fluctuations of personal and company's income. *Phys A* 324:352–358
113. Strudler M, Petska T, Petska R (2003) An analysis of the distribution of individual income and taxes, 1979–2001. The Internal Revenue Service, Washington DC. Available via DIALOG. <http://www.irs.gov/pub/irs-soi/03strudl.pdf>. Accessed 1 Jul 2008
114. Souma W, Nirei M (2005) Empirical study and model of personal income. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 34–42
115. Nirei M, Souma W (2007) A two factor model of income distribution dynamics. *Rev Income Wealth* 53:440–459
116. Ferrero JC (2005) The monomodal, polymodal, equilibrium and nonequilibrium distribution of money. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 159–167
117. Clementi F, Gallegati M (2005) Pareto's law of income distribution: evidence for Germany, the United Kingdom, the

- United States. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 3–14
118. Clementi F, Gallegati M, Kaniadakis G (2007) κ -generalized statistics in personal income distribution. *Europ Phys J B* 57:187–193
 119. Clementi F, Gallegati M (2005) Power law tails in the Italian personal income distribution. *Phys A* 350:427–438
 120. Clementi F, Di Matteo T, Gallegati M (2006) The power-law tail exponent of income distributions. *Phys A* 370:49–53
 121. Rawlings PK, Reguera D, Reiss H (2004) Entropic basis of the Pareto law. *Phys A* 343:643–652
 122. Banerjee A, Yakovenko VM, Di Matteo T (2006) A study of the personal income distribution in Australia. *Phys A* 370:54–59
 123. Gibrat R (1931) *Les Inégalités Economiques*. Sirely, Paris
 124. Kalecki M (1945) On the Gibrat distribution. *Econometrica* 13:161–170
 125. Champernowne DG (1953) A model of income distribution. *Econ J* 63:318–351
 126. Milaković M (2005) Do we all face the same constraints? In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 184–191
 127. Takayasu H, Sato AH, Takayasu M (1997) Stable infinite variance fluctuations in randomly amplified Langevin systems. *Phys Rev Lett* 79:966–969
 128. Kesten H (1973) Random difference equations and renewal theory for products of random matrices. *Acta Math* 131:207–248
 129. Fiaschi D, Marsili M (2007) Distribution of wealth: theoretical microfoundations and empirical evidence. Working paper. Available via DIALOG. <http://www.dse.ec.unipi.it/personale/docenti/fiaschi/Lavori/distributionWealthMicrofoundations.pdf>. Accessed 1 Jul 2008
 130. Levy M, Solomon S (1996) Power laws are logarithmic Boltzmann laws. *Int J Mod Phys C* 7:595–751
 131. Sornette D, Cont R (1997) Convergent multiplicative processes repelled from zero: power laws and truncated power laws. *J Phys I (France)* 7:431–444
 132. Lydall HF (1959) The distribution of employment incomes. *Econometrica* 27:110–115
 133. Feller W (1966) *An Introduction to Probability Theory and Its Applications*, vol 2. Wiley, New York, p 10
 134. Mimkes J, Aruka Y (2005) Carnot process of wealth distribution. In: Chatterjee A, Yarlagadda S, Chakrabarti BK (eds) *Econophysics of wealth distributions*. Springer, Milan, pp 70–78
 135. Mimkes J (2006) A thermodynamic formulation of economics. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) *Econophysics and sociophysics: trends and perspectives*. Wiley-VCH, Berlin, pp 1–33
 136. Schelling TC (1971) Dynamic models of segregation. *J Math Soc* 1:143–186
 137. Mimkes J (1995) Binary alloys as a model for the multicultural society. *J Therm Anal* 43:521–537
 138. Mimkes J (2006) A thermodynamic formulation of social science. In: Chakrabarti BK, Chakraborti A, Chatterjee A (eds) *Econophysics and sociophysics: trends and perspectives*. Wiley-VCH, Berlin
 139. Jeco C, Roehner BM (2007) A physicist's view of the notion of "racism". Available via DIALOG. <http://arxiv.org/abs/0704.2883>. Accessed 1 Jul 2008
 140. Stauffer D, Schulze C (2007) Urban and scientific segregation: the Schelling-Ising model. Available via DIALOG. <http://arxiv.org/abs/0710.5237>. Accessed 1 Jul 2008
 141. Dall'Asta L, Castellano C, Marsili M (2007) Statistical physics of the Schelling model of segregation. Available via DIALOG. <http://arxiv.org/abs/0707.1681>. Accessed 1 Jul 2008
 142. Lim M, Metzler R, Bar-Yam Y (2007) Global pattern formation and ethnic/cultural violence. *Science* 317:1540–1544
 143. Wright I (2005) The social architecture of capitalism. *Phys A* 346:589–620
 144. Defilla S (2007) A natural value unit – Econophysics as arbiter between finance and economics. *Phys A* 382:42–51
 145. McCauley JL (2006) Response to 'Worrying Trends in Econophysics'. *Phys A* 371:601–609
 146. Richmond P, Chakrabarti BK, Chatterjee A, Angle J (2006) Comments on 'Worrying Trends in Econophysics': income distribution models. In: Chatterjee A, Chakrabarti BK (eds) *Econophysics of stock and other markets*. Springer, Milan, pp 244–253
 147. Rosser JB (2006) Debating the Role of Econophysics. Working paper. Available via DIALOG. <http://cob.jmu.edu/rosserjb/>. Accessed 1 Jul 2008
 148. Rosser JB (2006) The nature and future of econophysics. In: Chatterjee A, Chakrabarti BK (eds) *Econophysics of stock and other markets*. Springer, Milan, pp 225–234
 149. Kuznets S (1955) Economic growth and income inequality. *Am Econ Rev* 45:1–28
 150. Levy F (1987) Changes in the distribution of American family incomes, 1947 to 1984. *Science* 236:923–927
 151. Internal Revenue Service (1999) *Statistics of Income–1997, Individual Income Tax Returns*. Publication 1304, Revision 12–99, Washington DC
 152. Hayes B (2002) Follow the money. *Am Sci* 90:400–405
 153. Ball P (2006) Econophysics: culture crash. *Nature* 441:686–688

Books and Reviews

- McCauley J (2004) *Dynamics of markets: econophysics and finance*. Cambridge University Press, Cambridge
- Farmer JD, Shubik M, Smith E (2005) Is economics the next physical science? *Phys Today* 58(9):37–42
- Samanidou E, Zschischang E, Stauffer D, Lux T (2007) Agent-based models of financial markets. *Rep Prog Phys* 70:409–450
- Econophysics forum. Available via DIALOG. <http://www.unifr.ch/econophysics/>. Accessed 1 Jul 2008

Elastic Percolation Networks

PHILLIP M. DUXBURY
Michigan State University, East Lansing, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Basic Theoretical Concepts

[Idealized Experiments](#)
[Chalcogenide Glasses](#)
[Gels and Semiflexible Rod Networks](#)
[Granular Media](#)
[Exact Solution on Bethe Lattices](#)
[Exact Algorithms and Percolative Geometries](#)
[Elastic Critical Behavior](#)
[Final Remarks and Future Directions](#)
[Bibliography](#)

Glossary

- Boson peak** The Boson peak is an excess of low frequency modes observed in glasses, as manifested, for example, in inelastic neutron scattering data. In rigidity percolation the Boson peak is related to the number of floppy modes.
- Constraint** Edges in a graph constrain the degrees of freedom of the nodes in the graph. If edges are independent, each edge acts as one constraint.
- Degrees of freedom** In d dimensions a point object has d degrees of freedom, while a body has $d(d + 1)/2$ degrees of freedom due to rotations and translations.
- Floppy mode** A floppy mode is a deformation of a structure which is soft and in ideal models is treated as a zero energy deformation.
- Generic rigidity** A network is generic if none of its edges are dependent due to the particular geometric arrangement of the nodes in the network. With high probability random networks are generic, while regular lattices are non-generic.
- Isostatic network** An isostatic network is rigid but has no redundant bonds. Isostatic networks are marginally rigid as removal of any edge induces a floppy mode. Ideal generic granular media are isostatic at jamming.
- Redundant bond** A redundant bond is not essential to the rigidity of a structure. Generic networks which contain redundant bonds are overconstrained and internally stressed.
- Rigidity percolation threshold** The rigidity threshold marks the transition from floppy networks which have zero elastic moduli to rigid networks with finite elastic moduli.

Definition of the Subject

Materials or structures with sufficiently low connectivity are floppy and have very low elastic moduli, while at high connectivity they are rigid and have relatively high elastic moduli. Elastic percolation networks describe the transition from floppy to rigid that occurs as the network

connectivity increases. The percolative geometry and elastic behavior near percolation are of particular interest. Conventional percolative geometries describe some experimental systems however the elastic critical behavior falls into several different universality classes. Moreover, distinct percolative geometries occur in systems with only central forces or which have soft torsional forces and in these cases both the geometry and elastic behavior may be distinct from conventional percolation. Granular media manifest a further distinct elastic percolation network, with the concept of an isostatic network underlying elastic behavior near jamming. This rich fundamental research framework is relevant to an enormous range of materials of scientific and technical interest [79], including physical and chemical gels [84,87], semiflexible networks in biology [6,42], proteins [79], chalcogenide glasses [10,79] and granular media [85]. This brief review outlines the broad underlying principles common to these diverse systems.

Introduction

Many materials exhibit a sharp change in elastic behavior as the degree of interconnection in the material is increased. One well studied example is the gel transition, in which a liquid becomes a gel (solid) as the number of short-range crosslinks increases [18,28,73,84,87]. A similar behavior occurs in epoxies where a polymer melt is irreversibly crosslinked by addition of crosslinking agents to form a stiff, hard solid [18,28,72,73,84]. Crosslinking can be invoked using a variety of stimuli, including chemical, microwave or thermal methods, given the appropriate precursors. Vulcanization in rubber formation is another example. Following the work by de Gennes [17], there has been considerable debate about the elastic scaling behavior occurring near the gelation point. Semiflexible rod networks in biology, including actin filaments in the cell and F filaments in blood clots, [6,42] are also gels. Moreover, paper is composed of cellulose rods and their crosslinking is critical to their mechanical performance [1]. As elucidated in Sects. “[Idealized Experiments](#)” – “[Granular Media](#)” the ideas of floppy modes and network rigidity are useful in understanding the elastic behavior of these materials, particularly near the gel point.

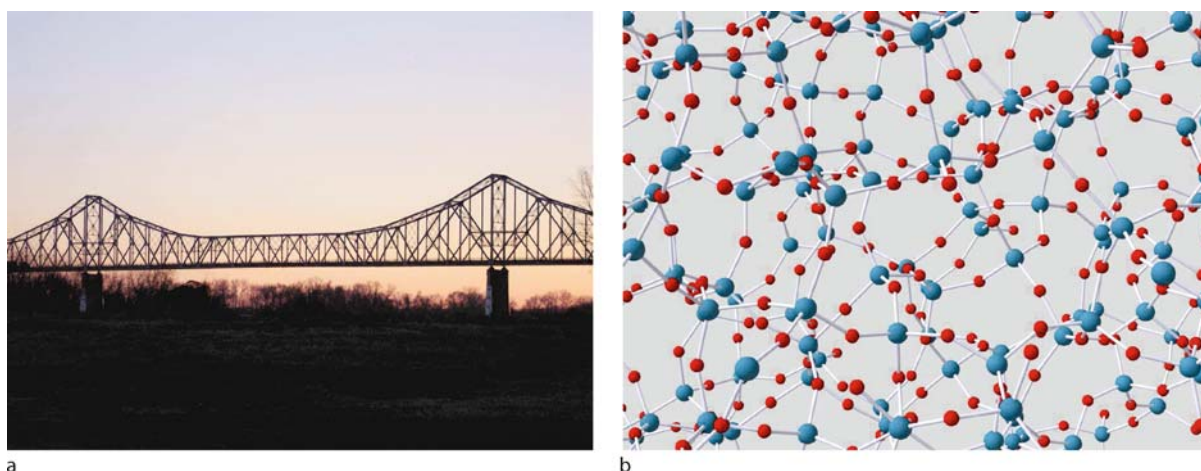
Elastic percolation is richer than conductivity percolation in that transmission of stress is a vector process while transmission of current only requires simple connectivity. Nevertheless, in some cases simple connectivity is sufficient to enable transmission of stress, while in others a more highly connected structure is required [67,79]. At the engineering level, the reasons for these distinctions have been known at least since the time of Maxwell [48],

though explicit models to elucidate the various types of elastic percolation processes in disordered media were only developed in the early 1980's [24,41]. Chalcogenide glasses, for example $Se_{1-x}Ge_x$, are miscible and provide a unique system in which to study the effect of increasing the crosslinking of low coordination networks. Selenium is relatively floppy as it is two-fold coordinated, while germanium is tetrahedrally bonded and rigid [59,78]. Though computational studies of these glasses indicate critical behavior near average co-ordination $r_c = 2.4$ [32], the experimental consensus is for rather smooth elastic moduli near threshold, presumably due to rounding effects such as dihedral forces or entropic elasticity. Nevertheless quantities such as the number of floppy modes in the Boson peak [40] and Raman scattering [26], do have a clear experimental signature near r_c .

The jamming transition of hard particles is different than the examples listed above as hard core repulsion provides a strong resistance to compression while there is no resistance to extension [2,30,55,85]. However in colloidal gels where there is a weak attractive potential, the behavior may be restored to that of rigidity percolation. Even in the absence of these attractive terms, some of the ideas of rigidity percolation are very useful in granular media, particularly the idea of an isostatic network [21,55]. In this picture, the onset of rigidity in a granular packing occurs at a stress free isostatic critical point. By considering systems with soft repulsive potentials, the elastic behavior in

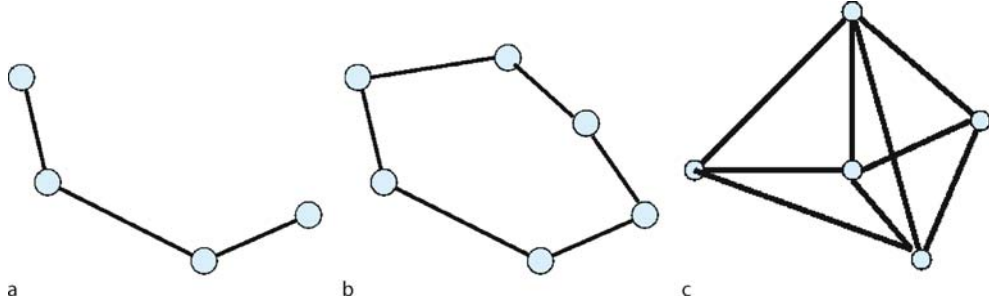
compression can be described by deviations from this critical point, as evident in recent large scale computational studies [57]. Experiments generally support the idea of an isostatic critical point [30,46,85].

The onset of elastic percolation can be determined by making realistic models and by studying their response to an applied stress. This strategy is important, however a deeper understanding has emerged through combinatorial methods which identify constraints that reduce the number of floppy modes [38,52,59,78]. These constraint counting methods enable determination of the ability of a structure or graph to transmit stress, as developed by Maxwell for engineering structures [48] (see Fig. 1). Understanding of stress bearing geometries enables a broader view of elasticity percolation, unifying a broad range of experimental examples and elucidating the aspects which are universal and those that are not. It also enables unification of perspectives from several disciplines (as illustrated in Fig. 1), including engineering, mathematics, material science, physics and more recently biological physics. Algorithms and concepts developed in the graph theory and topology communities have proved particularly rich [34,43,83]. These methods and their relation to elastic percolation are surveyed in Sect. "Exact Algorithms and Percolative Geometries". Constraint counting methods also enable an exact analysis on Bethe lattices [22] which demonstrate that rigidity percolation is often first order (see Sect. "Exact Solution on Bethe Lattices"). From



Elastic Percolation Networks, Figure 1

Constraint counting methods are similar at the scale of bridges or atoms. a Maxwell asked how many beams are required to make engineering structures rigid. Triangulation is the standard method for ensuring enough central force bonds (rods or beams) are available to support a load, as illustrated by the Cairo Mississippi River bridge from www.bridgehunter.com. b Phillips [59] asked how many higher co-ordination atoms are required to make a random network rigid, for example a structure such as silica from www.phys.uu.nl/~Barkema. Dohler et al. [20] showed that Maxwell counting implies that in bond-bending networks rigidity sets in when the average co-ordination is $r_c = 2.4$



Elastic Percolation Networks, Figure 2

Examples of networks which have just enough connectivity to be rigid. These structures are called isostatic and are stress free. They have a finite elastic constant and no floppy modes. **a** When central forces, bond angle forces and dihedral terms are all important (all three terms in Eq. (1) or (2)). **b** When dihedral terms are not important (only the first two terms in Eq. (1) or (2) are important). **c** When only central forces are important (only the first term in (1) or (2))

a theoretical perspective, replica symmetry breaking is not required in solving the rigidity percolation problem on Bethe lattices, though the system is frustrated [22,64].

There are several different types of elastic percolation, depending on the interaction energy, as illustrated by considering standard force fields, such as the CHARMM [12], AMBER [82] and the Dreiding [49] potentials. They contain three short range terms in their potential energy expressions, with a common form being,

$$E = K_b \sum_{\text{bonds}} (b - b_0)^2 + K_\theta \sum_{\text{angles}} (\theta - \theta_0)^2 + K_\phi \sum_{\text{torsions}} (1 + \cos(n\phi - \phi_0)) . \quad (1)$$

The first term in this expression describes the energy required to change the length of a nearest neighbor bond and is a central force term, while the second and third terms provide a restoring force when bond angles or dihedral (torsion) angles are deformed. For the arguments that follow we only need to be able to divide the energy into three terms so that,

$$E = E_{\text{bonds}} + E_{\text{angles}} + E_{\text{torsions}} . \quad (2)$$

Even long-range non-bonded interaction terms can be represented using this type of model, by adding further neighbor bonds. The theoretical problem is then: Given a graph or network where the nodes interact with a potential such as Eq. (1), what is the elastic response? Where is the rigidity percolation threshold located as we increase the average connectivity of the network? How many floppy modes are there?

The most interesting geometric aspect of elastic percolation is that the geometry which can support a stress

depends on the forces which are important, so that three different cases are illustrated by considering Eq. (1): (i) If all terms in Eq. (1) are important so that, $K_b \sim K_\theta \sim K_\phi$, the underlying geometry is the same as connectivity percolation as any connected network transmits stress (see Fig. 2a); (ii) If only the first two terms are important (i. e. $K_b \sim K_\theta \gg K_\phi$) then the underlying geometry is a special sort of rigidity percolation which can be solved using efficient combinatorial methods (see Fig. 2b). This applies to many Chalcogenide glasses as well as to many polymers and proteins, and is often called the bond-bending case; (iii) The hardest case is when only central force terms are present (i. e. $K_b \gg K_\theta, K_\phi$) in three dimensions (see Fig. 2c). This problem is called the central force rigidity problem and has no rigorous combinatorial characterization, so that we have to resort to direct simulations or approximations. Note that in two dimensions the torsional term is absent and the central force rigidity percolation problem can be solved using combinatorial methods [34,37,50].

It is intuitively evident that as the number of connections in a network increases, the system becomes more rigid or constrained. This intuitive observation is the basis of constraint counting methods introduced by James Clerk Maxwell in 1864 [48], who asked the question: “how many edges are required to make a graph internally rigid, so that it can support an applied stress?” He made the following simple constraint counting argument for three dimensional systems. Consider the case of central forces (only the first term in Eq. (1)) and a set of “ n ” nodes connected by “ b ” bonds. Each node has three degrees of freedom, its three translations. If a graph is internally rigid, it still has six degrees of freedom due to its global translations and rotations. Maxwell then stated that the minimal

number of bonds required to make a central force system rigid is given by the relation,

$$b = 3n - 6. \quad (3)$$

The example in Fig. 2c has $n = 5$ and $b = 9$ and so it satisfies this “Maxwell counting” condition. Maxwell counting is not exact but it provides a useful mean field model for most rigidity problems. It can also be applied to connectivity percolation, but it is a much poorer approximation in that case.

In 1970, Laman [43] proved a theorem showing that for a restricted class of rigidity problems, namely planar graphs, a Maxwell counting calculation on subgraphs can be used to infer the rigidity of a graph. This revived interest in combinatorial rigidity and several methods for implementing his theorem were developed, with Hendrickson’s [34] bond-testing procedure yielding the highly efficient algorithms in use today. Extensions of Laman’s theorem to body-bar networks and the conjecture of its extension to molecular frameworks [36,83] have yielded practical algorithms for a wide range of macromolecular systems and other networks. Maxwell’s work is also important in engineering where truss networks are analogous to central force systems and have a variety of applications (see e.g. Fig. 1a).

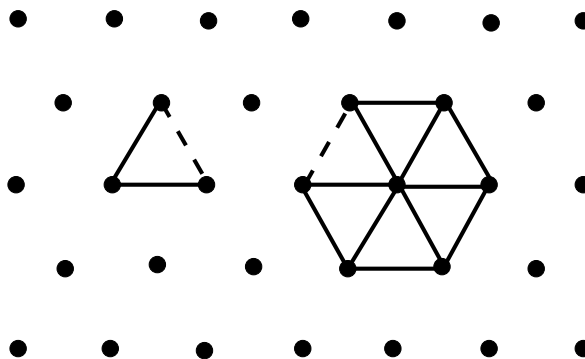
Basic Theoretical Concepts

Maxwell Counting in Random Networks

Maxwell counting [59,78] provides a quick, and frequently quite accurate, estimate of the number of floppy modes and the rigidity threshold. For example, consider bond-diluted triangular lattices where a fraction p of bonds are present. Following Thorpe [78], an unconstrained or flexible degree of freedom is called a floppy mode so that the number of floppy modes, F , remaining in a diluted triangular lattice is,

$$F = 2N - \frac{1}{2}pzN + R, \quad (4)$$

where N is the number of nodes in the lattice $z = 6$ is the coordination number of the lattice and R is the number of redundant bonds. In this expression $2N$ is the number of degrees of freedom in the absence of any bonds, while $pzN/2$ is the number of bonds in the network. Maxwell counting sets $R = 0$ so that every bond added to the lattice reduces the number of floppy modes by one. Redundant bonds are not essential for the rigidity of a structure, though they do increase the elastic moduli and they may also induce internal stresses in rigidity percolation problems. A comparison of the simplest examples of redundant



Elastic Percolation Networks, Figure 3

The simplest subgraphs on a triangular lattice which contain a redundant bond (dashed). Connectivity case (left), $g = 2$ rigidity case (right). From article by C. Moukarzel and P. M. Duxbury in [79]

bonds occurring in connected and rigid clusters on a triangular lattice is presented in Fig. 3.

In Maxwell counting we make the mean field approximation that the number of redundant bonds is zero, and that the rigidity transition occurs when the number of floppy modes goes to zero. In that case, Eq. (4) predicts that the rigidity transition for bond diluted triangular lattices occurs at, $p_r = 2/3$. This is surprisingly accurate as large scale simulations using methods which count the number of redundant bonds exactly, determine the threshold $p_r = 0.6602(3)$ [38] which is within 1% of the Maxwell counting estimate. The percolation threshold for bond bending systems is the same as that for connectivity percolation which is known exactly for bond diluted triangular lattices, $p_c = 2 \sin(\pi/18) = .347 \dots$ It is then clear that the central-force rigidity threshold occurs at a much higher bond concentration than the connectivity threshold [24]. Note that the Maxwell estimate of the connectivity percolation threshold is found from $F = N - pzN/2$, yielding $p_c = 0.33$, which is a much poorer approximation than for the central force rigidity case. In general Maxwell counting is a good approximation to the percolation threshold for all rigidity problems studied so far, except for cases where the connectivity percolation geometry applies!

Though Maxwell counting is very useful, the fact that it ignores redundant bond makes it incomplete. A Bethe lattice approach [22,53], and field theory methods [56,64] provide more complete theories. These approaches and also simulations on certain two [51] and three dimensional lattices [15], e.g. body-centered-cubic lattices, have demonstrated that the rigidity percolation transition is frequently first order with a large jump in the infinite cluster probability at the rigidity threshold. In fully random

networks, the onset of an infinite rigid cluster and the onset of an internally stressed cluster occur at the same threshold. However, in several important cases, the infinite cluster is unstressed or isostatic which is believed to occur in granular media [55,57] due to repulsive terms and in chalcogenide glasses due to self-organization [9,80]. In these cases, a mixed transition may occur where a first order jump and a continuous singularity occur at the same threshold. This behavior was first observed in Bethe lattice models of rigidity percolation [53].

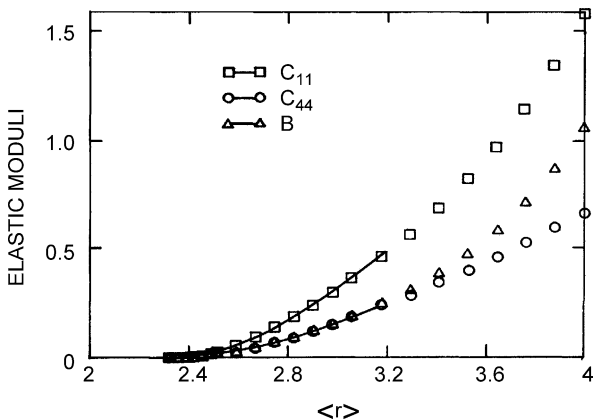
Elastic Behavior

In all random connectivity and rigidity percolation problems studied so far the elastic behavior is continuous near the rigidity threshold, even in cases where the infinite rigid cluster undergoes a first order jump [15,51,56]. However, the question remains open in models of granular media and in self-organizing glass systems where an isostatic critical point plays a role. Direct numerical studies of the elastic constants of bond diluted central force systems indicate that they are quite linear for $p > p_r$, except very near the rigidity threshold, and that mean field estimates of the elastic behavior are surprisingly accurate in these cases [24,25,70]. In contrast the elastic behavior of bond bending networks in three dimensions is non-linear over most of the concentration range (see Fig. 4). Nevertheless, in all cases, the number of floppy modes as a function of $r < r_c$ is close to linear as is consistent with the simple Maxwell estimate.

In ideal elastic percolation networks, the elastic behavior is singular with a behavior typical of a continuous sin-

gularity, $(p - p_r)^T$. The value of the critical exponent T has been a matter of debate. There is general agreement that in cases where bond-bending terms are dominant, the critical exponent T is larger than the conductivity exponent t [41]. For example in two dimensional lattices with bond-bending terms, $t = 1.31(1)$ while $T = 3.96(3)$ [90]. It is also well accepted that continuum systems may exhibit non-universal critical behavior due to the occurrence of necks of varying size in continuum systems [23]. Moreover, even in the absence of continuum effects there are several different elasticity models, with three well studied cases being: gel models where entropic effects are important [17,62] and following de Gennes, $T \approx t$; bonding bending networks where $T \approx t + 2\nu$ [8,24,69,90] and; central force networks where the elastic critical behavior is close to the bond bending case [31,68], at least on regular lattices.

However, in connecting network geometry to elastic behavior in central force networks and in three dimensional networks without torsional forces, it is important to distinguish between generic networks and non-generic networks. This distinction is noticed in granular media where perfectly monodisperse spheres can form regular packings in two dimensions with average contact number six, while random packings, through jamming of regular or polydisperse spheres, have average contact number near four. The latter case is the generic case, while the former is the non-generic case. Non-generic systems are characterized by constraints which are degenerate and occur in the dynamical matrix as dependent equations. In network structures they can occur as special sets of parallel bonds or series combinations of bonds pointing in the same direction (see Sect. "Exact Algorithms and Percolative Geometries"). These configurations do not occur on random networks which are therefore generic. Regular lattices such as the triangular lattice are thus non-generic and the consequences of this on elasticity is still poorly understood.



Elastic Percolation Networks, Figure 4

Elastic behavior of glass models as a function of average site coordination using continuous random network models like that of Fig. 1b, from [32]

Idealized Experiments

Idealized experiments have been carried out to test the prediction that, for a given geometric structure, the elastic moduli have different critical exponents than the those which apply to conductivity. A simple tabletop experiment to test this prediction was devised by Benguigui [7]. In his experiment, holes are drilled on a square grid in a metal sheet. The conductivity, σ and Young's modulus, C , were measured as a function of the remaining metal, ϕ . From these experiments the elastic exponent $T \approx 3.5$, and conductivity exponent $t \approx 1.3$ were extracted, which are in agreement with theoretical expectations in two dimen-

sions, within the experimental uncertainty. Conductivity and elasticity measurements on sintered sub micron silver powder aggregates [19] yielded $t = 2.15(25)$ and $T = 3.8(5)$, which agree with theoretical calculations on bond-bending networks in three dimensions which predict $T = t + 2\nu$ [8] and with simulations [69]. To our knowledge there have been no idealized experiments on central force systems, or on systems without torsional forces in three dimensions.

Chalcogenide Glasses

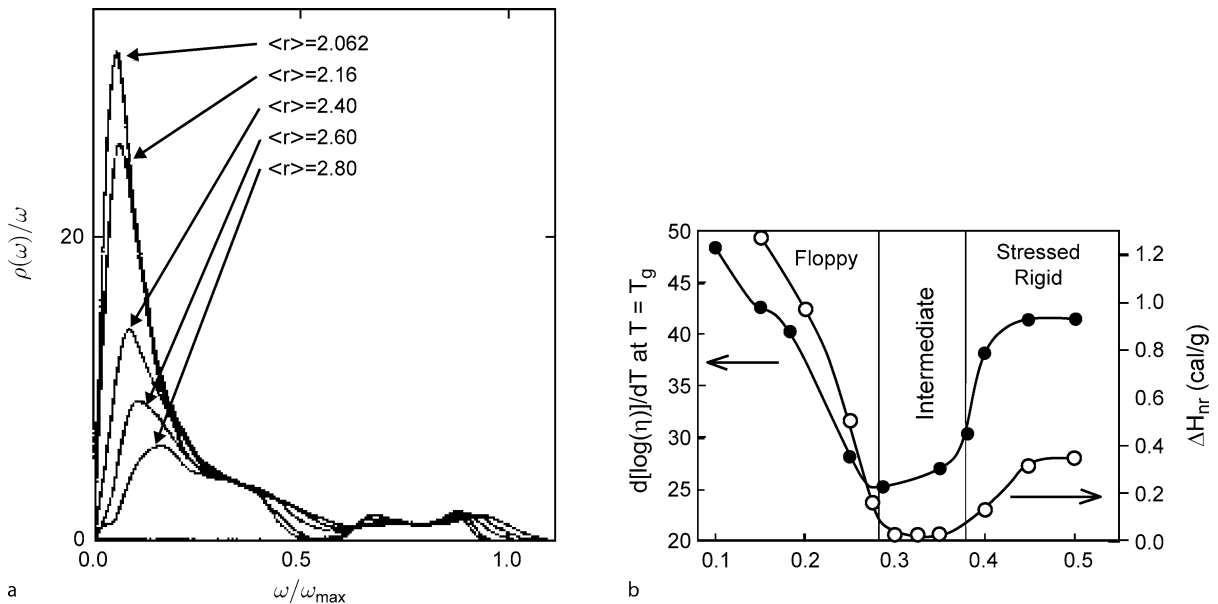
Following Phillips and Thorpe [20,59,78], Maxwell counting for the number of floppy modes in the $Se_{1-x}Ge_x$ system proceeds as follows. Each atom has three degrees of freedom while each r -fold co-ordinated atom imposes $r/2$ central-force constraints and $2r - 3$ bond bending constraints on the network. A two fold Se atom then imposes two constraints while a four fold Ge atom imposes seven constraints. Here we assume that the torsional forces are negligible and the network is continuous with all atoms in one connected cluster. The Maxwell counting estimate for the number of floppy modes is then,

$$F = 3N - 2(1 - x)N - 7xN \quad (5)$$

where N is the number of atoms in the network and $r = 2$ for Se while $r = 4$ for Ge so that the average co-ordina-

tion of a glass is $r = 4x + 2(1 - x)$. Equation (5) indicates that the number of floppy modes is linear in x and goes to zero at $x_c = 0.2$, which corresponds to the critical average co-ordination $r_c = 2.4$ [20]. This critical co-ordination also applies when a fraction y of three-fold coordinated atoms are added to the model, as is relevant to the ternary $Se_{1-x-y}As_yGe_x$. In that case $F = 3N - 2(1 - x - y)N - 9yN/2 - 7xN$, and the average site coordination, $r = 2(1 - x - y) + 3y + 4x$. Of course torsional forces are non-zero in these materials and have to be considered in comparisons with experiment.

The prediction of critical coordination $r_c = 2.4$ in chalcogenide glasses stimulated a search for experimental signatures of rigidity percolation. Though early experiments indicated a singular behavior in the elastic constants near r_c , these results were later found to be due to experimental artifacts. Later experiments indicated that the elastic moduli of chalcogenides are smooth near r_c . However the number of floppy modes [40] and an appropriately defined asymptotic glass transition temperature [3] do clearly indicate a rigidity threshold. The number of floppy modes is manifested in inelastic neutron scattering data [40] where a strong low frequency peak, called the Boson peak, exists for $r < r_c$ (see Fig. 5). The number of modes in the Boson peak gives a measure of the number of floppy modes and the location of the peak provides an estimate of the dihedral forces in the material. Theo-



Elastic Percolation Networks, Figure 5

Evidence for rigidity percolation in chalcogenides. **a** The boson peak near the critical coordination $r_c \approx 2.4$. $\rho(\omega)$ is the density of inelastic neutron scattering modes at frequency ω [40]. **b** Evidence for an intermediate phase in the chalcogenide glasses $Se_{1-x}As_x$. The intermediate phase observed in the derivative of the viscosity $d\eta/dT$ at T_g and in the non-reversible enthalpy ΔH_{nr} , from [9]

retical analysis of the dynamical response of glassy networks yields good agreement with this data [13]. Several other measurements, including Mössbauer spectra [11], Raman scattering [26], vibrational lifetimes [81], also support a critical value of r_c which is close to 2.4.

A stimulating, though still controversial, concept is the intermediate phase in chalcogenide glasses [9,10,80]. The intermediate phase lies between the traditional floppy and rigid phases and occurs by self-organization to avoid internal stress. There are then two co-ordination thresholds r_1 and r_2 , with the first heralding the onset of a rigid but stress-free network and the second the onset of a rigid but internally-stressed network [80]. The extent of the intermediate regime between r_1 and r_2 has been explored in a variety of models [80] and experimental support has come from studies of chalcogenide glasses over narrow composition regimes [9,10], as illustrated in Fig. 5b.

Constraint counting and rigidity concepts developed in chalcogenide glasses have been applied to a wide variety of covalently bonded materials, ranging from amorphous carbons to complex ternaries. Thermal effects along with local structural and chemical ordering often confound a simple interpretation of the data as a purely random rigidity percolation process, nevertheless rigidity and elastic percolation concepts are fundamental to much of the literature in this area [2,3,79,85].

Gels and Semiflexible Rod Networks

Soon after the development of percolation theory, gelation was recognized as a related process and the geometry of gels has been compared to percolation in many different systems [6,42,47,66,73,84,87,89]. Though percolation provides a useful framework for the description of the onset

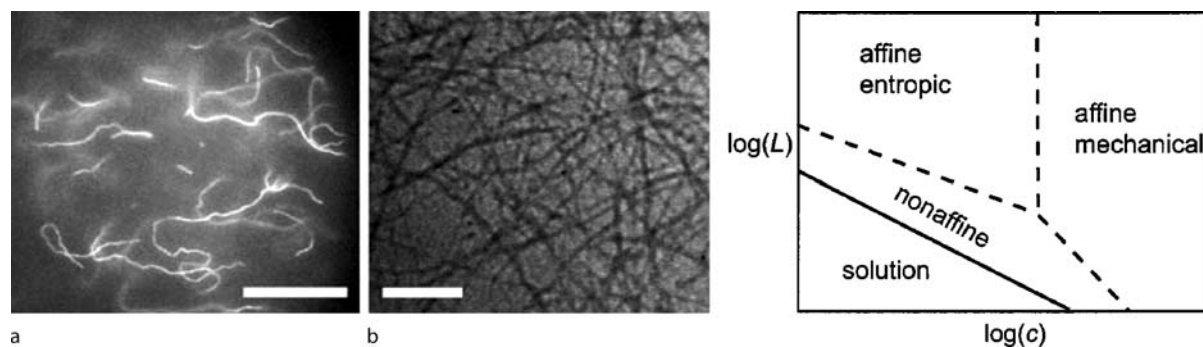
of rigidity in gels, it is often difficult to access the critical regime [65].

Gels are ubiquitous in science, nature and technology though there is no consensus definition of what constitutes a gel. Nevertheless we are all familiar with gels, ranging from jelly to clay dispersions and to chemically crosslinked systems such as rubber and epoxies. Gelation may occur through physical or chemical crosslinking, or through a combination of both processes. Crosslinking is a short range concept and long range forces can sometimes be critical. Nevertheless a broad range of physical and chemical gels can be understood using rigidity concepts based on the change in structure as the crosslinking in a network increases. Even at the single molecule level it is sometimes reasonable to use rigidity percolation ideas to evaluate flexibility as a function of internal crosslinking, for example in proteins [63]. Colloidal gels with strong repulsive terms and weak short-range attractive interactions have many features in common with granular media, as described in the next section.

The original work on gels related the gelation point to connectivity percolation [73], while recent work relates the gel point in rod networks to rigidity percolation [33,35,44,45], as illustrated in Fig. 6. Maxwell counting in two dimensional rod networks [44] states that the number of floppy modes in a network with N rods connected by P pivots is given by,

$$F = 3N - 2P - 3. \quad (6)$$

There are three degrees of freedom per rod, two translations and one rotation, while each pivot removes two translational degrees of freedom from the two rods which it connects. The Maxwell counting estimate, found by set-



Elastic Percolation Networks, Figure 6

Elastic percolation of biofibers. **a, b** Two views of the actin filament network in a human cell, from [77]. The right panel is a schematic phase diagram of semiflexible rod networks, from [33]. The solid line is associated with a rigidity threshold in a random rod network. The mean distance between crosslinks or entanglements is l_c , and $c \approx 1/l_c$. L is the molecular weight of the semiflexible polymers and the rigidity threshold occurs at $L \approx 1/c$ which defines the sol-gel transition in this model

ting $F = 0$ in Eq. (6), for the onset of rigidity in these two dimensional rod networks is then $P_r/N = 3/2$, so that there are three pivot points on each rod. Connectivity percolation occurs at $P_c/N = 1$. Elastic models have been simulated and good agreement with Eq. (6) was found. However experimental gels exhibit correlations in local structure [87], and in the case of fibers orientational ordering and bundling [6,42]. Moreover the elastic behavior of crosslinks is difficult to quantify. These factors are particularly important for non-universal parameters such as the percolation or gelation threshold, and they strongly affect the non-linear response of gels.

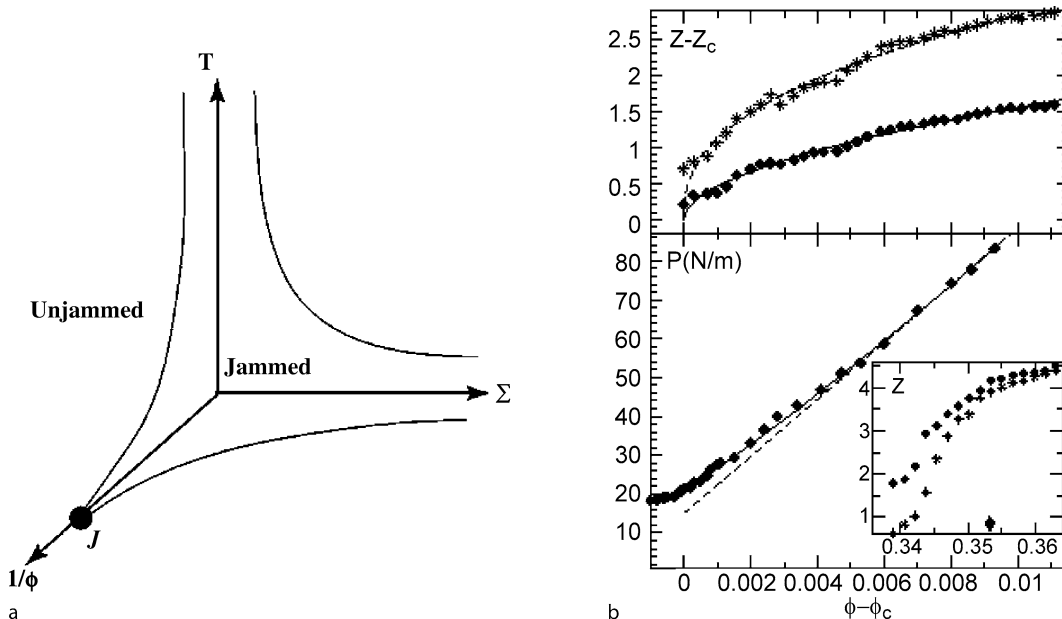
The elastic behavior near gelation is expected to be more universal provided long-range correlations in structure are absent. In a well known paper, de Gennes predicted that the elastic critical exponent, T , of gels should be the same as the conductivity exponent t [17]. However, as discussed in Sect. “Basic Theoretical Concepts” and elaborated upon in Sect. “Elastic Critical Behavior”, purely geometric models predict that the elasticity exponent is significantly larger than $t_{3D} = 2.01(1)$, and is well described by $T_{3D} = t_{3D} + 2\nu_{3D} = 3.77(3)$, where $\nu_{3D} = 0.88$ is the percolation correlation length exponent [8,24,41,69,90]. Over the years many experiments have measured T at the

gelation threshold, for example [4,29,47,65,84,87], with the consensus that $T = t$ in most gels as predicted by de Gennes. This issue will be discussed further in Sect. “Elastic Critical Behavior”.

Granular Media

Jamming occurs as the packing fraction or density of a colloidal or granular system increases. Though the relations between rigidity and jamming have been discussed for many years [30], it is only recently that the relationship has yielded to a clear quantitative analysis [2,55,57,85]. Alexander [2] discussed the presence of floppy modes or rattlers in granular systems and recent work has noted their relation to the Boson peak in glasses [86].

The presence of complex stress bearing networks has been imaged [30] and studied analytically and computationally and the relation to an isostatic network was noted by Moukarzel [55] who developed the idea of jamming as an isostatic critical point. Molecular dynamics simulations in systems with purely repulsive potentials provide a beautiful synthesis of the relations between rigidity, colloidal glasses and granular media [57] (see Fig. 7). These simulations demonstrate that the zero temperature jamming



Elastic Percolation Networks, Figure 7

Behavior of a theoretical and b experimental dense packings near jamming. a schematic of phase behavior found from simulations of purely repulsive models, from [57]. The jamming point J is an isostatic critical point. Σ is the applied shear stress, T is temperature and ϕ is packing fraction. b Behavior of the pressure, P and average co-ordination number $Z - Z_c$ on approach to jamming of photoelastic spheres. $\phi - \phi_c$ is the deviation from close packing, from [46]. In the top panel and in the inset, the diamonds exclude rattlers while the stars include them

point provides insight into the whole phase diagram and enables calculation of the way in which various properties approach zero on approach to jamming from above. These simulations avoid the non-generic critical point which is characteristic of the crystal phase of monodisperse packings, and instead focused upon non-crystalline random packing states which are characteristic of generic rigidity. The computed critical exponents for the shear modulus depend on the form of the repulsive potential but not on the spatial dimension. Later simulations suggest that the stress bearing backbone has a mixed behavior on approach to the jamming point, and its behavior is argued to be essentially the same as that observed in the Bethe lattice theory near an isostatic critical point [71]. Experiments on photoelastic spheres [46] are in general supportive of the physical picture emerging from these theoretical insights (see Fig. 7), including the idea of a jump in average coordination at the jamming point.

Exact Solution on Bethe Lattices

Maxwell counting provides a very useful mean field theory for the rigidity transition, nevertheless, it is not a complete theory as it ignores redundant bonds and does not provide a geometry for the rigidity percolation process. Moreover Maxwell counting predicts that the onset of rigidity occurs when all floppy modes are removed so that the whole network is in one giant isostatic cluster.

Bethe lattice theory is a more complete theory which resolves most of the problems with Maxwell counting methods, moreover it is simple and exactly solvable. In the following we describe the simplest Bethe lattice theory to illustrate the way in which the classic theory of connectivity percolation compares to rigidity percolation [27] and how Maxwell counting should be modified in light of the Bethe analysis. The extent of the intermediate phase [80] can also be simply calculated from the results of the Bethe lattice rigidity theory, as outlined at the end of this section.

The rigidity problem on Bethe lattices encompasses a wide range of different models [53], though here we focus on one generic case [22]. In d dimensions a point object has d degrees of freedom (d translations), while extended objects or bodies also have rotational modes and a total of $d(d+1)/2$ degrees of freedom. We define g as the number of degrees of freedom of a free or unbonded site, so that $g = 1$ corresponds to connectivity percolation, $g = d$ to point objects and $g = d(d+1)/2$ to bodies. A network with N sites (and no edges) has a total of $F = Ng$ degrees of freedom, or floppy (zero frequency) modes. Constraint counting notes that each time an *independent* edge is added to the network, the number floppy

modes is reduced by one, so that

$$F = Ng - E + R \quad (7)$$

where E is the number edges in the graph and R is the number of redundant edges. Note the additional term R on the right hand side. This term is key in understanding the relation between constraint counting and percolation, and in finding algorithms for rigidity percolation. An edge does not reduce the number of floppy modes if it is placed between two sites which are already mutually rigid, in which case this edge is redundant. The simplest examples of subgraphs containing a redundant bond on a triangular lattice are illustrated in Fig. 3 for the connectivity ($g = 1$) and $g = 2$ rigidity cases. Note that any one of the bonds in these structures could be labeled as the redundant one. However once any one of them is removed, all of the others are necessary to ensure the mutual rigidity of the structure. The set of all bonds which are mutually redundant form an overconstrained or stressed cluster. In rigidity theory each edge can be considered to be a central force spring, which means that there is a restoring force only in tension and compression. Then an overconstrained cluster of such springs (with random natural lengths) is internally stressed due to a redundant bond. In the connectivity case each bond is like a wire which can carry current or fluid flow. The simplest overconstrained cluster is then a loop which can support an internal eddy current. Rigid structures which contain no redundant bonds are minimally rigid or isostatic. In connectivity percolation isostatic structures are trees, whereas in ($g > 1$) rigidity percolation isostatic structures always contain many loops.

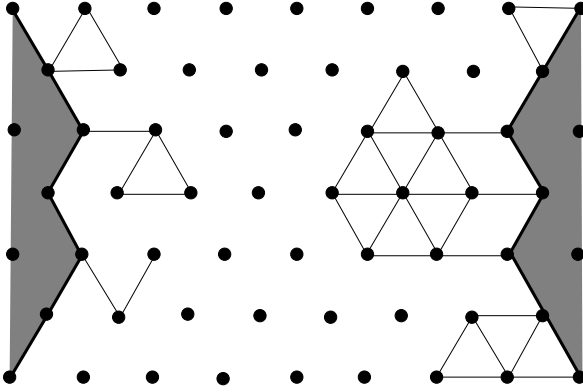
In percolation problems, we are interested in the asymptotic limit of very large graphs ($N \rightarrow \infty$), and it is more convenient to work with intensive quantities, so we define $f(p) = F/gN$ and $r(p) = 2R/gN$, so that

$$f(p) = 1 - \frac{z}{2g}(p - r(p)) \quad (8)$$

where, on average, the number of edges is $E/N = zp/2$. $g * r(p)$ is the probability that a bond is redundant [22,38]. The number of floppy modes, $f(p)$, acts as a free energy for both connectivity and rigidity problems, so that if $\partial f(p)/\partial p$ undergoes a jump discontinuity, the transition is first order [22]. The behavior of this quantity is directly related to the probability that a bond is overconstrained P_{ov} via the relation,

$$\frac{\partial f}{\partial p} = -\frac{z}{2g}(1 - P_{ov}). \quad (9)$$

If the transition is second order, the second derivative $\partial^2 f/\partial p^2 \sim (p - p_c)^{-\alpha}$, where α is the specific heat exponent [38].



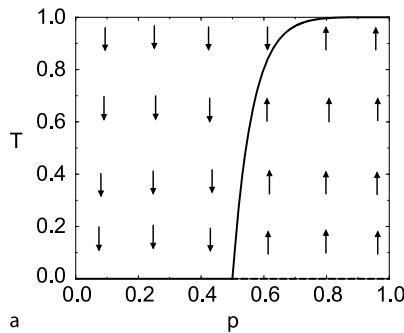
Elastic Percolation Networks, Figure 8

Dangling ends connected to a backbone (shaded). Connectivity case (left) and $g = 2$ rigidity case (right). Dangling ends don't contribute to current (left) or stress (right) which is applied in the vertical direction

The infinite cluster probability in rigidity problems is composed of the stress-bearing backbone plus dangling ends. Dangling ends are rigidly connected to the backbone but are not part of the stress bearing backbone. Examples of dangling ends in the connectivity and rigidity cases are illustrated in Fig. 8 for central force problems on the triangular lattice.

The quantities, $f(p)$, $\partial f(p)/\partial p$, $r(p)$, P_∞ , P_{ov} can be calculated exactly on Bethe lattices of general coordination z , where p is the bond probability of the Bethe lattice. The key quantity from which all the others is derived is the probability, T , that a site on a branch of a Bethe lattice is part of the infinite rigid cluster. This quantity has a steady state solution which is found from the equation,

$$T = \sum_{l=g}^{z-1} \binom{z-1}{l} (pT)^l (1-pT)^{z-1-l}, \quad (10)$$



where T is the probability that a site is connected to the infinite rigid cluster through one branch of the Bethe lattice. The case $g = 1$ recovers the well known Bethe lattice equation for connectivity percolation [27] while for general g , these equations are the same as those for k -core percolation on Bethe lattices [14], where $k = g + 1$. As illustrated in Fig. 9, the solutions to Eq. (10) depend sensitively on the value of g . The case $g = 1$ corresponds to connectivity percolation and shows a typical continuous behavior, while the case $g = 2$ is first order and is typical of all cases where $g > 1$. This means that central force percolation on random graphs is typically first order [22,54].

It is clearly seen from these figures that the rigidity transition is first order on Bethe lattices, while the connectivity transition is second order. We identify the point at which the stable solution becomes nonzero as p_s , the spinodal point, and in the connectivity case $p_s = p_c$ because the transition is second order.

Combining the stable solutions to branch probabilities, yields the infinite cluster probability of the Bethe lattice,

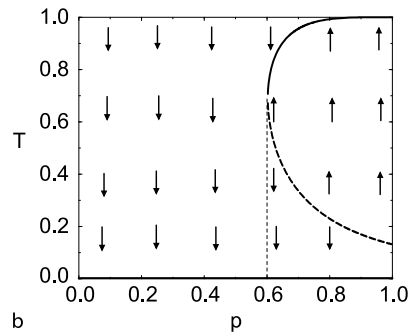
$$P_\infty = \sum_{l=g}^z \binom{z}{l} (pT)^l (1-pT)^{z-l}. \quad (11)$$

The other quantities of interest are found from the stable solution for T as follows. The probability that a bond is overconstrained is [22],

$$P_{ov} = T^2, \quad (12)$$

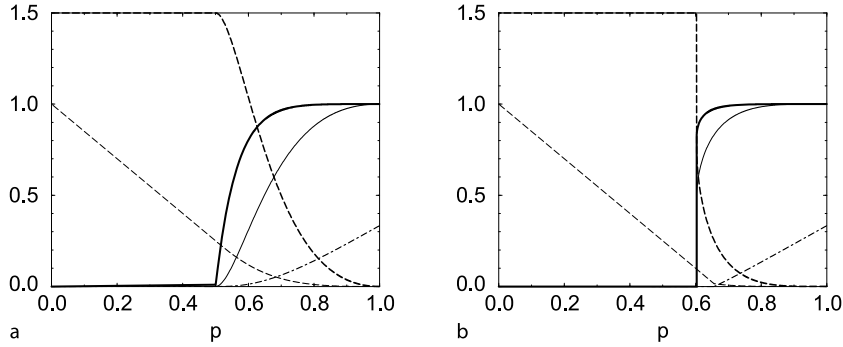
so that Eq. (9) yields,

$$\frac{\partial f}{\partial p} = -\frac{z}{2g} (1 - T^2). \quad (13)$$



Elastic Percolation Networks, Figure 9

Domains of attraction of the mean field equations, Eq. (10). **a** Typical connectivity percolation behavior (this example is $z = 3$); **b** a typical rigidity percolation behavior (this example is $z = 6$, $g = 2$). Dark lines are stable (attractive) solutions and the dashed line is unstable, from article by C. Moukarzel and P. Duxbury in [79]



Elastic Percolation Networks, Figure 10

The order parameters, floppy modes and its derivative as a function of p for a typical connectivity percolation case $z = 3, g = 1$, and **b** a typical rigidity percolation case, $z = 6, g = 2$. The quantities plotted are $f(p)$ (dashed), $r(p)$ (dot dashed), $-df/dp$ (heavy dashed), P_{ov} (thin solid), P_{inf} (heavy solid). In the connectivity case we find $p_c = p_s$, while in rigidity cases $p_s < p_c$. For the rigidity case, $p_s = 0.605$ and $p_c = 0.655$, from article by C. Moukarzel and P. M. Duxbury in [79]

The probability that a bond is on the infinite cluster is,

$$P_{inf} = T^2 + 2TT_1. \quad (14)$$

where T_1 is the probability that a site has one degree of freedom and can also be found straightforwardly [22]. We also want to find the total number of redundant bonds $r(p)$ and the total number of floppy modes $f(p)$. In order to find these quantities, we integrate Eq. (13) and then use Eq. (7). However, the integration of Eq. (13) leads to one free constant. To determine this constant, for a lattice of co-ordination z , we impose the constraint [22],

$$r(1) = 1 - \frac{2g}{z}. \quad (15)$$

We then find that $r(p)$ approaches zero at a critical point p_c , which lies above p_s , for any $g > 1$, and that p_c is close to the Maxwell estimate $p_c = 2g/z$. Results for five key quantities found from Bethe lattice theory are presented in Fig. 10 for typical conductivity and rigidity percolation cases.

The Bethe lattice theory outlined above has been carefully compared with exact simulations of random graphs which demonstrate that it is exact [22]. This is a symmetric theory as the order parameter does not need to be described by a non-trivial distribution. The fact that this symmetric theory is exact [64] and yet applies to glassy systems such as granular media and rod networks is intriguing as replica symmetry breaking would be expected in those cases, as occurs in problems such as KSAT, spin glasses, and lattice gases. It is worth noting that the Bethe lattice Eqs. (10)–(11) are the same as the Bethe lattice equations for k -core percolation [14] which is also being

used as a simple model for granular media [71]. Extensions of the Bethe lattice theory to chalcogenide glasses have been carried out and provide a more complete theory than Maxwell counting [79].

Extension to the Intermediate Phase

In the intermediate phase [80], the network is dominated by a giant rigid but isostatic cluster. To generate such a cluster, we consider adding bonds to a graph, with the constraint that any redundant bond is not added to the network. As described in the next section this is how the exact constraint counting algorithms proceed, so computation of geometric structures which are stress free is straightforward using Hendrickson's bond testing procedure [34].

In the Bethe lattice solution, the onset of a metastable solution occurs at p_s , however this solution contains a number, $r(p_s)$, of redundant bonds. Since bonds are added randomly, the lowest threshold at which a stress-free percolating cluster can occur is $p_1 = p_s - r(p_s)$. In first order rigidity cases $r(p_s)$ is small, so to a good approximation $p_1 \approx p_s$. The upper limit of the intermediate phase is an isostatic network to which no more isostatic bonds can be added. This is just the Maxwell counting estimate, so $p_2 \approx p_m$. Calculation of p_1 and p_2 for other graphs and lattices can be carried out in a similar way, using the exact algorithms outlined in the next subsection. In second order cases however, $p_s = p_c$, so that $p_1 = p_c - r(p_c)$ and $p_2 = p_m$. In glasses, it is unlikely that all stress inducing bonds can be avoided so we expect the observed values of the intermediate regime to begin at $p > p_1$ and to end at $p < p_2$.

Exact Algorithms and Percolative Geometries

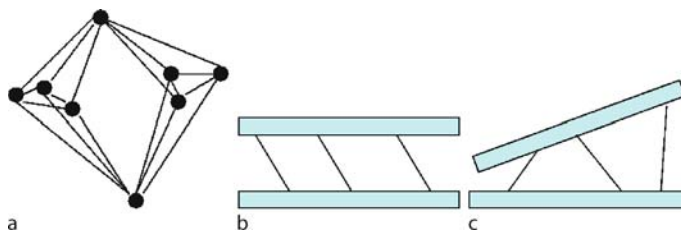
Exact constraint counting provides a procedure for figuring out if a structure can carry stress, without having to solve the elastic equations for the structure. The procedure is based on determining whether all internal degrees of freedom in the structure are fixed by the constraints. A small activity in this area has been ongoing for many years in the graph theory and topology branches of mathematics. The problem is called the graph rigidity problem and is a subset of the problem of graphic matroids. Laman's theorem in 1970 [43] provides the basis for practical algorithms and states that: A bar-joint graph in the plane is rigid *iff* it has no redundant bonds and $b = 2n - 3$, where b is the number of edges (bars) in the graph and n is the number of nodes. This is basically Maxwell counting (see Eq. (3)) in two dimensions, modified by the key requirement that there be no redundant bonds. Hendrickson [34] provided a practical algorithm which tests if a bond in a graph is redundant. In his algorithm, edges are added to the graph one at a time, and tested to see if they are redundant. If they are redundant, then they are not placed in the structure but are noted and stored elsewhere. In this way an isostatic graph obeying Maxwell counting is generated. A structure with no redundant bonds has the Maxwell counting value, $F = 2n - b - 3$, of floppy modes. The number of redundant bonds are also counted enabling a check of the relation Eq. (7). It is also possible to use this method to find the number of stressed bonds, as when a redundant bond is added to a network, an internally stressed cluster of bonds is generated. By cleverly adding a redundant bond across the whole network, this can also be used to identify the stressed backbone [37,50].

Laman's theorem has been extended to body-bar systems in arbitrary dimensions [75,76]. There is also a conjecture that it is exact for molecular frameworks which is

important in applications to chalcogenide glasses, polymers and proteins [36,83]. Molecular frameworks are systems where central forces and bond-bending terms are strong but torsional terms are neglected. However a key system where no combinatorial characterization is available is the central force rigidity problem in three dimensions where exceptions to Laman's theorem are known to exist, such as the famous double banana configuration of Fig. 11a.

A second issue, of significance to a variety of applications including granular media, is the issue of generic rigidity as compared to non-generic rigidity. The issue is illustrated in Fig. 11 where two examples of systems with two bodies and three bars in two dimensions are compared. In the generic rigidity case of Fig. 11c, the three bars have different lengths and are at different angles. Maxwell counting indicates that this system is rigid as there are three degrees of freedom per body and there are three bars. However the non-generic example in Fig. 11b is not rigid as the three bars are parallel enabling a shear distortion of the two bodies with respect to each other. The three bars, which are the constraints in this system, are not independent and this dependence is due to the particular way in which the bars are arranged. In this non-generic case, it is not sufficient to know only the connectivity of the graph as the particular geometry of the structure plays a key role. Clearly regular lattices and regular packings of granular media are non-generic while random lattices, polydisperse packings and random packings are typically generic. Combinatorial rigidity applies to generic cases so that disorder in some sense makes the generic problem more tractable.

To illustrate Hendrickson's redundant bond testing procedure, we consider percolation on diluted triangular lattices. The procedure applies to both rigidity and connectivity cases and it is useful to compare the way in which bond testing compares in the two cases. In the connectivity



Elastic Percolation Networks, Figure 11

Illustration of two key issues in combinatorial rigidity. **a** The double banana configuration which violates Laman's theorem and illustrates a difficulty in applying combinatorial rigidity to three dimensional central force networks. **b** Non-generic and **c** generic geometric structures. Laman's theorem and its extensions apply only to generic networks. The non-generic case depends on geometry and so a theory depending only on connectivity is insufficient. The probability of finding non-generic configurations in large random structures is negligible so generic rigidity applies

case Hendrickson's algorithm checks for loops. His redundant bond testing procedure is based on bipartite matching which is a well known problem in combinatorial optimization. It is implemented as follows [34,37,50]:

Start with an empty triangular lattice (no bonds) and assign to each node g degrees of freedom.

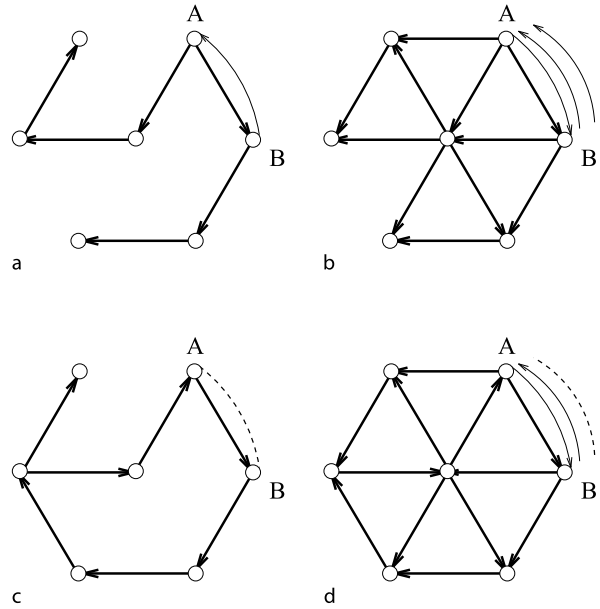
Then:

1. Randomly add a bond to the lattice.
2. Test whether this bond is redundant with respect to the bonds which are currently in the lattice.
3. If the bond is redundant *do not add it to the lattice*, but instead store its location in a different array.
4. Return to 1.

End

Step 2 is the key one and is implemented by matching constraints (bonds) to degrees of freedom, with the restriction that the number of constraints assigned to each node must be less than or equal to the number of degrees of freedom at a node ($g = 2$). It is natural to represent the assignment of constraints to nodes using arrows as illustrated in Fig. 12. When a new bond is added, we add $G + 1$ bonds to take account of the global translations and rotations of a body. For a triangular lattice a body has three degrees of freedom, so we add four bonds as indicated in the figure. Hendrickson noted that if the $G + 1$ added bonds can be matched to the degrees of freedom in the graph, then the bond is not redundant. However, if the arrows cannot be matched, added edge is *redundant*. Another useful way to think about the matching procedure is to consider associating pebbles with the degrees of freedom of the nodes in a graph. Then these pebbles may be placed on the edges of the graph with the constraint that they may only cover edges which enter the node. The matching of arrows to nodes then corresponds to placing pebbles on edges, with the constraint that pebbles can only sit on adjacent edges. The matching fails if it is not possible to cover all edges with pebbles [38]. A successful and a failed match are illustrated in Fig. 12 for a connectivity ($g = 1$) case and a rigidity ($g = 2$) case. Note that the bond that is being tested carries with it G additional copies which account for global degrees of freedom of a rigid cluster. In the connectivity case $G = 1$, while on central-force bar-joint networks in two dimensions $G = 3$.

When the bond test fails, the algorithm identifies all bonds which are overconstrained or stressed with respect to the redundant bond. This set of bonds is called a *Laman subgraph*. Note that if a redundant bond is already in

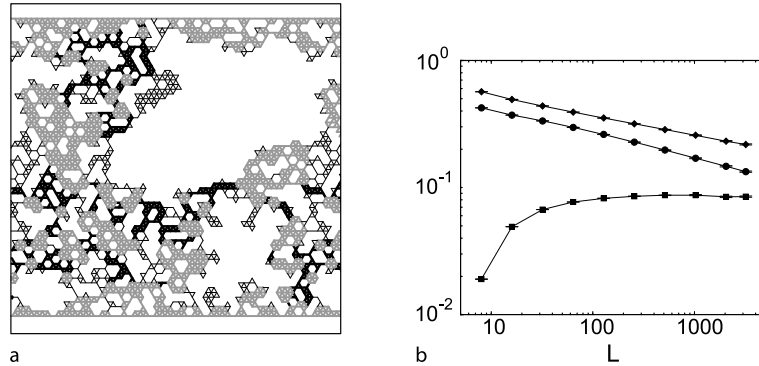


Elastic Percolation Networks, Figure 12

Examples of successful (top figures a and b) and failed (bottom figures, c and d) matches in the connectivity ($g = G = 1$, left figures, a and c) and joint-bar rigidity ($g = 2$, $G = 3$, right figures, b and d) cases on triangular lattices. AB is the new bond that is being tested. Each site has g degrees of freedom and therefore accepts at most g incoming arrows. Each new bond carries with it G auxiliary arrows, which must also be matched. If any of the arrows cannot be matched (dashed), the new bond is redundant. From article by C. Moukarzel and P. M. Duxbury in [79]

a graph, it is not possible to add a new bond and test its redundancy with this method. This is the reason that the algorithm proceeds by adding bonds one at a time starting with an empty lattice. Any error in testing the redundancy of a bond invalidates the rest of the addition sequence. However since this algorithm is an integer method there is no problem with roundoff. It is easy to see that the matching algorithm is quite efficient, however it requires quite a bit of effort to fully optimize these methods. A key step in this optimization procedure is to identify rigid clusters and to describe them in a more condensed way. Moukarzel [50] condenses rigid clusters to a body and hence renormalizes an original bar-joint network to an effective body-joint-bar network. Jacobs [37] instead uses a reduced bar-joint representation of rigid subgraphs. In either case the computational complexity is reduced significantly especially for $p > p_r$.

The geometry of rigidity percolation on a triangular lattice is presented in Fig. 13 [51]. Because of the fact that we add bonds one at a time until the percolation point is reached, it is possible to identify p_c or p_r exactly for



Elastic Percolation Networks, Figure 13

The geometry of rigidity percolation on site diluted triangular lattices. **a** The infinite-cluster geometry at threshold. *Dark wide lines* are cutting bonds, *wide lines* are non-critical backbone bonds (blobs) and *thin lines* are dangling ends. **b** The density of backbone bonds (circles), infinite cluster bonds (diamonds) and dangling bonds (squares) at p_r . A fit to the backbone density yields the exponent $P_B \approx L^{-\beta'/\nu}$, with $\beta' = 0.25(2)$. The dangling ends have not reached the asymptotic regime so it is not clear if the infinite cluster probability has either. Nevertheless, the transition is second order with correlation length exponent $\nu = 1.16(2)$ found from the finite size scaling of the number of cutting bonds and also from the scaling of finite size corrections to p_c [39,51,52], from [51]

each sample, and therefore measure the components of the spanning cluster exactly at p_c or p_r . This eliminates the error associated with measurements at fixed values of p , since estimated exponents are known to depend very sensitively on p . Just as in connectivity percolation, at p_r we identify three different types of bonds: *backbone bonds*, *dangling ends* and *cutting bonds*, as illustrated in Fig. 13a. These together form the *infinite cluster*. The cutting bonds are stressed (belong to the backbone), but they are “critical” because if one of them is removed, load is no longer transmitted across the infinite cluster.

In order to find the correlation length exponent, two relations were used [38,39,51,52]: the size dependence of the threshold behaves as $\delta p_c \sim L^{-1/\nu}$ and secondly, the number of cutting bonds varies as $n_c \sim L^{1/\nu}$, yielding $\nu = 1.16(2)$ [52] or $\nu = 1.20(3)$ [38]. Calculations on the rigidity of rod networks [44] find $\nu = 1.18(2)$. From data such as Fig. 13b the backbone density is found to decrease algebraically $P_B \sim L^{-\beta'/\nu}$, with $\beta' = 0.25 \pm 0.02$ [51,52,54]. It also appears that the infinite-cluster probability is decreasing algebraically, however that is difficult to reconcile with the behavior of the dangling ends which must also eventually decrease for this trend to be asymptotic.

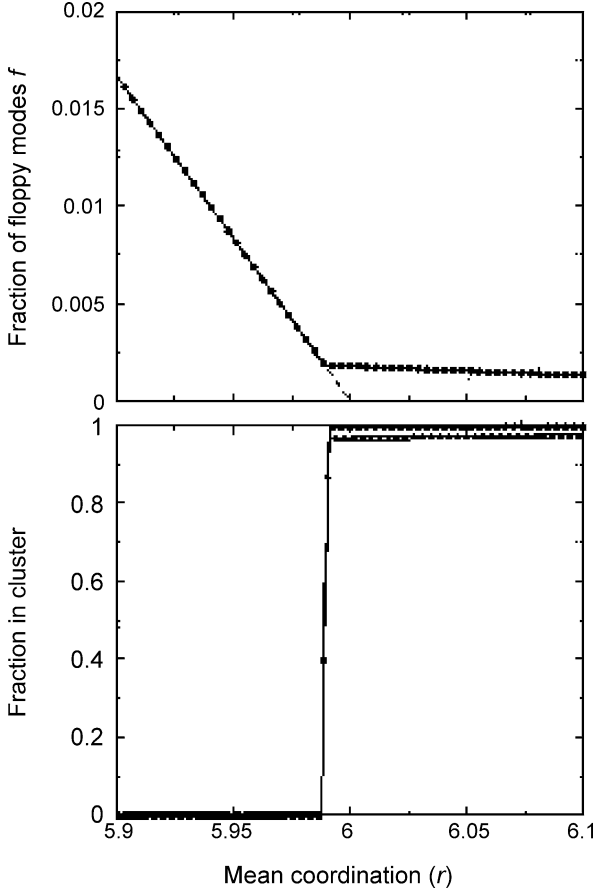
Since it has a diverging correlation length, rigidity percolation on triangular lattices is second order, however first order cases do occur in two dimensions, for example a square network to which diagonals are added at random locations exhibits a strong first order transition, even though the elastic behavior remains continuous [51,56]. Moreover, simulations of central force networks in three

dimensions [15] demonstrate a strong first order transition, as illustrated in Fig. 14.

Elastic Critical Behavior

After a flurry of activity in the 1980’s and early 1990’s study of elastic critical exponents has diminished, nevertheless many issues remain unresolved. The distinction between bond-bending networks and central force networks was brought into sharp focus in the early 1980’s [24], moreover the distinction between the critical exponents describing conductivity and elasticity was emphasized [41]. Many studies were purely energetic, so that thermal effects were ignored. In contrast, de Gennes [17] early argument concerning the equivalence of the conductivity and elasticity exponents of gels relies on entropic springs. The effect of entropic terms has been simulated more recently, supporting the early analysis by de Gennes [60,61].

In two dimensional systems with bond-bending terms or in three dimensional systems with both bond-bending and torsion or dihedral terms (see Eq. (1)), elastic percolation occurs at the conductivity percolation threshold. In the absence of entropic terms, it is well established that in this case the elastic critical behavior, $E \sim (\delta p)^T$ is different than the conductivity critical behavior, $\sigma \sim (\delta p)^t$. E is the Young’s modulus while σ is the conductivity, and $\delta p = p - p_c$ with p_c the conductivity percolation threshold. T is the elastic critical exponent and t is the conductivity critical exponent. A variety of methods have been applied to the calculation of t and T in two and three dimensions, with the currently



Elastic Percolation Networks, Figure 14

The number of floppy modes (*top panel*). The infinite-cluster probability and stressed backbone probability (*bottom panel*) on bond diluted bcc lattices. The transition is strongly first order, though the elastic behavior appears to be continuous, from [15]

accepted values being $t_{2D} = 1.30(1)$, $t_{3D} = 2.01(2)$ and $T_{2D} = 3.96(4)$ [90]. In three dimensional bond-bending networks $T_{3D} \approx 3.75$ [69], and both of these elastic critical exponents are consistent with the relation $T = t + 2\nu$, which is claimed to be exact [8]. Here ν is the percolation critical exponent and $\nu_{2D} = 4/3$ and $\nu_{3D} = 0.88(1)$. It is worth noting that a model where hard inclusions are placed in a soft elastic background has a different behavior. The critical point is the point at which an infinite hard cluster percolates. On approach to this point from below, the elastic constant diverges $E \sim |\Delta p|^{-S}$. The analogous electrical problem consists of perfectly conducting inclusions in a finite conductivity matrix where $\sigma \sim |\Delta p|^{-s}$. In this case Bergman [8] argues that $s = S$. The lattice results quoted above require modification in the case of continuum systems, for example in cases where holes are

punched randomly in a sheet so that narrow necks of material can exist [23]. This case is non-universal in both the conductivity and elasticity cases, though the geometric exponents, such as ν , remain universal.

Elastic percolation on central force networks occurs at different thresholds than conductivity percolation on the same networks. Maxwell counting gives a good first approximation to the elastic percolation thresholds, however precise values can only be found by direct simulation. Direct simulations of central force networks in two and three dimensions have been quite controversial, however the consensus seems to be that the behavior in two dimensions close to criticality is in the bond-bending universality class described in the paragraph above [31]. However central force systems in three dimensions [68] exhibit a different behavior, with $T/\nu \approx 2.1$ for body centered cubic lattices, which is much smaller than the bond bending value $T/\nu \approx 4.4$. Moreover, the geometry of these systems is now known to exhibit a strong first order jump (see Fig. 14) in three dimensions [15].

The elastic behavior of bond-diluted central force systems is remarkably simple away from p_r and is well described by a simple linear relation found using effective medium theory [25]. The critical regime appears to be quite narrow in these central force models. It is also important to note that the results for p_r found from combinatorial methods apply to triangular lattices where the sites are randomly displaced from their regular locations to remove degeneracies. This may be the reason that geometric calculations of p_r on triangular lattices are higher 0.6603 [38] than values found from direct solution of elastic response on regular lattices where $p_r \sim 0.64$ [16]. A further complication is that elastic calculations on displaced lattices lead to a non-linear stiffening of the lattice as bonds which are nearly co-linear are brought into alignment by the applied stress and lead to a stiffening of the network [52].

In geometric models of rigidity, as described above and in Sect. “Exact Solution on Bethe Lattices”, the rigidity percolation threshold, p_r , lies above the connectivity percolation threshold, p_c . However in real materials there is usually a finite elastic modulus even in the regime $p_c < p < p_r$, which may arise through a variety of different effects. In polymers, proteins and chalcogenide glasses, the torsional forces are smaller but significant leading to the onset of elasticity at p_c rather than p_r . Moreover entropic effects may be significant even in ideal central force systems, so that thermal fluctuations lead to a finite elastic modulus for all $p > p_c$. This has been demonstrated in simple central force systems such as diluted square, triangular and cubic lattices [62]. This effect is even more pronounced in crosslinked polymeric systems, such as rubber,

where entropic rubber elasticity dominates [5,17,88]. In these systems the regime $p_c < p < p_r$ is extremely broad so that even heavily crosslinked flexible polymeric systems are in this regime. A further mechanism which leads to a reduction of the percolation threshold from that predicted by geometric rigidity is the presence of tension in the network [74], or if the natural length of the springs in the network is taken to zero. Moreover the elastic critical exponent observed in entropic systems and in systems in tension is usually quite close to the conductivity value, as first predicted by de Gennes in 1976 [5,17,88] and as observed in many experiments on gels (see Sect. “Gels and Semiflexible Rod Networks”).

Final Remarks and Future Directions

The relevance of rigidity percolation to granular media and to semiflexible polymer networks is now well established, though many issues remain unresolved. Ideal experiments to test the recent theoretical predictions will greatly clarify the extent to which current rigidity concepts are sufficient to describe these systems. The intermediate phase in chalcogenide glasses is a stimulating, and physically very natural, concept, that perhaps could be observed in large scale molecular dynamics simulations with realistic potentials of systems such as $Se_{1-x}Ge_x$. Presumably there is a dynamical lengthscale at which self-organization is possible, which limits the extent of the intermediate phase.

Graph combinatorial methods and direct elastic model simulations provide complementary information about the mechanical response of complex materials. The combinatorial methods however remain limited due to the absence of an exact combinatorial characterization of rigid structures in three dimensional central force networks. Moreover, the physical significance of banana configurations, as illustrated in Fig. 11a, remains unexplored and is an intriguing outstanding problem. The common occurrence of a strong first order jump in the infinite rigid cluster along with a continuous elastic behavior [15,51,56] is surprising and also largely unexplored.

Rigidity percolation is emerging as a geometric model for structural glasses [85]. Clearly thermal effects are critical in many applications of rigidity percolation so that models which treat both geometric and thermal effects are valuable. Molecular dynamics models provide insight into the interplay of geometry and temperature and warrant further study. In particular the evolution of the boson peak in model glasses as a function of geometry and temperature remains unexplored and would help resolve

contrasting models which emphasize either temperature effects [58] or geometric effects [85].

The role of generic as compared to non-generic configurations is critical to granular media and to the elasticity of lattices. Even the most basic questions remain unresolved, for example are the elastic exponents of regular lattices the same as the elastic exponents of random lattices with the same co-ordination? Is the elasticity of random generic networks inherently non-linear due to alignment of bonds under stress?

Nevertheless, rigidity percolation provides a useful model for a broad range of physical phenomena, significantly generalizing the connectivity percolation model. Combinatorial methods drawn from graph theory have greatly extended our understanding of rigidity, providing both practical tools as well as conceptual novelties including the distinction between generic and non-generic rigidity. The applications of related geometric methods to granular media and to semiflexible gels is extremely rich and are harbingers of a broader use of rigidity concepts in the analysis of the mechanical behavior of complex materials.

Bibliography

1. Alava MJ, Niskanen KJ (2006) The physics of paper. Rep Prog Phys 69:669–724
2. Alexander S (1998) Amorphous materials: Their structure, lattice dynamics and elasticity. Phys Rep 296:65–236
3. Angell CA (2004) Boson peaks and floppy modes: Some relations between constraint and excitations phenomenology, and interpretation, of glasses and glass transition. J Phys Cond Mat 16:S5153–S5164
4. Axelos MAV, Kolb M (1990) Crosslinked biopolymers: Experimental evidence for scalar percolation theory. Phys Rev Lett 64:1457–1460
5. Barsky SJ, Plischke M (1996) Order and localization in randomly cross-linked polymer networks. Phys Rev E 53:871–876
6. Bausch AR, Kroy K (2006) A bottom-up approach to cell mechanics. Nat Phys 2:231–238
7. Benguigui L (1984) Experimental study of the elastic properties of a percolating system. Phys Rev Lett 53:2028–2030
8. Bergman DJ (2003) Exact relations between critical exponents for elastic stiffness and electrical conductivity of percolating networks. Physica B 338:240–246
9. Boolchand P, Georgiev DG, Goodman B (2001) Discovery of the intermediate phase in chalcogenide glasses. J Optoelectron Ad Mat 3:703–720
10. Boolchand P, Lucovsky G, Phillips JC, Thorpe MF (2005) Self-organization and the physics of glassy networks. Phil Mag 85:3823–3838
11. Bresser W, Boolchand P, Suranyi P (1986) Rigidity percolation and molecular clustering in network glasses. Phys Rev Lett 56:2493–2496
12. Brooks BR, Brucoleri RE, Olafson BD, States DJ, Swaminathan S et al (2004) Charrm: A program for macromolecular energy, minimization and dynamics calculations. J Comp Chem 4:187–217

13. Cai Y (1989) Floppy modes in network glasses. *Phys Rev B* 40:10535–10542
14. Chalupa J, Leath PL, Reich GR (1979) Bootstrap percolation on a bethe lattice. *J Phys C* 12:L31–L35
15. Chubynsky MV, Thorpe MF (2007) Algorithms for 3d-rigidity analysis and a first-order percolation transition. *Phys Rev E* 76:41135
16. Day AR, Tremblay RR, Tremblay AMS (1986) Rigid backbone: A new geometry for percolation. *Phys Rev Lett* 56:2501–2504
17. de Gennes PG (1976) On the relation between percolation theory and the elasticity of gels. *J Physique* 37:L1–L2
18. de Gennes PG (1979) Scaling concepts in polymer physics. Cornell University Press, New York
19. Deptuck D, Harrison JP, Zawadzki P (1985) Measurement of elasticity and conductivity of a three-dimensional percolation system. *Phys Rev Lett* 54:913–916
20. Dohler GH, Dandoloff R, Bilz H (1980) A topological-dynamical model of amorphycity. *J Non-Cryst Sol* 42:87–95
21. Donev A, Torquato S, Stillinger FH (2005) Pair correlation function characteristics of nearly jammed disordered and ordered hard-sphere packings. *Phys Rev E* 71:11105
22. Duxbury PM, Jacobs DJ, Thorpe MF, Moukarzel C (1999) Floppy modes and the free energy: Rigidity and connectivity percolation on bethe lattices. *Phys Rev E* 59:2084–2092
23. Feng S, Halperin B, Sen PN (1987) Transport properties of continuum systems near the percolation threshold. *Phys Rev B* 35:197–214
24. Feng S, Sen PN (1984) Percolation on elastic networks: New exponent and threshold. *Phys Rev Lett* 52:216–219
25. Feng S, Thorpe MF, Garboczi E (1985) Effective-medium theory of percolation on central-force networks. *Phys Rev B* 31:276–280
26. Feng X, Bresser WJ, Boolchand P (1997) Direct evidence for stiffness threshold in chalcogenide glasses. *Phys Rev Lett* 78:4422–4425
27. Fisher ME, Essam JW (1961) Some cluster size and percolation problems. *J Math Phys* 2:609–619
28. Flory PJ (1953) Principles of polymer chemistry. Cornell University Press, New York
29. Grant MC, Russell WB (1993) Volume fraction dependence of elastic moduli and transition temperatures for colloidal silica gels. *Phys Rev E* 47:2606–2614
30. Guyon E, Roux S, Hansen A, Bideau D, Troadec JP et al (1990) Non-local and non-linear problems in the mechanics of disordered systems: Application to granular media and rigidity problems. *Rep Prog Phys* 53:373–419
31. Hansen A, Roux S (1989) Universality class of central-force percolation. *Phys Rev B* 40:749–752
32. He H, Thorpe MF (1985) Elastic properties of glasses. *Phys Rev Lett* 54:2107–2110
33. Head DA, Levine AJ, MacKintosh FC (2003) Distinct regimes of elastic response and deformation modes of crosslinked cytoskeletal and semiflexible polymer networks. *Phys Rev E* 68:61907
34. Hendrickson B (1992) Conditions for unique graph realizations. *SIAM J Comput* 21:65–84
35. Heussinger C, Frey E (2006) Stiff polymers, foams and fiber networks. *Phys Rev Lett* 96:1–4
36. Jacobs DJ (1998) Generic rigidity in three-dimensional bond-bending networks. *J Phys A: Math Gen* 31:6653–6668
37. Jacobs DJ, Hendrickson B (1997) An algorithm for two-dimensional rigidity percolation: The pebble game. *J Comp Phys* 137:346–365
38. Jacobs DJ, Thorpe MF (1995) Generic rigidity percolation: The pebble game. *Phys Rev Lett* 75:4051–4054
39. Jacobs DJ, Thorpe MF (1996) Generic rigidity percolation in two dimensions. *Phys Rev E* 53:3682–3693
40. Kamitakahara WA, Cappelletti RL, Boolchand P, Halfpap B, Gompf F et al (1991) Vibrational densities of states and network rigidity in chalcogenide glasses. *Phys Rev B* 44:94–100
41. Kantor Y, Webman I (1984) Elastic properties of random percolating systems. *Phys Rev Lett* 52:1891–1894
42. Kasza KE, Rowat AC, Liu J, Angelini RE, Brangwynne CP et al (2007) The cell as a material. *Curr Opin Cell Bio* 19:101–107
43. Laman G (1970) On graphs and rigidity of plane skeletal structures. *J Eng Math* 4:331–340
44. LatvaKokko M, Makinen J, Timonen J (2001) Rigidity transition in two-dimensional random fiber networks. *Phys Rev E* 63:46113
45. LatvaKokko M, Timonen J (2001) Rigidity of random networks of stiff fibers in the low density limit. *Phys Rev E* 64:66117
46. Majmudar TS, Sperl M, Luding S, Behringer RP (2007) Jamming transition in granular systems. *Phys Rev Lett* 98:58001
47. Martin JE, Adolf D, Wilcox JP (1988) Viscoelasticity of near-critical gels. *Phys Rev Lett* 61:2620–2623
48. Maxwell JC (1864) On the calculation of the equilibrium stiffness of frames. *Phil Mag* 27:294–301
49. Mayo SL, Olafson BD, Goddard WA (1990) Dreiding: A generic force field for molecular simulations. *J Phys Chem* 94:8897–8909
50. Moukarzel C (1996) An efficient algorithm for testing the generic rigidity of graphs in the plane. *J Phys A: Math Gen* 29:8079–8098
51. Moukarzel C, Duxbury PM (1990) Comparison of rigidity and connectivity percolation in two dimensions. *Phys Rev E* 59:2614–2622
52. Moukarzel C, Duxbury PM (1995) Stressed backbone and elasticity of random central-force system. *Phys Rev Lett* 75:4055–4058
53. Moukarzel C, Duxbury PM, Leath PL (1997) First-order rigidity on cayley trees. *Phys Rev E* 55:5800–5811
54. Moukarzel C, Duxbury PM, Leath PL (1997) Infinite-cluster geometry in central-force networks. *Phys Rev Lett* 78:1480–1483
55. Moukarzel CF (1998) Isostatic phase transition and instability in stiff granular materials. *Phys Rev Lett* 81:1634–1637
56. Obukhov SP (1995) First order rigidity transition in random rod networks. *Phys Rev Lett* 74:4472–4475
57. Ohern CS, Silbert LE, Liu AJ, Nagel SR (2003) Jamming at zero temperature and zero applied stress: The epitome of disorder. *Phys Rev E* 68:11306
58. Parisi G (2003) On the origin of the boson peak. *J Phys Cond Mat* 15:S765–S774
59. Phillips JC (1979) Topology of covalent non-crystalline solids i: Short-range order in chalcogenide alloys. *J Non-Cryst Sol* 34:153–181
60. Plischke M (2006) Critical behavior of entropic shear rigidity. *Phys Rev E* 73:61406
61. Plischke M, Joos B (1998) Entropic elasticity of diluted central force networks. *Phys Rev Lett* 80:4907–4910
62. Plischke M, Vernon DC, Joos B, Zhou Z (1998) Entropic elasticity of diluted central force networks. *Phys Rev E* 60:3129–3135

63. Rader AJ, Hespenheide BM, Kuhn LA, Thorpe MF (2002) Protein unfolding: Rigidity lost. *PNAS* 99:3540–3545
64. Rivoire O, Barré J (2006) Exactly solvable models of adaptive networks. *Phys Rev Lett* 97:148701
65. Ross-Murphy SB (2007) Biopolymer gelation-exponents and critical exponents. *Polym Bull* 58:119–126
66. Rueb CJ, Zukoski CF (1997) Viscoelastic properties of colloidal gels. *J Rheol* 41:197–218
67. Sahimi M (1998) Non-linear and non-local transport processes in heterogeneous media: From long-range correlated percolation to fracture and materials breakdown. *Phys Rep* 306:214–395
68. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. i. percolation on elastic networks with central forces. *Phys Rev B* 47:695–702
69. Sahimi M, Arbabi S (1993) Mechanics of disordered solids. ii. percolation on elastic networks with bond-bending forces. *Phys Rev B* 47:703–712
70. Schwartz LM, Feng S, Thorpe MF, Sen PN (1985) Behavior of depleted elastic networks: Comparison of effective medium theory and numerical simulations. *Phys Rev B* 32:4607–4617
71. Schwarz JH, Liu AJ, Chayes LQ (2006) The onset of jamming as the sudden emergence of an infinite k-core cluster. *Europhys Lett* 73:560–566
72. Serrano D, Peyrelasse J, Boned C, Harran D, Monge P (1990) Application of the percolation model to gelation of an epoxy resin. *J Appl Poly Sci* 39:670–693
73. Stauffer D, Coniglio A, Adam M (1982) Gelation and critical phenomena. *Adv Poly Sci* 44:103–158
74. Tang W, Thorpe MF (1988) Percolation of elastic networks under tension. *Phys Rev B* 37:5539–5551
75. Tay TS (1984) Rigidity of multi-graphs. i. linking rigid bodies in n-space. *J Comb Theory B* 36:95–112
76. Tay TS, Whiteley W (1985) Generating isostatic frameworks. *Struct Topol* 11:21–68
77. Tharmann R, Claessens MMAE, Bausch AR (2007) Viscoelasticity of isotropically crosslinked actin networks. *Phys Rev Lett* 98:88103
78. Thorpe MF (1983) Continuous deformations in random networks. *J Non-Cryst Sol* 57:355–370
79. Thorpe MF, Duxbury PM (eds) (1999) *Rigidity theory and applications*. Plenum, New York
80. Thorpe MF, Jacobs DJ, Chubynsky MV, Phillips JC (2000) Self-organization in network glasses. *J Non-Cryst Sol* 266-269:859–866
81. Uebbing B, Sievers AJ (1996) Role of network topology on the vibrational lifetime of an h_2o molecule in the ge-as-se glass series. *Phys Rev Lett* 76:932–935
82. Wang J, Wolf RM, Caldwell JW, Kollman PA, Case DA (2004) Development and testing of a general amber force field. *J Comp Chem* 25:1157–1174
83. Whiteley W (2005) Counting out to the flexibility of molecules. *Phys Biol* 2:S116–S126
84. Winter HH, Mours M (1997) Rheology of polymers near the liquid-solid transition. *Adv Poly Sci* 134:165–234
85. Wyart M (2005) On the rigidity of amorphous solids. *Ann Phys Fr* 30:1–96
86. Wyart M, Nagel SR, Witten TA (2005) Geometric origin of excess low-frequency vibrational modes in weakly connected amorphous solids. *Europhys Lett* 72:486–492
87. Wyss HM, Tervoort EV, Gauckler LJ (2005) Mechanics and microstructures of concentrated particle gels. *J Am Ceram Soc* 88:2337–2348
88. Xing X, Mukhopadhyay S, Goldbart PM (2004) Scaling of entropic shear rigidity. *Phys Rev Lett* 93:225701
89. Xu J, Bohnsack DA, Mackay ME, Wooley KL (2007) Unusual mechanical performance of amphiphilic crosslinked polymer networks. *J Amer Chem Soc Communications* 129:506–507
90. Zabolitzky JG, Bergman DJ, Stauffer D (1985) Precision calculation of elasticity for percolation. *J Stat Phys* 54:913–916

Embodied and Situated Agents, Adaptive Behavior in

STEFANO NOLFI

Institute of Cognitive Sciences and Technologies,
National Research Council (CNR), Rome, Italy

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Embodiment and Situatedness](#)

[Behavior and Cognition as Complex Adaptive Systems](#)

[Adaptive Methods](#)

[Evolutionary Robotics Methods](#)

[Discussion and Conclusion](#)

[Bibliography](#)

Glossary

Phylogenesis Indicates the variations of the genetic characteristics of a population of artificial agents throughout generations.

Ontogenesis Indicates the variations which occur in the phenotypical characteristics of an artificial agent (i. e. in the characteristics of the control system or of the body of the agent) while it interacts with the environment.

Embodied agent Indicates an artificial system (simulated or physical) which has a body (characterized by physical properties such as shape, dimension, weight, etc), actuators (e. g. motorized wheels, motorized articulated joints), and sensors (e. g. touch sensors or vision sensors). For a more restricted definition see the concluding section of the paper.

Situated agent Indicates an artificial system which is located in a physical environment (simulated or real) with which it interacts on the basis of the law of physics. For a more restricted definition see the concluding section of the paper.

Morphological computation Indicates the ability of the body of an agent (with certain specific characteristics) to control its interaction with the environment so to produce a given desired behavior.

Definition of the Subject

Adaptive behavior concerns the study of how organisms develop their behavioral and cognitive skills through a synthetic methodology which consists in designing artificial agents which are able to adapt to their environment autonomously. These studies are important both from a modeling point of view (i. e. for making progress in our understanding of intelligence and adaptation in natural beings) and from an engineering point of view (i. e. for making progresses in our ability to develop artefacts displaying effective behavioral and cognitive skills).

Introduction

Adaptive behavior research concerns the study of how organisms can develop behavioral and cognitive skills by adapting to the environment and to the task they have to fulfill autonomously (i. e. without human intervention). This goal is achieved through a synthetic methodology, i. e. through the synthesis of artificial creatures which: (i) have a body, (ii) are situated in an environment with which they interact, and (iii) have characteristics which vary during an adaptation process. In the rest of the paper we will use the term “agent” to indicate artificial creatures which possess the first two features described above and the term “adaptive agent” to indicate artificial creatures which also possess the third feature.

The agents and the environment might be simulated or real. In the former case the characteristics of agents’ body, motor, and sensory system, the characteristics of the environment, and the rules that regulate the interactions between all the elements are simulated on a computer. In the latter case, the agents consist of physical entities (mobile robots) situated in a physical environment with which they interact on the basis of the physical laws.

The adaptive process which regulates how the characteristics of the agents (and eventually of the environment change) change might consist of a population-based evolutionary process and/or of a developmental/learning process. In the former case, the characteristics of the agents do not vary during their “lifetime” (i. e. during the time in which the agents interact with the environment) but phylogenetically, while individual agents “reproduce”. In the latter case, the characteristics of the agents vary ontogenetically, while they interact with the environment. The criteria which determine how variations are generated and/or

whether or not variations are retained can be task-dependent and/or task-independent, i. e. might be based on an evaluation of whether the variation increase or decrease agents’ ability to display a behavior which is adapted to the task/environment or might be based on task-independent criteria (i. e. general criteria which do not reward directly the exhibition of the requested skill).

The paper is organized as follows. In Sect. “**Introduction**” we briefly introduce the notion of *embodiment* and *situatedness* and their implications. In Sect. “**Embodiment and Situatedness**” we claim that behavior and cognition in embodied and situated adaptive agents should be characterized as a complex adaptive system. In Sect. “**Behavior and Cognition as Complex Adaptive Systems**” we briefly describe the methods which can be used to synthesize embodied and situated adaptive agents. Finally in Sect. “**Adaptive Methods**”, we draw our conclusions.

Embodiment and Situatedness

The notion of *embodiment* and *situatedness* has been introduced [8,9,12,34,48] to characterize systems (e. g. natural organism and robots) which have a physical body and which are situated in a physical environment with which they interact. In this and in the following sections we will briefly discuss the general implications of these two fundamental properties. This analysis will be further extended in the concluding section where we will claim on the necessity to distinguish between a weak and a strong notion of embodiment and situatedness.

One first important implication of being embodied and situated consists in the fact that these agents and their parts are characterized by their physical properties (e. g. weight, dimension, shape, elasticity etc.), are subjected to the laws of physics (e. g. inertia, friction, gravity, energy consumption, deterioration etc.), and interact with the environment through the exchange of energy and physical material (e. g. forces, sound waves, light waves etc.). Their physical nature also implies that they are quantitative in state and time [49]. The fact that these agents are quantitative in time implies, for example, that the joints which connect the parts of a robotic arm can assume any possible position within a given range. The fact that these agents are quantitative in time implies, for example, that the effects of the application of a force to a joint depend from the time duration of its application.

One second important implication is that the information measured by the sensors is not only a function of the environment but also of the relative position of the agent in the environment. This implies that the motor actions performed by an agent, by modifying the

agent/environmental relation or the environment, co-determine the agent sensory experiences.

One third important implication is that the information measured by the sensors provide information about the external environment which is egocentric (depends from the current position and the orientation of the agent in the environment), local (only provide information related to the local observable portion of the environment), incomplete (due to visual occlusion, for example), and subjected to noise. Similar characteristics apply to the motor actions produced by the agent's effectors.

It is important to notice that these characteristics do not only represent constraints but also opportunities to be exploited. Indeed, as we will see in the next section, the exploitation of some of these characteristics might allow embodied and situated agents to solve their adaptive problems through solutions which are robust and parsimonious (i. e. minimal) with respect to the complexity of the agent's body and control system.

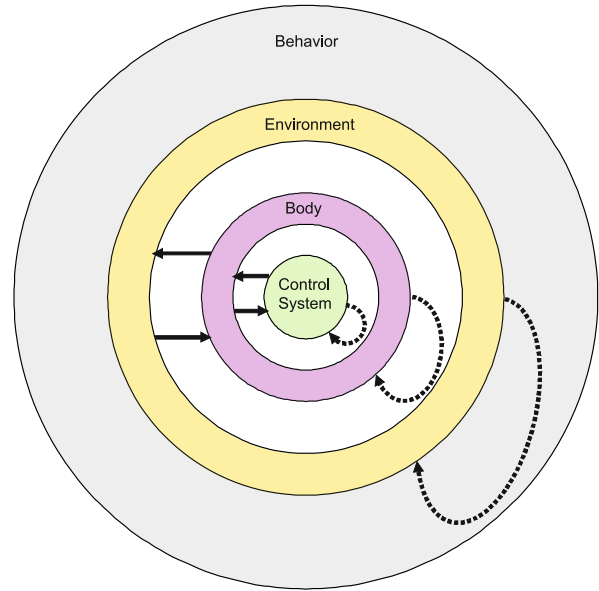
Behavior and Cognition as Complex Adaptive Systems

In embodied and situated agents, behavioral and cognitive skills are dynamical properties which unfold in time and which arise from the interaction between agents' nervous system, body, and the environment [3,11,19,29,31] and from the interaction between dynamical processes occurring within the agents' control system, the agents' body, and within the environment [4,15,45]. Moreover, behavioral and cognitive skills typically display a multi-level and multi-scale organization involving bottom-up and top-down influences between entities at different levels of organization. These properties imply that behavioral and cognitive skills in embodied and situated agents can be properly characterized as complex adaptive systems [29].

These aspects and the complex system nature of behavior and cognition will be illustrated in more details in the next subsections also with the help of examples. The theoretical and practical implication of these aspects for developing artificial agents able to exhibit effective behavioral and cognitive skills will be discussed in the forthcoming sections.

Behavior and Cognition as Emergent Dynamical Properties

Behavior and cognition are dynamical properties which unfold in time and which emerge from high-frequent non-linear interactions between the agent, its body, and the external environment [11].

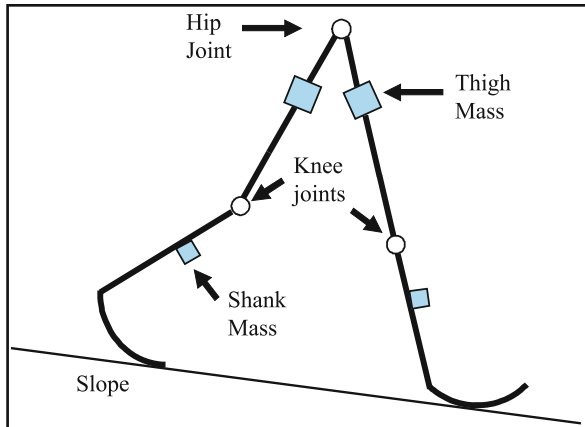


Embodied and Situated Agents, Adaptive Behavior in, Figure 1

A schematic representation of the relation between agent's control system, agent's body, and the environment. The behavioral and cognitive skills displayed by the agent are the emergent result of the bi-directional interactions (represented with *full arrows*) between the three constituting elements – agent's control system, agent's body, and environment. The *dotted arrows* indicate that the three constituting elements might be dynamical systems on their own. In this case, agents' behavioral and cognitive skills result of the dynamics originating from the agent/body/environmental interactions but also from the combination and the interaction between dynamical processes occurring within the agent's body, within the agent's control system, and within the environment (see Sect. "Embodiment and Situatedness")

At any time step, the environmental and the agent/environmental relation co-determine the body and the motor reaction of the agent which, in turn, co-determines how the environment and/or the agent/environmental relation vary. Sequences of these interactions, occurring at a fast time rate, lead to a dynamical process – behavior – which extends over significant larger time span than the interactions (Fig. 1).

Since interactions between the agent's control system, the agent's body, and the external environment are non-linear (i. e. small variations in sensory states might lead to significantly different motor actions) and dynamical (i. e. small variations in the action performed at time t might significantly impact later interactions at time $t+x$) the relation between the rules that govern the interactions and the behavioral and cognitive skills originating from the interactions tend to be very indirect. Behavioral and cogni-

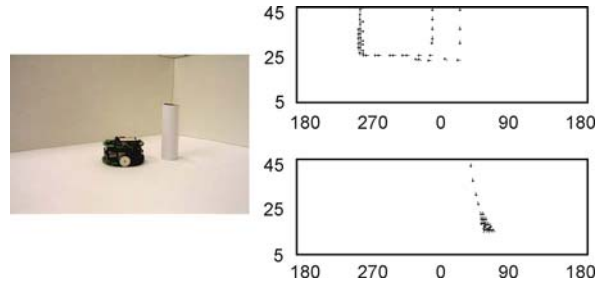


Embodied and Situated Agents, Adaptive Behavior in, Figure 2
A schematization of the passive walking machine developed by McGeer [24]. The machine includes two passive knee joints and a passive hip joint

tive skills thus emerge from the interactions between the three foundational elements and cannot be traced back to any of the three elements taken in isolation. Indeed, the behavior displayed by an embodied and situated agent can hardly be predicted or inferred from an external observer even on the basis of a complete knowledge of the interacting elements and of the rules governing the interactions.

A clear example of how behavioral skill might emerge from the interaction between the agents' body and the environment is constituted by the passive walking machines developed in simulation by McGeer [24] – a two-dimensional bipedal machines able to walk down a four-degree slope with no motors and no control system (Fig. 2). The walking behavior arises from the fact that the physical forces resulting from gravity and from the collision between the machine and the slope produce a movement of the robot and the fact that robot's movements produce a variation of the agent-environmental relation which in turn produce a modification of the physical forces to which the machine will be subjected in the next time step. The sequence of by-directional effects between the robot's body and the environment can lead to a stable dynamical process – the walking behavior.

The type of behavior which arises from the robot/environmental interaction depends from the characteristics of the environment, the physics law which regulate the interaction between the body and the environment, and the characteristics of the body. The first two factors can be considered as fixed but the third factor, the body structure, can be adapted to achieve a given function. Indeed, in the case of this biped robot, the author carefully selected



Embodied and Situated Agents, Adaptive Behavior in, Figure 3
Left: The agent situated in the environment. The agent is a Khepera robot [26]. The environment consists of an arena of 60×35 cm containing cylindrical objects placed in a randomly selected location. *Right:* Angular trajectories of an evolved robot close to a wall (*top graph*) and to a cylinder (*bottom graph*). The picture was obtained by placing the robot at a random position in the environment, leaving it free to move for 500 time steps each lasting 100 ms, and recording its relative movements with respect to the two types of objects for distances smaller than 45 mm. The x-axis and the y-axis indicate the relative angle (in degrees) and distance (in mm) between the robot and the corresponding object. For sake of clarity, *arrows* are used to indicate the relative direction, but not the amplitude of movements

the leg length, the leg mass, and the foot size to obtain the desired walking behavior. In more general term, this example shows how the role of regulating the interaction between the robot and the environment in the appropriate way can be played not only but the control system but also from the body itself providing that the characteristics of the body has been shaped so to favor the exhibition of the desired behavior. This property, i.e. the ability of the body to control its interaction with the environment, has been named with the term “morphological computation” [35]. For related work which demonstrate how effective walking machines can be obtained by integrating passive walking techniques with simple control mechanisms, see [6,13,50]. For related works which show the role of elastic material and elastic actuators for morphological computing see [23,40].

To illustrate of how behavioral and cognitive skills might emerge from agent's body, agent's control system, and environmental interactions we describe a simple experiment in which a small wheeled robot situated in an arena surrounded by walls has been evolved to find and to remain close to a cylindrical object. The Khepera robot [26] is provided with eight infrared sensors and two motors controlling the two corresponding wheels (Fig. 3).

From the point of view of an external observer, solving this problem requires robots able to: (a) explore the environment until an obstacle is detected, (b) discriminate whether the obstacle detected is a wall or a cylindrical ob-

ject, and (c) approach or avoid objects depending on the object type. Some of these behaviors (e. g. the wall-avoidance behavior) can be obtained through simple control mechanisms but others require non trivial control mechanisms. Indeed, a detailed analysis of the sensory patterns experienced by the robot indicated that the task of discriminating the two objects is far from trivial since the two classes of sensory patterns experienced by robots close to a wall and close to cylindrical objects largely overlap.

The attempt to solve this problem through an evolutionary adaptive method (see Sect. “[Behavior and Cognition as Complex Adaptive Systems](#)”) in which the free parameters (i. e. the parameters which regulate the fine-grained interaction between the robot and the environment) are varied randomly and in which variations are retained or discarded on the basis on an evaluation of the overall ability of the robot (i. e. on the basis of the time spent by the robot close to the cylindrical object) demonstrated how adaptive robots can find solutions which are robust and parsimonious in term of control mechanisms [28]. Indeed, in all replications of these experiment, evolved robot solve the problem by moving forward, by avoiding walls, and by oscillating back and fourth and left and right close to cylindrical objects (Fig. 3, right). All these behaviors result from sequences of interactions between the robot and the environment mediated by four types of simple control rules which consist in: turning left when the right infrared sensors are activated, turning right when the left infrared sensors are activated, moving back when the frontal infrared sensors are activated, and moving forward when the frontal infrared sensors are not activated.

To understand how these simple control rules can produce the required behaviors and the required arbitration between behaviors we should consider that the same motor responses produce different effects on different agent/environmental situations. For example, the execution of a left-turning action close to a cylindrical object and the subsequent modification of the robot/object relative position produce a new sensory state which triggers a right-turning action. Then, the execution of the latter action and the subsequent modification of the robot/object relative position produce a new sensory state which triggers a left-turning action. The combination and the alternation of these left and right-turning actions over time produce an attractor in the agent/environmental dynamics (Fig. 3, right, bottom graph) which allows the robot to remain close to the cylindrical object. On the other hand the execution of a left-turning behavior close to a wall object and the subsequent modification of the robot/wall position produce a new sensory state which triggers the reiteration

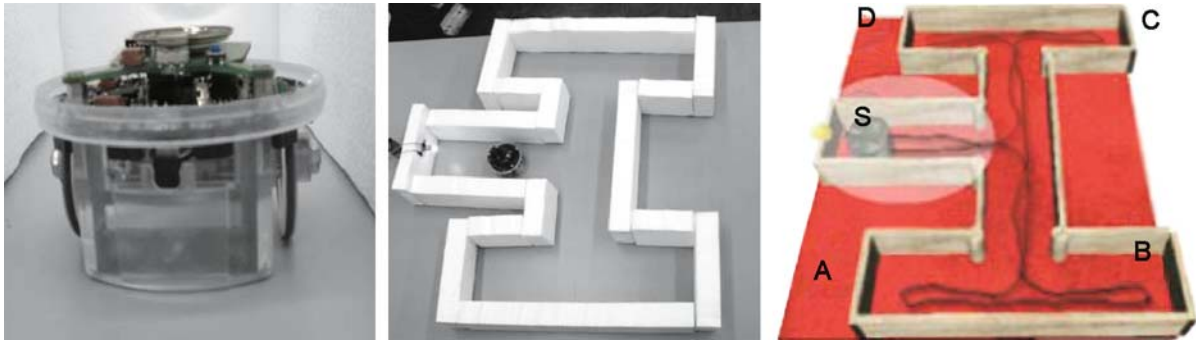
of the same motor action. The execution of a sequence of left-turning action then leads to the avoidance of the object and to a modification of the robot/environmental relation which finally lead to a perception of a sensory state which trigger a move-forward behavior (Fig. 4, right, top graph).

Before concluding the description of this experiment, it is important to notice that, although the rough classification of the robot motor responses into four different types of actions is useful to describe the strategy with which these robots solve the problem qualitatively, the quantitative aspects which characterize the robot motor reactions (e. g. how sharply a robot turns given a certain pattern of activation of the infrared sensors) are crucial for determining whether the robot will be able to solve the problem or not. Indeed, small differences in the robot’s motor response tend to cumulate in time and might prevent the robot for producing successful behavior (e. g. might prevent the robot to produce a behavioral attractor close to cylindrical objects).

This experiment clearly exemplifies some important aspects which characterize all adaptive behavioral system, i. e. systems which are embodied and situated and which have been designed or adapted so to exploit the properties that emerge from the interaction between their control system, their body, and the external environment. In particular, it demonstrates how required behavioral and cognitive skills (i. e. object categorization skills) might emerge from the fine-grained interaction between the robot’s control system, body, and the external environment without the need of dedicated control mechanisms. Moreover, it demonstrates how the relation between the control rules which mediate the interaction between the robot body and the environment and the behavioral skills exhibited by the agents are rather indirect. This means, for example, that an external human observer can hardly predict the behaviors which will be produced by the robot, before observing the robot interacting with the environment, even on the basis of a complete description of the characteristics of the body, of the control rules, and of the environment.

Behavior and Cognition as Phenomena Originating from the Interaction Between Coupled Dynamical Processes

Up to this point we restricted our analysis to the dynamics originating from the agent’s control system, agents’ body, and environmental interactions. However, the body of an agent, its control system, and the environment might have their own dynamics (dotted arrows in Fig. 1). For the sake of clarity, we will refer to the dynami-



Embodied and Situated Agents, Adaptive Behavior in, Figure 4

Left: The e-puck robot developed at EPFL, Switzerland <http://www.e-puck.org/>. **Center:** The environment which have a size of 52 cm by 60 cm. The light produced by the light bulb located on the *left side* of the central corridor cannot be perceived from the other two corridors. **Right:** The motor trajectory produced by the robot during a complete lap of the environment

cal processes occurring within the agent control system, within the agent body, or within the environment as *internal dynamics* and to the dynamics originating from the agent/body/environmental interaction as *external dynamics*. In cases in which agents' body, agents' control system, or the environment have their own dynamics, behavior should be characterized as a property emerging from the combination of several coupled dynamical processes.

The existence of several concurrent dynamical processes represents an important opportunity for the possibility to exploit emergent features. Indeed, behavioral and cognitive skills might emerge not only from the external dynamics, as we showed in the previous section, but also from the internal dynamical processes or from the interaction between different dynamical processes.

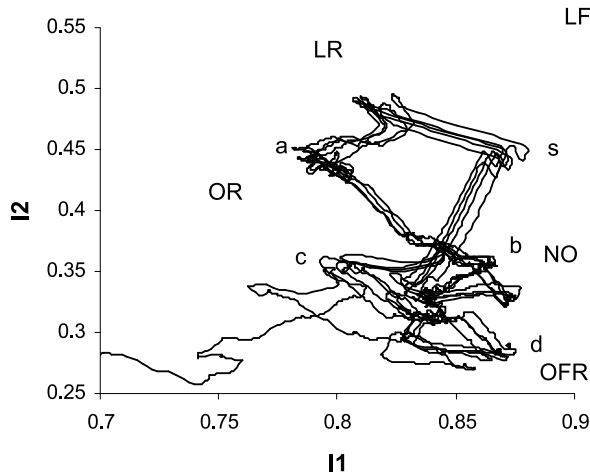
As an example which illustrates how complex cognitive skills can emerge from the interaction between a simple agent/body/environmental dynamic and a simple agent's internal dynamic consider the case of a wheeled robot placed in a maze environment (Fig. 4) which has been trained to: (a) produce a wall-following behavior which allows the robot to periodically visit and re-visit all environmental areas, (b) identify a target object constituted by a black disk which is placed in a randomly selected position of the environment for a limited time duration, and (c) recognize the location in which the target object was previously found every time the robot re-visit the corresponding location [15].

The robot has infrared sensors (which provide information about nearby obstacles), light sensors (which provide information about the light gradient generated by the light bulb placed in the central corridor), ground sensors (which detect the color of the ground), two motors

(which control the desired speed of the two corresponding wheels), and one additional output units which should be turned on when the robot re-visit the environmental area in which the black disk was previously found. The robot's controller consists of a three layers neural network which includes a layer of sensory neurons (which encode the state of the corresponding sensors), a layer of motor neurons which encode the state of the actuators, and a layer of internal neurons which consist of leaky integrators operating at tuneable time scale [3,15]. The free parameters of the robot's neural controllers (i. e. the connection weights, and the time constant of the internal neurons which regulate the time rate at which this neurons change their state over time) were adapted through an evolutionary technique [31].

By analyzing the evolved robot the authors observed how they are able to generate a spatial representation of the environment and of their location in the environment while they are situated in the environment itself. Indeed, while the robot travel by performing different laps of the environment (see Fig. 4, right), the states of the two internal neurons converge on a periodic limit cycle dynamic in which different states correspond to different locations of the robot in the environment (Fig. 5).

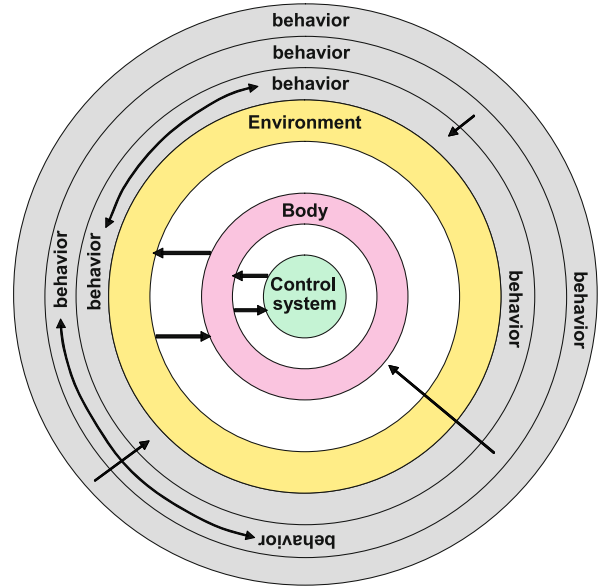
As we mentioned above, the ability to generate this form of representation which allow the robot to solve its adaptive problem originate from the coupling between a simple robot's internal dynamics and a simple robot/body/environmental dynamics. The former dynamics is characterized by the fact that the state of the two internal neurons tends to move slowly toward different fixed point attractors, in the robot's internal dynamics, which correspond to different type of sensory states exemplified in Fig. 5. The latter dynamics originate from the



Embodied and Situated Agents, Adaptive Behavior in, Figure 5
The state of the two internal neurons (i_1 and i_2) of the robot recorded for 330 s while the robot performs about 5 laps of the environment. The s , a , b , c , and d labels indicate the internal states corresponding to five different positions of the robot in the environment shown in Fig. 4. The other labels indicate the position of the fixed point attractors in the robot's internal dynamics corresponding to five types of sensory states experienced by the robot when it detects: a light in its frontal side (LF), a light on its rear side (LR), an obstacle on its right and frontal side (OFR), an obstacle on its right side (OR), no obstacles and no lights (NO)

fact that different types of sensory states last for different time durations and alternate with a given order while the robot move in the environment. The interaction between these two dynamical processes leads to a transient dynamics of agents' internal state which moves slowly toward the current fixed point attractor without never fully reaching it (thus preserving information about previously experienced sensory states, the time duration of these states, and the order with which they have been experienced). The coupling between the two dynamical processes originates from the fact that the free parameters which regulate the agent/environmental dynamics (e. g. the trajectory and the speed with which the robot moves in the environment) and the agent internal dynamics (e. g. the direction and the speed with which the internal neurons change their state) have been co-adapted and co-shaped during the adaptive process.

For related works which show how navigation and localization skills might emerge from the coupling between agent's internal and external dynamics, see [45]. For other works addressing other behavioral/cognitive capabilities see [4] for what concerns categorization, [16,41] for what concerns selective attention and [44] for what concern language and compositionality.

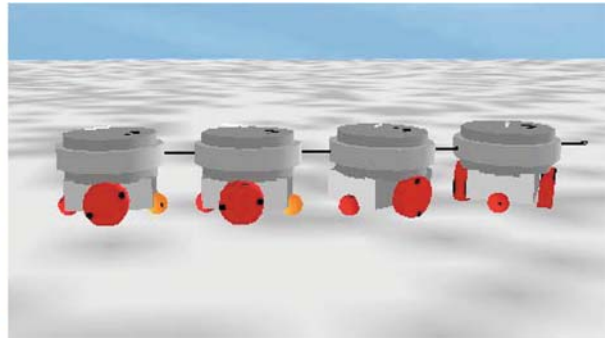


Embodied and Situated Agents, Adaptive Behavior in, Figure 6
A schematic representation of multi-level and multi-scale organization of behavior. The behaviors represented in the *inner circles* represent elementary behaviors which arise from fine-grained interactions between the control system, the body, and the environment, and which extend over limited time spans. The behaviors represented in the *external circles* represent higher level behaviors which arise from the combination and interaction between lower-level behaviors and which extend over longer time spans. The *arrows* which go from higher level behavior toward lower levels indicate the fact that the behaviors currently exhibited by the agents later affect the lower level behaviors and/or the fine-grained interaction between the constituting elements (agent's control system, agent's body, and the environment)

Behavior and Cognition as Phenomena with a Multi-Level and Multi-Scale Organization

Another fundamental feature that characterizes behavior is the fact that it is a multi-layer system with different levels of organizations extending at different time scales [2,19]. More precisely, as exemplified in Fig. 6, the behavior of an agent or of a group of agents involve both lower and higher level behaviors which extend for shorter or longer time spans, respectively. Lower level behaviors arise from few agent/environmental interactions and short term internal dynamical processes. Higher level behaviors, instead, arise from the combination and interaction of lower level behaviors and/or from long term internal dynamical processes.

The multi-level and multi-scale organization of agents' behavior play important roles: it is one of the factors which allow agents to produce functionally useful be-



Embodied and Situated Agents, Adaptive Behavior in, Figure 7

Left: Four robots assembled into a linear structure. **Right:** A simulation of the robots shown in the left part of the figure

havior without necessarily developing dedicated control mechanisms [8,9,29], it might favor the development of new behavioral and/or cognitive skills thanks to the recruitment of pre-existing capabilities [22], it allow agents to generalize their skills in new task/environmental conditions [29].

An exemplification of how the multi-level and multi-scale organization of behavior allow agents to generalize their skill in new environmental conditions is represented by the experiments carried out by Baldassarre et al. [2] in which the authors evolved the control system of a group of robots assembled into a linear structure (Fig. 7) for the ability to move in a coordinated manner and for the ability to display a coordinated light approaching behavior.

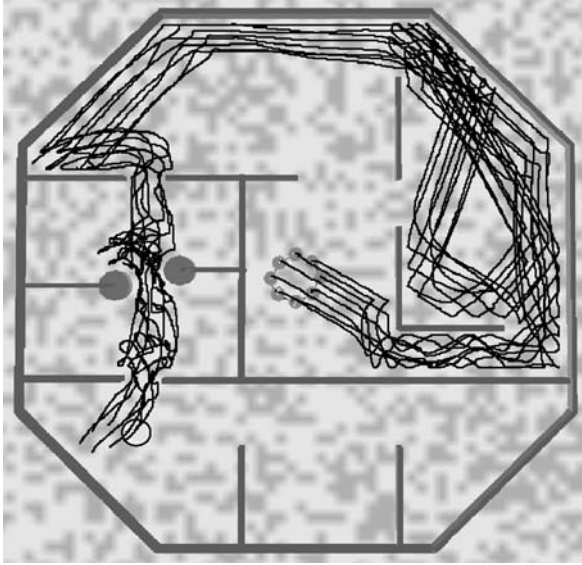
Each robot [27] consists of a mobile base (chassis) and a main body (turret) that can rotate with respect to the chassis along the vertical axis. The chassis has two drive mechanisms that control the two corresponding tracks and teathed wheels. The turret has one gripper, which allows robots to assemble together and to grasp objects, and a motor controlling the rotation of the turret with respect to the chassis. Robots are provided with a traction sensor, placed at the turret-chassis junction, that detects the intensity and the direction of the force that the turret exerts on the chassis (along the plane orthogonal to the vertical axis) and light sensors. Given that the orientations of individual robots might vary and given that the target light might be out of sight, robots need to coordinate to choose a common direction of movement and to change their direction as soon as one or few robots start to detect a light gradient.

Evolved individuals show the ability to negotiate a common direction of movement and by approaching light targets as soon as a light gradient is detected. By testing evolved robots in different conditions the authors observed that they are able to generalize their skills in new

conditions and also to spontaneously produce new behaviors which have not been rewarded during the evolutionary process. More precisely, groups of assembled robots display a capacity to generalize their skills with respect to the number of robots which are assembled together and to the shape formed by the assembled robots. Moreover, when the evolved controllers are embodied in eight robots assembled so to form a circular structure and situated in the maze environment shown in Fig. 8, the robots display an ability to collectively avoid obstacles, to rearrange their shape so to pass through narrow passages, and to explore the environment. The ability to display all these behavioral skills allow the robots to reach the light target even in large maze environments, i. e. even in environmental conditions which are rather different from the conditions that they experienced during the training process (Fig. 8).

By analyzing the behavior displayed by the evolved robots tested in the maze environment, a complex multi-level organization can be observed. The simpler behaviors that can be identified consist of low level individual behaviors which extend over short time spans:

1. A *move-forward behavior* which consists of the individuals' ability to move forward when the robot is coordinated with the rest of the team, is oriented toward the direction of the light gradient (if any), and does not collide with obstacles. This behavior results from the combination of: (a) a control rule which produces a move forward action when the perceived traction has a low intensity and when difference between the intensity of the light perceived on the left and the right side of the robot is low, and (b) the sensory effects of the execution of the move forward action selected mediated by the external environment which does not produce a variation of the state of the sensors until the conditions that should be satisfied to produce this behaviors hold.



Embodied and Situated Agents, Adaptive Behavior in, Figure 8

The behavior produced by eight robots assembled into a circular structure in a maze environment including walls and cylindrical objects (represented with gray lines and circles). The robots start in the central portion of the maze and reach the light target located in the bottom-left side of the environment (represented with an empty circle) by exhibiting a combination of coordinated-movement behaviors, collective obstacle-avoidance, and collective light-approaching behaviors. The irregular lines, that represent the trajectories of the individual robots, show how the shape of the assembled robots changes during motion by adapting to the local structure of the environment

2. A *conformistic behavior* which consists of the individuals' ability to conform its orientation with that of the rest of the team when the two orientations differ significantly. This behavior results from the combination of: (a) a control rule that makes the robot turns toward the direction of the traction when its intensity is significant, and (b) the sensory effects produced by the execution of this action mediated by the external environment that lead to a progressive reduction of the intensity of the traction until the orientation of the robot conform with the orientation of the rest of the group.
3. A *phototaxis behavior* which consists of the individuals' ability to orient toward the direction of the light target. This behavior results from the combination of: (a) a control rule that makes the robot turns toward the direction in which the intensity of the light gradient is higher, and (b) the sensory effects produced by the execution of this action mediated by the external environment that lead to a progressive reduction of the difference in the light intensity detected on the two side of the

robot until the orientation of the robot conforms with the direction of the light gradient.

4. An *obstacle-avoidance behavior* which consists of the individuals' ability to change direction of motion when the execution of a motor action produced a collision with an obstacle. This behavior results from the combination of: (a) the same control rule which lead to behavior #2 which make the robot turns toward the direction of the perceived traction (which in this case is caused by the collision with the obstacle while in the case of behavior #2 is caused by the forces exerted by the other assembled robots), and (b) the sensory effects produced by the execution of the turning action mediated by the external environment which make the robot turns until collisions do not prevent anymore the execution of a moving forward behavior.

The combination and the interaction between these three behaviors produce the following higher levels collective behaviors that extend over a longer time span:

5. A *coordinated-motion behavior* which consists in the ability of the robots to negotiate a common direction of movement and to keep moving along such direction by compensating further misalignments originating during motion. This behavior emerges from the combination and the interaction of the conformistic behavior (which plays the main role when robots are misaligned) and the move-forward behavior (which plays the main role when robots are aligned).
6. A *coordinated-light-approaching behavior* which consists in the ability of the robots to co-ordinately move toward a light target. This behavior emerges from the combination of the conformistic, the move-forward, and the phototaxis behaviors (which is triggered when the robots detect a light gradient). The relative importance of the three control rules which lead to the three corresponding behaviors depends both on the strength of the corresponding triggering condition (i. e. the extent of lack of traction forces, the intensity of traction forces, and the intensity of the light gradient, respectively) and on a priority relations among behaviors (i. e. the fact that the conformistic behavior tends to play a stronger role than the phototaxis behavior).
7. A *coordinated-obstacle-avoidance behavior* which consists in the ability of the robots to co-ordinately turn to avoid nearby obstacles. This behavior arises as the result of the combination of the obstacle avoidance-, the conformistic and the move-forward behaviors.

The combination and the interaction between these behaviors lead to the following higher levels collective behaviors that extend over still longer time spans:

8. A *collective-exploration-behavior* which consists in the ability of the robots to visit different area on the environment when the light target cannot be detected. This behavior emerges from the combination of the coordinated-motion behavior and the coordinate obstacle-avoidance behavior which ensures that the assembled robots can move in the environment without getting stuck and without entering into limit cycle trajectories.
9. A *shape-re-arrangement behavior* which consists in the ability of the assembled robots to dynamically adapt their shape to the current structure of the environment so to pass through narrow passages especially when the passages to be negotiated are in the direction of the light gradient. This behavior emerges from the combination and the interaction between coordinated motion and coordinated-light-approaching behaviors mediated by the effects produced by relative differences in motion between robots resulting from the execution of different motor actions and/or from differences in the collisions. The fact that the shape of the assembled robots adapt to the current environmental structure so to facilitate the overcoming of narrow passages can be explained by considering that collisions produce a modification of the shape which affect on particular the relative position of the colliding robots.

The combination and the interaction of all these behavior leads to a still higher level behavior:

10. A *collective-navigation-behavior* which consists in the ability of the assembled robots to navigate toward the light target by producing coordinated movements, exploring the environment, passing through narrow passages, and producing a coordinated-light-approaching behavior (Fig. 8).

This analysis illustrates two important mechanisms which explain the remarkable generalization abilities of these robots. The first mechanism consists in the fact that the control rules which regulate the interaction between the agents' and the environment so to produce certain behavioral skills in certain environmental conditions will produce different but related behavioral skills in other environmental conditions. In particular, the control rules which generate the behaviors #5 and #6 for which evolving robots have been evolved in an environment without obstacles also produce behavior #7 in an environment with obstacles. The second mechanism consists in the fact that the development of certain behaviors at a given level of organization which extend for a given time span will automatically lead to the exhibition of related higher-level

behaviors extending at longer time spans which originate from the interactions from the former behaviors (even if these higher level behaviors have not being rewarded during the adaptation process). In particular, the combination and the interaction of behaviors #5, #6, and #7 (which have been rewarded during the evolutionary process or which arise from the same control rules which lead to the generation of rewarded behaviors) automatically lead to the production of behaviors #8, #9, and #10 (which have not been rewarded). Obviously, there no warranty that the new behaviors obtained as a result of these generalization processes will play useful functions. However, the fact that these behaviors are related to the other functional behavioral skills implies that the probabilities that these new behavior will play useful functions is significant.

In principle, these generalization mechanisms can also be exploited by agents during their adaptive process to generate behavioral skills which play new functionalities and which emerge from the combination and the interaction between pre-existing behavioral skills playing different functions.

On the Top-Down Effect from Higher to Lower Levels of Organization

In the previous sections we have discussed how the interactions between the agents' body, the agents' control system, and the environment lead to behavioral and cognitive skills and how such skills have a multi-level and multi-scale organization in which the interaction between lower-level skills lead to the emergence of higher-level skills. However, higher level skills also affect lower level skills up to the fine-grained interaction between the constituting elements (agents' body, agents' control system, and environment). More precisely, the behaviors which originate from the interaction between the agent and the environment and from the interaction between lower levels behaviors, later affect the lower levels behaviors and the interaction from which they originate. These bi-directional influences between different levels of organization can lead to circular causality [20] where high level processes act as independent entities which constraint the lower level processes from which they originate.

One of the most important effects of this top-down influences consists in the fact that the behavior exhibited by an agent constraint the type of sensory patterns that the agent will experience later on (i. e. constraint the fine-grained agent/environmental interactions which determine the behavior that will be later exhibited by the agent). Since the complexity of the problem faced by an

agent depends on the sensory information experienced by the agent itself, this top down influences can be exploited in order to turn hard problems into simple ones.

One neat demonstration of this type of phenomena is given by the experiments conducted by Marocco and Nolfi [32] in which a simulated finger robot with six degree of freedom provided with sensors of its joint positions and with rough touch sensors is asked to discriminate between cubic and spherical objects varying in size. The problem is not trivial since, in general terms, the sensory patterns experienced by the robot do not provide clear regularities for discriminating between the two types of objects. However, the type of sensory states which are experienced by the agent also depend on the behavior previously exhibited by the agent itself – agents exhibiting different behavior might face simpler or harder problems. By evolving the robots in simulation for the ability to solve this problem and by analyzing the complexity of the problem faced by robots of successive generations, the authors observed that the evolved robot manage to solve their adaptive problem on the basis of simple control rules which allow the robot to approach the object and to move following the surface of the object from left to right, independently from the object shape. The exhibition of this behavior in interaction with objects characterized by a smooth or irregular surface (in the case of spherical or cubic objects, respectively) ensures that the same control rules lead to two types of behaviors depending on the type of the object. These behaviors consist in following the surface of the object and then moving away from the object in the case of spherical objects, and in following the surface of the object by getting stuck in a corner in the case of cubic objects. The exhibition of these two behaviors allows the agent to experience rather different proprioceptors states as a consequence of having had interacted with spherical or cubic object which nicely encode the regularities which are necessary to differentiate the two types of objects.

For other examples which shows how adaptive agents can exploit the fact that behavioral and cognitive processes which arise from the interaction between lower-level behaviors or between the constituting elements later affect these lower level processes see [4,28,39].

Adaptive Methods

In this section we briefly review the methods through which artificial embodied and situated agents can develop their skill autonomously while they interact at different levels of organization with the environment and eventually with other agents. These methods are inspired by the adap-

tive process observed in nature: evolution, maturation, development, and learning.

We will focus in particular on self-organized adaptive methodologies in which the role of the experimenter/designer is reduced to the minimum and in which the agents are free to develop their strategy to solve their adaptive problems within a large number of potentially alternative solutions. This choice is motivated by the following considerations:

- (a) These methods allow agents to identify the behavioral and cognitive skills which should be possessed, combined, and integrated so to solve the given problem. In other words, these methods can come up with effective ways of decomposing the overall required skill into a collection of simpler lower levels skills. Indeed, as we showed in the previous section, evolutionary adaptive techniques can discover ways of decomposing the high-level requested skill into lower-levels behavioral and cognitive skills so find solutions which are effective and parsimonious thanks to the exploitation of properties emerging from the interaction between lower-levels processes and skills and thanks to the recruitment of previously developed skills for performing new functions. In other words, these methods release the designer from the burden of deciding how the overall skill should be divided into a set of simpler skills and how these skills should be integrated. More importantly, these methods can come up with solutions exploiting emergent properties which would be hard to design [17,31].
- (b) These methods allow agents to identify how a given behavioral and cognitive skill can be produced, i.e. the appropriate fine-grained characteristics of agents' body structure and control rules regulating the agent/environmental interaction. As for the previous aspect, the advantage of using adaptive techniques lies not only in the fact that the experimenter is released from the burden of designing the fine-grained characteristics of the agents but also in the fact that adaptation might prove more effective than human design due to the inability of an external observer to foresee the effects of large number of non-linear interactions occurring at different levels of organization.
- (c) These methods allow agents to adapt to variations of the task, of the environment, and of the social conditions.

Current approaches, in this respect, can be grouped into two families which will be illustrated in the following subsections and which include Evolutionary Robotics methods and Developmental Robotics methods.

Evolutionary Robotics Methods

Evolutionary Robotics [14,31] is a method which allow to create embodied and situated agents able to adapt to their task/environment autonomously through an adaptive process inspired by natural evolution [18] and, eventually, through the combination of evolutionary, developmental, and learning processes.

The basic idea goes as follows (Fig. 9). An initial population of different artificial genotypes, each encoding the control system (and possibly the morphology) of an agent, is randomly created. Each genotype is translated into a corresponding phenotype (i. e. a corresponding agent) which is then left free to act (move, look around, manipulate the environment etc.) while its performance (fitness) with respect to a given task is automatically evaluated. In cases in which this methodology is applied to collective behaviors, agents are evaluated in groups which might be heterogeneous or homogeneous (i. e. might consist of agents which differ not with respect to their genetic and phenotypic characteristics). The fittest individuals (those having higher fitness) are allowed to reproduce by generating copies of their genotype with the addition of

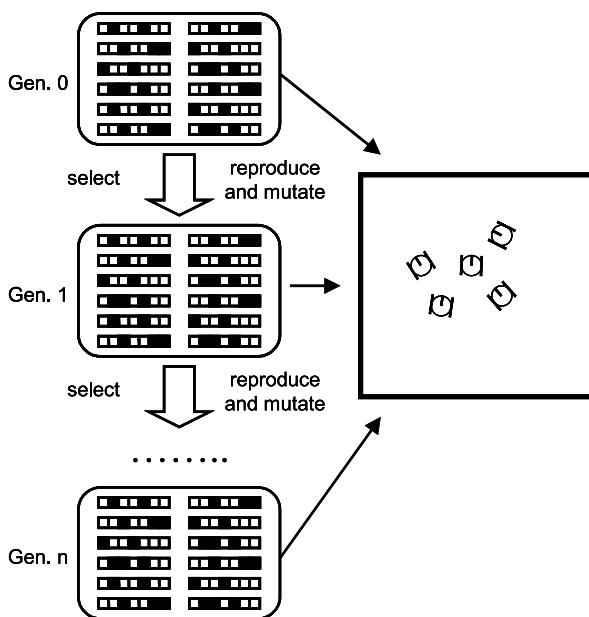
changes introduced by some genetic operators (e. g., mutations, exchange of genetic material). This process is repeated for a number of generations until an individual or a group of individuals is born which satisfies the performance level set by the user.

The process that determines how a genotype (i. e. typically a string of binary values) is turned into a corresponding phenotype (i. e. a robot with a given morphology and control system) might consist of a simple one-to-one mapping or of a complex developmental process. In the former case, many of the characteristics of the phenotypical individual (e. g. the shape of the body, the number and position of the sensors and of the actuators, and in some case the architecture of the neural controller) are pre-determined and fixed and the genotype encodes a vector of free parameters (e. g. the connection weights of the neural controller [31]). In the latter case, the genotype might encode a set of rules that determine how the body structure and the control system of the individual growth during an artificial developmental processes. Through these type of indirect developmental mappings most of the characteristics of the phenotypical robot can be encoded in the genotype and subjected to the evolutionary adaptive process [31,36]. Finally, in some cases the adaptation process might involve both an evolutionary process that regulates how the characteristics of the robots vary phylogenetically (i. e. throughout generations) and a developmental/learning process which regulates how the characteristics of the robots vary ontogenetically (i. e. during the phase in which the robots act in the environment [30]).

Evolutionary methods can be used to allow agents to develop the requested behavioral and cognitive skills from scratch (i. e. starting from agents which do not have any behavioral or cognitive capability) or in an incremental manner (i. e. starting from pre-evolved robots which already have some behavioral capability which consists, for example, in the ability to solve a simplified version of the adaptive problem).

The fitness function which determines whether an individual will be reproduced or not might also include, in the addition to a component that score the performance of the agent with respect to a given task, additional task-independent components. These additional components, in fact, can lead to the development of behavioral skills which are not necessarily functional but which can favor the development of functional skills later on [37].

Evolutionary methods can allow agents to develop low-levels behavioral and cognitive skills which have been previously identified by the designer/experimenter, which might later be combined and integrated in order to realize the high-level requested skill, or directly to develop the



Embodied and Situated Agents, Adaptive Behavior in, Figure 9
A schematic representation of the evolutionary process. The stripes with black and white squares represent individual genotypes. The rectangular boxes indicate the genome of a population of a certain generation. The small robots placed inside the square on the right part of the figure represent a group of robots situated in an environment which interact with the environment and between themselves

high-level requested skill. In the former case the adaptive process leads to the identification of the fine-grained features of the agent (e. g. number and type of sensors, body shape, architecture and connection weights of the neural controller) which by interacting between themselves and with the environment will produce the required skill. In the latter case, the adaptive process leads to the identification of the lower-levels skills (at different levels of organization) which are necessary to produce the required high-level skill, the identification of the way in which these lower levels skills should be combined and integrated, and (as for the former case) the identification of the fine grained features of the agent which, in interaction with the physical and social environment, will produce the required behavioral or cognitive skills.

Developmental Robotics Methods

Developmental Robotics [1,10,21], also known as *epigenetic robotics*, is a method for developing embodied and situated agents that adapt to their task/environment autonomously through processes inspired by biological developmental and learning processes.

Evolutionary and developmental robotics methods share the same fundamental assumptions but also present differences for what concerns the way in which they are realized and the type of situations in which they are typically applied. For what concerns the former aspect, unlike evolutionary robotics methods which operate on 'long' phylogenetic time scales, developmental methods typically operate on 'short' ontogenetic time scales. For what concerns the latter aspects, unlike evolutionary methods which are usually used to develop behavioral and cognitive skills from scratch, developmental methods are typically adopted to model the development of complex developmental and cognitive skills from simpler pre-existing skills which represents pre-requisites for the development of the required skills.

At the present stage, developmental robotics does not consist of a well defined methodology [1,21] but rather of a collection of approaches and methods often addressing complementary aspects which hopefully would be integrated in a single methodology in the future. Below briefly summarize some of the most important methodological aspects of the developmental robotics approach.

The Incremental Nature of the Developmental Process

Development should be characterized as an incremental process in which pre-existing structures and behavioral skills constitute important prerequisites and constraints for the development of more complex structures and be-

havioral skills and in which the complexity of the internal and external characteristics increases during developmental. One crucial aspect of developmental approach therefore consists in the identification of the initial characteristics and skills which should enable the bootstrapping of the developmental process: the layering of new skills on top of existing ones [10,25,38]. Another important aspect consists in shaping the developmental process so to ensure that the progressive increase in the complexity of the task matches the current competency of the system and so to drive the developmental process toward the progressive acquisition of the skills which represent the prerequisites for further developments. The progressive increase in complexity might concern not only the complexity of the task or of the required skills but also the complexity of single components of the robot/environmental interaction such as, for example, the number of freeze/unfreeze degrees of freedom [5].

The Social Nature of the Developmental Process

Development should involve social interaction with human subjects and with other developing robots. Social interactions (e. g. scaffolding, tutelage, mimicry, emulation, and imitation), in fact, play an important role not only for the development of social skills [7] but also as facilitators for the development of individual cognitive and behavioral skills [47]. Moreover, other types of social interactions (i. e. alignment processes or social games) might lead to the development of cognitive and/or behavioral skills which are generated by a collection of individuals and which could not be developed by a single individual robot [43].

Exploitation of the Interaction Between Concurrent Developmental Processes

Development should involve the exploitation of properties originating from the interaction and the integration of several co-occurring processes. Indeed, the co-development of different skills at the same time can favor the acquisition of the corresponding skills and of additional abilities arising from the combination and the integration of the developed skills. For example, the development of an ability to anticipate the sensory consequences of our own actions might facilitate the concurrent development of other skills such as categorical perception skills [46]. The development of an ability to pay attention to new situations (curiosity) and to look for new experiences after some time (boredom) might improve the learning of a given functional skill [33,42]. The co-development of behavioral and linguistic skills might favor the acquisition of the corresponding skills and the development of semantic combinatoriality skills Sugita and Tani [44].

Discussion and Conclusion

In this paper we described how artificial agents which are embodied and situated can develop behavioral and cognitive skills autonomously while they interact with their physical and social environment.

After having introduced the notion of embodiment and situatedness, we illustrated how the behavioral and cognitive skills displayed by adaptive agents can be properly characterized as complex system with multi-level and multi-scale properties resulting from a large number of interaction at different levels of organization and involving both bottom-up processes (in which the interaction between elements at lower levels of organization lead to higher levels properties) and top-down processes (in which properties at a certain level of organization later affect lower level properties or processes).

Finally, we briefly introduced the methods which can be used to synthesize adaptive embodied and situated agents.

The complex system nature of adaptive agents which are embodied and situated has important implications which constraint the organization of these systems and the dynamics of the adaptive process through which they develop their skills.

For what concerns the organization of these systems, it implies that agents' behavioral and/or cognitive skills (at any stage of the adaptive process) cannot be traced back to anyone of the three foundational elements (i. e. the body of the agents, the control system of the agents, and the environment) in isolation but should rather be characterized as properties which emerge from the interactions between these three elements and the interaction between behavioral and cognitive properties emerging from the former interactions at different levels of organizations. Moreover, it implies that 'complex' behavioral or cognitive skills might emerge from the interaction between simple properties or processes.

For what concerns agents' adaptive process, it implies that the development of new 'complex' skills does not necessarily require the development of new 'complex' morphological features or new 'complex' control mechanisms. Indeed, new 'complex' skills might arise from the addition of new 'simple' features or new 'simple' control rules which, in interaction with the pre-existing features and processes, might produce the required new behavioral or cognitive skills.

The study of adaptive behavior in artificial agents which has been reviewed in this paper has important implication both from an engineering point of view (i. e. for progressing in our ability to develop effective machines)

and from a modeling point of view (i. e. for understanding the characteristics of biological organisms).

In particular, from an engineering point of view, progresses in our ability to develop adaptive embodied and situated agents can lead to development of machines playing useful functionalities.

From a modeling point of view, progresses in our ability to model and analyze artificial adaptive agents can improve our understanding of the general mechanisms behind animal and human intelligence. For example, the comprehension of the complex system nature of behavioral and cognitive skills illustrated in this paper can allow us to better define the notion of embodiment and situatedness which represent two foundational concepts in the study of natural and artificial intelligence. Indeed, although possessing a body and being in a physical environment certainly represent a pre-requisite for considering an agent embodied and situated, a more useful definition of embodiment (or of truly embodiment) can be given in term of the extent to which a given agent exploits its body characteristics to solve its adaptive problem (i. e. the extent to which its body structure is adapted to the problem to be solved, or in other words, the extent to which its body performs morphological computation). Similarly, a more useful definition of situatedness (or truly situatedness) can be given in terms of the extent to which an agent exploits its interaction with the physical and social environment and the properties originating from this interaction to solve its adaptive problem. For sake of clarity we can refer to the former definition of the terms (i. e. possessing a physical body and being situated in a physical environment) as embodiment and situatedness as weak sense, and to the latter definition as embodiment and situatedness in a strong sense.

Bibliography

1. Asada M, MacDorman K, Ishiguro H, Kuniyoshi Y (2001) Cognitive developmental robotics as a new paradigm for the design of humanoid robots. *Robot Auton Syst* 37:185–193
2. Baldassarre G, Parisi D, Nolfi S (2006) Distributed coordination of simulated robots based on self-organisation. *Artif Life* 3(12):289–311
3. Beer RD (1995) A dynamical systems perspective on agent-environment interaction. *Artif Intell* 72:173–215
4. Beer RD (2003) The dynamics of active categorical perception in an evolved model agent. *Adapt Behav* 11:209–243
5. Berthouze L, Lungarella M (2004) Motor skill acquisition under environmental perturbations: on the necessity of alternate freezing and freeing. *Adapt Behav* 12(1):47–63
6. Bongard JC, Paul C (2001) Making evolution an offer it can't refuse: Morphology and the extradimensional bypass. In: Keleman J, Sosik P (eds) *Proceedings of the Sixth European Con-*

- ference on Artificial Life. Lecture Notes in Artificial Intelligence, vol 2159. Springer, Berlin
7. Breazeal C (2003) Towards sociable robots. *Robotics Auton Syst* 42(3–4):167–175
 8. Brooks RA (1991) Intelligence without reason. In: Mylopoulos J, Reiter R (eds) *Proceedings of 12th International Joint Conference on Artificial Intelligence*. Morgan Kaufmann, San Mateo
 9. Brooks RA (1991) Intelligence without reason. In: *Proceedings of 12th International Joint Conference on Artificial Intelligence*. Sydney, Australia, pp 569–595
 10. Brooks RA, Breazeal C, Irie R, Kemp C, Marjanovic M, Scassellati B, Williamson M (1998) Alternate essences of intelligence. In: *Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, Madison, Wisconsin, pp 961–976
 11. Chiel HJ, Beer RD (1997) The brain has a body: Adaptive behavior emerges from interactions of nervous system, body and environment. *Trends Neurosci* 20:553–557
 12. Clark A (1997) *Being there: Putting brain, body and world together again*. MIT Press, Cambridge
 13. Endo I, Yamasaki F, Maeno T, Kitano H (2002) A method for co-evolving morphology and walking patterns of biped humanoid robot. In: *Proceedings of the IEEE Conference on Robotics and Automation*, Washington, D.C.
 14. Floreano D, Husband P, Nolfi S (2008) Evolutionary Robotics. In: Siciliano B, Oussama Khatib (eds) *Handbook of Robotics*. Springer, Berlin
 15. Gigliotta O, Nolfi S (2008) On the coupling between agent internal and agent/environmental dynamics: Development of spatial representations in evolving autonomous robots. *Adapt Behav* 16:148–165
 16. Goldenberg E, Garcowski J, Beer RD (2004) May we have your attention: Analysis of a selective attention task. In: Schaal S, Ijspeert A, Billard A, Vijayakumar S, Hallam J, Meyer J-A (eds) *From Animals to Animats 8: Proceedings of the Eighth International Conference on the Simulation of Adaptive Behavior*. MIT Press, Cambridge
 17. Harvey I (2000) Robotics: Philosophy of mind using a screwdriver. In: Gomi T (ed) *Evolutionary Robotics: From Intelligent Robots to Artificial Life*, vol III. AAI Books, Ontario
 18. Holland J (1975) *Adaptation in natural and artificial systems*. University of Michigan Press, Ann Arbor
 19. Keijzer F (2001) *Representation and behavior*. MIT Press, London
 20. Kelso JAS (1995) *Dynamics patterns: The self-organization of brain and behaviour*. MIT Press, Cambridge
 21. Lungarella M, Metta G, Pfeifer R, Sandini G (2003) Developmental robotics: a survey. *Connect Sci* 15:151–190
 22. Marocco D, Nolfi S (2007) Emergence of communication in embodied agents evolved for the ability to solve a collective navigation problem. *Connect Sci* 19(1):53–74
 23. Massera G, Cangelosi A, Nolfi S (2007) Evolution of prehension ability in an anthropomorphic neurorobotic arm. *Front Neurobot* 1(4):1–9
 24. McGeer T (1990) Passive walking with knees. In: *Proceedings of the IEEE Conference on Robotics and Automation*, vol 2, pp 1640–1645
 25. Metta G, Sandini G, Natale L, Panerai F (2001) Development and Q30 robotics. In: *Proceedings of IEEE-RAS International Conference on Humanoid Robots*, pp 33–42
 26. Mondada F, Franzi E, lenne P (1993) Mobile robot miniaturisation: A tool for investigation in control algorithms. In: *Proceedings of the Third International Symposium on Experimental Robotics*, Kyoto, Japan
 27. Mondada F, Pettinaro G, Guirard A, Kwee I, Floreano D, Denebourg J-L, Nolfi S, Gambardella LM, Dorigo M (2004) Swarm-bot: A new distributed robotic concept. *Auton Robots* 17(2–3):193–221
 28. Nolfi S (2002) Power and limits of reactive agents. *Neurocomputing* 49:119–145
 29. Nolfi S (2005) Behaviour as a complex adaptive system: On the role of self-organization in the development of individual and collective behaviour. *Complexus* 2(3–4):195–203
 30. Nolfi S, Floreano D (1999) Learning and Evolution. *Auton Robots* 1:89–113
 31. Nolfi S, Floreano D (2000) *Evolutionary Robotics: The Biology, Intelligence, and Technology of Self-Organizing Machines*. MIT Press/Bradford Books, Cambridge
 32. Nolfi S, Marocco D (2002) Active perception: A sensorimotor account of object categorization. In: Hallam B, Floreano D, Hallam J, Hayes G, Meyer J-A (eds) *From Animals to Animats 7, Proceedings of the VII International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, pp 266–271
 33. Oudeyer P-Y, Kaplan F, Hafner V (2007) Intrinsic motivation systems for autonomous mental development. *IEEE Trans Evol Comput* 11(2):265–286
 34. Pfeifer R, Bongard J (2007) *How the body shape the way we think*. MIT Press, Cambridge
 35. Pfeifer R, Iida F, Gómez G (2006) Morphological computation for adaptive behavior and cognition. In: *International Congress Series*, vol 1291, pp 22–29
 36. Pollack JB, Lipson H, Funes P, Hornby G (2001) Three generations of coevolutionary robotics. *Artif Life* 7:215–223
 37. Prokopenko M, Gerasimov V, Tanev I (2006) Evolving spatiotemporal coordination in a modular robotic system. In: Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (eds) *Artificial Life X: Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*. MIT Press, Boston
 38. Scassellati B (2001) *Foundations for a Theory of Mind for a Humanoid Robot*. Ph D thesis, Department of Electrical Engineering and Computer Science, MIT, Boston
 39. Scheier C, Pfeifer R, Kunyoshi Y (1998) Embedded neural networks: exploiting constraints. *Neural Netw* 11:1551–1596
 40. Schmitz A, Gómez G, Iida F, Pfeifer R (2007) On the robustness of simple speed control for a quadruped robot. In: *Proceeding of the International Conference on Morphological Computation*, Venice, Italy
 41. Slocum AC, Downey DC, Beer RD (2000) Further experiments in the evolution of minimally cognitive behavior: From perceiving affordances to selective attention. In: Meyer J, Berthoz A, Floreano D, Roitblat H, Wilson S (eds) *From Animals to Animats 6. Proceedings of the Sixth International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge
 42. Schmidhuber J (2006) Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connect Sci* 18(2):173–187
 43. Steels L (2003) Evolving grounded communication for robots. *Trends Cogn Sci* 7(7):308–312
 44. Sugita Y, Tani J (2005) Learning semantic combinatoriality from the interaction between linguistic and behavioral processes. *Adapt Behav* 13(1):33–52

45. Tani J, Fukumura N (1997) Self-organizing internal representation in learning of navigation: A physical experiment by the mobile robot Yamabico. *Neural Netw* 10(1):153–159
46. Tani J, Nolfi S (1999) Learning to perceive the world as articulated: An approach for hierarchical learning in sensory-motor systems. *Neural Netw* 12:1131–1141
47. Tani J, Nishimoto R, Namikawa J, Ito M (2008) Co-developmental learning between human and humanoid robot using a dynamic neural network model. *IEEE Trans Syst Man Cybern B. Cybern* 38:1
48. Varela FJ, Thompson E, Rosch E (1991) *The Embodied mind: Cognitive science and human experience*. MIT Press, Cambridge
49. van Gelder TJ (1998) The dynamical hypothesis in cognitive science. *Behav Brain Sci* 21:615–628
50. Vaughan E, Di Paolo EA, Harvey I (2004) The evolution of control and adaptation in a 3D powered passive dynamic walker. In: Pollack J, Bedau M, Husband P, Ikegami T, Watson R (eds) *Proceedings of the Ninth International Conference on the Simulation and Synthesis of Living Systems*. MIT Press, Cambridge

Entropy

CONSTANTINO TSALLIS^{1,2}

¹ Centro Brasileiro de Pesquisas Físicas,
Rio de Janeiro, Brazil

² Santa Fe Institute, Santa Fe, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Some Basic Properties

Boltzmann–Gibbs Statistical Mechanics

On the Limitations of Boltzmann–Gibbs Entropy
and Statistical Mechanics

The Nonadditive Entropy S_q

A Connection Between Entropy and Diffusion

Standard and q -Generalized Central Limit Theorems

Future Directions

Acknowledgments

Bibliography

Glossary

Absolute temperature Denoted T .

Clausius entropy Also called *thermodynamic entropy*. Denoted S .

Boltzmann–Gibbs entropy Basis of Boltzmann–Gibbs statistical mechanics. This entropy, denoted S_{BG} , is *additive*. Indeed, for two probabilistically independent subsystems A and B , it satisfies $S_{BG}(A+B) = S_{BG}(A) + S_{BG}(B)$.

Nonadditive entropy It usually refers to the basis of nonextensive statistical mechanics. This entropy, denoted S_q , is *nonadditive* for $q \neq 1$. Indeed, for two probabilistically independent subsystems A and B , it satisfies $S_q(A+B) \neq S_q(A) + S_q(B)$ ($q \neq 1$). For historical reasons, it is frequently (but inadequately) referred to as *nonextensive entropy*.

q -logarithmic and q -exponential functions Denoted $\ln_q x$ ($\ln_1 x = \ln x$), and e_q^x ($e_1^x = e^x$), respectively.

Extensive system So called for historical reasons. A more appropriate name would be *additive system*. It is a system which, in one way or another, relies on or is connected to the (additive) Boltzmann–Gibbs entropy. Its basic dynamical and/or structural quantities are expected to be of the exponential form. In the sense of complexity, it may be considered a *simple system*.

Nonextensive system So called for historical reasons. A more appropriate name would be *nonadditive system*. It is a system which, in one way or another, relies on or is connected to a (nonadditive) entropy such as S_q ($q \neq 1$). Its basic dynamical and/or structural quantities are expected to asymptotically be of the power-law form. In the sense of complexity, it may be considered a *complex system*.

Definition of the Subject

Thermodynamics and statistical mechanics are among the most important formalisms in contemporary physics. They have overwhelming and intertwined applications in science and technology. They essentially rely on two basic concepts, namely *energy* and *entropy*. The mathematical expression that is used for the first one is well known to be *nonuniversal*; indeed, it depends on whether we are say in classical, quantum, or relativistic regimes. The second concept, and very specifically its connection with the microscopic world, has been considered during well over one century as essentially unique and *universal* as a physical concept. Although some mathematical generalizations of the entropy have been proposed during the last forty years, they have frequently been considered as mere practical expressions for disciplines such as cybernetics and control theory, with no particular physical interpretation. What we have witnessed during the last two decades is the growth, among physicists, of the belief that it is not necessarily so. In other words, the physical entropy would basically rely on the microscopic dynamical and structural properties of the system under study. For example, for systems microscopically evolving with strongly chaotic dynamics, the connection between the thermodynamical entropy and the thermostistical entropy would be the one

found in standard textbooks. But, for more complex systems (e.g., for weakly chaotic dynamics), it becomes either necessary, or convenient, or both, to extend the traditional connection. The present article presents the ubiquitous concept of entropy, useful even for systems for which no energy can be defined at all, within a standpoint reflecting a *nonuniversal* conception for the connection between the thermodynamic and the thermostistical entropies. Consequently, both the standard entropy and its recent generalizations, as well as the corresponding statistical mechanics, are here presented on equal footing.

Introduction

The concept of *entropy* (from the Greek $\epsilon\nu\tau\rho\epsilon\pi\omega$, *entrepo*, at turn, at transformation) was first introduced in 1865 by the German physicist and mathematician Rudolf Julius Emanuel Clausius, Rudolf Julius Emanuel in order to mathematically complete the formalism of classical thermodynamics [55], one of the most important theoretical achievements of contemporary physics. The term was so coined to make a parallel to *energy* (from the Greek $\epsilon\nu\epsilon\rho\gamma\omicron\varsigma$, *energos*, at work), the other fundamental concept of thermodynamics. Clausius connection was given by

$$dS = \frac{\delta Q}{T}, \quad (1)$$

where δQ denotes an infinitesimal transfer of heat. In other words, $1/T$ acts as an integrating factor for δQ . In fact, it was only in 1909 that thermodynamics was finally given, by the Greek mathematician Constantin Caratheodory, a logically consistent axiomatic formulation.

In 1872, some years after Clausius proposal, the Austrian physicist Ludwig Eduard Boltzmann introduced a quantity, that he noted H , which was defined in terms of microscopic quantities:

$$H \equiv \iiint f(\mathbf{v}) \ln[f(\mathbf{v})] d\mathbf{v}, \quad (2)$$

where $f(\mathbf{v})d\mathbf{v}$ is the number of molecules in the velocity space interval $d\mathbf{v}$. Using Newtonian mechanics, Boltzmann showed that, under some intuitive assumptions (*Stoßzahlansatz* or *molecular chaos hypothesis*) regarding the nature of molecular collisions, H does not increase with time. Five years later, in 1877, he identified this quantity with Clausius entropy through $-kH \equiv S$, where k is a constant. In other words, he established that

$$S = -k \iiint f(\mathbf{v}) \ln[f(\mathbf{v})] d\mathbf{v}, \quad (3)$$

later on generalized into

$$S = -k \iint f(\mathbf{q}, \mathbf{p}) \ln[f(\mathbf{q}, \mathbf{p})] d\mathbf{q} d\mathbf{p}, \quad (4)$$

where (\mathbf{q}, \mathbf{p}) is called the μ -space and constitutes the phase space (coordinate \mathbf{q} and momentum \mathbf{p}) corresponding to one particle.

Boltzmann's genius insight – the first ever mathematical connection of the macroscopic world with the microscopic one – was, during well over three decades, highly controversial since it was based on the hypothesis of the existence of atoms. Only a few selected scientists, like the English chemist and physicist John Dalton, the Scottish physicist and mathematician James Clerk Maxwell, and the American physicist, chemist and mathematician Josiah Willard Gibbs, believed in the reality of atoms and molecules. A large part of the scientific establishment was, at the time, strongly against such an idea. The intricate evolution of Boltzmann's lifelong epistemological struggle, which ended tragically with his suicide in 1906, may be considered as a neat illustration of Thomas Kuhn's *paradigm shift*, and the corresponding reaction of the scientific community, as described in *The Structure of Scientific Revolutions*. There are in fact two important formalisms in contemporary physics where the mathematical theory of probabilities enters as a central ingredient. These are statistical mechanics (with the concept of entropy as a functional of probability distributions) and quantum mechanics (with the physical interpretation of wave functions and measurements). In both cases, contrasting viewpoints and passionate debates have taken place along more than one century, and continue still today. This is no surprise after all. If it is undeniable that energy is a very deep and subtle concept, entropy is even more. Indeed, energy concerns the world of (microscopic) *possibilities*, whereas entropy concerns the world of the *probabilities* of those possibilities, a step further in epistemological difficulty.

In his 1902 celebrated book *Elementary Principles of Statistical Mechanics*, Gibbs introduced the modern form of the entropy for classical systems, namely

$$S = -k \int d\Gamma f(\mathbf{q}, \mathbf{p}) \ln[Cf(\mathbf{q}, \mathbf{p})], \quad (5)$$

where Γ represents the *full phase space* of the system, thus containing all coordinates and all momenta of its elementary particles, and C is introduced to take into account the finite size and the physical dimensions of the smallest admissible cell in Γ -space. The constant k is known today to be a universal one, called *Boltzmann constant*, and given by $k = 1.3806505(24) \times 10^{-23}$ Joule/Kelvin. The studies

of the German physicist Max Planck along Boltzmann and Gibbs lines after the appearance of quantum mechanical concepts, eventually led to the expression

$$S = k \ln W, \quad (6)$$

which he coined as *Boltzmann entropy*. This expression is carved on the stone of Boltzmann's grave at the Central Cemetery of Vienna. The quantity W is the total number of microstates of the system that are compatible with our macroscopic knowledge of it. It is obtained from Eq. (5) under the hypothesis of an uniform distribution or equal probabilities.

The Hungarian-American mathematician and physicist Johann von Neumann extended the concept of BG entropy in two steps – in 1927 and 1932 respectively –, in order to also cover quantum systems. The following expression, frequently referred to as the *von Neumann entropy*, resulted:

$$S = -k \operatorname{Tr} \rho \ln \rho, \quad (7)$$

ρ being the *density operator* (with $\operatorname{Tr} \rho = 1$).

Another important step was given in 1948 by the American electrical engineer and mathematician Claude Elwood Shannon. Having in mind the theory of digital communications he explored the properties of the discrete form

$$S = -k \sum_{i=1}^W p_i \ln p_i, \quad (8)$$

frequently referred to as *Shannon entropy* (with $\sum_{i=1}^W p_i = 1$). This form can be recovered from Eq. (5) for the particular case for which the phase space density $f(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^W p_i \delta(\mathbf{q} - \mathbf{q}_i) \delta(\mathbf{p} - \mathbf{p}_i)$. It can also be recovered from Eq. (7) when ρ is diagonal. We may generically refer to Eqs. (5), (6), (7) and (8) as the *BG entropy*, noted S_{BG} . It is a measure of the disorder of the system or, equivalently, of our degree of ignorance or lack of information about its state. To illustrate a variety of properties, the discrete form (8) is particularly convenient.

Some Basic Properties

Non-negativity It can be easily verified that, in all cases, $S_{BG} \geq 0$, the zero value corresponding to *certainty*, i. e., $p_i = 1$ for one of the W possibilities, and zero for all the others. To be more precise, it is exactly so whenever S_{BG} is expressed either in the form (7) or in the form (8). However, this property of non-negativity may be no longer true if it is expressed in the form (5).

This violation is one of the mathematical manifestations that, at the microscopic level, the state of any physical system exhibits its quantum nature.

Expansibility Also $S_{BG}(p_1, p_2, \dots, p_W, 0) = S_{BG}(p_1, p_2, \dots, p_W)$, i. e., zero-probability events do not modify our information about the system.

Maximal value S_{BG} is maximized at equal probabilities, i. e., for $p_i = 1/W$, $\forall i$. Its value is that of Eq. (6). This corresponds to the Laplace *principle of indifference* or *principle of insufficient reason*.

Concavity If we have two arbitrary probability distributions $\{p_i\}$ and $\{p'_i\}$ for the same set of W possibilities, we can define the *intermediate* probability distribution $p''_i = \mu p_i + (1 - \mu) p'_i$ ($0 < \mu < 1$). It straightforwardly follows that $S_{BG}(\{p''_i\}) \geq \mu S_{BG}(\{p_i\}) + (1 - \mu) S_{BG}(\{p'_i\})$. This property is essential for thermodynamics since it eventually leads to *thermodynamic stability*, i. e., to *robustness* with regard to energy fluctuations. It also leads to the tendency of the entropy to attain, as time evolves, its maximal value compatible with our macroscopic knowledge of the system, i. e., with the possibly known values for the macroscopic constraints.

Lesche stability or experimental robustness B. Lesche introduced in 1982 [107] the definition of an interesting property, which he called *stability*. It reflects the *experimental robustness* that a physical quantity is expected to exhibit. In other words, similar experiments should yield similar numerical results for the physical quantities. Let us consider two probability distributions $\{p_i\}$ and $\{p'_i\}$, assumed to be close, in the sense that $\sum_{i=1}^W |p_i - p'_i| < \delta$, $\delta > 0$ being a small number. An entropic functional $S(\{p_i\})$ is said *stable* or *experimentally robust* if, for any given $\epsilon > 0$, a $\delta > 0$ exists such that $|S(\{p_i\}) - S(\{p'_i\})|/S_{\max} < \epsilon$, where S_{\max} is the maximal value that the functional can attain ($\ln W$ in the case of S_{BG}). This implies that $\lim_{\delta \rightarrow 0} \lim_{W \rightarrow \infty} (S(\{p_i\}) - S(\{p'_i\}))/S_{\max} = 0$. As we shall see soon, this property is much stronger than it seems at first sight. Indeed, it provides a (necessary but not sufficient) criterion for classifying entropic functionals into physically admissible or not. It can be shown that S_{BG} is *Lesche-stable* (or *experimentally robust*).

Entropy production If we start the (deterministic) time evolution of a generic classical system from an arbitrarily chosen point in its Γ phase space, it typically follows a quite erratic trajectory which, in many cases, gradually visits the entire (or almost) phase space. By making partitions of this Γ -space, and counting the frequency of visits to the various cells (and related symbolic quantities), it is possible to define probabil-

ity sets. Through them, we can calculate a sort of time evolution of $S_{BG}(t)$. If the system is *chaotic* (sometimes called *strongly chaotic*), i.e., if its sensitivity to the initial conditions increases exponentially with time, then $S_{BG}(t)$ increases *linearly* with t in the appropriate asymptotic limits. This rate of increase of the entropy is called *Kolmogorov–Sinai entropy rate*, and, for a large class of systems, it coincides (*Pesin identity* or *Pesin theorem*) with the sum of the positive Lyapunov exponents. These exponents characterize the exponential divergences, along various directions in the Γ -space, of a small discrepancy in the initial condition of a trajectory.

It turns out, however, that the Kolmogorov–Sinai entropy rate is, in general, quite inconvenient for computational calculations for arbitrary nonlinear dynamical systems. In practice, another quantity is used instead [102], usually referred to as *entropy production per unit time*, which we note K_{BG} . Its definition is as follows. We first make a partition of the Γ -space into many W cells ($i = 1, 2, \dots, W$). In one of them, arbitrarily chosen, we randomly place M initial conditions (i.e., an *ensemble*). As time evolves, the occupancy of the W cells determines the set $\{M_i(t)\}$, with $\sum_{i=1}^W M_i(t) = M$. This set enables the definition of a probability set with $p_i(t) \equiv M_i(t)/M$, which in turn determines $S_{BG}(t)$. We then define the *entropy production per unit time* as follows:

$$K_{BG} \equiv \lim_{t \rightarrow \infty} \lim_{W \rightarrow \infty} \lim_{M \rightarrow \infty} \frac{S_{BG}(t)}{t}. \quad (9)$$

Up to date, no theorem guarantees that this quantity coincides with the Kolmogorov–Sinai entropy rate. However, many numerical approaches of various chaotic systems strongly suggest so. The same turns out to occur with what is frequently referred in the literature as a Pesin-like identity. For instance, if we have a one-dimensional dynamical system, its sensitivity to the initial conditions $\xi \equiv \lim_{\Delta x(0) \rightarrow 0} \Delta x(t)/\Delta x(0)$ is typically given by

$$\xi(t) = e^{\lambda t}, \quad (10)$$

where $\Delta x(t)$ is the discrepancy in the one-dimensional phase space of two trajectories initially differing by $\Delta x(0)$, and λ is the Lyapunov exponent ($\lambda > 0$ corresponds to *strongly sensitive to the initial conditions*, or *strongly chaotic*, and $\lambda < 0$ corresponds to *strongly insensitive to the initial conditions*). The so-called Pesin-like identity amounts, if $\lambda \geq 0$, to

$$K_{BG} = \lambda. \quad (11)$$

Additivity and extensivity If we consider a system $A + B$ constituted by two *probabilistically independent* subsystems A and B , i.e., if we consider $p_{ij}^{A+B} = p_i^A p_j^B$, we immediately obtain from Eq. (8) that

$$S_{BG}(A + B) = S_{BG}(A) + S_{BG}(B). \quad (12)$$

In other words, the BG entropy is *additive* [130]. If our system is constituted by N probabilistically independent identical subsystems (or *elements*), we clearly have $S_{BG}(N) \propto N$. It frequently happens, however, that the N elements are not exactly independent but only asymptotically so in the $N \rightarrow \infty$ limit. This is the usual case of many-body Hamiltonian systems involving only *short-range* interactions, where the concept of short-range will be focused in detail later on. For such systems, S_{BG} is only asymptotically additive, i.e.,

$$0 < \lim_{N \rightarrow \infty} \frac{S_{BG}(N)}{N} < \infty. \quad (13)$$

An entropy $S(\{p_i\})$ of a specific systems is said *extensive* if it satisfies

$$0 < \lim_{N \rightarrow \infty} \frac{S(N)}{N} < \infty, \quad (14)$$

where *no hypothesis at all* is made about the possible independence or weak or strong correlations between the elements of the system whose entropy S we are considering. Equation (13) amounts to say that the additive entropy S_{BG} is extensive for weakly correlated systems such as the already mentioned many-body short-range-interacting Hamiltonian ones. It is important to clearly realize that *additivity* and *extensivity* are independent properties. An additive entropy such as S_{BG} is extensive for simple systems such as the ones just mentioned, but it turns out to be nonextensive for other, more complex, systems that will be focused on later on. For many of these more complex systems, it is the nonadditive entropy S_q (to be analyzed later on) which turns out to be extensive for a non standard value of q (i.e., $q \neq 1$).

Boltzmann–Gibbs Statistical Mechanics

Physical systems (classical, quantum, relativistic) can be theoretically described in very many ways, through microscopic, mesoscopic, macroscopic equations, reflecting either stochastic or deterministic time evolutions, or even both types simultaneously. Those systems whose time evolution is completely determined by a well defined Hamiltonian with appropriate boundary conditions and ad-

missible initial conditions are the main purpose of an important branch of contemporary physics, named *statistical mechanics*. This remarkable *theory* (or *formalism*, as sometimes called), which for large systems satisfactorily matches classical thermodynamics, was primarily introduced by Boltzmann and Gibbs. The physical system can be in all types of situations. Two paradigmatic such situations correspond to *isolation*, and *thermal contact* with a large reservoir called *thermostat*. Their *stationary state* ($t \rightarrow \infty$) is usually referred to as *thermal equilibrium*. Both situations have been formally considered by Gibbs within his mathematical formulation of statistical mechanics, and they respectively correspond to the so-called *micro-canonical* and *canonical* ensembles (other ensembles do exist, such as the *grand-canonical* ensemble, appropriate for those situations in which the total number of elements of the system is not fixed; this is however out of the scope of the present article).

The stationary state of the micro-canonical ensemble is determined by $p_i = 1/W$ ($\forall i$, where i runs over all possible microscopic states), which corresponds to the extremization of S_{BG} with a single (and trivial) constraint, namely

$$\sum_{i=1}^W p_i = 1. \quad (15)$$

To obtain the stationary state for the canonical ensemble, the thermostat being at temperature T , we must (typically) add one more constraint, namely

$$\sum_{i=1}^W p_i E_i = U, \quad (16)$$

where $\{E_i\}$ are the energies of all the possible states of the system (i. e., eigenvalues of the Hamiltonian with the appropriate boundary conditions). The extremization of S_{BG} with the two constraints above straightforwardly yields

$$p_i = \frac{e^{-\beta E_i}}{Z} \quad (17)$$

$$Z \equiv \sum_{j=1}^W e^{-\beta E_j} \quad (18)$$

with the *partition function* Z , and the Lagrange parameter $\beta = 1/kT$. This is the celebrated *BG distribution* for thermal equilibrium (or *Boltzmann weight*, or *Gibbs state*, as also called), which has been at the basis of an enormous amount of successes (in fluids, magnets, superconductors, superfluids, Bose–Einstein condensation, conductors, chemical reactions, percolation, among many

other important situations). The connection with classical thermodynamics, and its Legendre-transform structure, occurs through relations such as

$$\frac{1}{T} = \frac{\partial S}{\partial U} \quad (19)$$

$$F \equiv U - TS = -\frac{1}{\beta} \ln Z \quad (20)$$

$$U = -\frac{\partial}{\partial \beta} \ln Z \quad (21)$$

$$C \equiv T \frac{\partial S}{\partial T} = \frac{\partial U}{\partial T} = -T \frac{\partial^2 F}{\partial T^2}, \quad (22)$$

where F , U and C are the *Helmholtz free energy*, the *internal energy*, and the *specific heat* respectively. The BG statistical mechanics historically appeared as the first connection between the microscopic and the macroscopic descriptions of the world, and it constitutes one of the cornerstones of contemporary physics. The Establishment resisted heavily before accepting the validity and power of Boltzmann's revolutionary ideas. In 1906 Boltzmann dramatically committed suicide, after 34 years that he had first proposed the deep ideas that we are summarizing here. At that early 20th century, few people believed in Boltzmann's proposal (among those few, we must certainly mention Albert Einstein), and most physicists were simply unaware of the existence of Gibbs and of his profound contributions. It was only half a dozen years later that the emerging new generation of physicists recognized their respective genius (thanks in part to various clarifications produced by Paul Ehrenfest, and also to the experimental successes related with Brownian motion, photoelectric effect, specific heat of solids, and black-body radiation).

On the Limitations of Boltzmann–Gibbs Entropy and Statistical Mechanics

Historical Background

As any other human intellectual construct, the applicability of the BG entropy, and of the statistical mechanics to which it is associated, naturally has restrictions. The understanding of present developments of both the concept of entropy, and its corresponding statistical mechanics, demand some knowledge of the historical background.

Boltzmann was aware of the relevance of the range of the microscopic interactions between atoms and molecules. He wrote, in his 1896 *Lectures on Gas Theory* [41], the following words:

When the distance at which two gas molecules interact with each other noticeably is vanishingly small relative to the average distance between a molecule

and its nearest neighbor—or, as one can also say, when the space occupied by the molecules (or their spheres of action) is negligible compared to the space filled by the gas—then the fraction of the path of each molecule during which it is affected by its interaction with other molecules is vanishingly small compared to the fraction that is rectilinear, or simply determined by external forces. [. . .] The gas is “ideal” in all these cases.

Also Gibbs was aware. In his 1902 book [88], he wrote:

In treating of the canonical distribution, we shall always suppose the multiple integral in equation (92) [the partition function, as we call it nowadays] to have a finite value, as otherwise the coefficient of probability vanishes, and the law of distribution becomes illusory. This will exclude certain cases, but not such apparently, as will affect the value of our results with respect to their bearing on thermodynamics. It will exclude, for instance, cases in which the system or parts of it can be distributed in unlimited space [. . .]. It also excludes many cases in which the energy can decrease without limit, as when the system contains material points which attract one another inversely as the squares of their distances. [. . .]. For the purposes of a general discussion, it is sufficient to call attention to the assumption implicitly involved in the formula (92).

The extensivity/additivity of S_{BG} has been challenged, along the last century, by many physicists. Let us mention just a few. In his 1936 *Thermodynamics* [82], Enrico Fermi wrote:

The entropy of a system composed of several parts is very often equal to the sum of the entropies of all the parts. This is true if the energy of the system is the sum of the energies of all the parts and if the work performed by the system during a transformation is equal to the sum of the amounts of work performed by all the parts. Notice that these conditions are not quite obvious and that in some cases they may not be fulfilled. Thus, for example, in the case of a system composed of two homogeneous substances, it will be possible to express the energy as the sum of the energies of the two substances only if we can neglect the surface energy of the two substances where they are in contact. The surface energy can generally be neglected only if the two substances are not very finely subdivided; otherwise, it can play a considerable role.

Laszlo Tisza wrote, in his *Generalized Thermodynamics* [178]:

The situation is different for the additivity postulate P a2, the validity of which cannot be inferred from general principles. We have to require that the interaction energy between thermodynamic systems be negligible. This assumption is closely related to the homogeneity postulate P d1. From the molecular point of view, additivity and homogeneity can be expected to be reasonable approximations for systems containing many particles, provided that the intermolecular forces have a short range character.

Corroborating the above, virtually all textbooks of quantum mechanics contain the mechanical calculations corresponding to a particle in a square well, the harmonic oscillator, the rigid rotator, a spin 1/2 in the presence of a magnetic field, and the Hydrogen atom. In the textbooks of statistical mechanics we can find the thermostatical calculations of all these systems . . . excepting the Hydrogen atom! Why? Because the long-range electron-proton interaction produces an energy spectrum which leads to a divergent partition function. This is but a neat illustration of the above Gibbs’ alert.

A Remark on the Thermodynamics of Short- and Long-Range Interacting Systems

We consider here a simple d -dimensional classical fluid, constituted by many N point particles, governed by the Hamiltonian

$$\mathcal{H} = K + V = \sum_{i=1}^N \frac{p_i^2}{2m} + \sum_{i \neq j} V(r_{ij}), \quad (23)$$

where the potential $V(r)$ has, if it is attractive at short distances, no singularity at the origin, or an integrable singularity, and whose asymptotic behavior at infinity is given by $V(r) \sim -B/r^\alpha$ with $B > 0$ and $\alpha \geq 0$. One such example is the $d = 3$ Lennard–Jones fluid, for which $V(r) = A/r^{12} - B/r^6$ ($A > 0$), i. e., repulsive at short distances and attractive at long distances. In this case $\alpha = 6$. Another example could be Newtonian gravitation with a phenomenological short-distance cutoff (i. e., $V(r) \rightarrow \infty$ for $r \leq r_0$ with $r_0 > 0$). In this case, $\alpha = 1$. The full Γ -space of such a system has $2dN$ dimensions. The total potential energy is expected to scale (assuming a roughly homogeneous distribution of the particles) as

$$\frac{U_{\text{pot}}(N)}{N} \propto -B \int_1^\infty dr r^{d-1} r^{-\alpha}, \quad (24)$$

where the integral starts appreciably contributing above a typical cutoff, here taken to be unity. This integral is finite

$[= -B/(\alpha - d)]$ for $\alpha/d > 1$ (short-range interactions), and diverges for $0 \leq \alpha/d \leq 1$ (long-range interactions). In other words, the energy cannot be generically characterized by Eq. (24), and we must turn onto a different and more powerful estimation. Given the finiteness of the size of the system, an appropriate one is, in all cases, given by

$$\frac{U_{\text{pot}}(N)}{N} \propto -B \int_1^{N^{1/d}} dr r^{d-1} r^{-\alpha} = -\frac{B}{d} N^{\star}, \quad (25)$$

where

$$N^{\star} \equiv \frac{N^{1-\alpha/d} - 1}{1 - \alpha/d} \sim \begin{cases} \frac{1}{\alpha/d - 1} & \text{if } \alpha/d > 1; \\ \ln N & \text{if } \alpha/d = 1; \\ \frac{N^{1-\alpha/d}}{1 - \alpha/d} & \text{if } 0 < \alpha/d < 1. \end{cases} \quad (26)$$

Notice that $N^{\star} = \ln_{\alpha/d} N$ where the q -log function $\ln_q x \equiv (x^{1-q} - 1)/(1 - q)$ ($x > 0$; $\ln_1 x = \ln x$) will be shown to play an important role later on. Satisfactorily enough, Eqs. (26) recover the characterization with Eq. (24) in the limit $N \rightarrow \infty$, but they have the great advantage of providing, for finite N , a finite value. This fact will be now shown to enable to properly scale the macroscopic quantities in the thermodynamic limit ($N \rightarrow \infty$), for all values of $\alpha/d \geq 0$.

Let us address the thermodynamical consequences of the microscopic interactions being short- or long-ranged. To present a slightly more general illustration, we shall assume from now on that our homogeneous and isotropic classical fluid is made by magnetic particles. Its Gibbs free energy is then given by

$$G(N, T, p, H) = U(N, T, p, H) - TS(N, T, p, H) + pV(N, T, p, H) - HM(N, T, p, H), \quad (27)$$

where (T, p, H) correspond respectively to the temperature, pressure and external magnetic field, V is the volume and M the magnetization. If the interactions are short-ranged (i. e., if $\alpha/d > 1$), we can divide this equation by N and then take the $N \rightarrow \infty$ limit. We obtain

$$g(T, p, H) = u(T, p, H) - Ts(T, p, H) + pv(T, p, H) - Hm(T, p, H), \quad (28)$$

where $g(T, p, H) \equiv \lim_{N \rightarrow \infty} G(N, T, p, H)/N$, and analogously for the other variables of the equation. If the interactions were instead long-ranged (i. e., if $0 \leq \alpha/d \leq 1$), all these quantities would be divergent, hence thermodynamically nonsense. Consequently, the generically correct

procedure, i. e. $\forall \alpha/d \geq 0$, must conform to the following lines:

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{G(N, T, p, H)}{NN^{\star}} &= \lim_{N \rightarrow \infty} \frac{U(N, T, p, H)}{NN^{\star}} \\ &- \lim_{N \rightarrow \infty} \frac{T}{N^{\star}} \frac{S(N, T, p, H)}{N} \\ &+ \lim_{N \rightarrow \infty} \frac{p}{N^{\star}} \frac{V(N, T, p, H)}{N} \\ &- \lim_{N \rightarrow \infty} \frac{H}{N^{\star}} \frac{M(N, T, p, H)}{N} \end{aligned} \quad (29)$$

hence

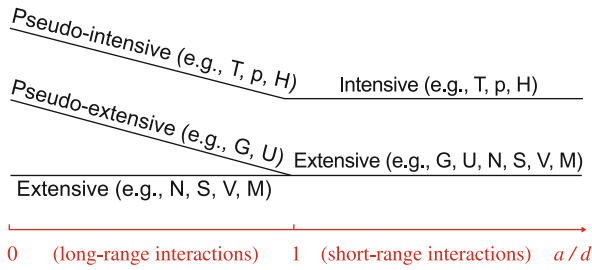
$$g(T^{\star}, p^{\star}, H^{\star}) = u(T^{\star}, p^{\star}, H^{\star}) - T^{\star} s(T^{\star}, p^{\star}, H^{\star}) + p^{\star} v(T^{\star}, p^{\star}, H^{\star}) - H^{\star} m(T^{\star}, p^{\star}, H^{\star}), \quad (30)$$

where the definitions of T^{\star} and all the other variables are self-explanatory (e. g., $T^{\star} \equiv T/N^{\star}$). In other words, in order to have finite thermodynamic equations of states, we must in general express them in the $(T^{\star}, p^{\star}, H^{\star})$ variables. If $\alpha/d > 1$, this procedure recovers the usual equations of states, and the usual extensive (G, U, S, V, M) and intensive (T, p, H) thermodynamic variables. But, if $0 \leq \alpha/d \leq 1$, the situation is more complex, and we realize that three, instead of the traditional two, classes of thermodynamic variables emerge. We may call them extensive (S, V, M, N), pseudo-extensive (G, U) and pseudo-intensive (T, p, H) variables. All the energy-type thermodynamical variables (G, F, U) give rise to pseudo-extensive ones, whereas those which appear in the usual Legendre thermodynamical pairs give rise to pseudo-intensive ones (T, p, H, μ) and extensive ones (S, V, M, N). See Figs. 1 and 2.

The possibly long-range interactions within Hamiltonian (23) refer to the dynamical variables themselves. There is another important class of Hamiltonians, where the possibly long-range interactions refer to the coupling constants between localized dynamical variables. Such is, for instance, the case of the following classical Hamiltonian:

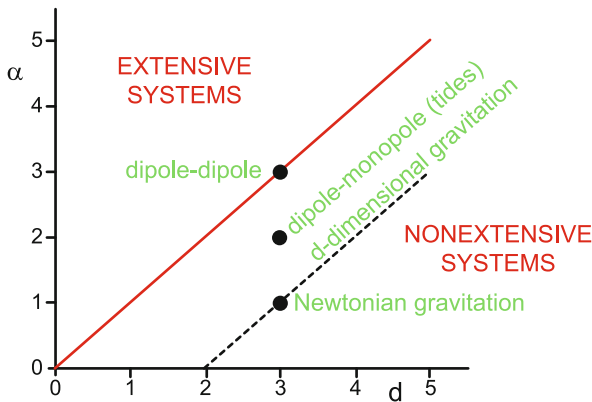
$$\begin{aligned} \mathcal{H} = K + V &= \sum_{i=1}^N \frac{L_i^2}{2I} \\ &- \sum_{i \neq j} \frac{J_x s_i^x s_j^x + J_y s_i^y s_j^y + J_z s_i^z s_j^z}{r_{ij}^{\alpha}} \quad (\alpha \geq 0), \end{aligned} \quad (31)$$

where $\{L_i\}$ are the angular momenta, I the moment of inertia, $\{(s_i^x, s_i^y, s_i^z)\}$ are the components of classical rotators, (J_x, J_y, J_z) are coupling constants, and r_{ij} runs



Entropy, Figure 1

For long-range interactions ($0 \leq \alpha/d \leq 1$) we have three classes of thermodynamic variables, namely the *pseudo-intensive* (scaling with N^*), *pseudo-extensive* (scaling with NN^*) and *extensive* (scaling with N) ones. For short range interactions ($\alpha/d > 1$) the pseudo-intensive variables become *intensive* (independent from N), and the pseudo-extensive merge with the extensive ones, all being now *extensive* (scaling with N), thus recovering the traditional two textbook classes of thermodynamical variables



Entropy, Figure 2

The so-called *extensive systems* ($\alpha/d > 1$ for the classical ones) typically involve *absolutely convergent series*, whereas the so-called *nonextensive systems* ($0 \leq \alpha/d < 1$ for the classical ones) typically involve *divergent series*. The marginal systems ($\alpha/d = 1$ here) typically involve *conditionally convergent series*, which therefore depend on the boundary conditions, i.e., typically on the external shape of the system. Capacitors constitute a notorious example of the $\alpha/d = 1$ case. The model usually referred to in the literature as the *Hamiltonian–Mean–Field* (HMF) one lies on the $\alpha = 0$ axis ($\forall d > 0$). The model usually referred to as the *d-dimensional α -XY model* [19] lies on the vertical axis at abscissa d ($\forall \alpha \geq 0$)

over all distances between sites i and j of a d -dimensional lattice. For example, for a simple hypercubic lattice with unit crystalline parameter we have $r_{ij} = 1, 2, 3, \dots$ if $d = 1$, $r_{ij} = 1, \sqrt{2}, 2, \dots$ if $d = 2$, $r_{ij} = 1, \sqrt{2}, \sqrt{3}, 2, \dots$ if $d = 3$, and so on. For such a case, we have that

$$N^* \equiv \sum_{i=2}^N r_{1i}^{-\alpha}, \quad (32)$$

which has in fact the same asymptotic behaviors as indicated in Eq. (26). In other words, here again $\alpha/d > 1$ corresponds to short-range interactions, and $0 \leq \alpha/d \leq 1$ corresponds to long-range ones.

The correctness of the present generalized thermodynamical scalings has already been specifically checked in many physical systems, such as a ferrofluid-like model [97], Lennard–Jones-like fluids [90], magnetic systems [16,19,59,158], anomalous diffusion [66], percolation [85,144].

Let us mention that, for the $\alpha = 0$ models (i.e., mean field models), it is largely spread in the literature to divide by N the potential term of the Hamiltonian in order to make it extensive *by force*. Although mathematically admissible (see [19]), this is obviously very unsatisfactory in principle since it implies a microscopic coupling constant which depends on N . What we have described here is the thermodynamically proper way of eliminating the mathematical difficulties emerging in the models in the presence of long-range interactions.

Last but not least, we verify a point which is crucial for the developments here below, namely that the entropy S is expected to be extensive *no matter the range of the interactions*.

The Nonadditive Entropy S_q

Introduction and Basic Properties

The possibility was introduced in 1988 [183] (see also [42,112,157,182]) to generalize the BG statistical mechanics on the basis of an entropy S_q which generalizes S_{BG} . This entropy is defined as follows:

$$S_q \equiv k \frac{1 - \sum_{i=1}^W p_i^q}{q - 1} \quad (q \in \mathbb{R}; S_1 = S_{BG}). \quad (33)$$

For equal probabilities, this entropy takes the form

$$S_q = k \ln_q W \quad (S_1 = k \ln W), \quad (34)$$

where the q -logarithmic function has already been defined.

Remark With the same or different prefactor, this entropic form has been successively and independently introduced in many occasions during the last decades. J. Havrda and F. Charvat [92] were apparently the first to ever introduce this form, though with a different prefactor (adapted to binary variables) in the context of cybernetics and information theory. I. Vajda [207], further studied this form, quoting Havrda and Charvat. Z. Daroczy [74] rediscovered this form (he quotes neither Havrda–Charvat

nor Vajda). J. Lindhard and V. Nielsen [108] rediscovered this form (they quote none of the predecessors) through the property of entropic composability. B.D. Sharma and D.P. Mittal [163] introduced a two-parameter form which reproduces both S_q and Renyi entropy [145] as particular cases. A. Wehrl [209] mentions the form of S_q in p. 247, quotes Daroczy, but ignores Havrda–Charvat, Vajda, Lindhard–Nielsen, and Sharma–Mittal. Myself I rediscovered this form in 1985 with the aim of generalizing Boltzmann–Gibbs statistical mechanics, but quote none of the predecessors in the 1988 paper [183]. In fact, I started knowing the whole story quite a few years later thanks to S.R.A. Salinas and R.N. Silver, who were the first to provide me with the corresponding informations. Such rediscoveries can by no means be considered as particularly surprising. Indeed, this happens in science more frequently than usually realized. This point is lengthily and colorfully developed by S.M. Stigler [167]. In p. 284, a most interesting example is described, namely that of the celebrated *normal distribution*. It was first introduced by Abraham De Moivre in 1733, then by Pierre Simon de Laplace in 1774, then by Robert Adrain in 1808, and finally by Carl Friedrich Gauss in 1809, nothing less than 76 years after its first publication! This distribution is universally called *Gaussian* because of the remarkable insights of Gauss concerning the theory of errors, applicable in all experimental sciences. A less glamorous illustration of the same phenomenon, but nevertheless interesting in the present context, is that of Renyi entropy [145]. According to I. Csiszar [64], p. 73, the Renyi entropy had already been essentially introduced by Paul-Marcel Schutzenberger [161].

The entropy defined in Eq. (33) has the following main properties:

- (i) S_q is nonnegative ($\forall q$);
- (ii) S_q is expansive ($\forall q > 0$);
- (iii) S_q attains its maximal (minimal) value $k \ln_q W$ for $q > 0$ (for $q < 0$);
- (iv) S_q is concave (convex) for $q > 0$ (for $q < 0$);
- (v) S_q is Lesche-stable ($\forall q > 0$) [2];
- (vi) S_q yields a finite upper bound of the entropy production per unit time for a special value of q , whenever the sensitivity to the initial conditions exhibits an upper bound which asymptotically increases as a *power* of time. For example, many $D = 1$ nonlinear dynamical systems have a vanishing maximal Lyapunov exponent λ_1 and exhibit a sensitivity to the initial conditions which is (upper) bounded by

$$\xi = e_q^{\lambda_q t}, \quad (35)$$

with $\lambda_q > 0$, $q < 1$, the q -exponential function e_q^x being the inverse of $\ln_q x$. More explicitly (see Fig. 3)

$$e_q^x \equiv \begin{cases} [1 + (1 - q)x]^{\frac{1}{1-q}} & \text{if } 1 + (1 - q)x > 0; \\ 0 & \text{otherwise.} \end{cases} \quad (36)$$

Such systems have a finite entropy production per unit time, which satisfies a q -generalized Pesin-like identity, namely, for the construction described in Sect. “Introduction”,

$$K_q \equiv \lim_{t \rightarrow \infty} \lim_{W \rightarrow \infty} \lim_{M \rightarrow \infty} \frac{S_q(t)}{t} = \lambda_q. \quad (37)$$

The situation is in fact sensibly much richer than briefly described here. For further details, see [27,28,29,30,93,116,117,146,147,148,149,150,151,152].

- (vii) S_q is nonadditive for $q \neq 1$. Indeed, for independent subsystems A and B , it can be straightforwardly proved

$$\frac{S_q(A + B)}{k} = \frac{S_q(A)}{k} + \frac{S_q(B)}{k} + (1 - q) \frac{S_q(A)}{k} \frac{S_q(B)}{k}, \quad (38)$$

or, equivalently,

$$S_q(A + B) = S_q(A) + S_q(B) + \frac{(1 - q)}{k} S_q(A) S_q(B), \quad (39)$$

which makes explicit that $(1 - q) \rightarrow 0$ plays the same role as $k \rightarrow \infty$. Property (38), occasionally referred to in the literature as *pseudo-additivity*, can be called *subadditivity* (*superadditivity*) for $q > 1$ ($q < 1$).

- (viii) $S_q = -k D_q \sum_{i=1}^W p_i^x|_{x=1}$, where the 1909 Jackson differential operator is defined as follows:

$$D_q f(x) \equiv \frac{f(qx) - f(x)}{qx - x} \quad (D_1 f(x) = df(x)/dx). \quad (40)$$

- (ix) An uniqueness theorem has been proved by Santos [159], which generalizes, for arbitrary q , that of Shannon [162].

Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:

- (a)

$$S(\{p_i\}) \quad \text{is a continuous function of } \{p_i\}; \quad (41)$$

(b)

$S(p_i = 1/W, \forall i)$ monotonically increases with the total number of possibilities W ;

(c)

$$\frac{S(A+B)}{k} = \frac{S(A)}{k} + \frac{S(B)}{k} + (1-q) \frac{S(A)}{k} \frac{S(B)}{k}$$

if $p_{ij}^{A+B} = p_i^A p_j^B \forall (i, j)$, with $k > 0$;

(d)

$$S(\{p_i\}) = S(p_L, p_M) + p_L^q S(\{p_i/p_L\}) + p_M^q S(\{p_i/p_M\})$$

with $p_L \equiv \sum_{L \text{ terms}} p_i, p_M \equiv \sum_{M \text{ terms}} p_i (L + M = W)$,

and $p_L + p_M = 1$.

(44)

Then and only then [159] $S(\{p_i\}) = S_q(\{p_i\})$.

(x) Another (equivalent) uniqueness theorem was proved by Abe [1], which generalizes, for arbitrary q , that of Khinchin [100].

Let us assume that an entropic form $S(\{p_i\})$ satisfies the following properties:

(a)

$S(\{p_i\})$ is a continuous function of $\{p_i\}$;

(b)

$S(p_i = 1/W, \forall i)$ monotonically increases with the total number of possibilities W ;

(46)

(c)

$$S(p_1, p_2, \dots, p_W, 0) = S(p_1, p_2, \dots, p_W);$$

(47)

(d)

$$\frac{S(A+B)}{k} = \frac{S(A)}{k} + \frac{S(B|A)}{k} + (1-q) \frac{S(A)}{k} \frac{S(B|A)}{k}$$

where $S(A+B) \equiv S(\{p_{ij}^{A+B}\})$,

$$S(A) \equiv S\left(\left\{\sum_{j=1}^{W_B} p_{ij}^{A+B}\right\}\right), \text{ and the conditional entropy}$$

$$S(B|A) \equiv \frac{\sum_{i=1}^{W_A} (p_i^A)^q S(\{p_{ij}^{A+B}/p_i^A\})}{\sum_{i=1}^{W_A} (p_i^A)^q} \quad (k > 0)$$

(48)

Then and only then [1] $S(\{p_i\}) = S_q(\{p_i\})$.

Additivity Versus Extensivity of the Entropy

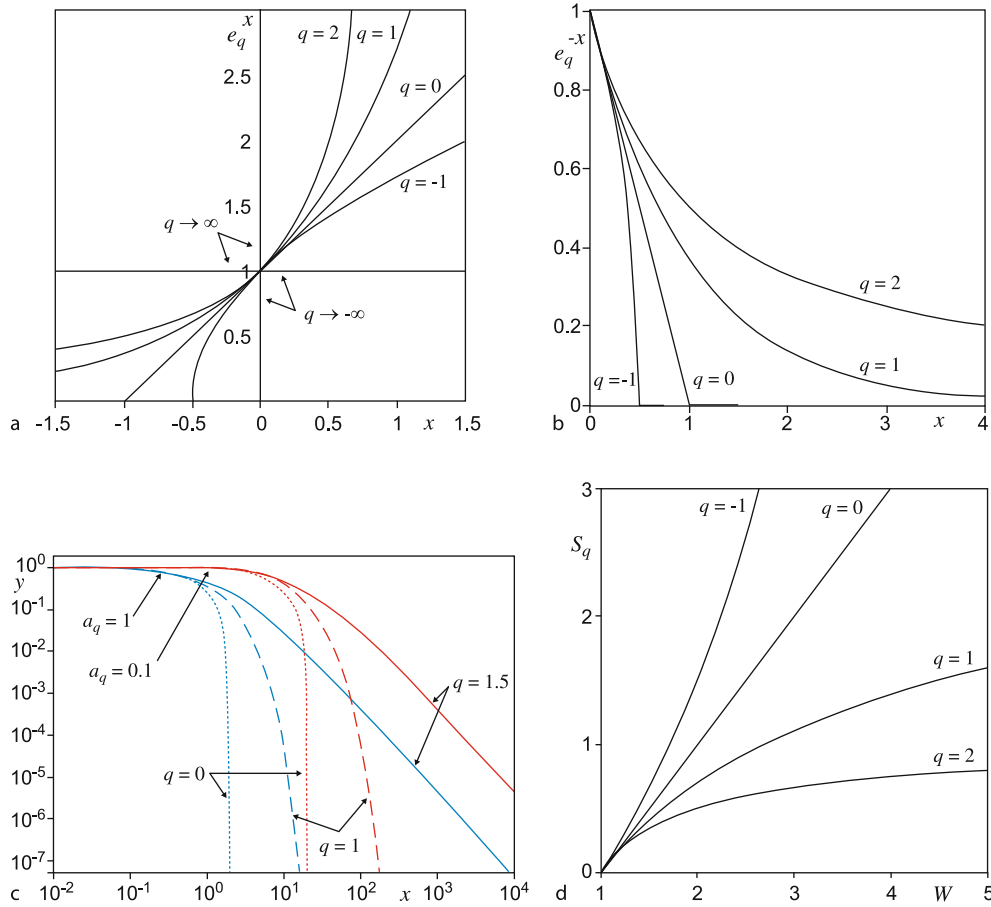
It is of great importance to distinguish *additivity* from *extensivity*. An entropy S is additive [130] if its value for a system composed by two independent subsystems A and B satisfies $S(A+B) = S(A) + S(B)$ (hence, for N independent equal subsystems or elements, we have $S(N) = NS(1)$). Therefore, S_{BG} is additive, and $S_q (q \neq 1)$ is nonadditive. A substantially different matter is whether a given entropy S is extensive for a given system. An entropy is extensive if and only if $0 < \lim_{N \rightarrow \infty} S(N)/N < \infty$. What matters for satisfactorily matching thermodynamics is extensivity *not* additivity. For systems whose elements are nearly independent (i.e., essentially weakly correlated), S_{BG} is extensive and S_q is nonextensive. For systems whose elements are strongly correlated in a special manner, S_{BG} is nonextensive, whereas S_q is extensive for a special value of $q \neq 1$ (and nonextensive for all the others).

Let us illustrate these facts for some simple examples of equal probabilities. If $W(N) \sim A\mu^N (A > 0, \mu > 1, \text{ and } N \rightarrow \infty)$, the entropy which is extensive is S_{BG} . Indeed, $S_{BG}(N) = k \ln W(N) \sim (\ln \mu)N \propto N$ (it is equally trivial to verify that $S_q(N)$ is nonextensive for any $q \neq 1$). If $W(N) \sim BN^\rho (B > 0, \rho > 0, \text{ and } N \rightarrow \infty)$, the entropy which is extensive is $S_{1-(1/\rho)}$. Indeed, $S_{1-(1/\rho)}(N) \sim k\rho B^{1/\rho} N \propto N$ (it is equally trivial to verify that $S_{BG}(N) \propto \ln N$, hence nonextensive). If $W(N) \sim C\mu^{N^\gamma} (C > 0, \mu > 1, \gamma \neq 1, \text{ and } N \rightarrow \infty)$, then $S_q(N)$ is nonextensive for any value of q . Therefore, in such a complex case, one must in principle refer to some other kind of entropic functional in order to match the extensivity required by classical thermodynamics.

Various nontrivial abstract mathematical models can be found in [113,160,186,198,199] for which $S_q (q \neq 1)$ is extensive. Moreover, a physical realization is also available now [60,61] for a many-body quantum Hamiltonian, namely the ground state of the following one:

$$\mathcal{H} = - \sum_{i=1}^{N-1} [(1+\gamma)S_i^x S_{i+1}^x + (1-\gamma)S_i^y S_{i+1}^y] - 2\lambda \sum_{i=1}^N S_i^z, \quad (49)$$

where λ is a transverse magnetic field, and (S_i^x, S_i^y, S_i^z) are Pauli matrices; for $|\gamma| = 1$ we have the Ising model, for $0 < |\gamma| < 1$, we have the anisotropic XY model, and, for $\gamma = 0$, we have the isotropic XY model. The two former share the same symmetry and consequently belong to the same critical universality class (the Ising uni-



Entropy, Figure 3

The q -exponential and q -logarithm functions in typical representations: **a** Linear-linear representation of e_q^x ; **b** Linear-linear representation of e_q^{-x} ; **c** Log-log representation of $y(x) = e_q^{-a_q x}$, solution of $dy/dx = -a_q y^q$ with $y(0) = 1$; **d** Linear-linear representation of $S_q = \ln_q W$ (value of the entropy for equal probabilities)

versality class, which corresponds to a so-called *central charge* $c = 1/2$), whereas the latter one belongs to a different universality class (the XX one, which corresponds to a central charge $c = 1$). At temperature $T = 0$ and $N \rightarrow \infty$, this model exhibits a second-order phase transition as a function of λ . For the Ising model, the critical value is $\lambda = 1$, whereas, for the XX model, the entire line $0 \leq \lambda \leq 1$ is critical. Since the system is at its ground state (assuming a vanishingly small magnetic field component in the $x - y$ plane), it is a *pure state* (i. e., its density matrix ρ_N is such that $\text{Tr} \rho_N^2 = 1, \forall N$), hence the entropy $S_q(N) (\forall q > 0)$ is strictly zero. However, the situation is drastically different for any L -sized block of the infinite chain. Indeed, $\rho_L \equiv \text{Tr}_{N-L} \rho_N$ is such that $\text{Tr} \rho_L^2 < 1$, i. e., it is a *mixed state*, hence it has a nonzero entropy. The block entropy $S_q(L) \equiv \lim_{N \rightarrow \infty} S_q(N, L)$ monotonically

increases with L for all values of q . And it does so *linearly* for

$$q = \frac{\sqrt{9 + c^2} - 3}{c}, \quad (50)$$

where c is the central charge which emerges in quantum field theory [54]. In other words, $0 < \lim_{L \rightarrow \infty} S_{(\sqrt{9+c^2}-3)/c}(L)/L < \infty$. Notice that q increases from zero to unity when c increases from zero to infinity; $q = \sqrt{37} - 6 \simeq 0.083$ for $c = 1/2$ (Ising model), $q = \sqrt{10} - 3 \simeq 0.16$ for $c = 1$ (isotropic XY model), $q = 1/2$ for $c = 4$ (dimension of space-time), and $q = (\sqrt{685} - 3)/26 \simeq 0.89$ for $c = 26$, related to string theory [89]. The possible physical interpretation of the limit $c \rightarrow \infty$ is still unknown, although it could correspond to some sort of mean field approach.

Nonextensive Statistical Mechanics

To generalize BG statistical mechanics for the canonical ensemble, we optimize S_q with constraint (15) and also

$$\sum_{i=1}^W P_i E_i = U_q, \quad (51)$$

where

$$P_i \equiv \frac{p_i^q}{\sum_{j=1}^W p_j^q} \quad \left(\sum_{i=1}^W P_i = 1 \right) \quad (52)$$

is the so-called *escort distribution* [33]. It follows that $p_i = P_i^{1/q} / \sum_{j=1}^W P_j^{1/q}$. There are various converging reasons for being appropriate to impose the energy constraint with the $\{P_i\}$ instead of with the original $\{p_i\}$. The full discussion of this delicate point is beyond the present scope. However, some of these intertwined reasons are explored in [184]. By imposing Eq. (51), we follow [193], which in turn reformulates the results presented in [71,183]. The passage from one to the other of the various existing formulations of the above optimization problem are discussed in detail in [83,193].

The entropy optimization yields, for the stationary state,

$$p_i = \frac{e_q^{-\beta_q(E_i - U_q)}}{\bar{Z}_q}, \quad (53)$$

with

$$\beta_q \equiv \frac{\beta}{\sum_{j=1}^W p_j^q}, \quad (54)$$

and

$$\bar{Z}_q \equiv \sum_i e_q^{-\beta_q(E_i - U_q)}, \quad (55)$$

β being the Lagrange parameter associated with the constraint (51). Equation (53) makes explicit that the probability distribution is, for fixed β_q , invariant with regard to the arbitrary choice of the zero of energies. The stationary state (or (meta)equilibrium) distribution (53) can be rewritten as follows:

$$p_i = \frac{e_q^{-\beta'_q E_i}}{Z'_q}, \quad (56)$$

with

$$Z'_q \equiv \sum_{j=1}^W e_q^{-\beta'_q E_j}, \quad (57)$$

and

$$\beta'_q \equiv \frac{\beta_q}{1 + (1-q)\beta_q U_q}. \quad (58)$$

The form (56) is particularly convenient for many applications where comparison with experimental or computational data is involved. Also, it makes clear that p_i asymptotically decays like $1/E_i^{1/(q-1)}$ for $q > 1$, and has a cut-off for $q < 1$, instead of the exponential decay with E_i for $q = 1$.

The connection to thermodynamics is established in what follows. It can be proved that

$$\frac{1}{T} = \frac{\partial S_q}{\partial U_q}, \quad (59)$$

with $T \equiv 1/(k\beta)$. Also we prove, for the free energy,

$$F_q \equiv U_q - TS_q = -\frac{1}{\beta} \ln_q Z_q, \quad (60)$$

where

$$\ln_q Z_q = \ln_q \bar{Z}_q - \beta U_q. \quad (61)$$

This relation takes into account the trivial fact that, in contrast with what is usually done in BG statistics, the energies $\{E_i\}$ are here referred to U_q in (53). It can also be proved

$$U_q = -\frac{\partial}{\partial \beta} \ln_q Z_q, \quad (62)$$

as well as relations such as

$$C_q \equiv T \frac{\partial S_q}{\partial T} = \frac{\partial U_q}{\partial T} = -T \frac{\partial^2 F_q}{\partial T^2}. \quad (63)$$

In fact the entire Legendre transformation structure of thermodynamics is q -invariant, which is both remarkable and welcome.

A Connection Between Entropy and Diffusion

We review here one of the main common aspects of entropy and diffusion. We shall present on equal footing both the BG and the nonextensive cases [13,138,192,216]. Let us extremize the entropy

$$S_q = k \frac{1 - \int_{-\infty}^{\infty} d(x/\sigma) [\sigma p(x)]^q}{q-1} \quad (64)$$

with the constraints

$$\int_{-\infty}^{\infty} dx p(x) = 1 \quad (65)$$

and

$$\langle x^2 \rangle_q \equiv \frac{\int_{-\infty}^{\infty} dx x^2 [p(x)]^q}{\int_{-\infty}^{\infty} dx [p(x)]^q} = \sigma^2, \quad (66)$$

$\sigma > 0$ being some fixed value having the same physical dimensions of the variable x . We straightforwardly obtain the following distribution:

$$p_q(x) = \begin{cases} \frac{1}{\sigma} \left[\frac{q-1}{\pi(3-q)} \right]^{1/2} \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2(q-1)}\right)} \frac{1}{\left[1 + \frac{q-1}{3-q} \frac{x^2}{\sigma^2}\right]^{1/(q-1)}} & \text{if } 1 < q < 3; \\ \frac{1}{\sigma} \frac{1}{\sqrt{2\pi}} e^{-x^2/2\sigma^2} & \text{if } q = 1; \\ \frac{1}{\sigma} \left[\frac{1-q}{\pi(3-q)} \right]^{1/2} \frac{\Gamma\left(\frac{5-3q}{2(1-q)}\right)}{\Gamma\left(\frac{2-q}{1-q}\right)} \left[1 - \frac{1-q}{3-q} \frac{x^2}{\sigma^2}\right]^{1/(1-q)} & \text{for } |x| < \sigma [(3-q)/(1-q)]^{1/2}, \text{ and zero otherwise,} \\ & \text{if } q < 1. \end{cases} \quad (67)$$

These distributions are frequently referred to as q -Gaussians. For $q > 1$, they asymptotically have a power-law tail ($q \geq 3$ is not admissible because the norm (65) cannot be satisfied); for $q < 1$, they have a compact support. For $q = 1$, the celebrated Gaussian is recovered; for $q = 2$, the Cauchy–Lorentz distribution is recovered; finally, for $q \rightarrow -\infty$, the uniform distribution within the interval $[-1, 1]$ is recovered. For $q = \frac{3+m}{1+m}$, m being an integer ($m = 1, 2, 3, \dots$), we recover the Student's t -distributions with m degrees of freedom [79]. For $q = \frac{n-4}{n-2}$, n being an integer ($n = 3, 4, 5, \dots$), we recover the so-called r -distributions with n degrees of freedom [79]. In other words, q -Gaussians are analytical extensions of Student's t - and r -distributions. In some communities they are also referred to as the Barenblatt form. For $q < 5/3$, they have a finite variance which monotonically increases for q varying from $-\infty$ to $5/3$; for $5/3 \leq q < 3$, the variance diverges.

Let us now make a connection of the above optimization problem with diffusion. We focus on the following quite general diffusion equation:

$$\frac{\partial^\delta p(x, t)}{\partial t^\delta} = \frac{\partial}{\partial x} \left[\frac{\partial U(x)}{\partial x} p(x, t) \right] + D \frac{\partial^\alpha [p(x, t)]^{2-q}}{\partial |x|^\alpha} \quad (0 < \delta \leq 1; 0 < \alpha \leq 2; q < 3; t \geq 0), \quad (68)$$

with a generic nonsingular potential $U(x)$, and a generalized diffusion coefficient D which is positive (negative) for $q < 2$ ($2 < q < 3$). Several particular instances of this equation have been discussed in the literature (see [40,86,106,131,188] and references therein).

For example, the stationary state for $\alpha = 2$, $\forall \delta$, and any confining potential (i.e., $\lim_{|x| \rightarrow \infty} U(x) = \infty$) is given by [106]

$$p(x, \infty)_q = \frac{e_q^{-\beta [U(x)-U(0)]}}{Z}, \quad (69)$$

$$Z \equiv \int_{-\infty}^{\infty} dx e_q^{-\beta [U(x)-U(0)]}, \quad (70)$$

$$1/\beta \equiv kT \propto |D|, \quad (71)$$

which precisely is the distribution obtained within nonextensive statistical mechanics through extremization of S_q .

Also, the solution for $\alpha = 2$, $\delta = 1$, $U(x) = -k_1 x + \frac{k_2}{2} x^2$ ($\forall k_1$, and $k_2 \geq 0$), and $p(x, 0) = \delta(x)$ is given by [188]

$$p_q(x, t) = \frac{e_q^{-\beta(t)[x-x_M(t)]^2}}{Z_q(t)}, \quad (72)$$

$$\frac{\beta(t)}{\beta(0)} = \left[\frac{Z_q(0)}{Z_q(t)} \right]^2 = \left[\left(1 - \frac{1}{K_2} e^{-t/\tau} \right) + \frac{1}{K_2} \right]^{-2/(3-q)}, \quad (73)$$

$$K_2 \equiv \frac{k_2}{2(2-q)D\beta(0)[Z_q(0)]^{q-1}}, \quad (74)$$

$$\tau \equiv \frac{1}{k_2(3-q)}, \quad (75)$$

$$x_M(t) \equiv \frac{k_1}{k_2} + \left[x_M(0) - \frac{k_1}{k_2} \right] e^{-k_2 t}. \quad (76)$$

In the limit $k_2 \rightarrow 0$, Eq. (73) becomes

$$Z_q(t) = \{ [Z_q(0)]^{3-q} + 2(2-q)(3-q)D\beta(0) [Z_q(0)]^2 t \}^{1/(3-q)}, \quad (77)$$

which, in the $t \rightarrow \infty$ limit, yields

$$\frac{1}{\beta(t)} \propto [Z_q(t)]^2 \propto t^{2/(3-q)}. \quad (78)$$

In other words, x^2 scales like t^γ , with

$$\gamma = \frac{2}{3-q}, \quad (79)$$

hence, for $q > 1$ we have $\gamma > 1$ (i.e., *superdiffusion*; in particular, $q = 2$ yields $\gamma = 2$, i.e., ballistic diffusion),

for $q < 1$ we have $\gamma < 1$ (i.e., *subdiffusion*; in particular, $q \rightarrow -\infty$ yields $\gamma = 0$, i.e., *localization*), and naturally, for $q = 1$, we obtain *normal diffusion*. Four systems are known for which results have been found that are consistent with prediction (79). These are the motion of *Hydra viridissima* [206], defect turbulence [73], simulation of a silo drainage [22], and molecular dynamics of a many-body long-range-interacting classical system of rotators (α -XY model) [143]. For the first three, it has been found $(q, \gamma) \simeq (3/2, 4/3)$. For the latter one, relation (79) has been verified for various situations corresponding to $\gamma > 1$.

Finally, for the particular case $\delta = 1$ and $U(x) = 0$, Eq. (68) becomes

$$\frac{\partial p(x, t)}{\partial t} = D \frac{\partial^\alpha [p(x, t)]^{2-q}}{\partial |x|^\alpha} \quad (0 < \alpha \leq 2; q < 3). \quad (80)$$

The diffusion constant D just rescales time t . Only two parameters are therefore left, namely α and q .

The linear case (i.e., $q = 1$) has two types of solutions: Gaussians for $\alpha = 2$, and Lévy- (or α -stable) distributions for $0 < \alpha < 2$. The case $\alpha = 2$ corresponds to the *Central Limit Theorem*, where the $N \rightarrow \infty$ attractor of the sums of N independent random variables with finite variance precisely is a Gaussian. The case $0 < \alpha < 2$ corresponds to the sometimes called *Levy-Gnedenko Central Limit Theorem*, where the $N \rightarrow \infty$ attractor of the sums of N independent random variables with infinite variance (and appropriate asymptotics) precisely is a Lévy distribution with index α .

The nonlinear case (i.e., $q \neq 1$) has solutions that are q -Gaussians for $\alpha = 2$, and one might conjecture that, similarly, interesting solutions exist for $0 < \alpha < 2$. Furthermore, in analogy with the $q = 1$ case, one expects corresponding q -generalized Central Limit Theorems to exist [187]. This is precisely what we present in the next Section.

Standard and q -Generalized Central Limit Theorems

The q -Product

It has been recently introduced (independently and virtually simultaneously) [43,125] a generalization of the product, which is called q -product. It is defined, for $x \geq 0$ and $y \geq 0$, as follows:

$$x \otimes_q y \equiv \begin{cases} [x^{1-q} + y^{1-q} - 1]^{1/(1-q)} & \text{if } x^{1-q} + y^{1-q} > 1; \\ 0 & \text{otherwise.} \end{cases} \quad (81)$$

It has, among others, the following properties:

it recovers the *standard product* as a particular instance, i.e.,

$$x \otimes_1 y = xy; \quad (82)$$

it is *commutative*, i.e.,

$$x \otimes_q y = y \otimes_q x; \quad (83)$$

it is *additive under q -logarithm*, i.e.,

$$\ln_q(x \otimes_q y) = \ln_q x + \ln_q y \quad (84)$$

(whereas we remind that $\ln_q(xy) = \ln_q x + \ln_q y + (1 - q)(\ln_q x)(\ln_q y)$;

it has a $(2 - q)$ -duality/inverse property, i.e.,

$$1/(x \otimes_q y) = (1/x) \otimes_{2-q} (1/y); \quad (85)$$

it is *associative*, i.e.,

$$x \otimes_q (y \otimes_q z) = (x \otimes_q y) \otimes_q z = x \otimes_q y \otimes_q z \\ = (x^{1-q} + y^{1-q} + z^{1-q} - 2)^{1/(1-q)}; \quad (86)$$

it admits *unity*, i.e.,

$$x \otimes_q 1 = x. \quad (87)$$

and, for $q \geq 1$, also a *zero*, i.e.,

$$x \otimes_q 0 = 0 \quad (q \geq 1). \quad (88)$$

The q -Fourier Transform

We shall introduce the q -Fourier transform of a quite generic function $f(x)$ ($x \in \mathcal{R}$) as follows [140,189,202,203,204,205]:

$$F_q[f](\xi) \equiv \int_{-\infty}^{\infty} dx e_q^{ix\xi} \otimes_q f(x) \\ = \int_{-\infty}^{\infty} dx e_q^{ix\xi[f(x)]^{q-1}} f(x), \quad (89)$$

where we have primarily focused on the case $q \geq 1$. In contrast with the $q = 1$ case (standard Fourier transform), this integral transformation is *nonlinear* for $q \neq 1$. It has a remarkable property, namely that the q -Fourier transform of a q -Gaussian is another q -Gaussian:

$$F_q \left[N_q \sqrt{\beta} e_q^{-\beta x^2} \right](\xi) = e_q^{-\beta_1 \xi^2}, \quad (90)$$

with

$$N_q \equiv \begin{cases} \left[\frac{q-1}{\pi} \right]^{1/2} \frac{\Gamma\left(\frac{1}{q-1}\right)}{\Gamma\left(\frac{3-q}{2(q-1)}\right)} & \text{if } 1 < q < 3, \\ \frac{1}{\sqrt{\pi}} & \text{if } q = 1, \\ \frac{3-q}{2} \left[\frac{1-q}{\pi} \right]^{1/2} \frac{\Gamma\left(\frac{3-q}{2(1-q)}\right)}{\Gamma\left(\frac{1}{1-q}\right)} & \text{if } q < 1, \end{cases} \quad (91)$$

and

$$q_1 = z(q) \equiv \frac{1+q}{3-q}, \quad (92)$$

$$\beta_1 = \frac{1}{\beta^{2-q}} \frac{N_q^{2(1-q)}(3-q)}{8}. \quad (93)$$

Equation (93) can be re-written as $\beta^{\sqrt{2-q}} \beta_1^{1/\sqrt{2-q}} = [(N_q^{2(1-q)}(3-q))/8]^{1/\sqrt{2-q}} \equiv K(q)$, which, for $q = 1$, recovers the well known Heisenberg-uncertainty-principle-like relation $\beta\beta_1 = 1/4$.

If we iterate n times the relation $z(q)$ in Eq. (92), we obtain the following algebra:

$$q_n(q) = \frac{2q + n(1-q)}{2 + n(1-q)} \quad (n = 0, \pm 1, \pm 2, \dots), \quad (94)$$

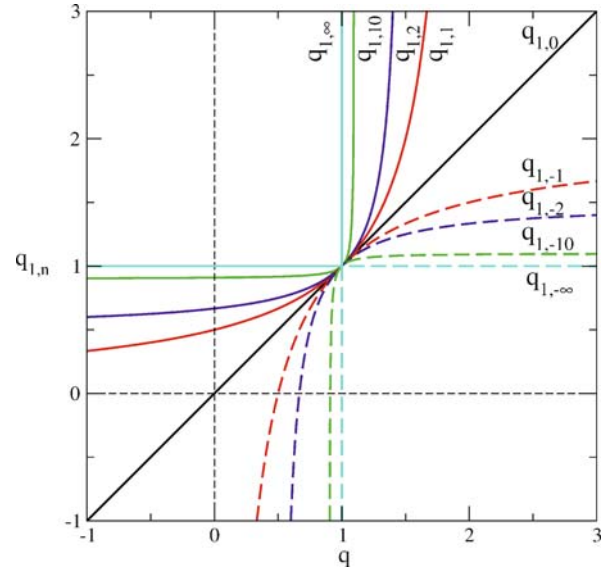
which can be conveniently re-written as

$$\frac{2}{1-q_n(q)} = \frac{2}{1-q} + n \quad (n = 0, \pm 1, \pm 2, \dots). \quad (95)$$

(See Fig. 4). We easily verify that $q_n(1) = 1 (\forall n)$, $q_{\pm\infty}(q) = 1 (\forall q)$, as well as

$$\frac{1}{q_{n+1}} = 2 - q_{n-1}. \quad (96)$$

This relation connects the so called *additive duality* $q \rightarrow (2-q)$ and *multiplicative duality* $q \rightarrow 1/q$, frequently emerging in all types of calculations in the literature. Moreover, we see from Eq. (95) that multiple values of q are expected to emerge in connection with diverse properties of nonextensive systems, i.e., in systems whose basic entropy is the nonadditive one S_q . Such is the case of the so called *q-triplet* [185], observed for the first time in the magnetic field fluctuations of the solar wind, as it has been revealed by the analysis of the data sent to NASA by the spacecraft Voyager 1 [48].



Entropy, Figure 4

The q -dependence of $q_n(q) \equiv q_{2,n}(q)$

q -Independent Random Variables

Two random variables X [with density $f_X(x)$] and Y [with density $f_Y(y)$] having zero q -mean values (e.g., if $f_X(x)$ and $f_Y(y)$ are even functions) are said *q-independent*, with q_1 given by Eq. (92), if

$$F_q[X + Y](\xi) = F_q[X](\xi) \otimes_{q_1} F_q[Y](\xi), \quad (97)$$

i.e., if

$$\int_{-\infty}^{\infty} dz e^{iz\xi} \otimes_q f_{X+Y}(z) = \left[\int_{-\infty}^{\infty} dx e^{ix\xi} \otimes_q f_X(x) \right]_{(1+q)/(3-q)} \left[\int_{-\infty}^{\infty} dy e^{iy\xi} \otimes_q f_Y(y) \right], \quad (98)$$

with

$$\begin{aligned} f_{X+Y}(z) &= \int_{-\infty}^{\infty} dx \int_{-\infty}^{\infty} dy h(x, y) \delta(x + y - z) \\ &= \int_{-\infty}^{\infty} dx h(x, z - x) \\ &= \int_{-\infty}^{\infty} dy h(z - y, y) \end{aligned} \quad (99)$$

where $h(x, y)$ is the joint density.

Clearly, q -independence means *independence* for $q = 1$ (i.e., $h(x, y) = f_X(x)f_Y(y)$), and implies a special correlation for $q \neq 1$. Although the full understanding of this correlation is still under progress, q -independence appears to be consistent with scale-invariance.

Entropy, Table 1

The attractors corresponding to the four basic cases, where the N variables that are being summed are q -independent (i. e., globally correlated) with $q_1 = (1 + q)/(3 - q)$; $\sigma_Q \equiv (\int_{-\infty}^{\infty} dx x^2 [f(x)]^Q) / (\int_{-\infty}^{\infty} dx [f(x)]^Q)$ with $Q \equiv 2q - 1$. The attractor for $(q, \alpha) = (1, 2)$ is a Gaussian $G(x) \equiv L_{1,2}$ (standard Central Limit Theorem); for $q = 1$ and $0 < \alpha < 2$, it is a Lévy distribution $L_\alpha \equiv L_{1,\alpha}$ (the so called Lévy-Gnedenko limit theorem); for $\alpha = 2$ and $q \neq 1$, it is a q -Gaussian $G_q \equiv L_{q,2}$ (the q -Central Limit Theorem; [203]); finally, for $q \neq 1$ and $0 < \alpha < 2$, it is a generic (q, α) -stable distribution $L_{q,\alpha}$ ([204,205]). See [140,189] for typical illustrations of the four types of attractors. The distribution $L_\alpha(x)$ remains, for $1 < \alpha < 2$, close to a Gaussian for $|x|$ up to about $x_c(1, \alpha)$, where it makes a crossover to a power-law. The distribution $G_q(x)$ remains, for $q > 1$, close to a Gaussian for $|x|$ up to about $x_c(q, 2)$, where it makes a crossover to a power-law. The distribution $L_{q,\alpha}(x)$ remains, for $q > 1$ and $\alpha < 2$, close to a Gaussian for $|x|$ up to about $x_c^{(1)}(q, \alpha)$, where it makes a crossover to a power-law (*intermediate regime*), which lasts further up to about $x_c^{(2)}(q, \alpha)$, where it makes a second crossover to another power-law (*distant regime*)

	$q = 1$ [independent]	$q \neq 1$ (i. e., $Q \neq 1$) [globally correlated]
$\sigma_Q < \infty$ ($\alpha = 2$)	$G(x)$ [with same σ_1 of $f(x)$]	$G_q(x) = G_{(3q_1-1)/(1+q_1)}(x)$ [with same σ_Q of $f(x)$] $G_q(x) \sim G(x)$ if $ x \ll x_c(q, 2)$ $G_q(x) \sim C_{q,2}/ x ^{2/(q-1)}$ if $ x \gg x_c(q, 2)$ for $q > 1$, with $\lim_{q \rightarrow 1} x_c(q, 2) = \infty$
$\sigma_Q \rightarrow \infty$ ($\alpha < 2$)	$L_\alpha(x)$ [with same $ x \rightarrow \infty$ behavior of $f(x)$] $L_\alpha(x) \sim G(x)$ if $ x \ll x_c(1, \alpha)$ $L_\alpha(x) \sim C_{1,\alpha}/ x ^{1+\alpha}$ if $ x \gg x_c(1, \alpha)$ with $\lim_{\alpha \rightarrow 2} q_c(1, \alpha) = \infty$	$L_{q,\alpha}(x)$ [with same $ x \rightarrow \infty$ behavior of $f(x)$] $L_{q,\alpha} \sim C_{q,\alpha}^{(\text{intermediate})}/ x ^{\frac{2(1-q)-\alpha(3-q)}{2(q-1)}}$ if $x_c^{(1)}(q, \alpha) \ll x \ll x_c^{(2)}(q, \alpha)$ $L_{q,\alpha} \sim C_{q,\alpha}^{(\text{distant})}/ x ^{\frac{1+\alpha}{1+\alpha(q-1)}}$ if $ x \gg x_c^{(2)}(q, \alpha)$

q -Generalized Central Limit Theorems

It is out of the scope of the present survey to provide the details of the complex proofs of the q -generalized central limit theorems. We shall restrict to the presentation of their structure. Let us start by introducing a notation which is important for what follows. A distribution is said (q, α) -stable distribution $L_{q,\alpha}(x)$ if its q -Fourier transform $L_{q,\alpha}(\xi)$ is of the form

$$L_{q,\alpha}(\xi) = a e^{-b|\xi|^\alpha} \quad [a > 0, b > 0, 0 < \alpha \leq 2, q_1 = (q + 1)/(3 - q)]. \quad (100)$$

Consistently, $L_{1,2}$ are Gaussians, $L_{1,\alpha}$ are Lévy distributions, and $L_{q,2}$ are q -Gaussians.

We are seeking for the $N \rightarrow \infty$ attractor associated with the sum of N identical and distinguishable random variables each of them associated with one and the same arbitrary symmetric distribution $f(x)$. The random variables are independent for $q = 1$, and correlated in a special manner for $q \neq 1$. To obtain the $N \rightarrow \infty$ invariant distribution, i. e. the attractor, the sum must be rescaled, i. e., divided by N^δ , where

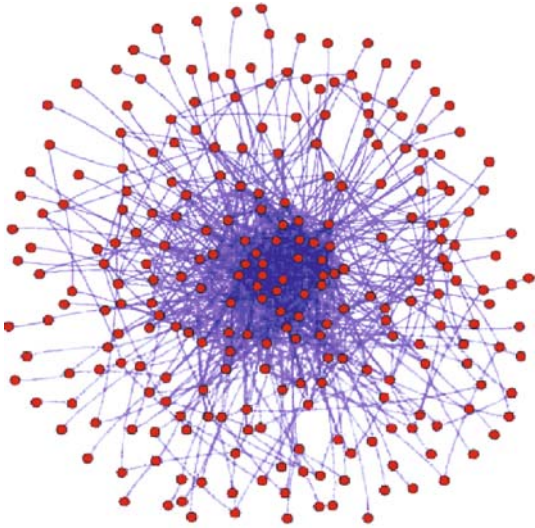
$$\delta = \frac{1}{\alpha(2 - q)}. \quad (101)$$

For $(\alpha, q) = (2, 1)$, we recover the traditional $1/\sqrt{N}$ rescaling of Brownian motion. At the present stage, the theorems have been established for $q \geq 1$ and are summarized in Table 1. The case $q < 1$ is still open at the time at which these lines are being written. Two $q < 1$ cases have been preliminarily explored numerically in [124] and in [171]. The numerics seemed to indicate that the $N \rightarrow \infty$ limits would be q -Gaussians for both models. However, it has been analytically shown [94] that it is not exactly so. The limiting distributions numerically are amazingly close to q -Gaussians, but they are in fact different. Very recently, another simple scale-invariant model has been introduced [153], whose attractor has been analytically shown to be a q -Gaussian.

These $q \neq 1$ theorems play for the nonadditive entropy S_q and nonextensive statistical mechanics the same grounding role that the well known $q = 1$ theorems play for the additive entropy S_{BG} and BG statistical mechanics. In particular, interestingly enough, the ubiquity of Gaussians and of q -Gaussians in natural, artificial and social systems may be understood on equal footing.

Future Directions

The concept of entropy permeates into virtually all quantitative sciences. The future directions could therefore be very varied. If we restrict, however, to the evidence



Entropy, Figure 5

Snapshot of a nongrowing dynamic network with $N = 256$ nodes (see details in [172], by courtesy of the author)

presently available, the main lines along which evolution occurs are:

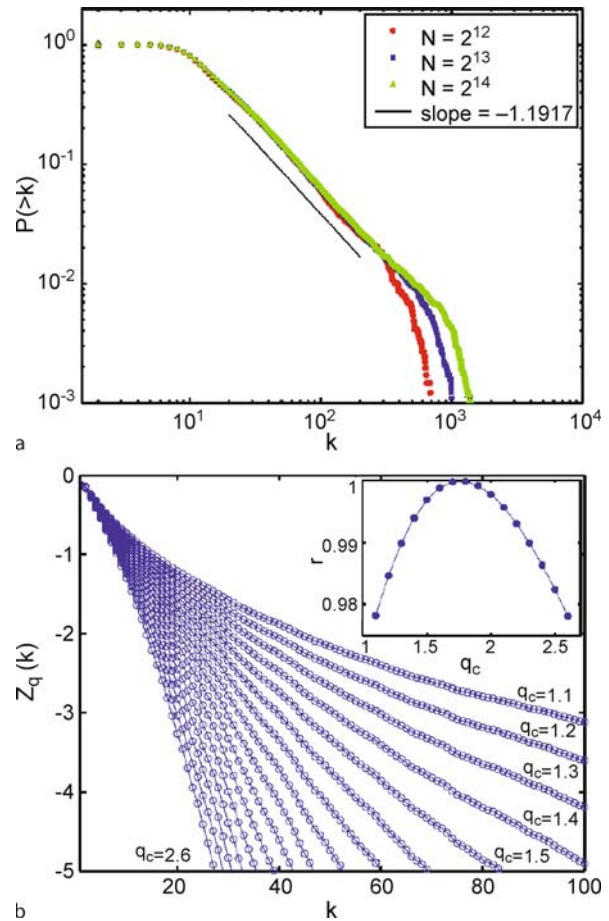
Networks Many of the so-called scale-free networks, among others, systematically exhibit a *degree distribution* $p(k)$ (probability of a node having k links) which is of the form

$$p(k) \propto \frac{1}{(k_0 + k)^\gamma} \quad (\gamma > 0; k_0 > 0), \quad (102)$$

or, equivalently,

$$p(k) \propto e_q^{-k/\kappa} \quad (q \geq 1; \kappa > 0), \quad (103)$$

with $\gamma = 1/(q-1)$ and $k_0 = \kappa/(q-1)$ (see Figs. 5 and 6). This is not surprising since, if we associate to each link an “energy” (or *cost*) and to each node half of the “energy” carried by its links (the other half being associated with the other nodes to which any specific node is linked), the distribution of energies optimizing S_q precisely coincides with the degree distribution. If, for any reason, we consider k as the modulus of a d -dimensional vector \mathbf{k} , the optimization of the functional $S_q[p(\mathbf{k})]$ may lead to $p(k) \propto k^\eta e_q^{-k/\kappa}$, where k^η plays the role of a density of states, $\eta(d)$ being either zero (which reproduces Eq. (103)) or positive or negative. Several examples [12,39,76,91,165,172,173,212,213] already exist in the literature; in particular, the Barabasi–Albert universality class $\gamma = 3$ corresponds to $q = 4/3$. A deeper understanding of this connection might enable the sys-

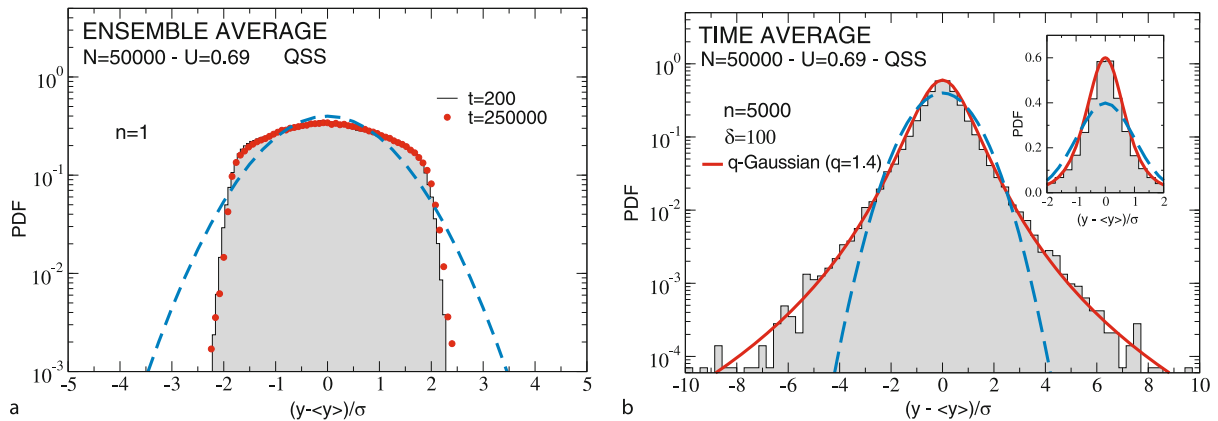


Entropy, Figure 6

Nongrowing dynamic network: **a** Cumulative degree distribution for typical values for the number N of nodes; **b** Same data of **a** in the convenient representation *linear q -log versus linear* with $Z_q(k) \equiv \ln_q[P_q(> k)] \equiv ([P_q(> k)]^{1-q} - 1)/(1 - q)$ (the optimal fitting with a q -exponential is obtained for the value of q which has the highest value of the linear correlation r as indicated in the *inset*; here this is $q_c = 1.84$, which corresponds to the slope -1.19 in **a**). See details in [172,173]

tematic calculation of several meaningful properties of networks.

Nonlinear dynamical systems, self-organized criticality, and cellular automata Various interesting phenomena emerge in both low- and high-dimensional weakly chaotic deterministic dynamical systems, either dissipative or conservative. Among these phenomena we have the sensitivity to the initial conditions and the entropy production, which have been briefly addressed in Eq. (37) and related papers. But there is much more, such as relaxation, escape, glassy states, and distributions associated with the station-



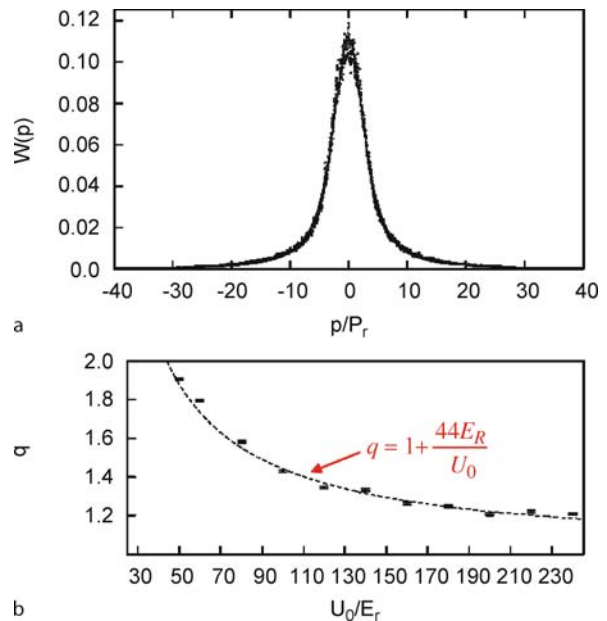
Entropy, Figure 7

Distribution of velocities for the HMF model at the quasi-stationary state (whose duration appears to diverge when $N \rightarrow \infty$). The blue curves indicate a Gaussian, for comparison. See details in [137]

ary state [14,15,31,46,62,67,68,77,103,111,122,123,154,170,174,176,177,179]. Also, recent numerical indications suggest the validity of a dynamical version of the q -generalized central limit theorem [175]. The possible connections between all these various properties is still in its infancy.

Long-range-interacting many-body Hamiltonians

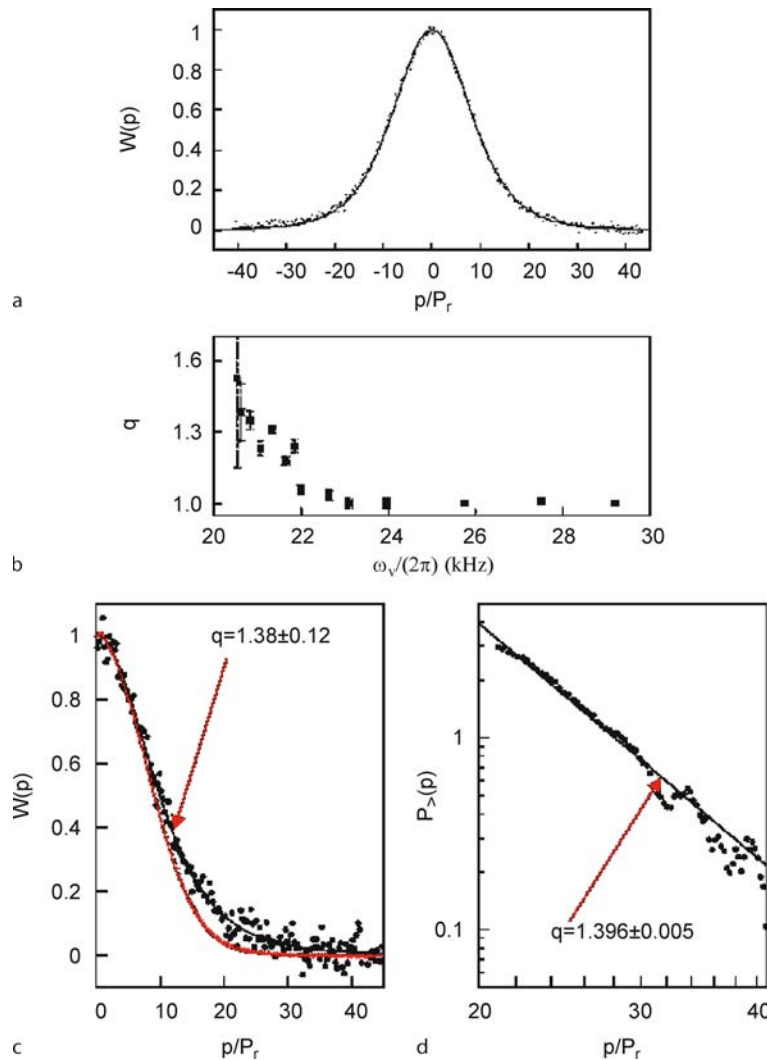
A wide class of long-range-interacting N -body classical Hamiltonians exhibits collective states whose Lyapunov spectrum has a maximal value that vanishes in the $N \rightarrow \infty$ limit. As such, they constitute natural candidates for studying whether the concepts derived from the nonadditive entropy S_q are applicable. A variety of properties have been calculated, through molecular dynamics, for various systems, such as Lennard-Jones-like fluids, XY and Heisenberg ferromagnets, gravitational-like models, and others. One or more long-standing quasi-stationary states (infinitely long-standing in the limit $N \rightarrow \infty$) are typically observed before the terminal entrance into thermal equilibrium. Properties such as distribution of velocities and angles, correlation functions, Lyapunov spectrum, metastability, glassy states, aging, time-dependence of the temperature in isolated systems, energy whenever thermal contact with a large thermostat at a given temperature is allowed, diffusion, order parameter, and others, are typically focused on. An ongoing debate exists, also involving Vlasov-like equations, Lynden-Bell statistics, among others. The breakdown of ergodicity that emerges in various situations makes the whole discussion rich and complex. The activity of the research nowadays in this area is illustrated in papers such as [21,26,45,53,56,57,63,104,119,121,



Entropy, Figure 8

Quantum Monte Carlo simulations in [81]: **a** Velocity distribution (superimposed with a q -Gaussian); **b** Index q (superimposed with Lutz prediction [110], by courtesy of the authors)

126,127,132,133,134,135,136,142,169,200]. A quite remarkable molecular-dynamical result has been obtained for a paradigmatic long-range Hamiltonian: the distribution of *time averaged* velocities sensibly differs from that found for the *ensemble-averaged* velocities, and has been shown to be numerically consistent with a q -Gaussian [137], as shown in Fig. 7. This result provides strong support to a conjecture made long ago: see Fig. 4 at p. 8 of [157].



Entropy, Figure 9

Experiments in [81]: **a** Velocity distribution (superimposed with a q -Gaussian); **b** Index q as a function of the frequency; **c** Velocity distribution (superimposed with a q -Gaussian; the red curve is a Gaussian); **d** Tail of the velocity distribution (superimposed with the asymptotic power-law of a q -Gaussian). [By courtesy of the authors]

Stochastic differential equations Quite generic Fokker–Planck equations are currently being studied. Aspects such as fractional derivatives, nonlinearities, space-dependent diffusion coefficients are being focused on, as well as their connections to entropic forms, and associated generalized Langevin equations [20,23,24,70,128,168,214]. Quite recently, computational (see Fig. 8) and experimental (see Fig. 9) verifications of Lutz’ 2003 prediction [110] have been exhibited [81], namely about the q -Gaussian form of the velocity distribution of cold atoms in dissipative optical lattices, with $q = 1 + 44E_R/U_0$ (E_R and U_0 being en-

ergy parameters of the optical lattice). These experimental verifications are in variance with some of those exhibited previously [96], namely double-Gaussians. Although it is naturally possible that the experimental conditions have not been exactly equivalent, this interesting question remains open at the present time. A hint might be hidden in the recent results [62] obtained for a quite different problem, namely the size distributions of avalanches; indeed, at a critical state, a q -Gaussian shape was obtained, whereas, at a noncritical state, a double-Gaussian was observed.

Quantum entanglement and quantum chaos The non-local nature of quantum physics implies phenomena that are somewhat analogous to those originated by classical long-range interactions. Consequently, a variety of studies are being developed in connection with the entropy S_q [3,36,58,59,60,61,155,156,195]. The same happens with some aspects of quantum chaos [11,180,210,211].

Astrophysics, geophysics, economics, linguistics, cognitive psychology, and other interdisciplinary applications Applications are available and presently searched in many areas of physics (plasmas, turbulence, nuclear collisions, elementary particles, manganites), but also in interdisciplinary sciences such astrophysics [38,47,48,49,78,84,87,101,109,129,196], geophysics [4,5,6,7,8,9,10,62,208], economics [25,50,51,52,80,139,141,197,215], linguistics [118], cognitive psychology [181], and others.

Global optimization, image and signal processing

Optimizing algorithms and related techniques for signal and image processing are currently being developed using the entropic concepts presented in this article [17,35,72,75,95,105,114,120,164,166,191].

Superstatistics and other generalizations The methods discussed here have been generalized along a variety of lines. These include Beck–Cohen superstatistics [32,34,65,190], crossover statistics [194,196], spectral statistics [201]. Also, a huge variety of entropies have been introduced which generalize in different manners the BG entropy, or even focus on other possibilities. Their number being nowadays over forty, we mention here just a few of them: see [18,44,69,98,99,115].

Acknowledgments

Among the very many colleagues towards which I am deeply grateful for profound and long-lasting comments along many years, it is a must to explicitly thank S. Abe, E.P. Borges, E.G.D. Cohen, E.M.F. Curado, M. Gell-Mann, R.S. Mendes, A. Plastino, A.R. Plastino, A.K. Rajagopal, A. Rapisarda and A. Robledo.

Bibliography

- Abe S (2000) Axioms and uniqueness theorem for Tsallis entropy. *Phys Lett A* 271:74–79
- Abe S (2002) Stability of Tsallis entropy and instabilities of Renyi and normalized Tsallis entropies: A basis for q -exponential distributions. *Phys Rev E* 66:046134
- Abe S, Rajagopal AK (2001) Nonadditive conditional entropy and its significance for local realism. *Physica A* 289:157–164
- Abe S, Suzuki N (2003) Law for the distance between successive earthquakes. *J Geophys Res (Solid Earth)* 108(B2):2113
- Abe S, Suzuki N (2004) Scale-free network of earthquakes. *Europhys Lett* 65:581–586
- Abe S, Suzuki N (2005) Scale-free statistics of time interval between successive earthquakes. *Physica A* 350:588–596
- Abe S, Suzuki N (2006) Complex network of seismicity. *Prog Theor Phys Suppl* 162:138–146
- Abe S, Suzuki N (2006) Complex-network description of seismicity. *Nonlinear Process Geophys* 13:145–150
- Abe S, Sarlis NV, Skordas ES, Tanaka H, Varotsos PA (2005) Optimality of natural time representation of complex time series. *Phys Rev Lett* 94:170601
- Abe S, Tirnakli U, Varotsos PA (2005) Complexity of seismicity and nonextensive statistics. *Europhys News* 36:206–208
- Abul AY-M (2005) Nonextensive random matrix theory approach to mixed regular-chaotic dynamics. *Phys Rev E* 71:066207
- Albert R, Barabasi AL (2000) *Phys Rev Lett* 85:5234–5237
- Aleman PA, Zanette DH (1994) Fractal random walks from a variational formalism for Tsallis entropies. *Phys Rev E* 49:R956–R958
- Ananos GFJ, Tsallis C (2004) Ensemble averages and nonextensivity at the edge of chaos of one-dimensional maps. *Phys Rev Lett* 93:020601
- Ananos GFJ, Baldovin F, Tsallis C (2005) Anomalous sensitivity to initial conditions and entropy production in standard maps: Nonextensive approach. *Euro Phys J B* 46:409–417
- Andrade RFS, Pinho STR (2005) Tsallis scaling and the long-range Ising chain: A transfer matrix approach. *Phys Rev E* 71:026126
- Andricioaei I, Straub JE (1996) Generalized simulated annealing algorithms using Tsallis statistics: Application to conformational optimization of a tetrapeptide. *Phys Rev E* 53:R3055–R3058
- Antenodo C, Plastino AR (1999) Maximum entropy approach to stretched exponential probability distributions. *J Phys A* 32:1089–1097
- Antenodo C, Tsallis C (1998) Breakdown of exponential sensitivity to initial conditions: Role of the range of interactions. *Phys Rev Lett* 80:5313–5316
- Antenodo C, Tsallis C (2003) Multiplicative noise: A mechanism leading to nonextensive statistical mechanics. *J Math Phys* 44:5194–5203
- Antoniazzi A, Fanelli D, Barre J, Chavanis P-H, Dauxois T, Ruffo S (2007) Maximum entropy principle explains quasi-stationary states in systems with long-range interactions: The example of the Hamiltonian mean-field model. *Phys Rev E* 75:011112
- Arevalo R, Garcimartin A, Maza D (2007) A non-standard statistical approach to the silo discharge. *Eur Phys J Special Topics* 143:191–197
- Assis PC Jr, da Silva LR, Lenzi EK, Malacarne LC, Mendes RS (2005) Nonlinear diffusion equation, Tsallis formalism and exact solutions. *J Math Phys* 46:123303
- Assis PC Jr, da Silva PC, da Silva LR, Lenzi EK, Lenzi MK (2006) Nonlinear diffusion equation and nonlinear external force: Exact solution. *J Math Phys* 47:103302
- Ausloos M, Ivanova K (2003) Dynamical model and nonextensive statistical mechanics of a market index on large time windows. *Phys Rev E* 68:046122

26. Baldovin F, Orlandini E (2006) Incomplete equilibrium in long-range interacting systems. *Phys Rev Lett* 97:100601
27. Baldovin F, Robledo A (2002) Sensitivity to initial conditions at bifurcations in one-dimensional nonlinear maps: Rigorous nonextensive solutions. *Europhys Lett* 60:518–524
28. Baldovin F, Robledo A (2002) Universal renormalization-group dynamics at the onset of chaos in logistic maps and nonextensive statistical mechanics. *Phys Rev E* 66:R045104
29. Baldovin F, Robledo A (2004) Nonextensive Pesin identity. Exact renormalization group analytical results for the dynamics at the edge of chaos of the logistic map. *Phys Rev E* 69:R045202
30. Baldovin F, Robledo A (2005) Parallels between the dynamics at the noise-perturbed onset of chaos in logistic maps and the dynamics of glass formation. *Phys Rev E* 72:066213
31. Baldovin F, Moyano LG, Majtey AP, Robledo A, Tsallis C (2004) Ubiquity of metastable-to-stable crossover in weakly chaotic dynamical systems. *Physica A* 340:205–218
32. Beck C, Cohen EGD (2003) Superstatistics. *Physica A* 322:267–275
33. Beck C, Schlogl F (1993) *Thermodynamics of Chaotic Systems*. Cambridge University Press, Cambridge
34. Beck C, Cohen EGD, Rizzo S (2005) Atmospheric turbulence and superstatistics. *Europhys News* 36:189–191
35. Ben A Hamza (2006) Nonextensive information-theoretic measure for image edge detection. *J Electron Imaging* 15: 013011
36. Batle J, Plastino AR, Casas M, Plastino A (2004) Inclusion relations among separability criteria. *J Phys A* 37:895–907
37. Batle J, Casas M, Plastino AR, Plastino A (2005) Quantum entropies and entanglement. *Intern J Quantum Inf* 3:99–104
38. Bernui A, Tsallis C, Villela T (2007) Deviation from Gaussianity in the cosmic microwave background temperature fluctuations. *Europhys Lett* 78:19001
39. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang D-U (2006) *Phys Rep* 424:175–308
40. Bologna M, Tsallis C, Grigolini P (2000) Anomalous diffusion associated with nonlinear fractional derivative Fokker-Planck-like equation: Exact time-dependent solutions. *Phys Rev E* 62:2213–2218
41. Boltzmann L (1896) *Vorlesungen über Gastheorie*. Part II, ch I, paragraph 1. Leipzig, p 217; (1964) *Lectures on Gas Theory* (trans: Brush S). Univ. California Press, Berkeley
42. Boon JP, Tsallis C (eds) (2005) *Nonextensive Statistical Mechanics: New Trends, New Perspectives*. Europhysics News 36(6):185–231
43. Borges EP (2004) A possible deformed algebra and calculus inspired in nonextensive thermostatics. *Physica A* 340:95–101
44. Borges EP, Roditi I (1998) A family of non-extensive entropies. *Phys Lett A* 246:399–402
45. Borges EP, Tsallis C (2002) Negative specific heat in a Lennard-Jones-like gas with long-range interactions. *Physica A* 305:148–151
46. Borges EP, Tsallis C, Ananos GFJ, Oliveira PMC (2002) Nonequilibrium probabilistic dynamics at the logistic map edge of chaos. *Phys Rev Lett* 89:254103
47. Burlaga LF, Vinas AF (2004) Multiscale structure of the magnetic field and speed at 1 AU during the declining phase of solar cycle 23 described by a generalized Tsallis PDF. *J Geophys Res Space – Phys* 109:A12107
48. Burlaga LF, Vinas AF (2005) Triangle for the entropic index q of non-extensive statistical mechanics observed by Voyager 1 in the distant heliosphere. *Physica A* 356:375–384
49. Burlaga LF, Ness NF, Acuna MH (2006) Multiscale structure of magnetic fields in the heliosheath. *J Geophys Res Space – Phys* 111:A09112
50. Borland L (2002) A theory of non-gaussian option pricing. *Quant Finance* 2:415–431
51. Borland L (2002) Closed form option pricing formulas based on a non-Gaussian stock price model with statistical feedback. *Phys Rev Lett* 89:098701
52. Borland L, Bouchaud J-P (2004) A non-Gaussian option pricing model with skew. *Quant Finance* 4:499–514
53. Cabral BJC, Tsallis C (2002) Metastability and weak mixing in classical long-range many-rotator system. *Phys Rev E* 66:065101(R)
54. Calabrese P, Cardy J (2004) *JSTAT – J Stat Mech Theory Exp* P06002
55. Callen HB (1985) *Thermodynamics and An Introduction to Thermostatistics*, 2nd edn. Wiley, New York
56. Chavanis PH (2006) Lynden-Bell and Tsallis distributions for the HMF model. *Euro Phys J B* 53:487–501
57. Chavanis PH (2006) Quasi-stationary states and incomplete violent relaxation in systems with long-range interactions. *Physica A* 365:102–107
58. Canosa N, Rossignoli R (2005) General non-additive entropic forms and the inference of quantum density operators. *Physica A* 348:121–130
59. Cannas SA, Tamarit FA (1996) Long-range interactions and nonextensivity in ferromagnetic spin models. *Phys Rev B* 54:R12661–R12664
60. Caruso F, Tsallis C (2007) Extensive nonadditive entropy in quantum spin chains. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. American Institute of Physics Conference Proceedings, vol 965. New York, pp 51–59
61. Caruso F, Tsallis C (2008) Nonadditive entropy reconciles the area law in quantum systems with classical thermodynamics. *Phys Rev E* 78:021101
62. Caruso F, Pluchino A, Latora V, Vinciguerra S, Rapisarda A (2007) Analysis of self-organized criticality in the Olami-Feder-Christensen model and in real earthquakes. *Phys Rev E* 75:055101(R)
63. Campa A, Giansanti A, Moroni D (2002) Metastable states in a class of long-range Hamiltonian systems. *Physica A* 305:137–143
64. Csizsar I (1978) Information measures: A critical survey. In: *Transactions of the Seventh Prague Conference on Information Theory, Statistical Decision Functions, Random Processes, and the European Meeting of Statisticians, 1974*. Reidel, Dordrecht
65. Cohen EGD (2005) Boltzmann and Einstein: Statistics and dynamics – An unsolved problem. Boltzmann Award Lecture at Statphys-Bangalore-2004. *Pramana* 64:635–643
66. Condat CA, Rangel J, Lamberti PW (2002) Anomalous diffusion in the nonasymptotic regime. *Phys Rev E* 65:026138
67. Coraddu M, Meloni F, Mezzorani G, Tonelli R (2004) Weak insensitivity to initial conditions at the edge of chaos in the logistic map. *Physica A* 340:234–239
68. Costa UMS, Lyra ML, Plastino AR, Tsallis C (1997) Power-law sensitivity to initial conditions within a logistic-like family

- of maps: Fractality and nonextensivity. *Phys Rev E* 56:245–250
69. Curado EMF (1999) General aspects of the thermodynamical formalism. *Braz J Phys* 29:36–45
 70. Curado EMF, Nobre FD (2003) Derivation of nonlinear Fokker-Planck equations by means of approximations to the master equation. *Phys Rev E* 67:021107
 71. Curado EMF, Tsallis C (1991) Generalized statistical mechanics: connection with thermodynamics. *Phys J A* 24:L69-L72; [Corrigenda: 24:3187 (1991); 25:1019 (1992)]
 72. Cvejic N, Canagarajah CN, Bull DR (2006) Image fusion metric based on mutual information and Tsallis entropy. *Electron Lett* 42:11
 73. Daniels KE, Beck C, Bodenschatz E (2004) Defect turbulence and generalized statistical mechanics. *Physica D* 193:208–217
 74. Daroczy Z (1970) *Information and Control* 16:36
 75. de Albuquerque MP, Esquef IA, Mello ARG, de Albuquerque MP (2004) Image thresholding using Tsallis entropy. *Pattern Recognition Lett* 25:1059–1065
 76. de Meneses MDS, da Cunha SD, Soares DJB, da Silva LR (2006) In: Sakagami M, Suzuki N, Abe S (eds) *Complexity and Nonextensivity: New Trends in Statistical Mechanics*. *Prog Theor Phys Suppl* 162:131–137
 77. de Moura FABF, Tirnakli U, Lyra ML (2000) Convergence to the critical attractor of dissipative maps: Log-periodic oscillations, fractality and nonextensivity. *Phys Rev E* 62:6361–6365
 78. de Oliveira HP, Soares ID, Tonini EV (2004) Role of the nonextensive statistics in a three-degrees of freedom gravitational system. *Phys Rev D* 70:084012
 79. de Souza AMC, Tsallis C (1997) Student's *t*- and *r*-distributions: Unified derivation from an entropic variational principle. *Physica A* 236:52–57
 80. de Souza J, Moyano LG, Queiros SMD (2006) On statistical properties of traded volume in financial markets. *Euro Phys J B* 50:165–168
 81. Douglas P, Bergamini S, Renzoni F (2006) Tunable Tsallis distributions in dissipative optical lattices. *Phys Rev Lett* 96:110601
 82. Fermi E (1936) *Thermodynamics*. Dover, New York, p 53
 83. Ferri GL, Martinez S, Plastino A (2005) Equivalence of the four versions of Tsallis' statistics. *J Stat Mech* P04009
 84. Ferro F, Lavagno A, Quarati P (2004) Non-extensive resonant reaction rates in astrophysical plasmas. *Euro Phys J A* 21:529–534
 85. Fulco UL, da Silva LR, Nobre FD, Rego HHA, Lucena LS (2003) Effects of site dilution on the one-dimensional long-range bond-percolation problem. *Phys Lett A* 312:331–335
 86. Frank TD (2005) *Nonlinear Fokker–Planck Equations – Fundamentals and Applications*. Springer, Berlin
 87. Gervino G, Lavagno A, Quarati P (2005) *CNO* reaction rates and chemical abundance variations in dense stellar plasma. *J Phys G* 31:S1865–S1868
 88. Gibbs JW (1902) *Elementary Principles in Statistical Mechanics – Developed with Especial Reference to the Rational Foundation of Thermodynamics*. C Scribner, New York; Yale University Press, New Haven, 1948; OX Bow Press, Woodbridge, Connecticut, 1981
 89. Ginsparg P, Moore G (1993) *Lectures on 2D Gravity and 2D String Theory*. Cambridge University Press, Cambridge; hep-th/9304011, p 65
 90. Grigera JR (1996) Extensive and non-extensive thermodynamics. A molecular dynamic test. *Phys Lett A* 217:47–51
 91. Hasegawa H (2006) Nonextensive aspects of small-world networks. *Physica A* 365:383–401
 92. Havrda J, Charvat F (1967) *Kybernetika* 3:30
 93. Hernandez H-S, Robledo A (2006) Fluctuating dynamics at the quasiperiodic onset of chaos, Tsallis *q*-statistics and Mori's *q*-phase thermodynamics. *Phys A* 370:286–300
 94. Hilhorst HJ, Schehr G (2007) A note on *q*-Gaussians and non-Gaussians in statistical mechanics. *J Stat Mech* P06003
 95. Jang S, Shin S, Pak Y (2003) Replica-exchange method using the generalized effective potential. *Phys Rev Lett* 91:058305
 96. Jersblad J, Ellmann H, Stochkel K, Kastberg A, Sanchez L-P, Kaiser R (2004) Non-Gaussian velocity distributions in optical lattices. *Phys Rev A* 69:013410
 97. Jund P, Kim SG, Tsallis C (1995) Crossover from extensive to nonextensive behavior driven by long-range interactions. *Phys Rev B* 52:50–53
 98. Kaniadakis G (2001) Non linear kinetics underlying generalized statistics. *Physica A* 296:405–425
 99. Kaniadakis G, Lissia M, Scarfone AM (2004) Deformed logarithms and entropies. *Physica A* 340:41–49
 100. Khinchin AI (1953) *Uspekhi Matem. Nauk* 8:3 (Silverman RA, Friedman MD, trans. *Math Found Inf Theory*. Dover, New York)
 101. Kronberger T, Leubner MP, van Kampen E (2006) Dark matter density profiles: A comparison of nonextensive statistics with *N*-body simulations. *Astron Astrophys* 453:21–25
 102. Latora V, Baranger M (1999) Kolmogorov-Sinai entropy rate versus physical entropy. *Phys Rev Lett* 82:520–523
 103. Latora V, Baranger M, Rapisarda A, Tsallis C (2000) The rate of entropy increase at the edge of chaos. *Phys Lett A* 273:97–103
 104. Latora V, Rapisarda A, Tsallis C (2001) Non-Gaussian equilibrium in a long-range Hamiltonian system. *Phys Rev E* 64:056134
 105. Lemes MR, Zacharias CR, Dal Pino A Jr (1997) Generalized simulated annealing: Application to silicon clusters. *Phys Rev B* 56:9279–9281
 106. Lenzi EK, Anteneodo C, Borland L (2001) Escape time in anomalous diffusive media. *Phys Rev E* 63:051109
 107. Lesche B (1982) Instabilities of Rényi entropies. *J Stat Phys* 27:419–422
 108. Lindhard J, Nielsen V (1971) *Studies in statistical mechanics*. Det Kongelige Danske Videnskabernes Selskab Matematisk-fysiske Meddelelser (Denmark) 38(9):1–42
 109. Lissia M, Quarati P (2005) Nuclear astrophysical plasmas: Ion distributions and fusion rates. *Europhys News* 36:211–214
 110. Lutz E (2003) Anomalous diffusion and Tsallis statistics in an optical lattice. *Phys Rev A* 67:051402(R)
 111. Lyra ML, Tsallis C (1998) Nonextensivity and multifractality in low-dimensional dissipative systems. *Phys Rev Lett* 80:53–56
 112. Mann GM, Tsallis C (eds) (2004) *Nonextensive Entropy – Interdisciplinary Applications*. Oxford University Press, New York
 113. Marsh JA, Fuentes MA, Moyano LG, Tsallis C (2006) Influence of global correlations on central limit theorems and entropic extensivity. *Physica A* 372:183–202
 114. Martin S, Morison G, Nailon W, Durrani T (2004) Fast and accurate image registration using Tsallis entropy and simultaneous perturbation stochastic approximation. *Electron Lett* 40(10):20040375

115. Masi M (2005) A step beyond Tsallis and Renyi entropies. *Phys Lett A* 338:217–224
116. Mayoral E, Robledo A (2004) Multifractality and nonextensivity at the edge of chaos of unimodal maps. *Physica A* 340:219–226
117. Mayoral E, Robledo A (2005) Tsallis' q index and Mori's q phase transitions at edge of chaos. *Phys Rev E* 72:026209
118. Montemurro MA (2001) Beyond the Zipf–Mandelbrot law in quantitative linguistics. *Physica A* 300:567–578
119. Montemurro MA, Tamarit F, Anteneodo C (2003) Aging in an infinite-range Hamiltonian system of coupled rotators. *Phys Rev E* 67:031106
120. Moret MA, Pascutti PG, Bisch PM, Mundim MSP, Mundim KC (2006) Classical and quantum conformational analysis using Generalized Genetic Algorithm. *Phys A* 363:260–268
121. Moyano LG, Anteneodo C (2006) Diffusive anomalies in a long-range Hamiltonian system. *Phys Rev E* 74:021118
122. Moyano LG, Majtey AP, Tsallis C (2005) Weak chaos in large conservative system – Infinite-range coupled standard maps. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. World Scientific, Singapore, pp 123–127
123. Moyano LG, Majtey AP, Tsallis C (2006) Weak chaos and metastability in a symplectic system of many long-range-coupled standard maps. *Euro Phys J B* 52:493–500
124. Moyano LG, Tsallis C, Gell-Mann M (2006) Numerical indications of a q -generalised central limit theorem. *Europhys Lett* 73:813–819
125. Nivanen L, Le Mehaute A, Wang QA (2003) Generalized algebra within a nonextensive statistics. *Rep Math Phys* 52:437–444
126. Nobre FD, Tsallis C (2003) Classical infinite-range-interaction Heisenberg ferromagnetic model: Metastability and sensitivity to initial conditions. *Phys Rev E* 68:036115
127. Nobre FD, Tsallis C (2004) Metastable states of the classical inertial infinite-range-interaction Heisenberg ferromagnet: Role of initial conditions. *Physica A* 344:587–594
128. Nobre FD, Curado EMF, Rowlands G (2004) A procedure for obtaining general nonlinear Fokker-Planck equations. *Physica A* 334:109–118
129. Oliveira HP, Soares ID (2005) Dynamics of black hole formation: Evidence for nonextensivity. *Phys Rev D* 71:124034
130. Penrose O (1970) *Foundations of Statistical Mechanics: A Deductive Treatment*. Pergamon Press, Oxford, p 167
131. Plastino AR, Plastino A (1995) Non-extensive statistical mechanics and generalized Fokker-Planck equation. *Physica A* 222:347–354
132. Pluchino A, Rapisarda A (2006) Metastability in the Hamiltonian Mean Field model and Kuramoto model. *Physica A* 365:184–189
133. Pluchino A, Rapisarda A (2006) Glassy dynamics and nonextensive effects in the *HMF* model: the importance of initial conditions. In: Sakagami M, Suzuki N, Abe S (eds) *Complexity and Nonextensivity: New Trends in Statistical Mechanics*. Prog Theor Phys Suppl 162:18–28
134. Pluchino A, Latora V, Rapisarda A (2004) Glassy dynamics in the *HMF* model. *Physica A* 340:187–195
135. Pluchino A, Latora V, Rapisarda A (2004) Dynamical anomalies and the role of initial conditions in the *HMF* model. *Physica A* 338:60–67
136. Pluchino A, Rapisarda A, Latora V (2005) Metastability and anomalous behavior in the *HMF* model: Connections to nonextensive thermodynamics and glassy dynamics. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. World Scientific, Singapore, pp 102–112
137. Pluchino A, Rapisarda A, Tsallis C (2007) Nonergodicity and central limit behavior in long-range Hamiltonians. *Europhys Lett* 80:26002
138. Prato D, Tsallis C (1999) Nonextensive foundation of Levy distributions. *Phys Rev E* 60:2398–2401
139. Queiros SMD (2005) On non-Gaussianity and dependence in financial time series: A nonextensive approach. *Quant Finance* 5:475–487
140. Queiros SMD, Tsallis C (2007) Nonextensive statistical mechanics and central limit theorems II – Convolution of q -independent random variables. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. American Institute of Physics Conference Proceedings, vol 965. New York, pp 21–33
141. Queiros SMD, Moyano LG, de Souza J, Tsallis C (2007) A nonextensive approach to the dynamics of financial observables. *Euro Phys J B* 55:161–168
142. Rapisarda A, Pluchino A (2005) Nonextensive thermodynamics and glassy behavior. *Europhys News* 36:202–206; Erratum: 37:25 (2006)
143. Rapisarda A, Pluchino A (2005) Nonextensive thermodynamics and glassy behaviour in Hamiltonian systems. *Europhys News* 36:202–206; Erratum: 37:25 (2006)
144. Rego HHA, Lucena LS, da Silva LR, Tsallis C (1999) Crossover from extensive to nonextensive behavior driven by long-range $d = 1$ bond percolation. *Phys A* 266:42–48
145. Renyi A (1961) In: *Proceedings of the Fourth Berkeley Symposium*, 1:547 University California Press, Berkeley; Renyi A (1970) *Probability theory*. North-Holland, Amsterdam
146. Robledo A (2004) Aging at the edge of chaos: Glassy dynamics and nonextensive statistics. *Physica A* 342:104–111
147. Robledo A (2004) Universal glassy dynamics at noise-perturbed onset of chaos: A route to ergodicity breakdown. *Phys Lett A* 328:467–472
148. Robledo A (2004) Criticality in nonlinear one-dimensional maps: RG universal map and nonextensive entropy. *Physica D* 193:153–160
149. Robledo A (2005) Intermittency at critical transitions and aging dynamics at edge of chaos. *Pramana-J Phys* 64:947–956
150. Robledo A (2005) Critical attractors and q -statistics. *Europhys News* 36:214–218
151. Robledo A (2006) Crossover from critical to chaotic attractor dynamics in logistic and circle maps. In: Sakagami M, Suzuki N, Abe S (eds) *Complexity and Nonextensivity: New Trends in Statistical Mechanics*. Prog Theor Phys Suppl 162:10–17
152. Robledo A, Baldovin F, Mayoral E (2005) Two stories outside Boltzmann-Gibbs statistics: Mori's q -phase transitions and glassy dynamics at the onset of chaos. In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. World Scientific, Singapore, p 43
153. Rodriguez A, Schwammle V, Tsallis C (2008) Strictly and asymptotically scale-invariant probabilistic models of N correlated binary random variables having q -Gaussians as $N \rightarrow$ infinity Limiting distributions. *J Stat Mech* P09006

154. Rohlf T, Tsallis C (2007) Long-range memory elementary 1D cellular automata: Dynamics and nonextensivity. *Physica A* 379:465–470
155. Rossignoli R, Canosa N (2003) Violation of majorization relations in entangled states and its detection by means of generalized entropic forms. *Phys Rev A* 67:042302
156. Rossignoli R, Canosa N (2004) Generalized disorder measure and the detection of quantum entanglement. *Physica A* 344:637–643
157. Salinas SRA, Tsallis C (eds) (1999) *Nonextensive Statistical Mechanics and Thermodynamics*. Braz J Phys 29(1)
158. Sampaio LC, de Albuquerque MP, de Menezes FS (1997) Nonextensivity and Tsallis statistic in magnetic systems. *Phys Rev B* 55:5611–5614
159. Santos RJV (1997) Generalization of Shannon's theorem for Tsallis entropy. *J Math Phys* 38:4104–4107
160. Sato Y, Tsallis C (2006) In: Bountis T, Casati G, Procaccia I (eds) *Complexity: An unifying direction in science*. Int J Bif Chaos 16:1727–1738
161. Schutzenberger PM (1954) Contributions aux applications statistiques de la theorie de l'information. *Publ Inst Statist Univ Paris* 3:3
162. Shannon CE (1948) A Mathematical Theory of Communication. *Bell Syst Tech J* 27:379–423; 27:623–656; (1949) *The Mathematical Theory of Communication*. University of Illinois Press, Urbana
163. Sharma BD, Mittal DP (1975) *J Math Sci* 10:28
164. Serra P, Stanton AF, Kais S, Bleil RE (1997) Comparison study of pivot methods for global optimization. *J Chem Phys* 106:7170–7177
165. Soares DJB, Tsallis C, Mariz AM, da Silva LR (2005) Preferential Attachment growth model and nonextensive statistical mechanics. *Europhys Lett* 70:70–76
166. Son WJ, Jang S, Pak Y, Shin S (2007) Folding simulations with novel conformational search method. *J Chem Phys* 126:104906
167. Stigler SM (1999) *Statistics on the table – The history of statistical concepts and methods*. Harvard University Press, Cambridge
168. Silva AT, Lenzi EK, Evangelista LR, Lenzi MK, da Silva LR (2007) Fractional nonlinear diffusion equation, solutions and anomalous diffusion. *Phys A* 375:65–71
169. Tamarit FA, Anteneodo C (2005) Relaxation and aging in long-range interacting systems. *Europhys News* 36:194–197
170. Tamarit FA, Cannas SA, Tsallis C (1998) Sensitivity to initial conditions and nonextensivity in biological evolution. *Euro Phys J B* 1:545–548
171. Thistleton W, Marsh JA, Nelson K, Tsallis C (2006) unpublished
172. Thurner S (2005) *Europhys News* 36:218–220
173. Thurner S, Tsallis C (2005) Nonextensive aspects of self-organized scale-free gas-like networks. *Europhys Lett* 72:197–204
174. Tirnakli U, Ananos GFJ, Tsallis C (2001) Generalization of the Kolmogorov–Sinai entropy: Logistic – like and generalized cosine maps at the chaos threshold. *Phys Lett A* 289:51–58
175. Tirnakli U, Beck C, Tsallis C (2007) Central limit behavior of deterministic dynamical systems. *Phys Rev E* 75:040106(R)
176. Tirnakli U, Tsallis C, Lyra ML (1999) Circular-like maps: Sensitivity to the initial conditions, multifractality and nonextensivity. *Euro Phys J B* 11:309–315
177. Tirnakli U, Tsallis C (2006) Chaos thresholds of the z-logistic map: Connection between the relaxation and average sensitivity entropic indices. *Phys Rev E* 73:037201
178. Tisza L (1961) *Generalized Thermodynamics*. MIT Press, Cambridge, p 123
179. Tonelli R, Mezzorani G, Meloni F, Lissia M, Coraddu M (2006) Entropy production and Pesin-like identity at the onset of chaos. *Prog Theor Phys* 115:23–29
180. Toscano F, Vallejos RO, Tsallis C (2004) Random matrix ensembles from nonextensive entropy. *Phys Rev E* 69:066131
181. Tsallis AC, Tsallis C, Magalhaes ACN, Tamarit FA (2003) Human and computer learning: An experimental study. *Complexus* 1:181–189
182. Tsallis C Regularly updated bibliography at <http://tsallis.cat.cbpf.br/biblio.htm>
183. Tsallis C (1988) Possible generalization of Boltzmann–Gibbs statistics. *J Stat Phys* 52:479–487
184. Tsallis C (2004) What should a statistical mechanics satisfy to reflect nature? *Physica D* 193:3–34
185. Tsallis C (2004) Dynamical scenario for nonextensive statistical mechanics. *Physica A* 340:1–10
186. Tsallis C (2005) Is the entropy S_q extensive or nonextensive? In: Beck C, Benedek G, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. World Scientific, Singapore
187. Tsallis C (2005) Nonextensive statistical mechanics, anomalous diffusion and central limit theorems. *Milan J Math* 73:145–176
188. Tsallis C, Bukman DJ (1996) Anomalous diffusion in the presence of external forces: exact time-dependent solutions and their thermostistical basis. *Phys Rev E* 54:R2197–R2200
189. Tsallis C, Queiros SMD (2007) Nonextensive statistical mechanics and central limit theorems I – Convolution of independent random variables and q -product. In: Abe S, Herrmann HJ, Quarati P, Rapisarda A, Tsallis C (eds) *Complexity, Metastability and Nonextensivity*. American Institute of Physics Conference Proceedings, vol 965. New York, pp 8–20
190. Tsallis C, Souza AMC (2003) Constructing a statistical mechanics for Beck-Cohen superstatistics. *Phys Rev E* 67:026106
191. Tsallis C, Stariolo DA (1996) Generalized simulated annealing. *Phys A* 233:395–406; A preliminary version appeared (in English) as *Notas de Fisica/CBPF* 026 (June 1994)
192. Tsallis C, Levy SVF, de Souza AMC, Maynard R (1995) Statistical-mechanical foundation of the ubiquity of Levy distributions in nature. *Phys Rev Lett* 75:3589–3593; Erratum: (1996) *Phys Rev Lett* 77:5442
193. Tsallis C, Mendes RS, Plastino AR (1998) The role of constraints within generalized nonextensive statistics. *Physica A* 261:534–554
194. Tsallis C, Bemsiki G, Mendes RS (1999) Is re-association in folded proteins a case of nonextensivity? *Phys Lett A* 257:93–98
195. Tsallis C, Lloyd S, Baranger M (2001) Peres criterion for separability through nonextensive entropy. *Phys Rev A* 63:042104
196. Tsallis C, Anjos JC, Borges EP (2003) Fluxes of cosmic rays: A delicately balanced stationary state. *Phys Lett A* 310:372–376
197. Tsallis C, Anteneodo C, Borland L, Osorio R (2003) Nonextensive statistical mechanics and economics. *Physica A* 324:89–100
198. Tsallis C, Mann GM, Sato Y (2005) Asymptotically scale-invariant occupancy of phase space makes the entropy S_q extensive. *Proc Natl Acad Sci USA* 102:15377–15382

199. Tsallis C, Mann GM, Sato Y (2005) Extensivity and entropy production. In: Boon JP, Tsallis C (eds) *Nonextensive Statistical Mechanics: New Trends, New perspectives*. Europhys News 36:186–189
200. Tsallis C, Rapisarda A, Pluchino A, Borges EP (2007) On the non-Boltzmannian nature of quasi-stationary states in long-range interacting systems. *Physica A* 381:143–147
201. Tsekouras GA, Tsallis C (2005) Generalized entropy arising from a distribution of q -indices. *Phys Rev E* 71:046144
202. Umarov S, Tsallis C (2007) Multivariate generalizations of the q -central limit theorem. *cond-mat/0703533*
203. Umarov S, Tsallis C, Steinberg S (2008) On a q -central limit theorem consistent with nonextensive statistical mechanics. *Milan J Math* 76. doi:10.1007/s00032-008-0087-y
204. Umarov S, Tsallis C, Gell-Mann M, Steinberg S (2008) Symmetric (q, α) -stable distributions. Part I: First representation. *cond-mat/0606038v2*
205. Umarov S, Tsallis C, Gell-Mann M, Steinberg S (2008) Symmetric (q, α) -stable distributions. Part II: Second representation. *cond-mat/0606040v2*
206. Upadhyaya A, Rieu J-P, Glazier JA, Sawada Y (2001) Anomalous diffusion and non-Gaussian velocity distribution of Hydra cells in cellular aggregates. *Physica A* 293:549–558
207. Vajda I (1968) *Kybernetika* 4:105 (in Czech)
208. Varotsos PA, Sarlis NV, Tanaka HK, Skordas ES (2005) Some properties of the entropy in the natural time. *Phys Rev E* 71:032102
209. Wehrl A (1978) *Rev Modern Phys* 50:221
210. Weinstein YS, Lloyd S, Tsallis C (2002) Border between between regular and chaotic quantum dynamics. *Phys Rev Lett* 89:214101
211. Weinstein YS, Tsallis C, Lloyd S (2004) On the emergence of nonextensivity at the edge of quantum chaos. In: Elze H-T (ed) *Decoherence and Entropy in Complex Systems*. Lecture notes in physics, vol 633. Springer, Berlin, pp 385–397
212. White DR, Kejzar N, Tsallis C, Farmer JD, White S (2005) A generative model for feedback networks. *Phys Rev E* 73:016119
213. Wilk G, Włodarczyk Z (2004) *Acta Phys Pol B* 35:871–879
214. Wu JL, Chen HJ (2007) Fluctuation in nonextensive reaction-diffusion systems. *Phys Scripta* 75:722–725
215. Yamano T (2004) Distribution of the Japanese posted land price and the generalized entropy. *Euro Phys J B* 38:665–669
216. Zanette DH, Alemany PA (1995) Thermodynamics of anomalous diffusion. *Phys Rev Lett* 75:366–369

Entropy in Ergodic Theory

JONATHAN L. F. KING

University of Florida, Gainesville, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Entropy example: How many questions?](#)

[Distribution Entropy](#)

[A Gander at Shannon's Noisy Channel Theorem](#)

[The Information Function](#)

[Entropy of a Process](#)

[Entropy of a Transformation](#)

[Determinism and Zero-Entropy](#)

[The Pinsker–Field and K-Automorphisms](#)

[Ornstein Theory](#)

[Topological Entropy](#)

[Three Recent Results](#)

[Exodos](#)

[Bibliography](#)

Glossary

Some of the following definitions refer to the “Notation” paragraph immediately below. Use *mpt* for ‘measure-preserving transformation’.

Measure space A *measure space* (X, \mathcal{X}, μ) is a set X , a *field* (that is, a σ -algebra) \mathcal{X} of subsets of X , and a countably-additive measure $\mu: \mathcal{X} \rightarrow [0, \infty]$. (We often just write (X, μ) , with the field implicit.) For a collection $\mathcal{C} \subset \mathcal{X}$, use $\text{Fld}(\mathcal{C}) \supset \mathcal{C}$ for the smallest field including \mathcal{C} . The number $\mu(B)$ is the “ μ -mass of B ”.

Measure-preserving map A *measure-preserving map* $\psi: (X, \mathcal{X}, \mu) \rightarrow (Y, \mathcal{Y}, \nu)$ is a map $\psi: X \rightarrow Y$ such that the inverse image of each $B \in \mathcal{Y}$ is in \mathcal{X} , and $\mu(\psi^{-1}(B)) = \nu(B)$. A (*measure-preserving*) *transformation* is a measure-preserving map $T: (X, \mathcal{X}, \mu) \rightarrow (X, \mathcal{X}, \mu)$. Condense this notation to $(T: X, \mathcal{X}, \mu)$ or $(T: X, \mu)$.

Probability space A *probability space* is a measure space (X, μ) with $\mu(X) = 1$; this μ is a *probability measure*. All our maps/transformations in this article are on probability spaces.

Factor map A *factor map*

$$\psi: (T: X, \mathcal{X}, \mu) \rightarrow (S: Y, \mathcal{Y}, \nu)$$

is a measure-preserving map $\psi: X \rightarrow Y$ which intertwines the transformations, $\psi \circ T = S \circ \psi$. And ψ is an *isomorphism* if – after deleting a nullset (a mass-zero set) in each space – this ψ is a bijection and ψ^{-1} is also a factor map.

Almost everywhere (a.e.) A measure-theoretic statement holds *almost everywhere*, abbreviated *a.e.*, if it holds off of a nullset. (Eugene Gutkin once remarked to me that the problem with Measure Theory is ... that you have to say “almost everywhere”, almost everywhere.) For example, $B \supset A$ means that $\mu(B \setminus A)$ is zero. The *a.e.* will usually be implicit.

Probability vector A *probability vector* $\vec{v} = (v_1, v_2, \dots)$ is a list of non-negative reals whose sum is 1. We generally assume that probability vectors and partitions (see

below) have *finitely* many components. Write “*countable probability vector/partition*”, when finitely or denumerably many components are considered.

Partition A *partition* $P = (A_1, A_2, \dots)$ splits X into pairwise disjoint subsets $A_i \in \mathcal{X}$ so that the disjoint union $\bigsqcup_i A_i$ is all of X . Each A_i is an *atom* of P . Use $|P|$ or $\#P$ for the number of atoms. When P partitions a probability space, then it yields a probability vector \vec{v} , where $v_j := \mu(A_j)$. Lastly, use $P(x)$ to denote the P -atom that owns x .

Fonts We use the font $\mathcal{H}, \mathcal{E}, \mathcal{I}$ for *distribution-entropy*, *entropy* and the *information function*. In contrast, the script font $\mathcal{ABC} \dots$ will be used for collections of sets; usually subfields of \mathcal{X} . Use $\mathbb{E}(\cdot)$ for the (conditional) expectation operator.

Notation \mathbb{Z} = integers, \mathbb{Z}_+ = positive integers, and \mathbb{N} = natural numbers = $0, 1, 2, \dots$ (Some well-meaning folk use \mathbb{N} for \mathbb{Z}_+ , saying ‘*Nothing could be more natural than the positive integers*’. And this is why $0 \in \mathbb{N}$. Use $\lceil \cdot \rceil$ and $\lfloor \cdot \rfloor$ for the *ceiling* and *floor* functions; $\lfloor \cdot \rfloor$ is also called the “*greatest-integer function*”. For an interval $J := [a, b) \subset [-\infty, +\infty]$, let $[a \dots b)$ denote the *interval of integers* $J \cap \mathbb{Z}$ (with a similar convention for closed and open intervals). E. g., $(e \dots \pi) = (e \dots \pi) = \{3\}$.

For subsets A and B of the same space, Ω , use $A \subset B$ for inclusion and $A \subsetneq B$ for *proper* inclusion. The difference set $B \setminus A$ is $\{\omega \in B \mid \omega \notin A\}$. Employ A^c for the complement $\Omega \setminus A$. Since we work in a probability space, if we let $x := \mu(A)$, then a convenient convention is to have

$$x^c \text{ denote } 1 - x,$$

since then $\mu(A^c)$ equals x^c .

Use $A \Delta B$ for the *symmetric difference* $[A \setminus B] \cup [B \setminus A]$. For a collection $\mathcal{C} = \{E_j\}_j$ of sets in Ω , let the *disjoint union* $\bigsqcup_j E_j$ or $\bigsqcup(\mathcal{C})$ represent the union $\bigcup_j E_j$ and also assert that the sets are pairwise disjoint.

Use “ $\forall_{\text{large } n}$ ” to mean: “ $\exists n_0$ such that $\forall n > n_0$ ”. To refer to left hand side of an Eq. (20), use LhS(20); do analogously for RhS(20), the right hand side.

Definition of the Subject

The word ‘*entropy*’ (originally German, *Entropie*) was coined by Rudolf Julius Emanuel Clausius circa 1865 [2,3], taken from the Greek $\epsilon\nu\tau\rho\omicron\pi\alpha$, ‘*a turning towards*’. This article thus begins (*Prolegomenon*, “introduction”) and ends (*Exodos*¹, “the path out”) in Greek.

¹This is the Greek spelling.

Clausius, in his coinage, was referring to the thermodynamic notion in physics. Our focus in this article, however, will be the concept in measurable and topological dynamics. (Entropy in differentiable dynamics² would require an article by itself.) Shannon’s 1948 paper [6] on Information Theory, then Kolmogorov’s [4] and Sinai’s [7] generalization to dynamical systems, will be our starting point. Our discussion will be of the one-dimensional case, where the acting-group is \mathbb{Z} .

“My greatest concern was what to call it. I thought of calling it ‘*information*’, but the word was overly used, so I decided to call it ‘*uncertainty*’. John von Neumann had a better idea, he told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function goes by that name in statistical mechanics. In the second place, and more important, **nobody knows what entropy really is**, so in a debate you will always have the advantage.’” (Shannon as quoted in [59])

Entropy example: How many questions?

Imagine a dartboard, Fig. 1, split in five regions A, \dots, E with known probabilities. Blindfolded, you throw a dart at the board. What is the expected number V of Yes/No questions needed to ascertain the region in which the dart landed?

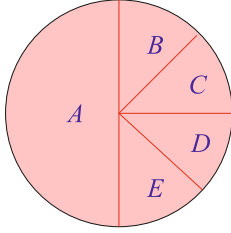
Solve this by always dividing the remaining probability in half. ‘*Is it A?*’ – if Yes, then $V = 1$. Else: ‘*Is it B or C?*’ – if Yes, then ‘*Is it B?*’ – if No, then the dart landed in C , and $V = 3$ was the number of questions. Evidently $V = 3$ also for regions B, D, E . Using “log” to denote base-2 logarithm³, the expected number of questions⁴ is thus

$$\begin{aligned} \mathbb{E}(V) &= \frac{1}{2} \cdot 1 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 \\ &= \sum_{j=0}^4 p_j \log \left(\frac{1}{p_j} \right) \stackrel{\text{note}}{=} 2. \end{aligned} \quad (1)$$

²For instance, see [15,18,24,25].

³In this paper, unmarked logs will be to base-2. In entropy theory, it does not matter much what base is used, but base-2 is convenient for computing entropy for messages described in bits. When using the natural logarithm, some people refer to the unit of information as a *nat*. In this paper, I have picked bits rather than nats.

⁴This holds when each probability p is a reciprocal power of two. For general probabilities, the “expected number of questions” interpretation holds in a weaker sense: Throw N darts independently at N copies of the dartboard. Efficiently ask Yes/No questions to determine where *all* N darts landed. Dividing by N , then sending $N \rightarrow \infty$, will be the $p \log(\frac{1}{p})$ sum of Eq. 1.



Entropy in Ergodic Theory, Figure 1

This dartboard is a probability space with a 5-set partition. The atoms have probabilities $\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}$. This probability distribution will be used later in Meshalkin's example

Letting $\vec{v} := (\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ be the probability vector, we can write this expectation as

$$\mathbb{E}(V) = \sum_{x \in \vec{v}} \eta(x).$$

Here, $\eta : [0, 1] \rightarrow [0, \infty)$ is the important function⁵

$$\eta(x) := x \log(1/x), \quad \text{so extending by continuity gives } \eta(0) = 0. \quad (2)$$

An interpretation of " $\eta(x)$ " is the number of questions needed to winnow down to an event of probability x .

Distribution Entropy

Given a probability vector \vec{v} , define its **distribution entropy** as

$$\mathcal{H}(\vec{v}) := \sum_{x \in \vec{v}} \eta(x). \quad (3)$$

This article will use the term **distropy** for 'distribution entropy', reserving the word **entropy** for the corresponding dynamical concept, when there is a notion of *time* involved. Getting ahead of ourselves, the *entropy* of a stationary process is the asymptotic average value that its distropy decays to, as we look at larger and larger finite portions of the process.

An equi-probable vector $\vec{v} := (\frac{1}{K}, \dots, \frac{1}{K})$ evidently has $\mathcal{H}(\vec{v}) = \log(K)$. On a probability space, the "**distropy of partition** \mathcal{P} ", written $\mathcal{H}(\mathcal{P})$ or $\mathcal{H}(A_1, A_2, \dots)$ shall mean the distropy of probability vector $j \mapsto \mu(A_j)$.

A (finite) partition necessarily has finite distropy. A *countable* partition can have finite distropy, e.g.

⁵There does not seem to be a standard name for this function. I use η , since an uppercase η looks like an H , which is the letter that Shannon used to denote what I am calling distribution-entropy.

$\mathcal{H}(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \dots) = 2$. One could also have infinite distropy: Consider a piece $B \subset X$ of mass $1/2^N$. Splitting B into 2^k many equal-mass atoms gives an η -sum of $2^k(k + N)/(2^k 2^N)$. Setting $k = k_N := 2^N - N$ makes this η -sum equal 1; so splitting the pieces of $X = \bigsqcup_{N=1}^{\infty} B_N$, with $\mu(B_N) = \frac{1}{2^N}$, yields an ∞ -distropy partition.

Function η

The $\eta(x) = x \log(1/x)$ function⁶ has vertical tangent at $x = 0$, maximum at $1/e$ and, when graphed in nats slope -1 at $x = 1$.

Consider partitions \mathcal{P} and \mathcal{Q} on the same space (X, μ) . Their **join**, written $\mathcal{P} \vee \mathcal{Q}$, has atoms $A \cap B$, for each pair $A \in \mathcal{P}$ and $B \in \mathcal{Q}$. They are **independent**, written $\mathcal{P} \perp \mathcal{Q}$ if $\mu(A \cap B) = \mu(A)\mu(B)$ for each A, B pair. We write $\mathcal{P} \succcurlyeq \mathcal{Q}$, and say that " \mathcal{P} **refines** \mathcal{Q} ", if each \mathcal{P} -atom is a subset of some \mathcal{Q} -atom. Consequently, each \mathcal{Q} -atom is a union of \mathcal{P} -atoms.

Recall, for δ a real number, our convention that δ^c means $1 - \delta$, in analogy with $\mu(B^c)$ equaling $1 - \mu(B)$ on a probability space.

Distropy Fact

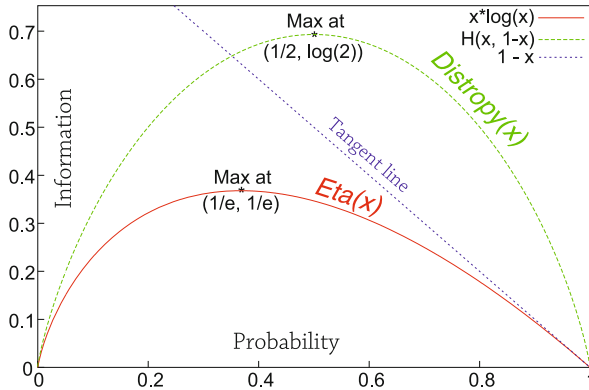
For partitions $\mathcal{P}, \mathcal{Q}, \mathcal{R}$ on probability space (X, μ) :

- (a) $\mathcal{H}(\mathcal{P}) \leq \log(\#\mathcal{P})$, with equality IFF \mathcal{P} is an equi-mass partition.
- (b) $\mathcal{H}(\mathcal{Q} \vee \mathcal{R}) \leq \mathcal{H}(\mathcal{Q}) + \mathcal{H}(\mathcal{R})$, with equality IFF $\mathcal{Q} \perp \mathcal{R}$.
- (c) For $\delta \in [0, \frac{1}{2}]$, the function $\delta \mapsto \mathcal{H}(\delta, \delta^c)$ is strictly increasing.
- (d) $\mathcal{R} \preccurlyeq \mathcal{P}$ implies $\mathcal{H}(\mathcal{R}) \leq \mathcal{H}(\mathcal{P})$, with equality IFF $\mathcal{R} \stackrel{a.e.}{=} \mathcal{P}$.

Proof Use the strict concavity of $\eta(\cdot)$, together with Jensen's inequality. \square

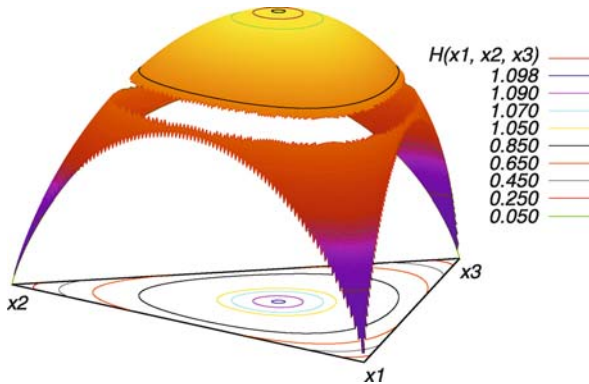
Remark 1 Although we will not discuss it in this paper, most distropy statements remain true with '*partition*' replaced by '*countable partition of finite distropy*'.

⁶Curiosity: Just in this paragraph we compute distropy in **nats**, that is, using natural logarithm. Given a small probability $p \in [0, 1]$ and setting $x := 1/p$, note that $\eta(p) = \frac{\log(x)}{x} \approx 1/\pi(x)$, where $\pi(x)$ denotes the number of prime numbers less-equal x . (This approximation is a weak form of the Prime Number Theorem.) Is there any actual connection between the '*approximate distropy*' function $\mathcal{H}_{\pi}(\vec{p}) := \sum_{p \in \vec{p}} 1/\pi(1/p)$ and Number Theory, other than a coincidence of growth rate?



Entropy in Ergodic Theory, Figure 2

Using natural log, here are the graphs of: $\eta(x)$ in solid red, $\mathcal{H}(x, x^c)$ in dashed green, $1-x$ in dotted blue. Both $\eta(x)$ and $\mathcal{H}(x, x^c)$ are strictly convex-down. The $1-x$ line is tangent to $\eta(x)$ at $x = 1/e$



Entropy in Ergodic Theory, Figure 3

Using natural log: The graph of $\mathcal{H}(x_1, x_2, x_3)$ in barycentric coordinates; a slice has been removed, between $z = 0.745$ and $z = 0.821$. The three arches are copies of the distropy curve from Fig. 2.

Binomial Coefficients

The dartboard gave an example where distropy arises in a natural way. Here is a second example.

For a small $\delta > 0$, one might guess that the binomial coefficient $\binom{n}{\delta n}$ grows asymptotically (as $n \rightarrow \infty$) like $2^{\varepsilon n}$, for some small ε . But what is the correct relation between ε and δ ?

Well, Stirling's formula $n! \approx [n/e]^n$ gives

$$\begin{aligned} \frac{n!}{[\delta n]! [\delta^c n]!} &\approx \frac{n^n}{[\delta n]^{\delta n} [\delta^c n]^{\delta^c n}} \\ &= \frac{1}{[\delta^{\delta n} \delta^c]^{\delta^c n}} \quad (\text{recall } \delta^c = 1 - \delta). \end{aligned}$$

Thus $\frac{1}{n} \log \binom{n}{\delta n} \approx \mathcal{H}(\delta, \delta^c)$. But by means of the above

distropy inequalities, we get an inequality true for *all* n , not just asymptotically.

Lemma 2 (Binomial Lemma) Fix a $\delta \in [0, \frac{1}{2}]$ and let $\mathbf{H} := \mathcal{H}(\delta, \delta^c)$. Then for each $n \in \mathbb{Z}_+$:

$$\sum_{j \in [0 \dots \delta n]} \binom{n}{j} \leq 2^{\mathbf{H}n}. \quad (4)$$

Proof Let $X \subset \{0, 1\}^n$ be the set of \vec{x} with

$$\#\{i \in [1 \dots n] \mid x_i = 1\} \leq \delta n.$$

On X , let P_1, P_2, \dots be the coordinate partitions; e.g. $P_7 = (A_7, A_7^c)$, where $A_7 := \{\vec{x} \mid x_7 = 1\}$. Weighting each point by $\frac{1}{|X|}$, the uniform distribution on X , gives that $\mu(A_7) \leq \delta$. So $\mathcal{H}(P_7) \leq \mathbf{H}$, by (c) in Sect. “Distropy Fact”.

Finally, the join $P_1 \vee \dots \vee P_n$ separates the points of X . So

$$\begin{aligned} \log(\#X) &= \mathcal{H}(P_1 \vee \dots \vee P_n) \\ &\leq \mathcal{H}(P_1) + \dots + \mathcal{H}(P_n) \\ &\leq \mathbf{H}n, \end{aligned}$$

making use of (a),(b) in “Distropy Fact”. And $\#X$ equals LhS (4). \square

A Gander at Shannon's Noisy Channel Theorem

We can restate the Binomial lemma using the *Hamming metric* on $\{0, 1\}^n$,

$$\text{Dist}(\vec{x}, \vec{y}) := \#\{i \in [1 \dots n] \mid x_i \neq y_i\}.$$

Use $\text{Bal}(\vec{x}, r)$ for the open radius- r ball centered at \vec{x} , and

$$\overline{\text{Bal}}(\vec{x}, r) := \{\vec{y} \mid \text{Dist}(\vec{x}, \vec{y}) \leq r\}$$

for the closed ball. The above lemma can be interpreted as saying that

$$|\overline{\text{Bal}}(\vec{x}, \delta n)| \leq 2^{\mathcal{H}(\delta, \delta^c)n}, \quad \text{for each } \vec{x} \in \{0, 1\}^n. \quad (5)$$

Corollary 3 Fix $n \in \mathbb{Z}_+$ and $\delta \in [0, \frac{1}{2}]$, and let

$$\mathbf{H} := \mathcal{H}(\delta, \delta^c).$$

Then there is a set $C \subset \{0, 1\}^n$, with $\#C \geq 2^{[1-\mathbf{H}]n}$, that is *strongly* δn -separated. I.e., $\text{Dist}(\vec{x}, \vec{y}) > \delta n$ for each distinct pair $\vec{x}, \vec{y} \in C$.

Noisy Channel

Shannon's theorem says that a noisy channel has a **channel capacity**. Transmitting *above* this speed, there is a minimum error-rate (depending how much “*above*”) that no error-correcting code can fix. Conversely, one can transmit *below* – but arbitrarily close to – the channel capacity, and encode the data so as to make the error-rate less than any given ε . We use Corollary 3 to show the existence of such codes, in the simplest case where the noise⁷ is a binary independent-process (a “Bernoulli” process, in the language later in this article).

We have a channel which can pass one bit per second. Alas, there is a fixed noise-probability $\nu \in [0, \frac{1}{2})$ so that a bit in the channel is perturbed into the other value. Each perturbation is independent of all others. Let $\mathbf{H} := \mathcal{H}(\nu, \nu^c)$. The value $[1 - \mathbf{H}]$ bits-per-second is the **channel capacity** of this noise-afflicted channel.

Encoding/Decoding Encode using an “ k, n -block-code”; an injective map $F: \{0, 1\}^k \rightarrow \{0, 1\}^n$. The source text is split into consecutive k -bit blocks. A block $\vec{x} \in \{0, 1\}^k$ is encoded to $F(\vec{x}) \in \{0, 1\}^n$ and then sent through the channel, where it comes out perturbed to $\vec{\alpha} \in \{0, 1\}^n$. The *transmission rate* is thus k/n bits per second.

For this example, we fix a radius $r > 0$ to determine the decoding map,

$$D_r: \{0, 1\}^n \rightarrow \{\text{Oops}\} \sqcup \{0, 1\}^k.$$

We set $D_r(\vec{\alpha})$ to \vec{z} if there is a *unique* \vec{z} with $F(\vec{z}) \in \text{Bal}(\vec{\alpha}, r)$; else, set $D_r(\vec{\alpha}) := \text{Oops}$.

One can think of the noise as a $\{0, 1\}$ -independent-process, with $\text{Prob}(1) = \nu$, which is added mod-2 to the signal-process. Suppose we can arrange that the set $\{F(\vec{x}) \mid \vec{x} \in \{0, 1\}^k\}$ of codewords, is a strongly r -separated-set. Then

The probability that a block is mis-decoded is the probability, flipping a ν -coin n times that we get more than r many Heads. (6)

Theorem 4 (Shannon) Fix a noise-probability $\nu \in [0, \frac{1}{2})$ and let $\mathbf{H} := \mathcal{H}(\nu, \nu^c)$. Consider a rate $R < [1 - \mathbf{H}]$ and an $\varepsilon > 0$. Then $\forall_{\text{large } n}$ there exists a k and a code $F: \{0, 1\}^k \rightarrow \{0, 1\}^n$ so that: The F -code transmits bits at faster than R bits-per-second, and with error-rate $< \varepsilon$.

⁷The noise-process is assumed to be *independent* of the signal-process. In contrast, when the perturbation is highly dependent on the signal, then it is sometimes called **distortion**.

Proof Let $\mathbf{H}' := \mathcal{H}(\delta, \delta^c)$, where $\delta > \nu$ was chosen so close to ν that

$$\delta < \frac{1}{2} \quad \text{and} \quad 1 - \mathbf{H}' > R. \quad (7)$$

Pick a large n for which

$$\frac{k}{n} > R, \quad \text{where} \quad k := \lfloor [1 - \mathbf{H}']n \rfloor. \quad (8)$$

By Corollary 3, there is a strongly δn -separated-set $C \subset \{0, 1\}^n$ with $\#C \geq 2^{[1 - \mathbf{H}']n}$. So C is big enough to permit an injection $F: \{0, 1\}^k \rightarrow C$. Courtesy Eq. (6), the probability of a decoding error is that of getting more than δn many Heads in flipping a ν -coin n times. Since $\delta > \nu$, the Weak Law of Large Numbers guarantees – once n is large enough – that this probability is less than the given ε . \square

The Information Function

Agree to use $\mathbf{P} = (A_1, \dots)$, $\mathbf{Q} = (B_1, \dots)$, $\mathbf{R} = (C_1, \dots)$ for partitions, and \mathcal{F}, \mathcal{G} for fields.

With \mathcal{C} a (finite or infinite) family of subfields of \mathcal{X} , their *join* $\bigvee_{\mathcal{G} \in \mathcal{C}} \mathcal{G}$ is the smallest field \mathcal{F} such that $\mathcal{G} \subset \mathcal{F}$, for each $\mathcal{G} \in \mathcal{C}$. A partition \mathbf{Q} can be interpreted also as a field; namely, the field of unions of its atoms. A join of denumerably many partitions will be interpreted as a field, but a join of *finitely* many, $\mathbf{P}_1 \vee \dots \vee \mathbf{P}_N$, will be viewed as a partition *or* as a field, depending on context.

Conditioning a partition \mathbf{P} on a positive-mass set B , let $\mathbf{P}|B$ be the probability vector $A \mapsto \frac{\mu(A \cap B)}{\mu(B)}$. Its distropy is

$$\mathcal{H}(\mathbf{P}|B) = \sum_{A \in \mathbf{P}} \log \left(\frac{1}{\mu(A \cap B)/\mu(B)} \right) \frac{\mu(A \cap B)}{\mu(B)}.$$

So conditioning \mathbf{P} on a *partition* \mathbf{Q} gives conditional distropy

$$\begin{aligned} \mathcal{H}(\mathbf{P}|\mathbf{Q}) &= \sum_{B \in \mathbf{Q}} \mathcal{H}(\mathbf{P}|B) \mu(B) \\ &= \sum_{A \in \mathbf{P}, B \in \mathbf{Q}} \log \left(\frac{1}{\mu(A \cap B)/\mu(B)} \right) \mu(A \cap B). \end{aligned} \quad (9)$$

A “dartboard” interpretation of $\mathcal{H}(\mathbf{P}|\mathbf{Q})$ is

The expected number of questions to ascertain the \mathbf{P} -atom that a random dart $x \in X$ fell in, given that we are told its \mathbf{Q} -atom.

For a set $A \subset X$, use $\mathbf{1}_A : X \rightarrow \{0, 1\}$ for its *indicator function*; $\mathbf{1}_A(x) = 1$ IFF $x \in A$. The **information function** of partition P , a map $\mathcal{I}_P : X \rightarrow [0, \infty)$, is

$$\mathcal{I}_P := \sum_{A \in P} \log \left(\frac{1}{\mu(A)} \right) \mathbf{1}_A(\cdot). \quad (10)$$

The information function has been defined so that its expectation is the distropy of P ,

$$\mathbb{E}(\mathcal{I}_P) = \int_X \mathcal{I}_P(\cdot) d\mu = \mathcal{H}(P).$$

Conditioning on a Field

For a subfield $\mathcal{F} \subset \mathcal{X}$, recall that each function $g \in \mathbb{L}^1(\mu)$ has a *conditional expectation* $\mathbb{E}(g|\mathcal{F}) \in \mathbb{L}^1(\mu)$. It is the *unique* \mathcal{F} -measurable function with

$$\forall B \in \mathcal{F}: \int_B \mathbb{E}(g|\mathcal{F}) d\mu = \int_B g d\mu.$$

Returning to distropy ideas, use $\mu(A|\mathcal{F})$ for the **conditional probability** function; it is the conditional expectation $\mathbb{E}(\mathbf{1}_A|\mathcal{F})$. So the **conditional information function** is

$$\mathcal{I}_{P|\mathcal{F}}(x) := \sum_{A \in P} \log \left(\frac{1}{\mu(A|\mathcal{F}(x))} \right) \mathbf{1}_A(x). \quad (11)$$

Its integral

$$\mathcal{H}(P|\mathcal{F}) := \int \mathcal{I}_{P|\mathcal{F}} d\mu,$$

is the **conditional distropy** of P on \mathcal{F} .

When \mathcal{F} is the field of unions of atoms from some partition Q , then the number $\mathcal{H}(P|\mathcal{F})$ equals the $\mathcal{H}(P|Q)$ from Eq. (9).

Write $\mathcal{G}_j \nearrow \mathcal{F}$ to indicate that fields $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \dots$ are nested, and that $\text{Fld}(\bigcup_1^\infty \mathcal{G}_j) = \mathcal{F}$, a.e. The Martingale Convergence Theorem (p. 103 in [20]) gives (c) below.

Conditional-Distropy Fact Consider partitions P, Q, R and fields \mathcal{F} and \mathcal{G}_j . Then

- (a) $0 \leq \mathcal{H}(P|\mathcal{F}) \leq \mathcal{H}(P)$, with equality IFF $P \stackrel{a.e.}{\subset} \mathcal{F}$, respectively, $P \perp \mathcal{F}$.
- (b) $\mathcal{H}(Q \vee R|\mathcal{F}) \leq \mathcal{H}(Q|\mathcal{F}) + \mathcal{H}(R|\mathcal{F})$.
- (c) Suppose $\mathcal{G}_j \nearrow \mathcal{F}$. Then $\mathcal{H}(P|\mathcal{G}_j) \searrow \mathcal{H}(P|\mathcal{F})$.
- (d) $\mathcal{H}(Q \vee R) = \mathcal{H}(Q|R) + \mathcal{H}(R)$.
- (d') $\mathcal{H}(Q \vee R_1|R_0) = \mathcal{H}(Q|R_1 \vee R_0) + \mathcal{H}(R_1|R_0)$.

Imagining our dartboard (Fig. 1) divided by superimposed partitions Q and R , equality (d) can be interpreted as saying: ‘You can efficiently discover where the dart landed

in both partitions, by first asking efficient questions about R , then – based on where it landed in R – asking intelligent questions about Q .’

Entropy of a Process

Consider a transformation $(T: X, \mu)$ and partition $P = (A_1, A_2, \dots)$. Each “time” n determines a partition $P_n := T^n P$, whose j th-atom is $T^{-n}(A_j)$. The **process** T, P refers to how T acts on the subfield $\bigvee_0^\infty P_n \subset \mathcal{X}$. (An alternative view of a process is as a stationary sequence V_0, V_1, \dots of random variables $V_n: X \rightarrow \mathbb{Z}_+$, where $V_n(x) := j$ because x is in the j th-atom of P_n .)

Write $\mathcal{E}(T, P)$ or $\mathcal{E}^T(P)$ for the “**entropy**” of the T, P process”. It is the limit of the **conditional-distropy-numbers**

$$c_n := \mathcal{H}(P_0|P_1 \vee P_2 \vee \dots \vee P_{n-1}).$$

This limit exists since $\mathcal{H}(P) = c_1 \geq c_2 \geq \dots \geq 0$.

Define the **average-distropy-number** $\frac{1}{n}h_n$, where

$$h_n := \mathcal{H}(P_0 \vee P_1 \vee \dots \vee P_{n-1}).$$

Certainly $h_n = c_n + \mathcal{H}(P_1 \vee \dots \vee P_{n-1}) = c_n + h_{n-1}$, since T is measure preserving. Induction gives $h_n = \sum_{j=1}^n c_j$. So the Cesàro averages $\frac{1}{n}h_n$ converge to the entropy.

Theorem 5 The entropy of process $(T, P: X, \mathcal{X}, \mu)$ equals

$$\begin{aligned} \lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{H}(P_0 \vee \dots \vee P_{n-1}) \\ = \lim_{n \rightarrow \infty} \mathcal{H} \left(P_0 \middle| \bigvee_1^n P_j \right) = \mathcal{H} \left(P_0 \middle| \bigvee_1^\infty P_j \right). \end{aligned}$$

Both limits are non-increasing. The entropy $\mathcal{E}^T(P) \geq 0$, with equality IFF $P \subset \bigvee_1^\infty P_j$. And $\mathcal{E}^T(P) \leq \mathcal{H}(P)$, with equality IFF T, P is an independent process.

Generators

We henceforth only discuss *invertible* mpts, that is, when T^{-1} is itself an mpt. Viewing the atoms of P as “letters”, then, each $x \in X$ has a T, P -**name**

$$\dots x_{-2}x_{-1}x_0x_1x_2x_3\dots,$$

where x_n is $P(T^n(x))$, the P -letter owning $T^n(x)$.

A partition P **generates** (the whole field) under $(T: X, \mu)$, if $\bigvee_{-\infty}^\infty T^n P =_\mu \mathcal{X}$. It turns out⁸ that

⁸I am now at liberty to reveal that our X has always been a **Lebesgue space**, that is, measure-isomorphic to an interval of \mathbb{R} to-

P generates IFF P **separates points**. That is, after deleting a (T -invariant) nullset, distinct points of X have distinct T, P -names.

A finite set $[1 \dots L]$ of integers, interpreted as an **alphabet**, yields the **shift space** $X := [1 \dots L]^{\mathbb{Z}}$ of doubly-infinite sequences $x = (\dots x_{-1}x_0x_1 \dots)$. The **shift** $T: X \rightarrow X$ acts on X by

$$T(x) := [n \mapsto x_{n+1}] .$$

The time-zero partition P separates points, under the action of the shift. This L -atom **time-zero partition** has $P\langle x \rangle = P\langle y \rangle$ IFF $x_0 = y_0$. So no matter what shift-invariant measure is put on X , the time-zero partition will generate under the action of T .

Time Reversibility

A transformation need not be isomorphic to its inverse. Nonetheless, the average-distropy-numbers show that $\mathcal{E}(T^{-1}, P) = \mathcal{E}(T, P)$; although this is not obvious from the conditioning-definition of entropy. Alternatively,

$$\begin{aligned} \mathcal{H}\left(P_0 \middle| \bigvee_1^n P_j\right) &= \mathcal{H}(P_0 \vee \dots \vee P_n) - \mathcal{H}(P_1 \vee \dots \vee P_n) \\ &= \mathcal{H}(P_{-n} \vee \dots \vee P_0) - \mathcal{H}(P_{-n} \vee \dots \vee P_{-1}) \\ &= \mathcal{H}\left(P_0 \middle| \bigvee_{-1}^{-n} P_j\right) . \end{aligned} \quad (12)$$

Bernoulli Processes

A probability vector $\vec{v} := (v_1, \dots, v_L)$ can be viewed as a measure on alphabet $[1 \dots L]$. Let $\mu_{\vec{v}}$ be the resulting product measure on $X := [1 \dots L]^{\mathbb{Z}}$, with T the shift on X and P the time-zero partition. The independent process $(T, P: X, \mu_{\vec{v}})$ is called, by ergodic theorists, a **Bernoulli process**. Not necessarily consistently, we tend to refer to the underlying transformation as a **Bernoulli shift**.

The $(\frac{1}{2}, \frac{1}{2})$ -Bernoulli and the $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ -Bernoulli have different process-entropies, but perhaps their underlying transformations are isomorphic? Prior to the Kol-

gether with countably many point-atoms (points with positive mass). The equivalence of *generating* and *separating* is a technical theorem, due to Rokhlin.

Assuming μ to be Lebesgue is not much of a limitation. For instance, if μ is a finite measure on any Polish space, then μ extends to a Lebesgue measure on the μ -completion of the Borel sets. To not mince words: All spaces are Lebesgue spaces unless you are actively looking for trouble.

mogorov–Sinai definition of entropy⁹ of a transformation, this question remained unanswered.

Entropy of a Transformation

The Kolmogorov–Sinai definition of the entropy of an mpt is

$$\mathcal{E}(T) := \sup\{\mathcal{E}^T(Q) \mid Q \text{ a partition on } X\} .$$

Certainly entropy is an isomorphism invariant – but is it useful? After all, the supremum of *distropies* of partitions is always infinite (on non-atomic spaces) and one might fear that the same holds for entropies. The key observation (re-stated in Lemma 8c and proved below) was this, from [4] and [7].

Theorem 6 (Kolmogorov–Sinai Theorem) *If P generates under T , then $\mathcal{E}(T) = \mathcal{E}(T, P)$.*

Thereupon the $(\frac{1}{2}, \frac{1}{2})$ and $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ Bernoulli-shifts are *not* isomorphic, since their respective entropies are $\log(2) \neq \log(3)$.

Wolfgang Krieger later proved a converse to the Kolmogorov–Sinai theorem.

Theorem 7 (Krieger Generator Theorem, 1970) *Suppose T ergodic. If $\mathcal{E}(T) < \infty$, then T has a generating partition. Indeed, letting K be the smallest integer $K > \mathcal{E}(T)$, there is a K -atom generator.¹⁰*

Proof See Rudolph [21], or § 5.1 in ► [Joinings in Ergodic Theory](#), where Krieger’s theorem is stated in terms of joinings. \square

Entropy Is Continuous

Given *ordered* partitions $Q = (B_1, \dots)$ and $Q' = (B'_1, \dots)$, extend the shorter by null-atoms until $|Q| = |Q'|$. Let $Fat := \bigsqcup_j [B_j \cap B'_j]$; this set should have mass close to 1 if Q and Q' are almost the same partition. Define a new partition

$$Q \triangle Q' := \{Fat\} \sqcup \{B_i \cap B'_j \mid \text{with } i \neq j\} .$$

(In other words, take $Q \vee Q'$ and coalesce, into a single atom, all the $B_k \cap B'_k$ sets.) Topologize the space of parti-

⁹This is sometimes called **measure(-theoretic) entropy** or (perhaps unfortunately) **metric entropy**, to distinguish it from topological entropy. Tools known prior to entropy, such as *spectral* properties, did *not* distinguish the two Bernoulli-shifts; see ► [Spectral Theory of Dynamical Systems](#) for the definitions.

¹⁰It is an easier result, undoubtedly known much earlier, that every ergodic T has a *countable* generating partition – possibly of ∞ -distropy.

tions by saying¹¹ that $Q^{(L)} \rightarrow Q$ when $\mathcal{H}(Q \Delta Q^{(L)}) \rightarrow 0$. Then Lemma 8b says that process-entropy varies continuously with varying the partition.

Lemma 8 Fix a mpt $(T: X, \mu)$. For partitions P, Q, Q' , define $R := Q \Delta Q'$ and let $\delta := \mathcal{H}(R)$. Then

- (a) $|\mathcal{H}(Q) - \mathcal{H}(Q')| \leq \delta$. (Distropy varies continuously with the partition.)
- (b) $|\mathcal{E}^T(Q) - \mathcal{E}^T(Q')| \leq \delta$. (Process-entropy varies continuously with the partition.)
- (c) For all partitions $Q \subset \text{Fld}(T, P)$: $\mathcal{E}^T(Q) \leq \mathcal{E}^T(P)$.

Proof (of (a)) Evidently $Q' \vee R = Q' \vee Q = Q \vee R$. So $\mathcal{H}(Q') \leq \mathcal{H}(Q \vee R) \leq \mathcal{H}(Q) + \delta$. \square

Proof (of (b)) As above,

$$\mathcal{H}\left(\bigvee_1^N Q'_j\right) \leq \mathcal{H}\left(\bigvee_1^N Q_j\right) + \mathcal{H}\left(\bigvee_1^N R_j\right).$$

Sending $N \rightarrow \infty$ gives $\mathcal{E}^T(Q') \leq \mathcal{E}^T(Q) + \mathcal{E}^T(R)$. Finally, $\mathcal{E}^T(R) \leq \mathcal{H}(R)$ and so $\mathcal{E}^T(Q') \leq \mathcal{E}^T(Q) + \delta$. \square

Proof (of (c)) Let $K := |Q|$. Then there is a sequence of K -set partitions $Q^{(L)} \rightarrow Q$ with $Q^{(L)} \leq \bigvee_{-L}^L P_\ell$. By above, $\mathcal{E}^T(Q^{(L)}) \rightarrow \mathcal{E}^T(Q)$, so showing that

$$\mathcal{E}^T\left(\bigvee_{-L}^L P_\ell\right) \stackrel{?}{\leq} \mathcal{E}^T(P)$$

will suffice. Note that

$$h_N := \mathcal{H}\left(\bigvee_{n=0}^{N-1} T^n\left(\bigvee_{-L}^L P_\ell\right)\right) = \mathcal{H}\left(\bigvee_{j=-L}^{N-1+L} P_j\right).$$

So $\frac{1}{N} h_N \leq \frac{1}{N} \mathcal{H}\left(\bigvee_0^{N-1} P_j\right) + \frac{1}{N} 2L\mathcal{H}(P)$. Now send $N \rightarrow \infty$. \square

Entropy Is Not Continuous

The most common topology placed on the space Ω of mpts is the *coarse topology*¹² that Halmos discusses in his “little red book” [14].

¹¹On the set of ordered K -set partitions (with K fixed) this convergence is the same as: $Q^{(L)} \rightarrow Q$ when $\mu(\text{Fat}(Q^{(L)}, Q)) \rightarrow 1$.

An alternative approach is the *Rokhlin metric*, $\text{Dist}(P, Q) := \mathcal{H}(P|Q) + \mathcal{H}(Q|P)$, which has the advantage of working for *un-ordered* partitions.

¹²i.e., $S_n \rightarrow T$ IFF $\forall A \in \mathcal{X}: \mu(S_n^{-1}(A) \Delta T^{-1}(A)) \rightarrow 0$; this is a metric-topology, since our probability space is countably generated. This can be restated in terms of the unitary operator U_T on $\mathbb{L}^2(\mu)$, where $U_T(f) := f \circ T$. Namely, $S_n \rightarrow T$ in the coarse topology IFF $U_{S_n} \rightarrow U_T$ in the strong operator topology.

The Rokhlin lemma (see p. 33 in [21]) implies that the isomorphism-class of *each* ergodic mpt is *dense* in Ω , (e.g., see p. 77 in [14]) disclosing that the $S \mapsto \mathcal{E}(S)$ map is ex-orbitantly discontinuous.

Indeed, the failure happens already for process-entropy with respect to a fixed partition. A Bernoulli process T, P has positive entropy. Take mpts $S_n \rightarrow T$, each isomorphic to an irrational rotation. Then each $\mathcal{E}(S_n, P)$ is zero, as shown in the later section “[Determinism and Zero-Entropy](#)”.

Further Results When \mathcal{F} is a T -invariant subfield, agree to use $T \upharpoonright_{\mathcal{F}}$ for “ T restricted to \mathcal{F} ”, which is a factor (see Glossary) of T . Transformations T and S are **weakly isomorphic** if each is isomorphic to a factor of the other.

The foregoing entropy tools make short shrift of the following.

Lemma 9 (Entropy Lemma) Consider T -invariant subfields \mathcal{G}_j and \mathcal{F} .

- (a) Suppose $\mathcal{G}_j \nearrow \mathcal{F}$. Then $\mathcal{E}T \upharpoonright_{\mathcal{G}_j} \nearrow \mathcal{E}(T \upharpoonright_{\mathcal{F}})$. In particular, $\mathcal{G} \subset \mathcal{F}$ implies that $\mathcal{E}(T \upharpoonright_{\mathcal{G}}) \leq \mathcal{E}(T \upharpoonright_{\mathcal{F}})$, so entropy is an invariant of weak-isomorphism.
- (b) $\mathcal{E}(T \upharpoonright_{\mathcal{G}_1 \vee \mathcal{G}_2 \vee \dots}) \leq \sum_j \mathcal{E}(T \upharpoonright_{\mathcal{G}_j})$.
And $\mathcal{E}(T, Q_1 \vee Q_2 \vee \dots) \leq \sum_j \mathcal{E}(T, Q_j)$.
- (c) For mpts $(S_j: Y_j, \nu_j)$: $\mathcal{E}(S_1 \times S_2 \times \dots) = \sum_j \mathcal{E}(S_j)$.
- (d) $\mathcal{E}(T^{-1}) = \mathcal{E}(T)$. More generally, $\mathcal{E}(T^n) = |n| \cdot \mathcal{E}(T)$.

Meshalkin’s Map

In the wake of Kolmogorov’s 1958 entropy paper, for two Bernoulli-shifts to be isomorphic one now knew that they had to have equal entropies. Meshalkin provided the first non-trivial example in 1959 [45].

Let $S: Y \rightarrow Y$ be the Bernoulli-shift over the “letter” alphabet $\{E, D, P, N\}$, with probability distribution $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. The letters E, D, P, N stand for *Even, oDd, Positive, Negative*, and will be used to describe the code (isomorphism) between the processes.

Use $T: X \rightarrow X$ for the Bernoulli-shift over “digit” alphabet $\{0, +1, -1, +2, -2\}$, with probability distribution $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$. Both distributions $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ and $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ have distropy $\log(4)$.

After deleting invariant nullsets from X and Y , the construction will produce a measure-preserving isomorphism $\psi: X \rightarrow Y$ so that $T \circ \psi = \psi \circ S$.

The Code In X , consider this point x :

... 0 0 0 −1 0 0 +1 +2 −1 +1 0 ...

Regard each 0 as a left-parenthesis, and each non-zero as a right-parenthesis. Link them according to the legal way of matching parentheses, as shown in the top row, below:

$$\begin{array}{cccccccccccc} \hline 0 & 0 & 0 & -1 & 0 & 0 & +1 & +2 & -1 & +1 & 0 \\ \hline P & N & N & D & P & P & D & E & D & D & ? \end{array}$$

The leftmost 0 is linked to the rightmost +1, as indicated by the longest-overbar. The left/right-parentheses form a $(\frac{1}{2}, \frac{1}{2})$ -random-walk. Since this random walk is recurrent, every position in x will be linked (except for a nullset of points x).

Below each 0, write “P” or “N” as the 0 is linked to a *positive* or *negative* digit. And below the other digits, write “E” or “D” as the digit is *even* or *odd*. So the upper name in X is mapped to the lower name, a point $y \in Y$.

This map $\psi: X \rightarrow Y$ carries the upstairs $(\frac{1}{2}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8}, \frac{1}{8})$ distribution to $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, downstairs. It takes some arguing to show that independence is preserved.

The inverse map ψ^{-1} views D and E as right-parentheses, and P and N as left. Above D, write the odd digit +1 or -1, as this D is linked to Positive or Negative.

Markov Shifts

A Bernoulli process T, P has independence $P_{(-\infty \dots 0]} \perp P_1$ whereas a **Markov process** is a bit less aloof:

The infinite Past $P_{(-\infty \dots 0]}$ doesn’t provide any more information about Tomorrow than Today did.

That is, the conditional distribution $P_1|P_{(-\infty \dots 0]}$ equals $P_1|P_0$. Equivalently,

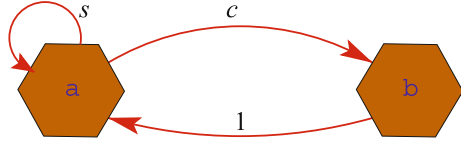
$$\mathcal{H}(P_1|P_0) = \mathcal{H}(P_1|P_{(-\infty \dots 0]}) \stackrel{\text{note}}{=} \mathcal{E}(T, P). \quad (13)$$

The simplest non-trivial Markov process $(T, P: X, \mu)$ is over a two-letter alphabet $\{a, b\}$, and has transition graph Fig. 4, for some choice of transition probabilities s and c . The graph’s Markov matrix is

$$M = [m_{i,j}]_{i,j} = \begin{bmatrix} s & c \\ 1 & 0 \end{bmatrix},$$

where $c = 1 - s$, and $m_{i,j}$ denotes the probability of going from state i to state j .

If Today’s distribution on the two states is the probability-vector $\vec{v} := [p_a \ p_b]$, then Tomorrow’s is the product $\vec{v}M$. So a *stationary* process needs $\vec{v}M = \vec{v}$. This equation has the unique solution $p_a = \frac{1}{1+c}$ and $p_b = \frac{c}{1+c}$. An example of computing the probability of a word or



Entropy in Ergodic Theory, Figure 4

Call the transition probabilities $s := \text{Prob}(a \rightarrow a)$ for stay, and $c := \text{Prob } a \rightarrow b$ for change. These are non-negative reals, and $s + c = 1$

cylinder set; (see Sect. “The Carathéodory Construction” in [► Measure Preserving Systems](#)) in the process, is

$$\begin{aligned} \mu_s(\text{baaaba}) &= p_b m_{ba} m_{aa} m_{aa} m_{ab} m_{ba} \\ &= \frac{c}{1+c} \cdot 1 \cdot s \cdot s \cdot c \cdot 1. \end{aligned}$$

The subscript on μ_s indicates the dependence on the transition probabilities; let’s also mark the mpt and call it T_s . Using Eq. (13), the entropy of our Markov map is

$$\begin{aligned} \mathcal{E}(T_s) &= p_a \cdot \mathcal{H}(s, c) + p_b \cdot \underbrace{\mathcal{H}(1, 0)}_{=0} \\ &= \frac{-1}{1+c} \cdot [s \log(s) + c \log(c)]. \end{aligned} \quad (14)$$

Determinism and Zero-Entropy

Irrational rotations have zero-entropy; let’s reveal this in two different ways.

Equip $X := [0, 1)$ with “length” (Lebesgue) measure and wrap it into a circle. With “ \oplus ” denoting addition mod-1, have $T: X \rightarrow X$ be the rotation $T(x) := x \oplus \alpha$, where the rotation number α is irrational. Pick distinct points $y_0, z_0 \in X$, and let P be the partition whose two atoms are the intervals $[y_0, z_0)$ and $[z_0, y_0)$, wrapping around the circle.

The T -orbit of each point x is dense¹³ in X . In particular, y_0 has dense orbit, so P separates points – hence generates – under T . Our goal, thus, is $\mathcal{E}(T, P) \stackrel{?}{=} 0$.

Rotations are Deterministic

The forward T -orbit of each point is dense. This is true for y_0 , and so the *backward* T, P -name of each x actually tells us which point x is. I. e., $P \subset \bigvee_{-\infty}^{-1} T^n P$, which is our definition of “process T, P is **deterministic**”. Our P being finite, this determinism implies that $\mathcal{E}(T, P)$ is zero, by Theorem 5.

¹³Fix an $\varepsilon > 0$ and an $N > 1/\varepsilon$. Points $x, T(x), \dots, T^N(x)$ have some two at distance less than $\frac{1}{N}$; say, $\text{Dist}(T^i(x), T^j(x)) < \varepsilon$, for some $0 \leq i < j \leq N$. Since T is an isometry, $\varepsilon > \text{Dist}(x, T^k(x)) > 0$, where $k := j - i$. So the T^k -orbit of x is ε -dense.

Counting Names in a Rotation The $P_0 \vee \dots \vee P_{n-1}$ partition places n translates of points y_0 and of z_0 , cutting the circle into at most $2n$ intervals. Thus $\mathcal{H}(P_0 \vee \dots \vee P_{n-1}) \leq \log(2n)$. And $\frac{1}{n} \log(2n) \rightarrow 0$.

Alternatively, the below SMB-theorem implies, for an ergodic process T, P , that the number of length- n names is approximately $2^{\mathcal{E}(T,P) \cdot n}$, this, after discarding small mass from the space. But the growth of $n \mapsto 2n$ is sub-exponential and so, for our rotation, $\mathcal{E}(T, P)$ must be zero.

Theorem 10 (Shannon–McMillan–Breiman Theorem (SMB-Theorem)) Set $E := \mathcal{E}(T, P)$, where the tuple $(T, P: X, \mu)$ is an ergodic process. Then the average information function

$$\frac{1}{n} \mathcal{I}_{P_{[0..n)}}(x) \xrightarrow{n \rightarrow \infty} E, \quad \text{for a.e. } x \in X. \quad (15)$$

The functions $f_n := \mathcal{I}_{P_{[0..n)}}(x)$ converge to the constant function E both in the \mathbb{L}^1 -norm and in probability.¹⁴

Proof See the texts of Karl Petersen (p. 261 in [20]), or Dan Rudolph (p. 77 in [21]). \square

Consequences Recall that $P_{[0..n)}$ means $P_0 \vee P_1 \vee \dots \vee P_{n-1}$, where $P_j := T^j P$. As usual, $P_{[0..n)}\langle x \rangle$ denotes the $P_{[0..n)}$ -atom owning x .

Having deleted a nullset, we can restate Eq. (15) to now say that $\forall \varepsilon, \forall x, \forall_{\text{large } n}$:

$$1/2^{[E+\varepsilon]n} \leq \mu(P_{[0..n)}\langle x \rangle) \leq 1/2^{[E-\varepsilon]n}. \quad (16)$$

This has the following consequence. Fixing a number $\delta > 0$, we consider any set with $\mu(B) \geq \delta$ and count the number of n -names of points in B . The SMB-Thm implies

$$\forall \varepsilon, \forall_{\text{large } n}, \forall B \geq \delta: |\{n\text{-names in } B\}| \geq 2^{[E-\varepsilon]n}. \quad (17)$$

Rank-1 Has Zero-Entropy

There are several equivalent definitions for “rank-1 transformation”, several of which are discussed in the introduction of [28]. (See Chap. 6 in [13] as well as [51] and [27] for examples of stacking constructions.)

A **rank-1 transformation** $(T: X, \mu)$ admits a generating partition P and a sequence of Rokhlin stacks $S_n \subset X$, with heights going to ∞ , and with $\mu(S_n) \rightarrow 1$. Moreover, each of these Rokhlin stacks is P -monochromatic, that is, each level of the stack lies entirely in some atom of P .

Taking a stack of some height $2n$, let $B = B_n$ be the union of the bottom n levels of the stack. There are at most

n many length- n names starting in B_n , by monochromaticity. Finally, $\mu(B_n)$ is almost $\frac{1}{2}$, so is certainly larger than $\delta := \frac{1}{3}$. Thus Eq. (17) shows that our rank-1 T has zero entropy.

Cautions on Determinism’s Relation to Zero-Entropy

A finite-valued process T, P has zero-entropy iff $P \subset \bigvee_{-\infty}^{-1} P_j$. Iterating gives

$$\bigvee_0^\infty P_j \subset \bigvee_{-\infty}^{-1} P_j,$$

i. e., the future is measurable with respect to the past.

This was the case with the rotation, where a point’s past uniquely identified the point, thus telling us its future.

While determinism and zero-entropy mean the same thing for finite-valued processes, this fails catastrophically for real-valued (i. e., continuum-valued) processes, as shown by an example of the author’s. A stationary real-valued process $V = \dots V_{-1} V_0 V_1 V_2 \dots$ is constructed in [40] which is simultaneously

strongly deterministic: The two values V_0, V_1 determine all of V , future and past.

and **non-consecutively independent**. This latter means that for each bi-infinite increasing integer sequence $\{n_j\}_{j=-\infty}^\infty$ with no consecutive pair (always $1 + n_j < n_{j+1}$), then the list of random variables $\dots V_{n_{-1}} V_{n_0} V_{n_1} V_{n_2} \dots$ is an independent process.

Restricting the random variables to be countably-valued, how much of the example survives? Joint work with Kalikow [39] produced a countably-valued stationary V which is non-consecutively independent as well as deterministic. (Strong determinism is ruled out, due to cardinality considerations.) A side-effect of the construction is that V ’s time-reversal $n \mapsto V_{-n}$ is **not** deterministic.

The Pinsker–Field and K-Automorphisms

Consider the collection of **zero-entropy sets**,

$$\mathcal{Z} = \mathcal{Z}_T := \{A \in \mathcal{X} \mid \mathcal{E}(T, (A, A^c)) = 0\}. \quad (18)$$

Courtesy of Lemma 9b, \mathcal{Z} is a T -invariant field, and

$$\forall Q \subset \mathcal{Z}: \mathcal{E}(T, Q) = 0. \quad (19)$$

The **Pinsker field**¹⁵ of T is this \mathcal{Z} . It is maximal with respect to Eq. (19). Unsurprisingly, the **Pinsker factor** $T \upharpoonright_{\mathcal{Z}}$ has

¹⁴In engineering circles, this is called the Almost-everywhere equipartition theorem.

¹⁵Traditionally, this called the *Pinsker algebra* where, in this context, “algebra” is understood to mean “ σ -algebra”.

zero entropy, that is, $\mathcal{E}(T \upharpoonright_{\mathcal{Z}}) = 0$. A transformation T is said to have **completely-positive entropy** if it has no (non-trivial) zero-entropy factors. That is, its Pinsker field \mathcal{Z}_T is the trivial field, $\emptyset := \{\emptyset, X\}$.

K-Processes

Kolmogorov introduced the notion of a **K-process** or **Kolmogorov process**, in which the present becomes asymptotically independent of the distant past. The asymptotic past of the T, P process is called its **tail field**, where

$$\text{Tail}(T, P) := \bigcap_{M=1}^{\infty} \bigvee_{j=-\infty}^{-M} P_j.$$

This T, P is a K-process if $\text{Tail}(T, P) = \emptyset$. This turns out to be equivalent to what we might call a strong form of “sensitive dependence on initial conditions”: For each fixed length L , the distant future

$$\bigvee_{j \in (G \dots G+L]} P_j$$

becomes more and more independent of

$$\bigvee_{j \in (-\infty \dots 0]} P_j,$$

as the gap $G \rightarrow \infty$.

A transformation T is a **Kolmogorov automorphism** if it possesses a generating partition P for which T, P is a K-process.

Here is a theorem that relates the “asymptotic forgetfulness” of $\text{Tail}(T, P) = \emptyset$, to the lack of determinism implied by having no zero-entropy factors (See Walters, [23], p. 113. Related results appear in Berg [44]).

Theorem 11 (Pinsker-Algebra Theorem) *Suppose P is a generating partition for an ergodic T . Then $\text{Tail}(T, P)$ equals \mathcal{Z}_T .*

Since \mathcal{Z}_T does not depend on P , this means that all generating partitions have the same tail field, and therefore K-ness of T can be detected from any generator.

Another non-evident fact follows from the above. The **future field** of T, P is defined to be $\text{Tail}(T^{-1}P)$. It is not obvious that if the present is independent of the distant past, then it is automatically independent of the distant future. (Indeed, the precise definitions are important; witness the *Cautions on determinism* section.) But since the entropy of a process, $\mathcal{E}(T, (A, A^c))$, equals the entropy of the time-reversed process $\mathcal{E}(T^{-1}, (A, A^c))$, it follows that \mathcal{Z}_T equals $\mathcal{Z}_{T^{-1}}$.

Ornstein Theory

In 1970, Don Ornstein [46] solved the long-standing problem of showing that entropy was a complete isomorphism-invariant of Bernoulli transformations; that is, that two independent processes with same entropy necessarily have the same underlying transformation. (Earlier, Sinai [52]) had shown that two such Bernoulli maps were **weakly isomorphic**, that is, each isomorphic to a factor of the other.)

Ornstein introduced the notion of a process being **finitely determined**, see [46] for a definition, proved that a transformation T was Bernoulli IFF it had a finitely-determined generator IFF every partition was finitely-determined with respect to T , and showed that entropy completely classified the finitely-determined processes up to isomorphism.

This seminal result led to a vast machinery for proving transformations to be Bernoulli, as well as classification and structure theorems [47, 51, 53]. Showing that the class of K-automorphisms far exceeds the Bernoulli maps, Ornstein and Shields produced in [48] an uncountable family of non-isomorphic K-automorphisms all with the same entropy.

Topological Entropy

Adler, Konheim and McAndrew, in 1965, published the first definition of *topological entropy* in the eponymous article [33]. Here, $T: X \rightarrow X$ is a continuous self-map of a compact topological space. The role of atoms is played by open sets. Instead of a finite partition, one uses a finite¹⁶ **open-cover** $\mathcal{V} = \{U_j\}_{j=1}^L$, i. e. each **patch** U_j is open, and their union $\bigcup(\mathcal{V}) = X$. (Henceforth, ‘cover’ means “open cover”.)

Let $\text{Card}(\mathcal{V})$ be the *minimum* cardinality over all subcovers.

$$\text{Card}(\mathcal{V}) := \text{Min} \{ \# \mathcal{V}' \mid \mathcal{V}' \subset \mathcal{V} \text{ and } \bigcup(\mathcal{V}') = X \},$$

and let

$$\mathcal{H}(\mathcal{V}) = \mathcal{H}_{\text{top}}(\mathcal{V}) := \log(\text{Card}(\mathcal{V})).$$

Analogous to the definitions for partitions, define

$$\mathcal{V} \vee \mathcal{W} := \{V \cap W \mid V \in \mathcal{V} \text{ and } W \in \mathcal{W}\};$$

$$T\mathcal{V} := \{T^{-1}(U) \mid U \in \mathcal{V}\}$$

$$\text{and } \mathcal{V}_{[0 \dots n]} := \mathcal{V}_0 \vee \mathcal{V}_1 \vee \dots \vee \mathcal{V}_{n-1};$$

$$\mathcal{W} \succcurlyeq \mathcal{V}, \text{ if each } \mathcal{W}\text{-patch is a subset of some } \mathcal{V}\text{-patch.}$$

¹⁶Because we only work on a compact space, we can omit “finite”. Some generalizations of topological entropy to non-compact spaces require that only *finite* open-covers be used [37].

The T, \mathcal{V} -**entropy** is

$$\begin{aligned} \mathcal{E}^T(\mathcal{V}) &= \mathcal{E}(T, \mathcal{V}) = \mathcal{E}_{\text{top}}(T, \mathcal{V}) \\ &:= \limsup_{n \rightarrow \infty} \frac{1}{n} \mathcal{H}_{\text{top}}(\mathcal{V}_{[0 \dots n]}) . \end{aligned} \quad (20)$$

And the **topological entropy** of T is

$$\mathcal{E}_{\text{top}}(T) := \sup_{\mathcal{V}} \mathcal{E}_{\text{top}}(T, \mathcal{V}) , \quad (21)$$

taken over all open covers \mathcal{V} .

Thus \mathcal{E}_{top} counts, in some sense, the growth rate in the number of T -orbits of length n .

Evidently, topological entropy is an isomorphism invariant. Two continuous maps $T: X \rightarrow X$ and $S: Y \rightarrow Y$ are **topologically conjugate** (as *isomorphism* is called in this category) if there exists a homeomorphism $\psi: X \rightarrow Y$ with $\psi T = S\psi$.

Lemma 12 (Subadditive Lemma) *Consider a sequence $\mathbf{s} = (s_l)_1^\infty \subset [-\infty, \infty]$ satisfying $s_{k+l} \leq s_k + s_l$, for all $k, l \in \mathbb{Z}$. Then the following limit exists in $[-\infty, \infty]$, and $\lim_{n \rightarrow \infty} \frac{s_n}{n} = \inf_n \frac{s_n}{n}$.*

Topological entropy, or “**top ent**” for short, satisfies many of the relations of measure-entropy.

Lemma 13

(a) $\mathcal{V} \preceq \mathcal{W}$ implies

$$\mathcal{H}(\mathcal{V}) \leq \mathcal{H}(\mathcal{W}) \quad \text{and} \quad \mathcal{E}(T, \mathcal{V}) \leq \mathcal{E}(T, \mathcal{W}) .$$

(b) $\mathcal{H}(\mathcal{V} \vee \mathcal{W}) \leq \mathcal{H}(\mathcal{V}) + \mathcal{H}(\mathcal{W})$.

(c) $\mathcal{H}(T(\mathcal{V})) \leq \mathcal{H}(\mathcal{V})$, with equality if T is surjective. Also, $\mathcal{E}(T, \mathcal{V}) \leq \mathcal{H}(\mathcal{V})$.

(d) In Eq. (20), the $\lim_{n \rightarrow \infty} \frac{1}{n} \mathcal{H}(\mathcal{V}_{[0 \dots n]})$ exists.

(e) Suppose T is a homeomorphism. Then

$$\mathcal{E}(T^{-1}, \mathcal{V}) = \mathcal{E}(T, \mathcal{V}) ,$$

for each cover \mathcal{V} . Consequently, $\mathcal{E}_{\text{top}}(T^{-1}) = \mathcal{E}_{\text{top}}(T)$.

(f) Suppose \mathcal{C} is a collection of covers such that: For each cover \mathcal{W} , there exists a $\mathcal{V} \in \mathcal{C}$ with $\mathcal{V} \preceq \mathcal{W}$. Then $\mathcal{E}_{\text{top}}(T)$ equals the supremum of $\mathcal{E}_{\text{top}}(T, \mathcal{V})$, just taken over those $\mathcal{V} \in \mathcal{C}$.

(g) For all $\ell \in \mathbb{N}$:

$$\mathcal{E}_{\text{top}}(T^\ell) = \ell \mathcal{E}_{\text{top}}(T) .$$

Proof (of (c)) Let $\mathcal{C} \preceq \mathcal{V}$ be a min-cardinality subcover. Then $T\mathcal{C}$ is a subcover of $T\mathcal{V}$. So $\text{Card } T\mathcal{V} \leq |T\mathcal{C}| = |\mathcal{C}|$.

As for entropy, inequality (b) and the foregoing give $\mathcal{H}(\mathcal{V}_{[0 \dots n]}) \leq \mathcal{H}(\mathcal{V})n$. \square

Proof (of (d)) Set $s_n := \mathcal{H}(\mathcal{V}_{[0 \dots n]})$. Then

$$s_{k+l} \leq s_k + \mathcal{H}(T^k(\mathcal{V}_{[0 \dots l]})) \leq s_k + s_l ,$$

by (b) and (c), and so the Subadditive Lemma 12, applies. \square

Proof (of (g)) WLOG, $\ell = 3$. Given \mathcal{V} a cover, triple it to $\widehat{\mathcal{V}} := \mathcal{V} \cap T\mathcal{V} \cap T^2\mathcal{V}$; so

$$\bigvee_{j \in [0 \dots N]} [T^3]^j(\widehat{\mathcal{V}}) = \bigvee_{i \in [0 \dots 3N]} T^i(\mathcal{V}) .$$

Thus $\mathcal{H}(T^3, \widehat{\mathcal{V}}, N) = \mathcal{H}(T, \mathcal{V}, 3N)$, extending notation.

Part (d) and sending $N \rightarrow \infty$, gives $\mathcal{E}(T^3, \widehat{\mathcal{V}}) = 3\mathcal{H}(T, \mathcal{V})$.

Lastly, take covers such that

$$\mathcal{E}(T^3, \mathcal{C}^{(k)}) \rightarrow \mathcal{E}_{\text{top}}(T^3) \quad \text{and}$$

$$\mathcal{E}(T, \mathcal{D}^{(k)}) \rightarrow \mathcal{E}_{\text{top}}(T) ,$$

as $k \rightarrow \infty$. Define $\mathcal{V}^{(k)} := \mathcal{C}^{(k)} \vee \mathcal{D}^{(k)}$. Apply the above to $\mathcal{V}^{(k)}$, then send $k \rightarrow \infty$. \square

Using a Metric

From now on, our space is a compact metric space (X, d) .

Dinaburg [36] and Bowen [34,35], gave alternative, equivalent, definitions of topological entropy, in the compact metric-space case, that are often easier to work with than covers. Bowen gave a definition also when X is not compact¹⁷ (see [35] and Chap. 7 in [23]).

Metric Preliminaries An **ε -ball-cover** comprises finitely many balls, all of radius ε . Since our space is compact, every cover \mathcal{V} has a **Lebesgue number** $\varepsilon > 0$. I. e., for each $z \in X$, the $\text{Bal}(z, \varepsilon)$ lies entirely inside at least one \mathcal{V} -patch. (In particular, there is an ε -ball-cover which *refines* \mathcal{V} .) Let $\text{LEB}(\mathcal{V})$ be the supremum of the Lebesgue numbers. Courtesy of Lemma 13f we can

Fix a “universal” list $\mathcal{V}^{(1)} \preceq \mathcal{V}^{(2)} \preceq \dots$, with $\mathcal{V}^{(k)}$ a $\frac{1}{k}$ -ball-cover. For every $T: X \rightarrow X$, then, the $\lim_k \mathcal{E}(T, \mathcal{V}^{(k)})$ computes $\mathcal{E}_{\text{top}}(T)$.

¹⁷When X is not compact, the definitions need not coincide; e. g. [37]. And topologically-equivalent metrics, but which are not uniformly equivalent, may give the same T different entropies (see p. 171 in [23]).

An ε -Microscope Three notions are useful in examining a metric space (X, m) at scale ε . Subset $A \subset X$ is an ε -**separated-set**, if $m(z, z') \geq \varepsilon$ for all distinct $z, z' \in A$. Subset $F \subset X$ is ε -**spanning** if $\forall x \in X, \exists z \in F$ with $m(x, z) < \varepsilon$.

Lastly, a cover \mathcal{V} is ε -**small** if $\text{Diam}(U) < \varepsilon$, for each $U \in \mathcal{V}$.

You Take the High Road and I'll Take the Low Road

There are several routes to computing top-ent, some via maximization, others, minimization. Our foregoing discussion computed $\mathcal{E}_{\text{top}}(T)$ by a family of **sizes** $f_k(n) = f_k^T(n)$, depending on a parameter k which specifies the fineness of scale. (In Sect. "Metric Preliminaries", this k is an integer; in the original definition, an open cover.) Define two numbers:

$$\begin{aligned}\widehat{L}^f(k) &:= \limsup_{n \rightarrow \infty} \frac{1}{n} \log f_k(n) \quad \text{and} \\ \underline{L}^f(k) &:= \liminf_{n \rightarrow \infty} \frac{1}{n} \log f_k(n).\end{aligned}\quad (22)$$

Finally, let $\mathcal{E}^f(T) := \sup_k \widehat{L}^f(k)$. If the limit exists in Eq. (22) then agree to write $L^f(k)$ for the common value.

The A-K-M definition used the size $f_{\mathcal{V}}(n) := \text{Card}(\mathcal{V}_{[0 \dots n]})$, where

$\text{Card}(\mathcal{W}) :=$ Minimum cardinality of a subcover from \mathcal{W} .

Here are three metric-space sizes $f_{\varepsilon}(n)$:

$\text{Sep}(n, \varepsilon) :=$ Maximum cardinality of a d_n - ε -separated set.

$\text{Spn}(n, \varepsilon) :=$ Minimum cardinality of a d_n - ε -spanning set.

$\text{Cov}(n, \varepsilon) :=$ Minimum cardinality of a d_n - ε -small cover.

These use a list $(d_n)_{n=1}^{\infty}$ of progressively finer metrics on X , where

$$d_N(x, y) := \max_{j \in [0 \dots N]} d(T^j(x), T^j(y)).$$

Theorem 14 (All-Roads-Lead-to-Rome Theorem) Fix ε and let \mathcal{W} be any d - ε -small cover. Then

- (i) $\forall n$: $\text{Cov}(n, 2\varepsilon) \leq \text{Spn}(n, \varepsilon) \leq \text{Sep}(n, \varepsilon) \leq \text{Card}(\mathcal{W}_{[0 \dots n]})$.
- (ii) Take a cover \mathcal{V} and a $\delta < \text{LEB}(\mathcal{V})$. Then $\forall n$: $\text{Card}(\mathcal{V}_{[0 \dots n]}) \leq \text{Cov}(n, \delta)$.
- (iii) The limit $L^{\text{Cov}}(\varepsilon) = \lim_n \frac{1}{n} \log(\text{Cov}(n, \varepsilon))$ exists in $[0 \dots \infty)$.

(iv)

$$\mathcal{E}^{\text{Sep}}(T) = \mathcal{E}^{\text{Spn}}(T) = \mathcal{E}^{\text{Cov}}(T) = \mathcal{E}^{\text{Card}}(T) \stackrel{\text{by defn}}{=} \mathcal{E}_{\text{top}}(T).$$

Proof (of (i)) Take $F \subset X$, a min-cardinality d_n - ε -spanning set. So $\bigcup_{z \in F} D_z = X$, where

$$D_z := d_n\text{-Bal}(z, \varepsilon) \stackrel{\text{note}}{=} \bigcap_{j=0}^{n-1} T^{-j} \left(\text{Bal}(T^j z, \varepsilon) \right).$$

This $\mathcal{D} := \{D_z\}_z$ is a cover, and it is d_n - 2ε -small. Thus $\text{Cov}(n, 2\varepsilon) \leq |\mathcal{D}| = |F|$.

For any metric, a *maximal* ε -separated-set is automatically ε -spanning; adjoin a putative unspanned point to get a larger separated set.

Let A be a max-cardinality d_n - ε -separated set. Take \mathcal{C} , a min-cardinality subcover of $\mathcal{W}_{[0 \dots n]}$. For each $z \in A$, pick a \mathcal{C} -patch $C_z \ni z$. Could some pair $x, y \in A$ pick the same C ? Well, write $C = \bigcap_{j=0}^{n-1} T^{-j}(W_j)$, with each $W_j \in \mathcal{W}$. For every $j \in [0 \dots n]$, then,

$$d(T^j(x), T^j(y)) \leq \text{Diam}(W_j) < \varepsilon.$$

Hence $d_n(x, y) < \varepsilon$; so $x = y$. Accordingly, the $z \mapsto C_z$ map is injective, whence $|A| \leq |\mathcal{C}|$. \square

Proof (of (ii)) Choose a min-cardinality d_n - δ -small cover \mathcal{C} . For each $C \in \mathcal{C}$ and $j \in [0 \dots n]$, the $d\text{-Diam}(T^j C) < \delta$. So there is a \mathcal{V} -patch $V_{C,j} \supset T^j(C)$. Hence

$$\mathcal{V}_{[0 \dots n]} \stackrel{\text{note}}{\ni} \bigcap_{j=0}^{n-1} T^{-j}(V_{C,j}) \supset C.$$

Thus $\mathcal{V}_{[0 \dots n]} \preccurlyeq \mathcal{C}$. So

$$\text{Card}(\mathcal{V}_{[0 \dots n]}) \leq \text{Card}(\mathcal{C}) \leq |\mathcal{C}| = \text{Cov}(n, \delta). \quad \square$$

Proof (of (iii)) To upper-bound $\text{Cov}(k+l, \varepsilon)$ let \mathcal{V} and \mathcal{W} be min-cardinality ε -small covers, respectively, for metrics d_k and d_l . Then $\mathcal{V} \cap T^l(\mathcal{W})$ is a ε -small for d_{k+l} . Consequently $\text{Cov}(k+l, \varepsilon) \leq \text{Cov}(k, \varepsilon) \cdot \text{Cov}(l, \varepsilon)$. Thus $n \mapsto \log(\text{Cov}(n, \varepsilon))$ is subadditive. \square

Proof (of (iv)) Pick a \mathcal{V} from the list in Sect. "Metric Preliminaries", choose some $2\varepsilon < \text{LEB}(\mathcal{V})$ followed by an ε -small \mathcal{W} from Sect. "Metric Preliminaries". Pushing $n \rightarrow \infty$ gives

$$\begin{aligned}L^{\text{Card}}(\mathcal{V}) &\leq L^{\text{Cov}}(2\varepsilon) \leq \frac{\widehat{L}^{\text{Spn}}(\varepsilon) \leq \widehat{L}^{\text{Sep}}(\varepsilon)}{\underline{L}^{\text{Spn}}(\varepsilon) \leq \underline{L}^{\text{Sep}}(\varepsilon)} \leq L^{\text{Card}}(\mathcal{W}).\end{aligned}\quad (23)$$

Now send \mathcal{V} and \mathcal{W} along the list in Sect. "Metric Preliminaries". \square

Pretension Topological entropy takes its values in $[0, \infty]$. A useful corollary of Eq. (23) can be stated in terms of any Distance (\cdot, \cdot) which topologizes $[0, \infty]$ as a compact interval.

For each continuous $T: X \rightarrow X$ on a compact metric-space, the Distance $(\widehat{L}^{\text{Sep}(\varepsilon)}, \underline{L}^{\text{Sep}(\varepsilon)})$ goes to zero as $\varepsilon \searrow 0$. Consequently, we can pretend that the

$$\underline{L}^{\text{Sep}(\varepsilon)} = \lim_{n \rightarrow \infty} \frac{1}{n} \log(\text{Sep}(n, \varepsilon)) \quad (24)$$

limit exists, in arguments that subsequently send $\varepsilon \searrow 0$. Ditto for $\underline{L}^{\text{Spn}(\varepsilon)}$.

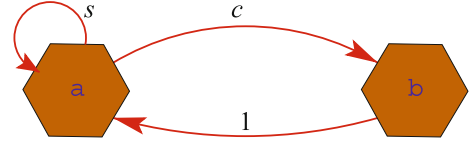
This will be used during the proof of the **Variational Principle**. But first, here are two entropy computations which illustrate the efficacy in having several characterizations of topological entropy.

$\mathcal{E}_{\text{top}}(\text{Isometry}) = 0$ Suppose $(T: X, d)$ is a distance-preserving map of a compact metric-space. Fixing ε , a set is d_n - ε -separated IFF it is d - ε -separated. Thus $\text{Sep}(n, \varepsilon)$ does *not* grow with n . So each $\widehat{L}^{\text{Sep}(\varepsilon)}$ is zero.

Topological Markov Shifts Imagine ourselves back in the days when computer data is stored on large reels of fast-moving magnetic tape. One strategy to maximize the density of binary data stored is to *not* put timing-marks (which take up space) on the tape. This has the defect that when the tape-writer writes, say, 577 consecutive 1-bits, then the tape-reader may erroneously count 578 copies of 1. We sidestep this flaw by first encoding our data so as to avoid the 11⁵⁷⁷1 word, then writing to tape.

Generalize this to a finite alphabet Q and a finite list \mathcal{F} of disallowed Q -words. Extend each word to a common length $K + 1$; now $\mathcal{F} \subset Q^{K+1}$. The resulting “ K -step TMS (topological Markov shift)” is the shift on the set of doubly- ∞ Q -names having no substring in \mathcal{F} . In the above magnetic-tape example, $K = 576$. Making it more realistic, suppose that some string of zeros, say 00⁵⁷⁴0, is also forbidden¹⁸. Extending to length 577, we get $2^3 = 8$ new disallowed words of form 00⁵⁷⁴0 $b_1b_2b_3$.

We **recode** to a 1-step TMS (just called a TMS or a **subshift of finite type**) over the alphabet $P := Q^K$. Each outlawed Q -word $w_0w_1 \cdots w_K$ engenders a length-2 forbidden P -word $(w_0, \dots, w_{K-1})(w_1, \dots, w_K)$. The resulting TMS is topologically conjugate to the original K -step. The *allowed* length-2 words can be viewed as the edges in



Entropy in Ergodic Theory, Figure 5

Ignoring the labels on the edges, for the moment, the **Golden shift**, T , acts on the space of doubly-infinite paths through this graph. The space can be represented as a subset $X_{\text{Gold}} \subset \{a, b\}^{\mathbb{Z}}$, namely, the set of sequences with no two consecutive b letters.

a directed-graph and the set of points $x \in X$ is the set of doubly- ∞ paths through the graph. Once trivialities removed, this X is a Cantor set and the shift $T: X \rightarrow X$ is a homeomorphism.

The Golden Shift As the simplest example, suppose our magnetic-tape is constrained by the Markov graph, Fig. 5 that we studied measure-theoretically in Fig. 4.

We want to store the text of *The Declaration of Independence* on our magnetic tape. Imagining that English is a stationary process, we'd like to encode English into this Golden TMS as efficiently as possible. We seek a shift-invariant measure μ on X_{Gold} of *maximum entropy*, should such exist.

View $P = \{a, b\}$ as the time-zero partition on X_{Gold} ; that is, name $x = \dots x_{-1}x_0x_1x_2 \dots$, is in atom b IFF letter x_0 is “ b ”. Any measure μ gives conditional probabilities

$$\begin{aligned} \mu(a|a) &= s, & \mu(b|a) &= c, \\ \mu(a|b) &\stackrel{\text{note}}{=} 1, & \mu(b|b) &\stackrel{\text{note}}{=} 0. \end{aligned}$$

But recall, $\mathcal{E}(T) = \mathcal{H}(P_1|P_{[-\infty, \dots, 0]}) \leq \mathcal{H}(P_1|P_0)$. So among all measures that make the conditional distribution $P|a$ equal (s, c) , the *unique* one maximizing entropy is the (s, c) -Markov-process. Its entropy, derived in Eq. (14), is

$$\begin{aligned} f(s) &:= \frac{1}{2-s} \mathcal{H}(s, 1-s) \\ &= \frac{-1}{2-s} [s \log(s) + (1-s) \log(1-s)]. \end{aligned} \quad (25)$$

Certainly $f(0) = f(1) = 0$, so f 's maximum occurs at the (it turns out) *unique* point \hat{s} where the derivative $f'(\hat{s})$ equals zero. This $\hat{s} = (-1 + \sqrt{5})/2$. Plugging in, the maximum entropy supportable by the Golden Shift is

$$\begin{aligned} \text{MaxEnt} &= \frac{2}{5 - \sqrt{5}} \left[\frac{-1 + \sqrt{5}}{2} \log \left(\frac{2}{-1 + \sqrt{5}} \right) \right. \\ &\quad \left. + \frac{3 - \sqrt{5}}{2} \log \left(\frac{2}{3 - \sqrt{5}} \right) \right]. \end{aligned} \quad (26)$$

¹⁸Perhaps the \emptyset -bad-length, 574, is shorter than the 1-bad-length because, say, \emptyset s take less tape-space than 1s and so – being written more densely – cause ambiguity sooner.

Exponentiating, the number of μ -typical n -names grows like G^n , where

$$G := \left[\frac{2}{-1 + \sqrt{5}} \right]^{\frac{-1+\sqrt{5}}{5-\sqrt{5}}} \left[\frac{2}{3 - \sqrt{5}} \right]^{\frac{3-\sqrt{5}}{5-\sqrt{5}}} . \quad (27)$$

This expression¹⁹ looks unpleasant to simplify – it isn't even obviously an algebraic number – and yet topological entropy will reveal its familiar nature. This, because the Variational Principle (proved in the next section) says that the top-ent of a system is the supremum of measure-entropies supportable by the system.

Top-ent of the Golden Shift For a moment, let's work more generally on an arbitrary subshift (a closed, shift-invariant subset) $X \subset Q^{\mathbb{Z}}$, where Q is a finite alphabet. Here, the transformation is always the shift – but the *space* is varying – so agree to refer to the top-ent as $\mathcal{E}_{\text{top}}(X)$. Let $\text{Names}_X(n)$ be the number of distinct words in the set $\{x \upharpoonright_{[0..n)} \mid x \in X\}$. Note that a metric inducing the product-topology on $Q^{\mathbb{Z}}$ is

$$d(x, x') := \frac{1}{1 + |m|} , \quad (28)$$

for the smallest $|m|$ with $x_m \neq x'_m$.

Lemma 15 *Consider a subshift X . Then the*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log(\text{Names}_X(n))$$

exists in $[0, \infty]$, and equals $\mathcal{E}_{\text{top}}(X)$.

Proof With $\varepsilon \in (0, 1)$ fixed, two n -names are d_n - ε -separated IFF they are not the same name. Hence $\text{Sep}(n, \varepsilon) = \text{Names}_X(n)$. \square

To compute $\mathcal{E}_{\text{top}}(X_{\text{Gold}})$, declare that a word is “golden” if it appears in some $x \in X_{\text{Gold}}$. Each $[n + 1]$ -golden word ending in a has form wa , where w is n -golden. An $[n + 1]$ -golden word ending in b , must end in ab and so has form wab , where w is $[n - 1]$ -golden. Summing up,

$$\begin{aligned} \text{Names}_{X_{\text{Gold}}}(n + 1) \\ = \text{Names}_{X_{\text{Gold}}}(n) + \text{Names}_{X_{\text{Gold}}}(n - 1) . \end{aligned}$$

This is the Fibonacci recurrence, and indeed, these are the Fibonacci numbers, since $\text{Names}_{X_{\text{Gold}}}(0) = 1$ and $\text{Names}_{X_{\text{Gold}}}(1) = 2$. Consequently, we have that

$$\text{Names}_{X_{\text{Gold}}}(n) \sim \text{Const} \cdot \lambda^n ,$$

¹⁹A popular computer-algebra-system was not, at least under my inexpert tutelage, able to simplify this. However, once top-ent gave the correct answer, the software was able to detect the equality.

where $\lambda = \frac{1+\sqrt{5}}{2}$ is the Golden Ratio. So the sesquipedalian number G from Eq. (27) is simply λ , and $\mathcal{E}_{\text{top}}(X_{\text{Gold}}) = \log(\lambda)$.

Since $\log(\lambda) \approx 0.694$, each thousand bits written on tape (subject to the “no bb substrings” constraint) can carry at most 694 bits of information.

Top-ent of a General TMS A (finite) digraph G engenders a **TMS** $T: X_G \rightarrow X_G$, as well as a $\{0, 1\}$ -valued adjacency matrix $\mathbf{A} = \mathbf{A}_G$, where $a_{i,j}$ is the number of directed-edges from state i to j . (Here, each $a_{i,j}$ is 0 or 1.) The (i, j) -entry in power \mathbf{A}^n is automatically the number of length- n paths from i to j . Employing the matrix-norm $\|\mathbf{M}\| := \sum_{i,j} |m_{i,j}|$, then,

$$\|\mathbf{A}\|^n = \text{Names}_X(n) .$$

Happily Gelfand's formula (see 10.13 in [58] or [Spectral radius](#) in [60]) applies: For an arbitrary (square) complex matrix,

$$\lim_{n \rightarrow \infty} \|\mathbf{A}^n\|^{\frac{1}{n}} = \text{SpecRad}(\mathbf{A}) . \quad (29)$$

This right hand side, the **spectral radius** of \mathbf{A} , means the maximum of the absolute values of \mathbf{A} 's eigenvalues. So the top-ent of a **TMS** is thus the

$$\begin{aligned} \mathcal{E}_{\text{top}}(X_G) &= \text{SpecRad}(\mathbf{A}_G) \\ &:= \text{Max} \{ |e| \mid e \text{ is an eigenvalue of } \mathbf{A}_G \} . \end{aligned} \quad (30)$$

The (a, b) -adjacency matrix of Fig. 5 is

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} ,$$

whose eigenvalues are λ and $\frac{1}{\lambda}$.

Labeling Edges Interpret $(s, c, 1)$ simply as edge-labels in Fig. 5. The set of doubly- ∞ paths can also be viewed as a subset $Y_{\text{Gold}} \subset \{s, c, 1\}^{\mathbb{Z}}$, and it too is a **TMS**. The shift on Y_{Gold} is conjugate (topologically isomorphic) to the shift on X_{Gold} , so they *a fortiori* have the same top-ent, $\log(\lambda)$. The $(s, c, 1)$ -adjacency matrix is

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} .$$

Its $|\cdot|$ -largest eigenvalue is still λ , as it must.

Now we make a new graph. We modify Fig. 5 by manufacturing a total of two s -edges, seven c -edges, and three edges l_1, l_2, l_3 . Give these $2 + 7 + 3$ edges twelve *distinct* labels. We could compute the resulting **TMS**-entropy

from the corresponding 12×12 adjacency matrix. Alternatively, look at the (a, b)-adjacency matrix

$$\mathbf{A} := \begin{bmatrix} 2 & 7 \\ 3 & 0 \end{bmatrix}.$$

The roots of its characteristic polynomial are $1 \pm \sqrt{22}$. Hence \mathcal{E}_{top} of this 12-symbol TMS is $\log(1 + \sqrt{22})$.

The Variational Principle

Let $\mathcal{M} := \mathcal{M}(X, d)$ be the set of Borel probability measures, and $\mathcal{M}(T) := \mathcal{M}(T: X, d)$ the set of T -invariant $\mu \in \mathcal{M}$. Assign

$$\text{EntSup}(T) := \sup \{ \mathcal{E}_{\mu}(T) \mid \mu \in \mathcal{M}(T) \}.$$

Theorem 16 (Variational principle (Goodson))
 $\text{EntSup}(T) = \mathcal{E}_{\text{top}}(T)$.

This says that top-ent is the top entropy – if there is a measure μ which realizes the supremum. There doesn't have to be. Choose a sequence of metric-systems $(S_k: Y_k, m_k)$ whose entropies *strictly* increase $\mathcal{E}_{\text{top}}(S_k) \nearrow L$ to some limit in $(0, \infty]$. Let $(S_{\infty}: Y_{\infty}, m_{\infty})$ be the identity-map on a 1-point space. Define a new system $(T: X, d)$, where $X := \bigsqcup_{k \in [1 \dots \infty]} Y_k$. Have $T(x) := S_k(x)$, for the unique k with $Y_k \ni x$. As for the metric, on Y_k let d be a scaled version of m_k , so that the d -Diam(Y_k) is less than $1/2^k$. Finally, for points in *distinct* components, $x \in Y_k$ and $z \in Y_{\ell}$, decree that $d(x, z) := |2^{-k} - 2^{-\ell}|$. Our T is continuous, and is a homeomorphism if each of the S_k is. Certainly $\mathcal{E}_{\text{top}}(T) = L > \mathcal{E}_{\text{top}}(S_k)$, for every $k \in [1 \dots \infty]$.

If L is *finite* then there is *no* measure μ of maximal entropy; for μ must give mass to *some* Y_k ; this pulls the entropy below L , since there are no compensatory components with entropy exceeding L .

In contrast, when $L = \infty$ then there is a maximal-entropy measure (put mass $1/2^j$ on some component Y_{k_j} , where $k_j \nearrow \infty$ swiftly); indeed, there are continuum-many maximal-entropy measures. *But* there is no ²⁰ *ergodic* measure of maximal entropy.

For a concrete $L = \infty$ example, let S_k be the shift on $[1 \dots k]^{\mathbb{Z}}$.

²⁰The ergodic measures are the extreme points of $\mathcal{M}(T)$; call them $\mathcal{M}_{\text{Erg}}(T)$. This $\mathcal{M}(T)$ is the set of barycenters obtained from Borel probability measures on $\mathcal{M}_{\text{Erg}}(T)$ (see [Krein-Milman theorem](#), [Choquet theory](#) in [60]). In this instance, what explains the failure to have an *ergodic* maximal-entropy measure? Let μ_k be an invariant ergodic measure on Y_k . These measures *do* converge to the one-point (ergodic) probability measure μ_{∞} on Y_{∞} . But the map $\mu \mapsto \mathcal{E}_{\mu}(T)$ is not continuous at μ_{∞} .

Topology on \mathcal{M} Let's arrange our tools for establishing the Variational Principle. The argument will follow Misiurewicz's proof, adapted from the presentations in [23] and [11].

Equip \mathcal{M} with the *weak*-* topology.²¹ An $A \subset X$ is μ -*nice* if its topological boundary $\partial(A)$ is μ -null. And a *partition* is μ -*nice* if each atom is.

Proposition 17 If $\alpha_L \rightarrow \mu$ and $A \subset X$ is μ -nice, then $\alpha_L(A) \rightarrow \mu(A)$.

Proof Define operator $\mathcal{U}(D) := \limsup_L \alpha_L(D)$. It suffices to show that $\mathcal{U}(A) \leq \mu(A)$. For since A^c is μ -nice too, then $\mathcal{U}(A^c) \leq \mu(A^c)$. Thus $\lim_L \alpha_L(A)$ exists, and equals $\mu(A)$.

Because $C := \overline{A}$ is closed, the continuous functions $f_N \searrow \mathbf{1}_C$ pointwise, where

$$f_N(x) := 1 - \text{Min}(N \cdot d(x, C), 1).$$

By the Monotone Convergence theorem, then,

$$\int f_N d\mu \xrightarrow{N} \mu(C).$$

And $\mu(C) = \mu(A)$, since A is nice. Fixing N , then, it suffices to establish $\mathcal{U}(A) \leq \int f_N d\mu$. But f_N is continuous, so

$$\begin{aligned} \int f_N d\mu &= \limsup_{L \rightarrow \infty} \int f_N d\alpha_L \\ &\geq \limsup_{L \rightarrow \infty} \int \mathbf{1}_A d\alpha_L = \mathcal{U}(A). \quad \square \end{aligned}$$

Corollary 18 Suppose $\alpha_L \rightarrow \mu$, and partition P is μ -nice. Then $\mathcal{H}_{\alpha_L}(P) \rightarrow \mathcal{H}_{\mu}(P)$.

The *diameter* of partition P is $\text{Max}_{A \in P} \text{Diam}(A)$.

Proposition 19 Take $\mu \in \mathcal{M}$ and $\varepsilon > 0$. Then there exists a μ -nice partition with $\text{Diam}(P) < \varepsilon$.

Proof Centered at an x , the uncountably many balls $\{\text{Bal}(x, r) \mid r \in (0, \varepsilon)\}$ have disjoint boundaries. So all but countably many are μ -nice; pick one and call it B_x . Compactness gives a finite nice cover, say, $\{B_1, \dots, B_7\}$, at different centers. Then the partition $P := (A_1, \dots, A_7)$ is nice,²² where $A_k := B_k \setminus \bigcup_{j=1}^{k-1} B_j$. \square

Here is a consequence of Jensen's inequality.

²¹Measures $\alpha_L \rightarrow \mu$ IFF $\int f d\alpha_L \rightarrow \int f d\mu$, for each continuous $f: X \rightarrow \mathbb{R}$. This metrizable topology makes \mathcal{M} compact. Always, $\mathcal{M}(T)$ is a non-void compact subset (see [Measure Preserving Systems](#).)

²²For any two sets $B, B' \subset X$, the union $\partial B \cup \partial B'$ is a superset of the three boundaries $\partial(B \cup B')$, $\partial(B \cap B')$, $\partial(B \setminus B')$.

Lemma 20 (Distropy-Averaging Lemma) For $\mu, \nu \in \mathcal{M}$, a partition R , and a number $t \in [0, 1]$,

$$t\mathcal{H}_\mu(R) + t^c\mathcal{H}_\nu(R) \leq \mathcal{H}_{t\mu+t^c\nu}(R).$$

Strategy for $\text{EntSup}(T) \geq \varepsilon_{\text{top}}(T)$. Choose an $\varepsilon > 0$. For $L = 1, 2, 3, \dots$, take a maximal (L, ε) -separated-set $F_L \subset X$, then define

$$\mathbf{F} = \mathbf{F}_\varepsilon := \limsup_{L \rightarrow \infty} \frac{1}{L} \log(|F_L|).$$

Let $\varphi_L()$ be the equi-probable measure on F_L ; each point has weight $1/|F_L|$. The desired invariant measure μ will come from the Cesàro averages

$$\alpha_L := \frac{1}{L} \sum_{\ell \in [0 \dots L]} T^\ell \varphi_L,$$

which get more and more invariant.

Lemma 21 Let μ be any weak-* accumulation point of the above $\{\alpha_L\}_1^\infty$. (Automatically, μ is T -invariant.) Then $\varepsilon_\mu(T) \geq \mathbf{F}$. Indeed, if Q is any μ -nice partition with $\text{Diam}(Q) < \varepsilon$, then $\varepsilon_\mu(T, Q) \geq \mathbf{F}$.

Tactics As usual, $Q_{[0 \dots N]}$ means $Q_0 \vee Q_1 \vee \dots \vee Q_{N-1}$. Our goal is

$$\forall N: \quad \mathbf{F} \stackrel{?}{\leq} \frac{1}{N} \mathcal{H}_\mu(Q_{[0 \dots N]}). \quad (31)$$

Fix N and $\mathbf{P} := Q_{[0 \dots N]}$, and a $\delta > 0$. It suffices to verify: $\forall_{\text{large } L} L \gg N$,

$$\frac{1}{L} \log(|F_L|) \stackrel{?}{\leq} \delta + \frac{1}{N} \mathcal{H}_{\alpha_L}(\mathbf{P}), \quad (32)$$

since this and Corollary 17 will prove Eq. (31): Pushing $L \rightarrow \infty$ along the sequence that produced μ essentially sends LhS(32) to \mathbf{F} , courtesy Eq. (24). And RhS(32) goes to $\delta + \frac{1}{N} \mathcal{H}_\mu(\mathbf{P})$, by Corollary 17, since \mathbf{P} is μ -nice. Descending $\delta \searrow 0$, hands us the needed Eq. (31).

Remark 22 The idea in the following proof is to mostly fill interval $[0..L]$ with N -blocks, starting with a offset $K \in [0..N)$. Averaging over the offset will create a Cesàro average over each N -block. Averaging over the N -blocks will allow us to compute distropy with respect to the averaged measure, α_L .

Proof (of Eq. (32)) Since L is frozen, agree to use φ for the φ_L probability measure.

Our d_L - ε -separated set F_L has at most *one* point in any given atom of $Q_{[0 \dots L]}$, thereupon

$$\log(|F_L|) = \mathcal{H}_\varphi(Q_{[0 \dots L]}).$$

Regardless of the “offset” $K \in [0 \dots N)$, we can always fit $C := \lfloor \frac{L-N}{N} \rfloor$ many N -blocks into $[0 \dots L]$. Denote by $\mathcal{G}(K) := [K \dots K + CN]$, this union of N -blocks, the **good** set of indices. Unsurprisingly, $\mathcal{B}(K) := [0 \dots L] \setminus \mathcal{G}(K)$ is the **bad** index-set. Therefore,

$$\mathcal{H}_\varphi(Q_{[0 \dots L]}) \leq \overbrace{\mathcal{H}_\varphi\left(\bigvee_{j \in \mathcal{B}(K)} Q_j\right)}^{\text{Bad}(K)} + \overbrace{\mathcal{H}_\varphi\left(\bigvee_{j \in \mathcal{G}(K)} Q_j\right)}^{\text{Good}(K)}. \quad (33)$$

Certainly $\text{Bad}(K) \leq 3N \log(|Q|)$. So

$$\frac{1}{NL} \sum_{K \in [0 \dots N)} \text{Bad}(K) \leq \frac{3N}{L} \log(|Q|).$$

This is less than δ , since L is large. Applying $\frac{1}{NL} \sum_{K \in [0 \dots N)}$ to Eq. (28) now produces

$$\frac{1}{L} \log(|F_L|) \leq \delta + \frac{1}{NL} \sum_K \text{Good}(K). \quad (34)$$

Note

$$\bigvee_{j \in \mathcal{G}(K)} T^j(Q) = \bigvee_{c \in [0 \dots C) T^{K+cN}(\mathbf{P})}.$$

So

$$\text{Good}(K) \leq \sum_c \mathcal{H}_\varphi(T^{K+cN}(\mathbf{P})).$$

This latter, by definition, equals $\sum_c \mathcal{H}_{T^{K+cN}(\varphi)}(\mathbf{P})$. We conclude that

$$\frac{1}{NL} \sum_K \text{Good}(K) \leq \frac{1}{NL} \sum_K \sum_c \mathcal{H}_{T^{K+cN}(\varphi)}(\mathbf{P})$$

$$\leq \frac{1}{NL} \sum_{\ell \in [0 \dots L)} \mathcal{H}_{T^\ell \varphi}(\mathbf{P}),$$

by adjoining a few translates of \mathbf{P} ,

$$\leq \frac{1}{N} \mathcal{H}_{\alpha_L}(\mathbf{P}),$$

by the Distropy-averaging lemma 18,

since α_L is the average $\frac{1}{L} \sum_\ell T^\ell \varphi$. Thus Eq. (34) implies Eq. (32), our goal. \square

Proof (of $\text{EntSup}(T) \leq \mathcal{E}_{\text{top}}(T)$) Fix a T -invariant μ . For partition $Q = (B_1, \dots, B_K)$, choose a compact set $A_k \subset B_k$ with $\mu(B_k \setminus A_k)$ small. (This can be done, since μ is automatically a regular measure [58].) Letting $D := [\bigsqcup_i A_i]^c$ and $P := (D, A_1, \dots, A_K)$, we can have made $\mathcal{H}(P|Q)$ as small as desired. Courtesy of Lemma 8b, then, we only need consider partitions of the form that P has.

Open-cover $\mathcal{V} = (U_1, \dots, U_K)$ has patches $U_k := D \cup A_k$. What atoms of, say, $P_{[0 \dots 3]}$, can the intersection $U_9 \cap T^{-1}(U_2) \cap T^{-2}(U_5)$ touch? Only the eight atoms

$$(D \text{ or } A_9) \cap T^{-1}(D \text{ or } A_2) \cap T^{-2}(D \text{ or } A_5).$$

Thus $\#P_{[0 \dots n]} \leq 2^n \cdot \#\mathcal{V}_{[0 \dots n]}$. (Here, $\#()$ counts the number of non-void atoms/patches.) So

$$\begin{aligned} \frac{1}{n} \mathcal{H}_\mu(P_{[0 \dots n]}) &\leq 1 + \frac{1}{n} \log(\#\mathcal{V}_{[0 \dots n]}) \\ &\leq 1 + 1 + \mathcal{E}_{\text{top}}(T); \end{aligned}$$

this last inequality, when n is large. The upshot: $\mathcal{E}_\mu(T) \leq 2 + \mathcal{E}_{\text{top}}(T)$.

Applied to a power T^ℓ , this asserts that $\mathcal{E}_\mu(T^\ell) \leq 2 + \mathcal{E}_{\text{top}}(T^\ell)$. Thus

$$\mathcal{E}_\mu(T) \leq \frac{2}{\ell} + \mathcal{E}_{\text{top}}(T),$$

using Lemma 9d and using Lemma 13g. Now coax $\ell \rightarrow \infty$. \square

Three Recent Results

Having given an survey of older results in measure-theoretic entropy and in topological entropy, let us end this survey with a brief discussion of a few recent results, chosen from many.

Ornstein–Weiss: Finitely-Observable Invariant

In a landmark paper [10], Ornstein and Weiss show that all “finitely observable” properties of ergodic processes are secretly entropy; indeed, they are continuous functions of entropy. This was generalized by Gutman and Hochman [9]; some of the notation below is from their paper.

Here is the setting. Consider an ergodic process, on a non-atomic space, taking on only finitely many values in \mathbb{N} ; let \mathcal{C} be some family of such processes. An **observation scheme** is a metric space (Ω, d) and a sequence of functions $\mathbf{S} = (S_n)_1^\infty$, where S_n maps $\mathbb{N} \times \dots \times \mathbb{N}$ into Ω . On a point $\vec{x} \in \mathbb{N}^\infty$, the scheme **converges** if

$$n \mapsto S_n(x_1, x_2, \dots, x_n) \quad (35)$$

converges in Ω . And on a particular process X , say that \mathbf{S} **converges**, if \mathbf{S} converges on a.e. \vec{x} in X .

A function $J: \mathcal{C} \rightarrow \Omega$ is isomorphism invariant if, whenever the underlying transformations of two processes $X, X' \in \mathcal{C}$ are isomorphic, then $J(X) = J(X')$. Lastly, say that \mathbf{S} “converges to J ”, if for each $X \in \mathcal{C}$, scheme \mathbf{S} converges to the value $J(X)$.

The work of David Bailey [38], a student of Ornstein, produced an observation scheme for entropy. The Lempel–Ziv algorithm [43] was another entropy observer, with practical application.

Ornstein and Weiss provided entropy schemes in [41] and [42]. Their recent paper “Entropy is the only finitely-observable invariant” [10], gives a converse, a uniqueness result.

Theorem 23 (Ornstein, Weiss) Suppose J is a finitely observable function, defined on all ergodic finite-valued processes. If J is an isomorphism invariant, then J is a continuous function of the entropy.

Gutman–Hochman: Finitely-Observable Extension

Extending the Ornstein–Weiss result, Yonatan Gutman and Michael Hochman, in [9] proved that it holds even when the isomorphism invariant, J , is well-defined only on certain subclasses of the set of all ergodic processes. In particular they obtain the following result on three classes of zero-entropy transformations.

Theorem 24 (Gutman, Hochman) Suppose $J()$ is a finitely observable invariant on one of the following classes:

- (i) The Kronecker systems; the class of systems with pure point spectrum.
- (ii) The zero-entropy mild mixing processes.
- (iii) The zero-entropy strong mixing processes.

Then $J()$ is constant.

Entropy of Actions of Free Groups

Consider (G, \mathcal{G}) , a topological group and its Borel field (sigma-algebra). Let $\mathcal{X} \times \mathcal{X}$ be the field on $G \times X$ generated by the two coordinate-subfields. A map

$$\phi: G \times X \rightarrow X \quad (36)$$

is **measurable** if

$$\psi^{-1}(X) \subset \mathcal{G} \times \mathcal{X}.$$

Use $\psi^g(x)$ for $\psi(g, x)$.

This map in Eq. (36) is a (measure-preserving) **group action** if $\forall g, h \in G: \psi^g \circ \psi^h = \psi^{gh}$, and each $\psi^g: X \rightarrow X$ is measure preserving.

This encyclopedia article has only discussed entropy for \mathbb{Z} -actions, i. e., when $G = \mathbb{Z}$. The ergodic theorem, our definition of entropy, and large parts of ergodic theory, involve taking averages (of some quantity of interest) over larger and larger “pieces of time”. In \mathbb{Z} , we typically use the intervals $I_n := [0 \dots n]$. When G is $\mathbb{Z} \times \mathbb{Z}$, we might average over squares $I_n \times I_n$.

The *amenable groups* are those which possess, in a certain sense, larger and larger averaging sets. Parts of ergodic theory have been carried over to actions of amenable groups, e. g. [49] and [55]. Indeed, much of the Bernoulli theory was extended to certain amenable groups by Ornstein and Weiss, [50].

The stereotypical example of a *non*-amenable group, is a free group (on more than one generator). But recently, Lewis Bowen [8] succeeded in extending the definition of entropy to actions of finite-rank free groups.

Theorem 25 (Lewis Bowen) *Let G be a finite-rank free group. Then two Bernoulli G -actions are isomorphic IFF they have the same entropy.*

The paper introduces a new isomorphism invariant, the “ f invariant”, and shows that, for Bernoulli actions, the f invariant agrees with entropy, that is, with the distropy of the independent generating partition.

Exodos

Ever since the pioneering work of Shannon, and of Kolmogorov and Sinai, entropy has been front and center as a major tool in Ergodic Theory. Simply *mentioning* all the substantial results in entropy theory would dwarf the length of this encyclopedia article many times over. And, as the above three results (cherry-picked out of many) show, Entropy shows no sign of fading away...

Bibliography

Historical

1. Adler R, Weiss B (1967) Entropy, a complete metric invariant for automorphisms of the torus. *Proc Natl Acad Sci USA* 57:1573–1576
2. Clausius R (1864) *Abhandlungen ueber die mechanische Wärmetheorie*, vol 1. Vieweg, Braunschweig
3. Clausius R (1867) *Abhandlungen ueber die mechanische Wärmetheorie*, vol 2. Vieweg, Braunschweig
4. Kolmogorov AN (1958) A new metric invariant of transitive automorphisms of Lebesgue spaces. *Dokl Akad Nauk SSSR* 119(5):861–864

5. McMillan B (1953) The basic theorems of information theory. *Ann Math Stat* 24:196–219
6. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423, 623–656
7. Sinai Y (1959) On the Concept of Entropy of a Dynamical System, *Dokl Akad Nauk SSSR* 124:768–771

Recent Results

8. Bowen L (2008) A new measure-conjugacy invariant for actions of free groups. <http://www.math.hawaii.edu/%7Elpbowen/notes11.pdf>
9. Gutman Y, Hochman M (2006) On processes which cannot be distinguished by finitary observation. <http://arxiv.org/pdf/math/0608310>
10. Ornstein DS, Weiss B (2007) Entropy is the only finitely-observable invariant. *J Mod Dyn* 1:93–105; <http://www.math.psu.edu/jmd>

Ergodic Theory Books

11. Brin M, Stuck G (2002) *Introduction to dynamical systems*. Cambridge University Press, Cambridge
12. Cornfeld I, Fomin S, Sinai Y (1982) *Ergodic theory*. Grundlehren der Mathematischen Wissenschaften, vol 245. Springer, New York
13. Friedman NA (1970) *Introduction to ergodic theory*. Van Nostrand Reinhold, New York
14. Halmos PR (1956) *Lectures on ergodic theory*. The Mathematical Society of Japan, Tokyo
15. Katok A, Hasselblatt B (1995) *Introduction to the modern theory of dynamical systems*. (With a supplementary chapter by Katok and Leonardo Mendoza). *Encyclopedia of Mathematics and its Applications*, vol 54. Cambridge University Press, Cambridge
16. Keller G, Greven A, Warnecke G (eds) (2003) *Entropy*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton
17. Lind D, Marcus B (1995) *An introduction to symbolic dynamics and coding*. Cambridge University Press, Cambridge
18. Mañé R (1987) *Ergodic theory and differentiable dynamics*. *Ergebnisse der Mathematik und ihrer Grenzgebiete*, ser 3, vol 8. Springer, Berlin
19. Parry W (1969) *Entropy and generators in ergodic theory*. Benjamin, New York
20. Petersen K (1983) *Ergodic theory*. Cambridge University Press, Cambridge
21. Rudolph DJ (1990) *Fundamentals of measurable dynamics*. Clarendon Press, Oxford
22. Sinai Y (1994) *Topics in ergodic theory*. Princeton Mathematical Series, vol 44. Princeton University Press, Princeton
23. Walters P (1982) *An introduction to ergodic theory*. Graduate Texts in Mathematics, vol 79. Springer, New York

Differentiable Entropy

24. Ledrappier F, Young L-S (1985) The metric entropy of diffeomorphisms. *Ann Math* 122:509–574

25. Pesin YB (1977) Characteristic Lyapunov exponents and smooth ergodic theory. *Russ Math Surv* 32:55–114
26. Young L-S (1982) Dimension, entropy and Lyapunov exponents. *Ergod Theory Dyn Syst* 2(1):109–124

Finite Rank

27. Ferenczi S (1997) Systems of finite rank. *Colloq Math* 73(1):35–65
28. King JLF (1988) Joining-rank and the structure of finite rank mixing transformations. *J Anal Math* 51:182–227

Maximal-Entropy Measures

29. Buzzi J, Ruelle S (2006) Large entropy implies existence of a maximal entropy measure for interval maps. *Discret Contin Dyn Syst* 14(4):673–688
30. Denker M (1976) Measures with maximal entropy. In: Conze J-P, Keane MS (eds) *Théorie ergodique, Actes Journées Ergodiques, Rennes, 1973/1974. Lecture Notes in Mathematics*, vol 532. Springer, Berlin, pp 70–112
31. Misiurewicz M (1973) Diffeomorphism without any measure with maximal entropy. *Bull Acad Polon Sci Sér Sci Math Astron Phys* 21:903–910

Topological Entropy

32. Adler R, Marcus B (1979) Topological entropy and equivalence of dynamical systems. *Mem Amer Math Soc* 20(219)
33. Adler RL, Konheim AG, McAndrew MH (1965) Topological entropy. *Trans Am Math Soc* 114(2):309–319
34. Bowen R (1971) Entropy for group endomorphisms and homogeneous spaces. *Trans Am Math Soc* 153:401–414. Errata 181:509–510 (1973)
35. Bowen R (1973) Topological entropy for noncompact sets. *Trans Am Math Soc* 184:125–136
36. Dinaburg EI (1970) The relation between topological entropy and metric entropy. *Sov Math Dokl* 11:13–16
37. Hasselblatt B, Nitecki Z, Propp J (2005) Topological entropy for non-uniformly continuous maps. <http://www.citebase.org/abstract?id=oai:arXiv.org:math/0511495>

Determinism and Zero-Entropy, and Entropy Observation

38. Bailey D (1976) Sequential schemes for classifying and predicting ergodic processes. Ph D Dissertation, Stanford University
39. Kalikow S, King JLF (1994) A countably-valued sleeping stockbroker process. *J Theor Probab* 7(4):703–708
40. King JLF (1992) Dilemma of the sleeping stockbroker. *Am Math Monthly* 99(4):335–338
41. Ornstein DS, Weiss B (1990) How sampling reveals a process. *Ann Probab* 18(3):905–930
42. Ornstein DS, Weiss B (1993) Entropy and data compression schemes. *IEEE Trans Inf Theory* 39(1):78–83
43. Ziv J, Lempel A (1977) A universal algorithm for sequential data compression. *IEEE Trans Inf Theory* 23(3):337–343

Bernoulli Transformations, K-Automorphisms, Amenable Groups

44. Berg KR (1975) Independence and additive entropy. *Proc Am Math Soc* 51(2):366–370; <http://www.jstor.org/stable/2040323>
45. Meshalkin LD (1959) A case of isomorphism of Bernoulli schemes. *Dokl Akad Nauk SSSR* 128:41–44
46. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. *Adv Math* 5:337–352
47. Ornstein DS (1974) *Ergodic theory randomness and dynamical systems*, Yale Math Monographs, vol 5. Yale University Press, New Haven
48. Ornstein DS, Shields P (1973) An uncountable family of K-automorphisms. *Adv Math* 10:63–88
49. Ornstein DS, Weiss B (1983) The Shannon–McMillan–Birman theorem for a class of amenable groups. *Isr J Math* 44(3):53–60
50. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. *J Anal Math* 48:1–141
51. Shields P (1973) *The theory of Bernoulli shifts*. University of Chicago Press, Chicago
52. Sinai YG (1962) A weak isomorphism of transformations having an invariant measure. *Dokl Akad Nauk SSSR* 147:797–800
53. Thouvenot J-P (1975) Quelques propriétés des systèmes dynamiques qui se décomposent en un produit de deux systèmes dont l'un est un schéma de Bernoulli. *Isr J Math* 21:177–207
54. Thouvenot J-P (1977) On the stability of the weak Pinsker property. *Isr J Math* 27:150–162

Abramov Formula

55. Ward T, Zhang Q (1992) The Abramov–Rohlin entropy addition formula for amenable group actions. *Monatshefte Math* 114:317–329

Miscellaneous

56. Newhouse SE (1989) Continuity properties of entropy. *Ann Math* 129:215–235
57. <http://www.isib.cnr.it/control/entropy/>
58. Rudin W (1973) *Functional analysis*. McGraw-Hill, New York
59. Tribus M, McIrvine EC (1971) Energy and information. *Sci Am* 224:178–184
60. Wikipedia, <http://en.wikipedia.org/wiki/>. pages: http://en.wikipedia.org/wiki/Spectral_radius, http://en.wikipedia.org/wiki/Information_entropy

Books and Reviews

- Boyle M, Downarowicz T (2004) The entropy theory of symbolic extensions. *Invent Math* 156(1):119–161
- Downarowicz T, Serafin J (2003) Possible entropy functions. *Isr J Math* 135:221–250
- Hassner M (1980) A non-probabilistic source and channel coding theory. Ph D Dissertation, UCLA
- Katok A, Sinai YG, Stepin AM (1977) *Theory of dynamical systems and general transformation groups with an invariant measure*. J Sov Math 7(6):974–1065

Entropy Maximization and Species Abundance

BILL SHIPLEY

Département de Biologie, Université de Sherbrooke,
Sherbrooke, Canada

Article Outline

Glossary

Definition of the Subject

Introduction

Information Theory Basis of Entropy Maximization

Fixed Species Pools: Predicting Community Composition

Species Abundance Distributions

Future Directions

Bibliography

Glossary

Species richness the number of species per unit area; a synonym is species density.

Species abundance the number of individuals or amount of living biomass per species per unit area.

Per capita instantaneous growth rate the net number of new individuals produced by an average individual in a population at time t .

Deterministic chaos a pattern of change over time that is perfectly determined by initial conditions, but for which tiny changes in initial conditions result in divergences in the dynamics such that the pattern appears random.

Attractor a set of values (a point, a curve or higher – including a fractal – dimensional object) towards which a dynamic trajectory will move.

Basin of attraction the set of values from which any dynamic trajectory will move towards the same attractor.

Maximum entropy formalism a formal method for producing probability estimates that agree with certain constraints specifying available information about a system but that are otherwise maximally uninformative.

Maximally uninformative (or ignorance) prior a probability distribution that encodes the basic structure of a logical problem as given by the definition of the variable, but that contains no other information.

Definition of the Subject

An understanding of the determinants of biodiversity is an important goal of academic ecology and the protection

of biodiversity is an important goal in conservation. Biodiversity has two components: the number (or density) of species and the relative abundance of species. Both components vary over different spatial and temporal scales. The allocation of limiting resources to different species in an ecological community is reflected in the abundance of each species. The notion of a “community” in ecology (as in more general usages of this word) is poorly defined but will here mean the total number of organisms found within a fixed area of space (a site). Abundance is measured in different ways but the most exact measure is the total mass of living tissues (i.e. biomass) found in a given species at the site. Relative abundance is the proportional abundance of each species relative to the total abundance of all species at the site. The species “pool” is the set of species that can potentially disperse into the site, including those that are not found within the community because they have been excluded due either to biotic interactions or to an inability to survive the abiotic conditions of the site. The vector of abundances, $\mathbf{a} = \{a_i\}$, or relative abundances, $\mathbf{ra} = \{ra_i\}$, of each species in the species pool that is observed at the site is a description of the community structure. The change in this vector over time is a description of the community dynamics.

As such, community structure and dynamics are fundamental properties of ecological communities and relate to basic questions posed by ecologists: Which species will be found at a site? Which of these will be rare? Which species will dominate? How many species will be found? How is the flow of matter and energy through the organisms affected by community structure? How will the community structure change if the environmental conditions of the site change? Every one of these questions involves complex dynamics of large numbers of interactions of organisms. Because of this, answers to these questions will have to deal with this complexity. Most of the results that follow are based on plant communities, since these are the only ones that have been studied to date using entropy maximization, but animal communities will be discussed when looking at future directions. The answers to such questions have important applied implications for forestry, agriculture and conservation.

Introduction

Four aspects of ecological communities are obvious to even a casual observer of nature. First, the organisms found at any given site are distributed into different species. Second, the distribution of individuals and biomass, and therefore of resources, is not equal between species. It is not simply that some species are common and

some are rare, but that a small number of species in every community make up the majority of the biomass and most species in every community make up a minority of the biomass. Third, which species are common, which are rare, and which are absent, differ across sites having different physical (abiotic) environmental conditions. Fourth, which species are common, which are rare, and which are absent, change systematically at the same site over time following a major disturbance (i. e. ecological succession). Explaining and predicting these four properties of ecological communities is a major goal of ecological science.

Although much progress has been made at the empirical level, a general theoretical explanation of these properties with good predictive ability under field conditions is still largely missing despite almost a century of effort by influential theoretical ecologists. All of these theoretical efforts were based on dynamic models derived from population-level demography and based on an analogy with Newtonian mechanics. The purpose of this essay is to contrast this majority tradition in ecological modeling of communities with a more recent statistical mechanistic approach based on maximization of entropy or relative entropy.

Population dynamic models begin with two undeniable properties of any biological population. First, because each mature member of the population can potentially reproduce, population growth is a multiplicative process defining a geometric series. This is the first fundamental insight of Malthus that inspired Darwin [1] and Wallace [2] to propose the theory of evolution by natural selection: every biological population has the *potential* for exponential growth. If $n_i(t)$ is the number of individuals of species i alive at time t , then the per capita instantaneous growth rate at time t , $r_i(t)$, is defined as:

$$\frac{1}{n_i(t)} \frac{dn_i(t)}{dt} = r_i(t). \quad (1)$$

If $r_i(t)$ is constant over time then integration of Eq. (1) leads to exponential growth: $n_i(t) = n_i(0)e^{r_i t}$. Per capita growth rates can also be expressed as difference equations by an obvious modification of Eq. (1). The second undeniable property of any biological population is that exponential growth cannot persist indefinitely without resources becoming exhausted. For instance a single bacterium with a mass of 1 picogram that divides by binary fission once every 20 minutes would, if this rate of growth continued without interruption, produce offspring weighting the mass of the Earth in a little under two days. Therefore there must be a negative feedback such that $r_i(t)$ decreases in some fashion as population size, $n_i(t)$, increases. The simplest type of feedback leads to the well-known logistic (or Lotka–Volterra) equation, shown in Eq. (2). Lotka [3]

published his model in 1926 although the original idea is due to Pierre–François Verhulst [4]. Volterra [5] published the genesis of this model in 1926. Hutchinson [4] gives a good historical description of the origins of this equation.

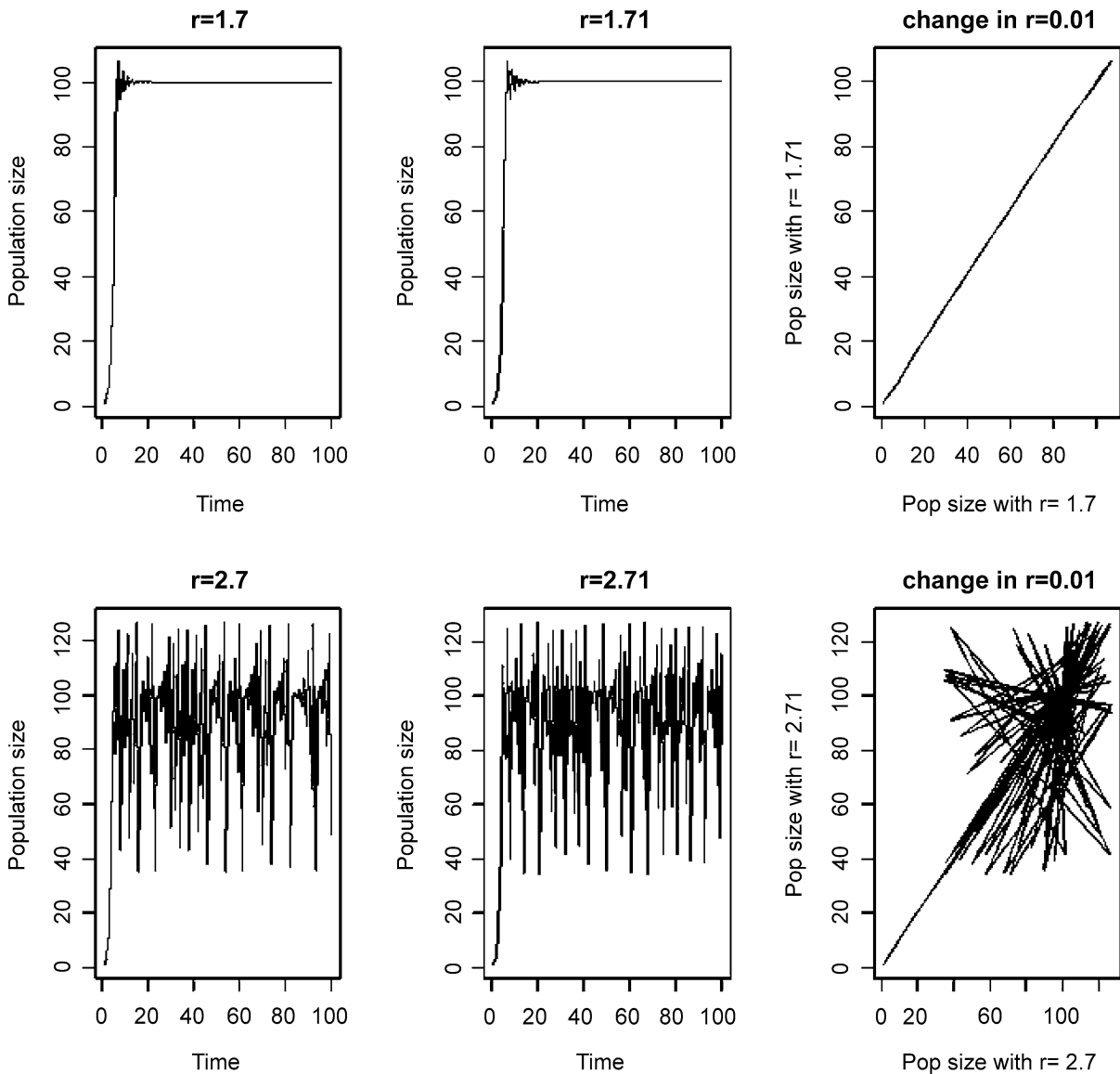
$$\frac{1}{n_i(t)} \frac{dn_i(t)}{dt} = r_i \left(\frac{K_i - \alpha_{ii} n_i(t)}{K_i} \right). \quad (2)$$

Here r_i is the maximum per capita growth rate, attained when $n_i \rightarrow 0$ and therefore when all competition disappears. K_i is the carrying capacity, attained when negative feedback reduces birth rates and increases death rates such that per capita birth and death rates are equal and leading to a constant population size with $r_i(t) = 0$. The coefficient α_{ii} (the per capita competition coefficient) is the amount by which each additional individual of species i decreases the per capita growth rate below the maximum (r_i). The extension to a community of S species leads to a system of simultaneous differential equation that are shown in Eq. (3), where the $S \times S$ matrix of competition coefficients is called the community competition matrix.

$$\frac{1}{n_i(t)} \frac{dn_i(t)}{dt} = r_i \left(\frac{K_i - \sum_{j=1}^S \alpha_{ij} n_j(t)}{K_i} \right). \quad (3)$$

Equation (3) has been modified in many ways, including the addition of time lags, of nonlinear terms, of replacing constants by stochastic terms [6], and replacing the generalized logistic equation by other feedback functions. A more recent modification has been to replace the phenomenological competition coefficients by a dynamic coupling of organisms having a population feedback following a Monod-type equation with an explicit description of dynamic changes in resource levels [7]. This essentially extends the demographic approach to resources as well as organisms. Despite this body of theoretical work, the ability of such demographically-based models to actually predict the structure and dynamics of real ecological communities under field conditions is very weak. In fact, even simplified multispecies communities of phytoplankton [8] or fruit flies reared under controlled laboratory conditions [9] are not well described by these equations despite their intuitive logic. Of course, given the large number of parameters that must be estimated in such models, they are not applicable to real ecological communities in practice.

Recent theoretical and empirical work, involving very simple one-species systems in controlled laboratory conditions, has suggested one reason why such demographic models have low predictive ability even in simplified sys-



Entropy Maximization and Species Abundance, Figure 1

Time series of the deterministic discrete logistic equation showing extreme sensitivity to initial conditions. *Top row:* Time series with the maximum per capita growth rate (r) set at either 1.7 or 1.71 and the relationship of population sizes at each time t . *Bottom row:* Time series with the maximum per capita growth rate (r) set at either 2.7 or 2.71 and the relationship of population sizes at each time t . Note that the same change in the maximum per capita growth rate ($\Delta = 0.01$) will produce changes in population sizes that are very small (*top right*) or very large and essentially random (*bottom right*)

tems. It has been known since May [10] that the simple logistic equation (Eq. (1)) can display deterministic chaos over some ranges of the three parameters (r , α , K). In fact, the basic structure of all such dynamic demographic models in ecology involves nonlinearity and feedback and these two properties are necessary for deterministic chaos. Thus, such models can be sensitive to initial conditions because of the presence of chaotic attractors in phase space.

To illustrate the problem, consider the logistic equation for a single population written as a difference equation: $N(t+1) = N(t) + N(t)r((K - N(t))/(K))$. There are no random variables in this equation and, given the values of r , K and $N(0)$, the subsequent population sizes, $N(t)$, over time are completely deterministic and predictable. The first row of Fig. 1 (top row) shows, from left to right, the changes in population size over time when

$r = 1.7$, when $r = 1.71$ and the relationship between the population sizes at each time when comparing these two dynamics. We see the classic discrete version of the logistic equation and the tiny change (0.01) in the value of r is mirrored by the tiny change in the population sizes over time. The second row of Fig. 1 shows the same thing except that the values of r are 2.7 versus 2.71. The behavior of the population size over time looks completely random. More importantly, the same tiny change in r (0.01) results in trajectories of population size that rapidly diverge and that quickly become essentially independent. In other words, making an error of 0.01 in the estimation of the per capita growth rate would result in completely incorrect predictions of population size. Yet, if r were measured in a real field population, a difference of 0.01 would be far less than the measurement error even for a huge sample size. In such a case the population dynamics is, for all practical purposes, unpredictable. Depending on the equation and the part of parameter space that is being explored, even changes in parameter values of 10^{-6} can produce equally spectacular differences; this is an ecological example of Edward Lorenz' famous "butterfly effect" [11].

A good real-life example of this is the series of modeling and experimental studies described in Constantino et al. [12]. The experimental system was extremely simple: Flour beetles were grown over many generations in standardized containers in the laboratory. Each container received a fixed amount of food at regular intervals. Temperature, humidity and other environmental conditions are maintained at fixed values. A simple nonlinear dynamic model was developed to capture the essential elements of the population dynamics of this species under these carefully controlled conditions. An analysis of the model showed which areas of parameter space resulted in chaotic dynamics (as well as other dynamic properties) and the experimental conditions were then manipulated according to the model predictions in order to bring the measured parameters into these regions of parameter space. The experiments reproduced the various predictions of chaotic dynamics and showed that even very small differences in parameter values could quickly lead to dynamic trajectories that rapidly diverge and soon become indistinguishable from independent random variables.

These experimental results are very bad news for models attempting to model community assembly from this reductionist perspective of population dynamics. If even such a simplified system – a single species growing in controlled conditions in the laboratory – can exhibit such complicated dynamics, then it seems very likely that much more complicated systems, involving hundreds of competing species as well as their predators, prey, parasites and

pathogens existing in constantly fluctuating environmental conditions over many different temporal and spatial scales, will also show at least such complicated dynamics. If sensitivity to initial conditions, a signature of deterministic stochasticity, exists then it seems certain that accurate predictions are beyond the reach of such models in field conditions where parameter measurements are always estimated with a substantial degree of error.

These results, and the inherent difficulty in measuring the large number of parameter values required by dynamic demographic models, suggest that a new approach is needed to link properties of species to community structure.

Information Theory Basis of Entropy Maximization

Conceptually, the entire enterprise of modeling community structure is one of describing the relationship between macroscopic properties of communities (i.e. abundances of each species or the distribution of abundances) and microscopic properties of resource allocation when there are a large number of interacting entities. Stated in this way it is tempting to draw an analogy between, on the one hand, the patterns in ecological communities and the individuals that make up such communities and, on the other hand, the relationship between the macroscopic properties of a gas (pressure, volume, temperature) and the microscopic properties of individual molecules. Historically, the relationship between macroscopic properties of gases (for instance, temperature) and microscopic properties of molecules, for instance, the amount of kinetic and potential energy possessed by each one, was not obtained by solving a system of differential equations for the community of molecules, but rather by a statistical mechanistic approach based on macroscopic constraints [13,14]. Statistical mechanics tries to understand the macroscopic behavior of complex systems, involving large numbers of interacting microscopic components, from a probabilistic viewpoint. Although Boltzmann sought a dynamic explanation, this explanation was refuted by Jaynes [15]. More recent theoretical [16] and experimental studies [17] have confirmed Jaynes claim. Maxwell [14] saw that the second law of thermodynamics was a statistical, not a dynamic and deterministic, law; the seemingly deterministic behavior of macroscopic systems was due to the fact that such behavior was the most likely behavior by a very large margin.

Historically, Boltzmann derived his distribution from microstate counting arguments, but Gibbs [18] showed how this, and other distributions from statistical mechanics, could be derived by maximizing entropy subject to

certain physical constraints. As a way of illustration, consider N entities that can exist in one of S different states. The number of different ways (W) that these N entities can be distributed into these S states such that the macrostate distribution (i. e. the number n_i of entities expressing each state i) is the same is given by the multinomial formula (Eq. (4a)). Taking logarithms, applying Sterling's approximation ($\ln(x!) \approx x \ln(x) - x$ for large values of x), and re-expressing the numbers (n_i) as proportions ($p_i = n_i/N$), we get Shannon's information entropy (Eq. (4b)). We then see that maximizing the number of ways (W) that different microstates will produce the same macroscopic structure, consistent with certain physical constraints, is equivalent to maximizing the information entropy of $\mathbf{p} = \{p_i\}$ subject to these constraints. This link between counting microstates and entropy breaks down when the states vary continuously rather than being discrete with a fixed number, and a more general link is needed; this is explained below. In what follows we will attempt to develop a macroscopic description of ecological communities following this information theoretic approach to statistical mechanics. Although this approach began with Gibbs it was developed into its modern and generalized form, in the context of the Maximum Entropy Formalism, by Jaynes [19,20,21].

$$W = \frac{N!}{n_1!n_2!\dots n_S!} \quad (4a)$$

$$\frac{\ln(W)}{N} \approx H = - \sum_{i=1}^S p_i \ln(p_i) \quad (4b)$$

One begins by defining the average information content of a probability distribution using a Bayesian interpretation of probabilities, recognizing that the set of relative abundances $\mathbf{ra} = \{ra_i\}$ has the formal properties of a probability distribution. To emphasize this equivalence, we will express relative abundances as Bayesian probabilities $\mathbf{ra} = \{p_i\}$ where p_i is the probability that a unit of resource will be allocated to species i given certain information about macroscopic properties of the community. The uncertainty of the information contained in p_i is defined as $I_i = -\ln(p_i)$ and the average uncertainty of information content of a probability distribution, $\sum p_i I_i$, is the information (or Shannon) entropy (Eq. (4b)). For instance, if our available information is sufficient to perfectly predict which of S different states will be expressed by an entity before it happens, then we would have zero uncertainty, we would ascribe $p_i = 1$ to the correct state and ascribe $p_j = 0$ ($i \neq j$) to all others, and the Information entropy would be zero. If, on the other hand, our available information was simply that there are S different and

mutually exclusive states *and nothing else*, then we would ascribe $p_i = 1/S$ for all these states and the Information entropy would be maximal: $-\ln(S)$.

Equation (4b) is valid when the different possible states are discrete and fixed in number. A more general expression, valid for continuous states or when the number of states is not fixed, is the relative entropy expressed with reference to a maximally uninformative (or ignorance) prior probability distribution \mathbf{q} (Eq. (5)). The ignorance prior \mathbf{q} is chosen using transformation groups such that \mathbf{q} is invariant under changes in scale or location; this is equivalent to a scale-invariant measure on the probability space [21]. For a fixed number of unordered discrete states, \mathbf{q} is the discrete uniform distribution and, in this special case, the relative entropy (Eq. (5) equals the entropy (Eq. (4b)). The relative entropy is the negative of the Kullback–Leibler divergence [22] in Bayesian inference, which measures the divergence between two probability distributions (\mathbf{p}, \mathbf{q}).

$$RH = - \sum p_i \ln \left(\frac{p_i}{q_i} \right) \quad (5a)$$

$$RH = - \int p_i \ln \left(\frac{p_i}{q_i} \right) . \quad (5b)$$

The macroscopic constraints on the behavior of the system are expressed as linear functions of products of \mathbf{p} and traits of each of the i states (x_{ij}) in the form of j macroscopic constraints on means (Eq. (6)); examples will be given shortly.

$$\bar{X}_j = \sum x_{ij} p_i . \quad (6)$$

From basic axioms of inductive reasoning [21] we want to ascribe probability values (p_i) that agree with the information encoded in the constraint equations, but that do not imply any additional information beyond this available macroscopic information. In other words, they agree with what we know (the macroscopic constraints), but that is otherwise maximally uncertain (i. e. that doesn't assume further information for which we are lacking). Since Eqs. (4) and (5) each measure the average uncertainty of information content of \mathbf{p} , maximizing uncertainty subject to the constraints is equivalent to maximizing the (relative) entropy subject to the constraints. This is done using the method of Lagrange multipliers. We first specify our objective function (Q , Eq. (7)) which consists of the relative entropy (RH) plus each of the j constraint equations multiplied by its Lagrange multiplier (λ_j) and solve for the values of \mathbf{p} that maximize Q (Eq. (8)). The solution is a generalized exponential distribution (Eq. (9)). Analytical solutions to the values of λ are possible when there

are only a few constraint equations and numerical methods [23,24] are used for more complicated situations.

$$Q = RH + \lambda_0 \left(1 - \sum p_i\right) + \lambda_1 \left(\bar{X}_1 - \sum p_i x_{i1}\right) + \cdots + \lambda_j \left(\bar{X}_j - \sum p_i x_{ij}\right) \quad (7)$$

$$\frac{\partial Q}{\partial p_i} = -(\ln p_i + 1) + \ln q_i - \lambda_0 - \sum_{j=1}^K x_{ij} \lambda_j = 0 \quad (8)$$

$$p_i = \frac{q_i e^{-\sum_{j=1}^K x_{ij} \lambda_j}}{\sum_i q_i e^{-\sum_{j=1}^K x_{ij} \lambda_j}}. \quad (9)$$

The reason why entropy maximization works (when the correct physical constraints are identified) is that if the same macroscopic behavior arises every time we apply the same physical constraints, then these constraints must be sufficient in the explanation and the myriad complexities of the microscopic dynamics must be irrelevant: the overwhelming majority of microstates that are compatible with the macroscopic constraints look the same at the macroscopic level. At the macroscopic level it does not matter which of the macroscopically redundant microstates the system finds itself.

Fixed Species Pools: Predicting Community Composition

This section deals with the case in which the total number of species that can potentially arrive at a given site is known, as are the values of certain morphological, physiological or phenological traits of each species. This list of potential species is called the species pool. In this case, the goal is to predict the relative abundance of each of these species in the pool. A more general, but less precise problem arises when one does not know the number (and therefore the specific identity) of the species making up such a pool. This more general problem will be developed later.

Imagine a collection of genotypes possessing different values of some heritable morphological, physiological or phenological trait x . If different values of this trait cause differences in per capita growth rates of these genotypes then natural selection will result in increases in relative abundance of these genotypes according to the general replicator equation [25] of natural selection (Eq. (10)).

$$\frac{dp_i(t)}{dt} = p_i(t) (r_i(t) - \bar{r}(t)). \quad (10)$$

If we assume for simplicity that the relationship between per capita growth rates and the trait is linear plus a random error term (ε) with zero mean, i.e. $r_i(t) = a(t) + b(t)x_i + \varepsilon_i$, then the general replicator equation is:

$$\frac{dp_i(t)}{dt} = p_i(t) b(t) (x_i - \bar{X}(t)). \quad (11)$$

Equation (11), or more complicated versions if the relationship between per capita growth rates and traits is non-linear, defines a dynamic for community composition. The dynamics of such equations can be very complex, including the presence of chaotic attractors. However complex the dynamic might be, it is a function of the difference between the trait of each genotype, x_i , and the average trait value at a given time, $\bar{X}(t) = \sum p_i(t) X_i$. The trajectory of \bar{X} over time is a consequence of this dynamic and is constrained by it.

The breeder's equation of quantitative genetics (Eqs. (12a), (12b)), or its multivariate equivalent [26], gives the amount by which \bar{X} changes in one time unit (formally, before and after a single selection event). In this equation $h^2(t)$ is the heritability of the trait, varying between 1 (trait values are perfectly transmitted to offspring) and 0 (trait values are independent between parents and offspring) and $S(t)$ is the selection gradient at time t , i.e. the amount by which birth/death probabilities bias genotypes possessing different values of the trait. This equation therefore describes how the average trait value will behave in a given environment and is subject to natural selection. Because these average trait values change systematically over time and across different environmental conditions, they provide information about the trait distribution and are constraints (Eq. (6)). Therefore, if we know the values of these average trait values, and we know the trait values of each species, then we have the same formal mathematical problem as given in the Maximum Entropy Formalism.

$$\bar{X}(t+1) = h^2(t) S(t) + \bar{X}(t) \quad (12a)$$

$$\bar{X}(t) = \sum_{j=0}^t h^2(j) S(j) + \bar{X}(0) \quad (12b)$$

Natural selection, which describes the degree to which genotypes possessing particular traits in a given environmental setting have biased probabilities of reproduction and death, therefore represents a physical constraint on the dynamics of the system. We don't have to know the details of this dynamic so long as we can predict, or measure, these average trait values at the level of the entire community. Natural selection among potentially interbreeding genotypes (i.e. genotypes belonging

to the same species) leads to evolution as genotypes possessing the favored trait replace those of genotypes selected against. However, natural selection among genotypes that are reproductively isolated (i.e. genotypes belonging to different species) cannot lead to evolution since genes causing the particular phenotypic trait cannot be transmitted across the species barrier. In this case, natural selection leads to changes in the relative abundance of different species. In other words, natural selection among genotypes of different species leads to community assembly.

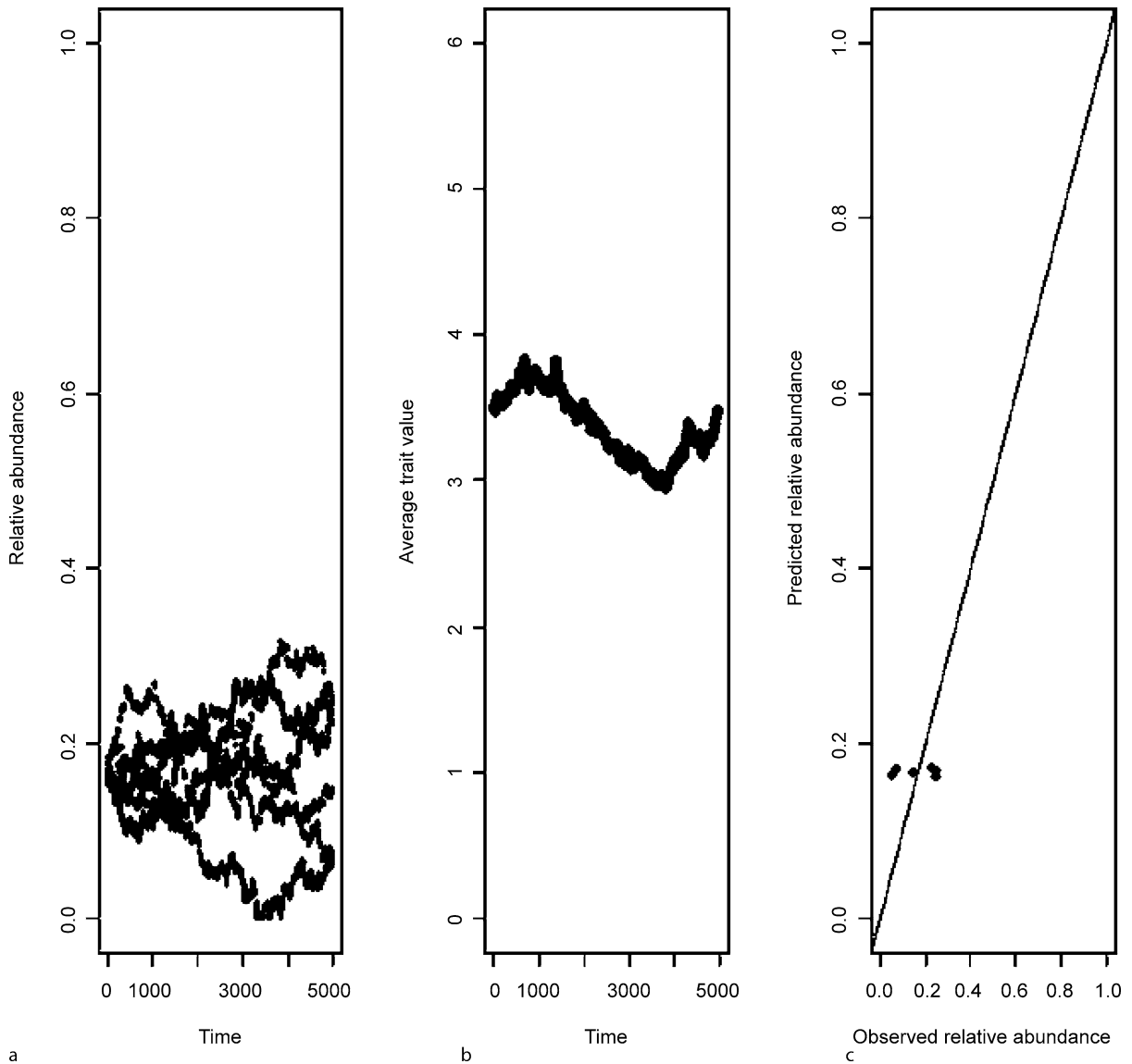
Before describing a field test of this statistical mechanistic approach, it is useful to look at a simple simulation. Consider a species pool consisting of six species, each species having one relevant trait whose value for species i is ($x_i = 1, 2, \dots, 6$). A neutral process of community assembly [27] would consist of each individual of each species having exactly the same probabilities of survival and reproduction irrespective of the trait value, so that any actual differences in population growth rates (i.e. $r_i(t)$, the per capita growth rates) are due purely to sampling fluctuations. We begin with 60 individuals per species and, in each generation, each species adds one new individual (reproduces) with a probability of $1/6$ and loses one existing individual (death) with a probability of $1/6$. Figure 2a shows the results of a simulation over 5000 generations and Fig. 2b shows the change in the average trait value (a community aggregated trait) over time in the community. Because probabilities of birth and death, and thus the effect of natural selection, are independent of the phenotypic trait, the average trait value fluctuates randomly. Notice that, since the trait value has no causal relationship to the probabilities of survival and reproduction and therefore does not constrain the population dynamics, the average trait value is very close to the value predicted by the maximally uninformative prior for this problem which is a discrete uniform distribution ($p_i = 1/6$). Using the average trait value at generation 5000 as a constraint and maximizing the entropy conditional on this average trait value gives, as expected, a maximum entropy distribution which is almost a discrete uniform distribution (Fig. 2c). In other words, since the phenotypic traits did not constrain the dynamics, the average trait value provided no new information that was not already present in the maximally uninformative prior.

Now, we repeat the simulation except that the probabilities of birth and death are functions of the trait x_i : larger values of x_i increase the probability of an individual of species i dying, while larger values of x_i decrease the probability of an individual of species i reproducing. Figure 3 shows the result of a typical simulation run.

Now, since there is a causal connection between the trait value and the probabilities of birth and death, these trait values constrain the population dynamics of each species and therefore the average trait value reflects the direction of natural selection as specified in the breeder's equation. Since natural selection favors individuals with low values of trait x , the average trait value (the community aggregated trait value) decreases over time. Since the trait constrains the dynamics, the average trait value provides information. Using the average trait value at the end of this simulation (generation 1000) and maximizing the entropy conditional on this average trait value correctly predicts the relative abundance of each species at this time (Fig. 3c).

It is possible to make these simulations much more complicated and realistic but the basic result is the same. If trait values are causally connected to probabilities of survival, reproduction, immigration and emigration then the population dynamics of the species will be constrained by them as specified in the general replicator equation (Eq. (11)) and the changes in these average trait values over time will be described by the breeder's equation or its multivariate version if (as there always is) there is more than one trait involved in determining fitness. In fact, since the general replicator equation is a mathematical expression of the notion of evolutionary fitness that links heritable phenotypic traits to probabilities of survival and reproduction relative to competing genotypes, such community average trait values are tracking the constraint of fitness over time and this is why they provide information on relative abundances.

An empirical example can be found in the context of secondary succession in an herbaceous plant community. Secondary succession is the well-known process of re-colonization and subsequent vegetation change following a major disturbance that kills off an existing plant community. A classic example is the pattern of vegetation change over time following abandonment of an agricultural field. If one lists the species and their relative abundances each year following abandonment in different regions and compares these lists then the patterns are unintelligible. Different species will be found in each list. On the other hand, if one lists the values of key morphological, physiological and phenological traits related to the ability of species to survive and reproduce that are found over time then very consistent patterns emerge. If one concentrates on the traits possessed by the more abundant species in each list over time, the patterns are even more consistent. Immediately following abandonment the common species will possess traits that allow them to rapidly colonize the newly opened space [28]. For instance, species

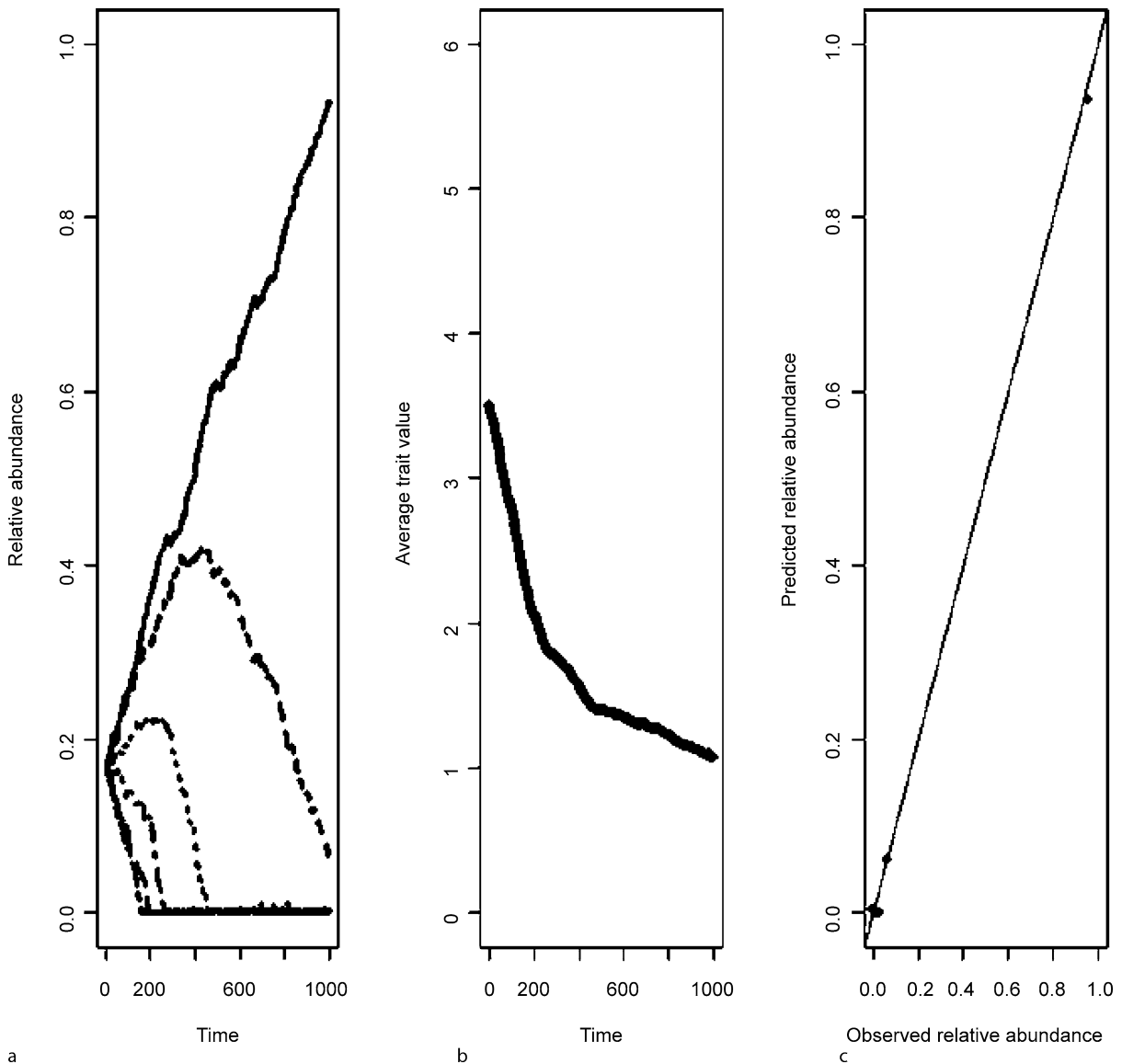


Entropy Maximization and Species Abundance, Figure 2

a An example of a neutral dynamic involving six species. Probabilities of birth and death are equal and all fluctuations in population numbers are due to sampling fluctuations. **b** Changes in the community averaged (aggregated) trait value over time that is produced by such a neutral dynamic. **c** Predicted relative abundances, given the community aggregated trait value at the end of the series, using the Maximum Entropy Formalism

having rapid growth rate and having small seeds, and especially seeds with appendages that allow for long distance dispersal, are very common. Species whose individuals produce large numbers of seeds are common since larger numbers of seeds increase the chances of some arriving at the site. Species adapted to sites having frequent large mortality events reproduce very rapidly following germination and this limits their size. Small plants that are required to

produce many seeds must further decrease the size of each one, and so on. Such ruderal species have physiologies that are adapted to quickly acquire available resources and allocate then to production rather than maintenance and this leads to a correlated series of traits (specific leaf area, maximum photosynthetic rate, tissue nutrient supplies, tissue turnover rates) that maximize production of new tissues over maintenance of existing ones [29,30]. As a vegetation



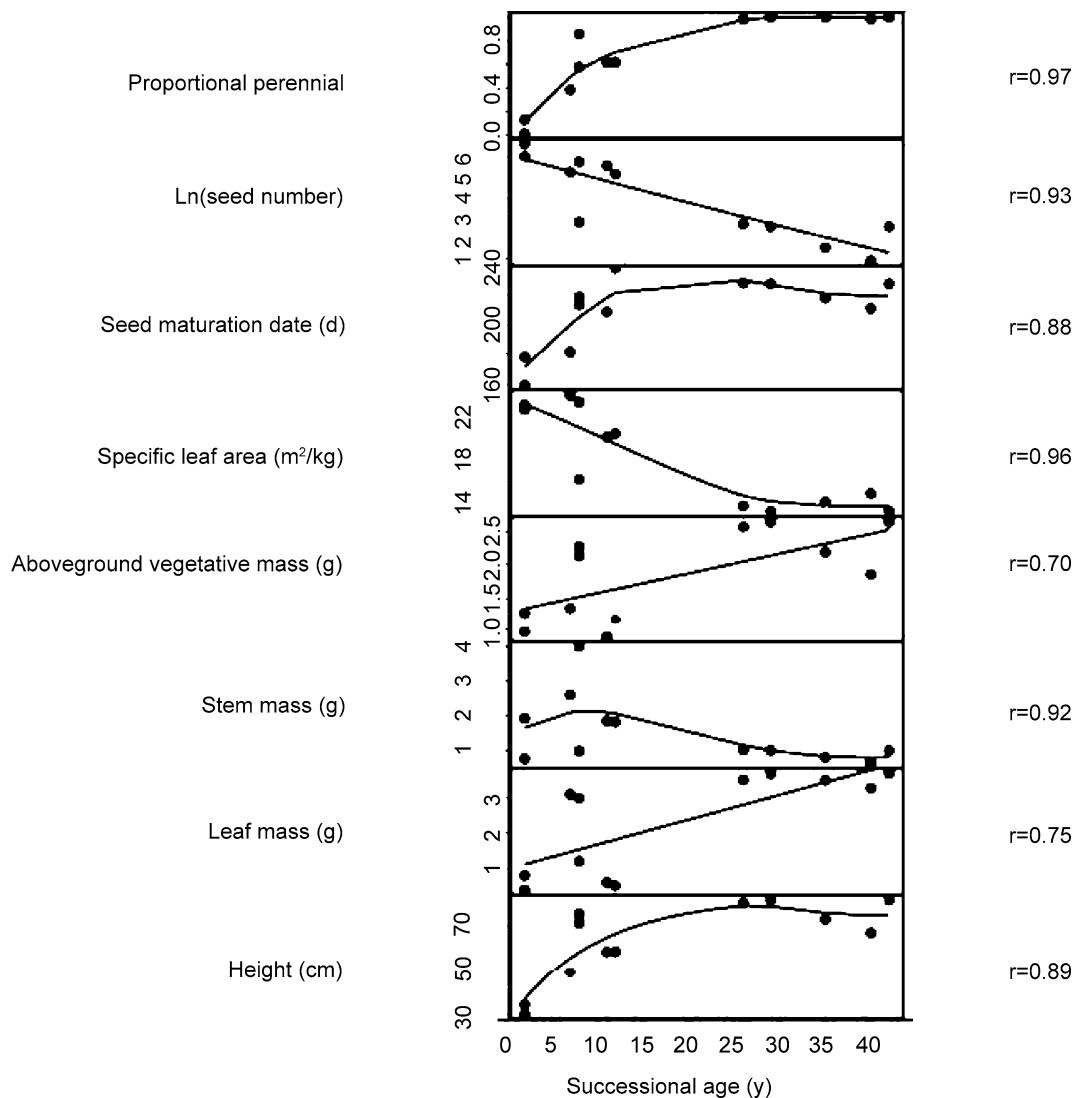
Entropy Maximization and Species Abundance, Figure 3

a An example of a dynamic involving six species in which one trait determines the differential probabilities of birth and death plus fluctuations in population numbers due to sampling fluctuations. **b** Changes in the community averaged (aggregated) trait value over time that is produced by such a dynamic. **c** Predicted relative abundances, given the community aggregated trait value at the end of the series, using the Maximum Entropy Formalism

cover develops, these ruderal species are replaced by a series of other species, larger and better able to competitively suppress those coming before, and this leads to changes in the suite of traits possessed by such species.

A recent structural equation model linking such plant traits to time following abandonment of vineyards in southern France was published by Vile et al. [31]. To test the ability of a maximum entropy approach to pre-

dicting secondary succession, Shipley et al. [32] used the same series of 10 sites. The sites varied from two to 42 years post-abandonment and estimates of the above-ground biomass of each species at each site were obtained from two $0.5 \times 0.5 \text{ m}^2$ quadrants in each. All were within a 4 km^2 distance and all had the same history of land use before abandonment. Using the 30 species in these that were more than transients in these 12 sites as the species



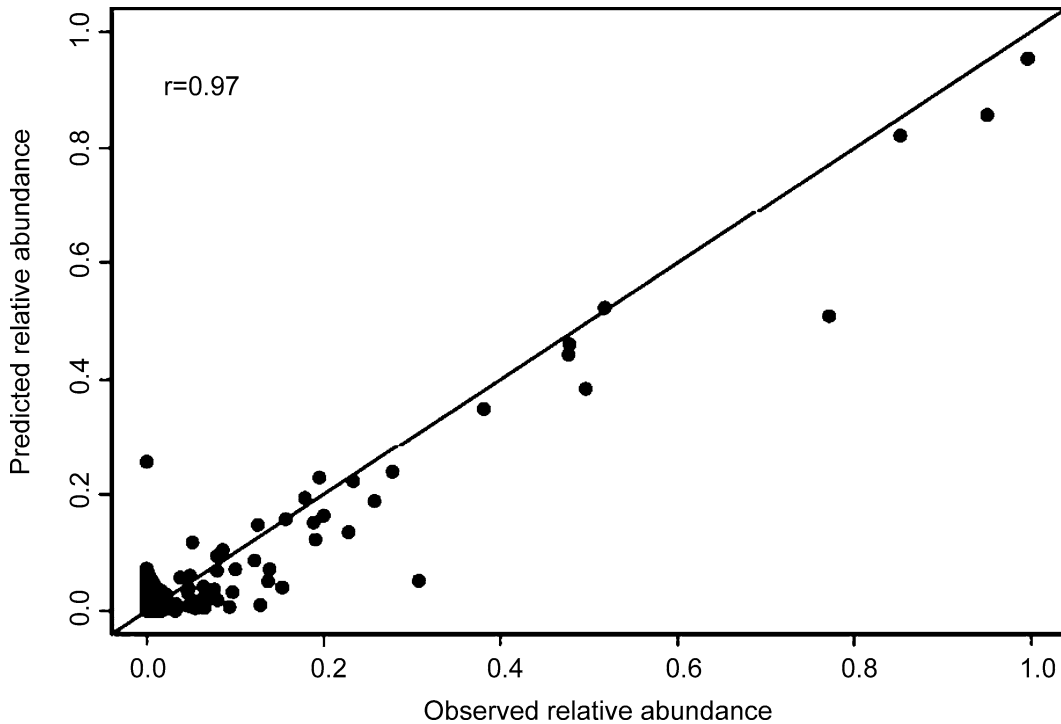
Entropy Maximization and Species Abundance, Figure 4
Each point represents the community aggregated value of the trait measured in one field, whose successional age (years since abandonment) is given on the ordinate. Lines are regression (cubic spline) smoothers. Also shown is the Pearson correlation between observed and predicted values. Figure taken from [32]

pool, and using eight traits measured on each of these 30 species, we calculated the community average (aggregated) values for each trait in each site. Figure 4 shows the results. As can be seen, there are clear systematic trends over time in these community averaged traits and these are consistent with trends reported by others.

If these traits are causally linked to fitness of these 30 species over this successional sequence then they will constrain the actual population dynamics as individual plants possessing different values of these traits immigrate into the sites, compete and reproduce. Using these eight com-

munity averaged trait values for each site, we then calculated the relative abundances assuming that these traits are sufficient to determine the population dynamics except for random sampling fluctuation. We did this by finding the distributions of relative abundance having maximum entropy after constraining then to agree with the community averaged traits. This was done using an algorithm published by Della Pietra et al. [24]. Figure 5 shows the result.

As can be seen, the predicted and observed values are close and highly significant. Clearly, there are prediction errors and this is probably be due to a combination of



Entropy Maximization and Species Abundance, Figure 5

Observed and predicted relative abundances of each of 30 species in each of 10 different fields. Predicted relative abundances were obtained by maximizing the entropy of the distribution of relative abundances for each field conditional on the eight community aggregated traits shown in Fig. 4

missing constraints (i. e. unmeasured traits that contribute to fitness), measurement error of the trait values and random demographic fluctuations due to the fact that the numbers of individuals per site are sufficiently small that probabilities of survival and reproduction and actual frequencies of survival and reproduction differ due to sampling effects.

Species Abundance Distributions

The previous section considered the case in which we know which species occur in the species pool and in which interest resides in predicting the relative abundance of each species given certain traits. In such a situation the number of species (i. e. states in which biomass can be allocated) is fixed. However, one can also imagine situations in which the number and identity of species in the species pool is unknown. In this case we cannot predict the abundance of any particular species but we do know the distribution of abundances of the species that are present. This leads to the species abundance distribution (SAD) which describes the probability that a species will have an abundance of n individuals.

The empirical study of species abundance distributions has a long history in ecology. The first known treatment was published in Japanese by Motomira [33], as cited in Hutchinson [4]. The proposed distribution was the geometric series (Eq. (13)). Corbet [34] published a species abundance distribution based on species of butterflies and this distribution, like all others since, shows that a few species are very common and most species are very rare. C.B. Williams had collected similar data for moths and so these two authors, in collaboration with the statistician and evolutionary theorist R.A. Fisher, put forward the log-series distribution (Eq. (14)) as a statistical description of the species abundance relationship [35]. Pueyo [36], applying a Taylor series expansion to the log-series distribution, found that small deviations from a log-series gave a bounded power law (Eq. (15)). F.W. Preston, working with birds, plotted the number of species having different numbers of individuals (i. e. abundance) by binning values on a logarithmic scale and showed that, in this form, the species abundance distribution was a log-normal distribution [37,38], although any variable that must be greater than zero and is strongly right-skewed will tend to show a curve similar to a log-normal distribution after binning.

More recently, so-called neutral models of population dynamics [27,39] have been shown to recreate these patterns. Neutral models assume that all species have equal fitnesses (thus probabilities of survival, reproduction and migration) and thus all changes in population size are due to sampling fluctuations. Hubbell [27], in particular, has proposed a slightly different SAD, the zero-sum multinomial. However, differentiating between these different distributions using empirical data has proved problematic because they are very similar over most of the range of observed values [40]. Even more problematic is that authors searching for specifically biological explanations for these patterns tend to ignore that such statistical patterns are commonly found in non-biological systems. Significantly for our purposes, Preston [41] noted the similarity of species abundance distributions and gas laws as well as the distribution of wealth. Limpert et al. [42] and Nekola and Brown [43] document other cases. If the patterns transcend ecology then the explanation for the patterns must also transcend ecology.

$$P(n) = \frac{\Psi}{n} \quad (13)$$

$$p(n) = \frac{\phi x^n}{n} = \frac{\phi (e^{-\beta})^n}{n} = \frac{\phi e^{-\beta n}}{n} \quad (14)$$

$$p(n) = \frac{\phi e^{-\beta_1 n}}{n \beta_2} \quad (15)$$

Very recently, two different groups have proposed that the common patterns of species abundance distributions can be more parsimoniously explained by the same statistical mechanistic approach as described above. The first hurdle in applying a maximum entropy approach to this area is to obtain the maximally uninformative prior. In the case of a fixed species pool the maximally uninformative prior is well-known: a discrete uniform distribution. However, when the number of species in the species pool is unknown and abundance is measured in numbers of individuals (a discrete variable) the prior is not clear. If abundance was measured as biomass, a continuous variable, then the maximally uninformative prior is known based on a consideration of invariance under transformations of scale and location [21] and is a Jeffrey's prior. Pueyo et al. [44], again based on a consideration of invariance under transformations of scale and location have shown that when abundance is measured as counts of individuals the same maximally uninformative prior is obtained: $q(n) = 1/n$.

The first constraint that one might propose is a limit on the total number of individuals (N) that can exist at a site, at least when averaged over time. This follows from

the fact that the total amount of resources within a fixed area is itself fixed but also follows from basic postulates of per capita growth, namely that as populations increase the increased competition for resources leads to a decrease in per capita reproductive rates and an increase in per capita death rates, until the per capita growth rate is zero. The second constraint is that the total number of species (S) that can exist at a site is also fixed, at least when averaged over time. This follows from the theory of island biogeography [45] where, as the available resources at a site decrease, the rate of introduction of new species decreases and the rate of local extinction of species already present increases, each due to increased competition. If we accept these two macroscopic constraints, then the average number of individuals per species must also be constrained (Eq. (16)).

$$\bar{n} = \frac{N}{S} \approx k_1. \quad (16)$$

To find the predicted proportion of species having abundance n in the community, and assuming that the only constraint acting on the system is to limit the average number of individuals per species (Eq. (16)), we must maximize the relative entropy conditional on this one constraint. In other words, find the values of \mathbf{p} that maximize Eq. (5a), subject to the constraint given in Eq. (16) plus the constraint on normalization ($\sum p_i = 1$). The solution, given in Eq. (17), is Fisher's log-series distribution where Z is a constant.

$$p_i = \frac{q_i e^{-\lambda_1 n_i}}{\sum_i q_i e^{-\lambda_1 n_i}} = \frac{e^{-\lambda_1 n_i}}{n_i Z}. \quad (17)$$

We know that, in any real ecological community, the process of community assembly involves a myriad of complicated interactions between the individuals of each species with each other and with those of other species plus the interactions of each with the physical properties of the environment. Equation (17) says that the statistical pattern of species relative abundance that was fit to empirical data by Fisher, Corbet and Williams is simply a consequence of two constraints: the total number of individuals at the site is (approximately) constant and the total number of species at the site is (approximately) constant. Equation (17) doesn't deny the existence of the complicated biological interactions that are involved in community assembly but does claim that these interactions are irrelevant with respect to the macroscopic pattern, *except* in how they constrain the total number of individuals and species. Pueyo et al. [44] call this an idiosyncratic model of community assembly: the exact process (or model) determining the abundance of species i is independent of the

process determining the abundance of every other species, therefore knowing the abundance of species i provides no information about the abundance of species j . Thus, from a macroscopic perspective, the abundances of each species at any point in time can be viewed as independent and identically distributed variables even though independent does not imply no interactions and identically distributed does not imply ecologically equivalent.

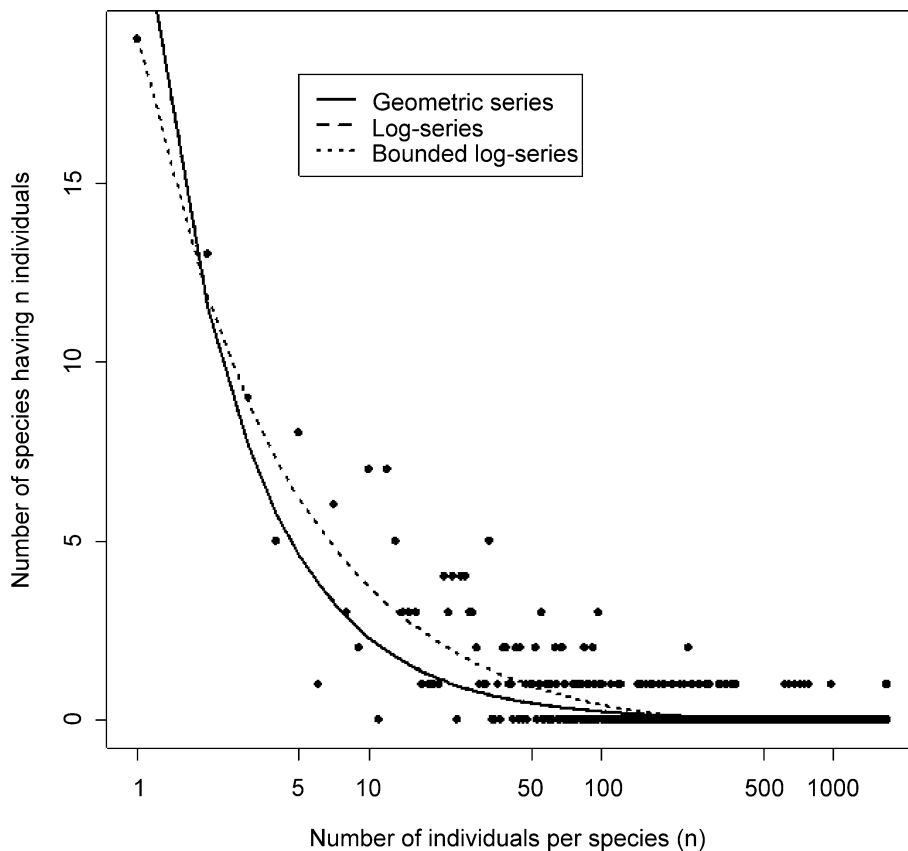
Can we imagine other macroscopic constraints that might exist in real communities? The per capita growth rate of species i over one small unit of time (Δt) is given as $r_i(t + \Delta t) = \ln(n_i(t + 1)) - \ln(n_i(t))$. If each species in the community is at a stable equilibrium by time t then $r_i(t) \approx 0$ for all species and so the following macroscopic constraint must hold:

$$\frac{\sum \ln(n(t+1))}{S} = \frac{\sum \ln(n(t))}{S} = k_2. \quad (18)$$

Given the per capita growth rate describes exponential growth at time t , and since we are imagining a community

that is filled with individuals, it seems quite reasonable that Eq. (18) will hold; if it did not hold over any extended time then there would be huge swings in total biomass, and this is not observed. Maximizing the relative entropy conditional on the constraints in Eqs. (16) and (18), we recover the bounded power law of Pueyo [36], as given in Eq. (15).

Finally, we could imagine a case in which the community dynamics are such that the population trajectories are in a basin of attraction but in which random external perturbations result in total resource levels fluctuations around some average values so that the per capita growth rates, $r_i(t)$, are zero over a longer time scale but can randomly deviate from zero at any instant. The fact that the trajectories are in a basin of attraction means that there will be a constraint on how far the $r_i(t)$ values can deviate from zero and this will place a constraint on the variance of the $r_i(t)$ values and this would produce a bounded log-normal distribution when maximizing the relative entropy. Since this last constraint is quite speculative at this point, we won't explore this point further.



Entropy Maximization and Species Abundance, Figure 6

Distribution of abundances of tree species in a 50 hectare plot in Barro Colorado Island (Panama). Three theoretical curves, derived the maximization of relative entropy, are shown. The Geometric and Log-series curves are indistinguishable at this scale

To place these results in an empirical context, I have fit the geometric series of Motomura, the log-series of Fisher, and the bounded log-series of Pueyo to the extensive data set of tree species on Barro Colorado Island (Panama); these data come from the appendix of Volkov [46]. This data set consists of a complete census of each tree whose stem diameter at breast height is greater than 1 cm within a 50 hectare area [47]. The resulting curves are given in Eq. (19) and the result is shown in Fig. 6. The fitted curves for the geometric series and the log-series are not distinguishable and are not significantly different ($F_{1,1715} = 2.78$, $p = 0.10$) but the fitted curves for the log-series and the bounded log-series are highly significant ($F_{1,1715} = 534.43$, $p \ll 0.001$).

$$p_i = \frac{1}{933.32n} \quad (19a)$$

$$p_i = \frac{e^{-(4.263e-4)n}}{935.2n} \quad (19b)$$

$$p_i = \frac{e^{-(6.78e-3)n}}{1121n^{0.68}} \quad (19c)$$

Future Directions

Before discussing future directions, it is important to appreciate how little published work has been done in ecology using entropy maximization. Plank studied statistical mechanistic properties of the Lotka–Volterra equations [48,49,50]. A few studies have considered some theoretical aspects of entropy maximization in food webs [51, 52,53,54]. Levich and co-workers [55,56,57] applied entropy maximization to algal communities but without including maximally uninformative priors. Besides [44], discussed above, I am not aware of any other work in this area. Clearly, future directions are many and varied.

At a theoretical level it will be important to more precisely link the dynamics of natural selection and the allocation of energy and resources to individuals to microstate properties of communities. In this context Dewar's fluctuation theorem [58,59,60] might prove fruitful since it links entropy maximization with those dynamic trajectories of microstates that maximize the production of entropy. If possible this could open up a thermodynamic interpretation of community assembly and possibly even of evolution by natural selection.

Those few empirical results that have been obtained so far come from plant communities. In part, this is because it is easier to census plant communities, and therefore to get reasonably accurate determinations of abundance for all species present, but the main reason is probably idiosyncratic: the application of these ideas to ecology is very

recent and those very few ecologists who have explored them have been trained in plant ecology. The extension of these methods to animal communities is an obvious future direction. The use of constraints based on food webs seems particularly promising since such constraints can be linked to basic thermodynamic principles of energy transfer across trophic levels. As an example, homeotherms and poikilotherms have different basal metabolic rates and so the proportion of these two types of animals should constrain the length of food chains. Metabolic rate scales allometrically with body size [61] and so the average body size in a community should be constrained by the total amount of available energy and this, in turn, will be related to potential evapotranspiration (a function of average temperatures and precipitation). The ratio of predators to prey also scales with body size and so this might also serve as a constraint on the assembly of animal communities.

At an empirical level, it is clear that the success or failure of this approach resides in our ability to properly identify those physical constraints that actually control community dynamics and how such constraints might change as a function of environmental variables. This requires further work in order to make statistical inferences in the context of maximum entropy models and, especially, in detecting lack of fit. When the entities being allocated to different states are discrete and mutually independent and the states themselves are also discrete and fixed in number then the problem is a classical one. Given such assumptions then the statistic $-2N \sum p_i \ln p_i$ follows a chi-squared distribution with $N - k - 1$ degrees of freedom (k = the number of constraints). However, such conditions rarely apply to ecological problems either because the sequential allocations are not mutually independent or because the total number (N) is unknown.

Roderick Dewar and Annabel Porté (INRA Centre de Bordeaux–Aquitaine, France) have been exploring other directions (pers. comm.). For instance, if we consider that the total amount of available resources (R) in an area is limited, that the organisms are using all of these resources, and that each species has a characteristic per capita rate of resource use, then the following constraint must hold: $\sum p_i r_i = R$. If this is combined with some well-known scaling laws in ecology relating body size with per capita resource use then, even without knowing the number of identity of species in a species pool, one can derive the species abundance distribution. Using the exponent of the Shannon entropy as an expression of the expected number of species, they explore how species richness (\hat{S} , the number of species per unit area) should vary as a function of R at different scales. At local spatial scales it is well known that \hat{S} has a so-called humped-back distribution [62] with

respect to productivity, and therefore R . At such local spatial scales, it is reasonable to consider R fixed and S (the number of species in the species pool) to be finite. At spatial scales large enough to include substantial variation in mean annual temperature and precipitation \hat{S} scales allometrically with potential evapotranspiration [63] and it is perhaps reasonable to consider R free. Can these contrasting patterns of biodiversity at different spatial scales be accounted for by a statistical mechanistic approach in which constraints on R and S change?

Bibliography

1. Darwin C (1859) On the origin of species. John Murray, London
2. Wallace AR (1858) On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. III. On the tendency of varieties to depart indefinitely from the original type. *J Proc Linnean Soc Lond* 3:53–62
3. Lotka AJ (1925) Elements of physical biology. Williams and Williams, Baltimore
4. Hutchinson GE (1978) An introduction to population ecology. Yale University Press, New Haven
5. Volterra V (1926) Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. *Mem R Acad dei Lincei* (Ser. 6) 2:31–113
6. Roughgarden J (1979) Theory of population genetics and evolutionary ecology: An introduction. Macmillan Publishing, New York
7. Tilman D (1982) Resource competition and community structure. Princeton University Press, Princeton
8. Vandermeer JH (1969) The competitive structure of communities: An experimental approach with protozoa. *Ecology* 50: 362–371
9. Gilpin ME, Carpenter MP, Pomerantz MJ (1986) The assembly of a laboratory community: Multispecies competition in *Drosophila*. In: Diamond JM, Case T (eds) Community ecology. Harper and Row, Cambridge
10. May RA (1974) Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science* 186:645–647
11. Hilborn RC (2004) Seagulls, butterflies, and grasshoppers: A brief history of the butterfly effect in nonlinear dynamics. *Am J Phys* 72:425–427
12. Costantino RF et al (2005) Nonlinear stochastic population dynamics: The flour beetle *Tribolium* as an effective tool of discovery. *Advances in ecological research: Population dynamics and laboratory ecology*, vol 37. Academic Press, London, pp 101–141
13. Boltzmann L (1898) Lectures on gas theory. Dover Publications, New York
14. Maxwell JC (1871) Theory of heat. Dover Publications, New York
15. Jaynes ET (1971) Violations of Boltzmann's h theorem in real gases. *Phys Rev A* 4:747–751
16. Evans DJ, Searles DJ (2002) The fluctuation theorem. *Adv Phys* 51:1529–1585
17. Wang GM, Sevcik EM, Mittag E, Searles DJ, Evans DJ (2002) Experimental demonstration of violations of the second law of thermodynamics for small systems and short time scales. *Phys Rev Lett* 89:50601
18. Gibbs JW (1902) Elementary principles in statistical mechanics. Yale University Press, New Haven
19. Jaynes ET (1957) Information theory and statistical mechanics i. *Phys Rev* 106:620–630
20. Jaynes ET (1957) Information theory and statistical mechanics ii. *Phys Rev* 108:171–190
21. Jaynes ET (2003) Probability theory. The logic of science. Cambridge University Press, Cambridge
22. Kullback S, Leibler RA (1951) On information and sufficiency. *Ann Math Statist* 22:79–86
23. Agmon N, Alhassid Y, Levine RD (1979) An algorithm for finding the distribution of maximal entropy. *J Comput Phys* 30:250–258
24. Della Pietra S, Della Pietra V, Lafferty J (1997) Inducing features of random fields. *IEEE Trans Pattern Anal Mach Intell* 19: 1–13
25. Schuster P, Sigmund K (1983) Replicator dynamics. *J Theor Biol* 100:533–538
26. Roff DA (1997) Evolutionary quantitative genetics. Chapman and Hall, NY
27. Hubbell SP (2001) The unified neutral theory of biodiversity and biogeography. Princeton University Press, Princeton
28. Grime JP (2007) Comparative plant ecology. Castlepoint Press, Colvend
29. Shipley B, Lechowicz MJ, Wright I, Reich PB (2006) Fundamental trade-offs generating the worldwide leaf economics spectrum. *Ecology* 87:535–541
30. Wright IJ et al (2004) The worldwide leaf economics spectrum. *Nature* 428:821–827
31. Vile D, Shipley B, Garnier E (2006) A structural equation model to integrate changes in functional strategies during old-field succession. *Ecology* 87:504–517
32. Shipley B, Vile D, Garnier E (2006) From plant traits to plant communities: A statistical mechanistic approach to biodiversity. *Science* 314:812–814
33. Motomura I (1932) A statistical treatment of associations [in Japanese]. *Jpn J Zool* 44:379–383
34. Corbet AS (1941) The distribution of butterflies in the Malay peninsula (Lepid.). *Proc Roy Entomol Soc Ser A* 16:101–116
35. Fisher RA, Corbet AS, Williams CB (1943) The relation between the number of species and the number of individuals in a random sample from an animal population. *J Animal Ecol* 12: 42–58
36. Pueyo S (2006) Diversity: Between neutrality and structure. *Oikos* 112:392–405
37. Preston FW (1948) The commonness, and rarity, of species. *Ecology* 29:254–283
38. Preston FW (1962) The canonical distribution of commonness and rarity. *Ecology* 43:185–215
39. Bell G (2000) The distribution of abundance in neutral communities. *Am Nat* 155:606–617
40. McGill BJ, Maurer BA, Weiser MD (2006) Empirical evaluation of neutral theory. *Ecology* 87:1411–1423
41. Preston FW (1950) Gas laws and wealth laws. *Sci Mon* 71:309–311
42. Limpert E, Stahel WA, Abbt M (2001) Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51:341–352

43. Nekola JC, Brown JH (2007) The wealth of species: Ecological communities, complex systems and the legacy of frank preston. *Ecol Lett* 10:188–196
44. Pueyo S, He F, Zillio T (2007) The maximum entropy formalism and the idiosyncratic theory of biodiversity. *Ecol Lett* 10:1017–1028
45. MacArthur RH (1972) *Geographical ecology: Patterns in the distribution of species*. Harper and Row, NY
46. Volkov I, Banavar JR, He FL, Hubbell SP, Maritan A (2005) Density dependence explains tree species abundance and diversity in tropical forests. *Nature* 438:658–661
47. Hubbell SP, Foster RB (1990) Structure, dynamics and equilibrium status of old-growth forest on barro colorado island. In: Gentry A (ed) *Four neotropical forests*. Yale University Press, New Haven
48. Plank M (1995) Hamiltonian structures for the n -dimensional Lotka–Volterra equations. *J Math Phys* 36:3520–3534
49. Plank M (1996) Bi-hamiltonian systems and Lotka–Volterra equations: A three-dimensional classification. *Nonlinearity* 9:887–896
50. Plank M (1999) On the dynamics of Lotka–Volterra equations having an invariant hyperplane. *Siam J Appl Math* 59:1540–1551
51. Kerner EH (1964) *Gibbs ensemble: Biological ensemble*. Gordon and Breach, New York
52. Kerner EH (1979) The gibbs grand ensemble and the eco-genetic gap. In: Levine RD, Tribus M (eds) *The maximum entropy formalism*. Massachusetts Institute of Technology, Massachusetts, pp 468–476
53. Wagensberg J, Valls J (1987) The [extended] maximum entropy formalism and the statistical structure of ecosystems. *Bull Math Biol* 48:531–538
54. Lurie D, Wasenburger J (1985) An extremal principle for biomass diversity in ecology. In: Lamprecht I, Zotin AI (eds) *Thermodynamics and regulation of biological processes*. De Gruyter, Berlin, pp 577–271
55. Levich AP (1988) What are the possible theoretical principles in the ecology of communities? In: Kull K, Tiivel T (eds) *Lectures in theoretical biology*. Valgus, Tallinn, pp 121–127
56. Alexeyev VL, Levich AP (1997) A search for maximum species abundances in ecological communities under conditional diversity optimization. *Bull Math Biol* 59:649–677
57. Levich AP (2000) Variational modelling theorems and algorithmic functioning principles. *Ecol Model* 131:207–227
58. Dewar R (2003) Information theory explanation of the fluctuation theorem, maximum entropy production and self-organized criticality in non-equilibrium stationary states. *J Phys Math Gen* 36:631–641
59. Dewar R (2004) Maximum entropy production and non-equilibrium statistical mechanics. In: Klinedon A, Lorenz RD (eds) *Non-equilibrium thermodynamics and the production of entropy: Life, earth and beyond*. Springer, Berlin, pp 41–56
60. Dewar R (2005) Maximum entropy production and the fluctuation theorem. *J Phys Math Gen* 38:L371–L381
61. Peters RH (1983) *The ecological implications of body size*. Cambridge University Press, Cambridge
62. Al-Mufti MM, Sydes CL, Furness SB, Grime JP, Band SR (1977) A quantitative analysis of shoot phenology and dominance in herbaceous vegetation. *J Ecol* 65:759–791
63. Currie DJ, Paquin V (1987) Large-scale biogeographic patterns of species richness of trees. *Nature* 329:326–327

Ergodicity and Mixing Properties

ANTHONY QUAS

Department of Mathematics and Statistics,
University of Victoria, Victoria, Canada

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Basics and Examples](#)

[Ergodicity](#)

[Ergodic Decomposition](#)

[Mixing](#)

[Hyperbolicity and Decay of Correlations](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Bernoulli shift Mathematical abstraction of the scenario in statistics or probability in which one performs repeated independent identical experiments.

Markov chain A probability model describing a sequence of observations made at regularly spaced time intervals such that at each time, the probability distribution of the subsequent observation depends only on the current observation and not on prior observations.

Measure-preserving transformation A map from a measure space to itself such that for each measurable subset of the space, it has the same measure as its inverse image under the map.

Measure-theoretic entropy A non-negative (possibly infinite) real number describing the complexity of a measure-preserving transformation.

Product transformation Given a pair of measure-preserving transformations: T of X and S of Y , the product transformation is the map of $X \times Y$ given by $(T \times S)(x, y) = (T(x), S(y))$.

Definition of the Subject

Many physical phenomena in equilibrium can be modeled as measure-preserving transformations. Ergodic theory is the abstract study of these transformations, dealing in particular with their long term average behavior.

One of the basic steps in analyzing a measure-preserving transformation is to break it down into its simplest possible components. These simplest components are its ergodic components, and on each of these components, the system enjoys the ergodic property: the long-term time

average of any measurement as the system evolves is equal to the average over the component. Ergodic decomposition gives a precise description of the manner in which a system can be split into ergodic components.

A related (stronger) property of a measure-preserving transformation is mixing. Here one is investigating the correlation between the state of the system at different times. The system is mixing if the states are asymptotically independent: as the times between the measurements increase to infinity, the observed values of the measurements at those times become independent.

Introduction

The term ergodic was introduced by Boltzmann [8,9] in his work on statistical mechanics, where he was studying Hamiltonian systems with large numbers of particles. The system is described at any time by a point of *phase space*, a subset of \mathbb{R}^{6N} where N is the number of particles. The configuration describes the 3-dimensional position and velocity of each of the N particles. It has long been known that the Hamiltonian (i. e. the overall energy of the system) is invariant over time in these systems. Thus, given a starting configuration, all future configurations as the system evolves lie on the same *energy surface* as the initial one.

Boltzmann's *ergodic hypothesis* was that the trajectory of the configuration in phase space would fill out the entire energy surface. The term ergodic is thus an amalgamation of the Greek words for work and path. This hypothesis then allowed Boltzmann to conclude that the long-term average of a quantity as the system evolves would be equal to its average value over the phase space.

Subsequently, it was realized that this hypothesis is rarely satisfied. The ergodic hypothesis was replaced in 1911 by the *quasi-ergodic hypothesis* of the Ehrenfests [17] which stated instead that each trajectory is dense in the energy surface, rather than filling out the entire energy surface. The modern notion of ergodicity (to be defined below) is due to Birkhoff and Smith [7]. Koopman [44] suggested studying a measure-preserving transformation by means of the associated isometry on Hilbert space, $U_T: L^2(X) \rightarrow L^2(X)$ defined by $U_T(f) = f \circ T$. This point of view was used by von Neumann [91] in his proof of the mean ergodic theorem. This was followed closely by Birkhoff [6] proving the pointwise ergodic theorem. An ergodic measure-preserving transformation enjoys the property that Boltzmann first intended to deduce from his hypothesis: that long-term averages of an observable quantity coincide with the integral of that quantity over the phase space.

These theorems allow one to deduce a form of independence on the average: given two sets of configurations A and B , one can consider the volume of the phase space consisting of points that are in A at time 0 and in B at time t . In an ergodic measure-preserving transformation, if one computes the average of the volumes of these regions over time, the ergodic theorems mentioned above allow one to deduce that the limit is simply the product of the volume of A and the volume of B . This is the weakest mixing-type property. In this article, we will outline a rather full range of mixing properties with ergodicity at the weakest end and the Bernoulli property at the strongest end.

We will set out in some detail the various mixing properties, basing our study on a number of concrete examples sitting at various points of this hierarchy. Many of the mixing properties may be characterized in terms of the Koopman operators mentioned above (i. e. they are *spectral properties*), but we will see that the strongest mixing properties are not spectral in nature.

We shall also see that there are connections between the range of mixing properties that we discuss and measure-theoretic entropy. In measure-preserving transformations that arise in practice, there is a correlation between strong mixing properties and positive entropy, although many of these properties are logically independent.

One important issue for which many questions remain open is that of higher-order mixing. Here, the issue is if instead of asking that the observations at two times separated by a large time T be approximately independent, one asks whether if one makes observations at more times, each pair suitably separated, the results can be expected to be approximately independent. This issue has an analogue in probability theory, where it is well-known that it is possible to have a collection of random variables that are pairwise independent, but not mutually independent.

Basics and Examples

In this article, except where otherwise stated, the measure-preserving transformations that we consider will be defined on probability spaces.

More specifically, given a measurable space (X, \mathcal{B}) and a probability measure μ defined on \mathcal{B} , a *measure-preserving transformation* of (X, \mathcal{B}, μ) is a \mathcal{B} -measurable map $T: X \rightarrow X$ such that $\mu(T^{-1}B) = \mu(B)$ for all $B \in \mathcal{B}$.

While this definition makes sense for arbitrary measures, not simply probability measures, most of the results and definitions below only make sense in the probability measure case. Sometimes it will be helpful to make the assumption that the underlying probability space is

a Lebesgue space (that is, the space together with its completed σ -algebra agrees up to a measure-preserving bijection with the unit interval with Lebesgue measure and the usual σ -algebra of Lebesgue measurable sets). Although this sounds like a strong restriction, in practice it is barely a restriction at all, as almost all of the spaces that appear in the theory (and all of those that appear in this article) turn out to be Lebesgue spaces. For a detailed treatment of the theory of Lebesgue spaces, the reader is referred to Rudolph's book [76]. The reader is referred also to the chapter on [► Measure Preserving Systems](#).

While many of the definitions that we shall present are valid for both invertible and non-invertible measure-preserving transformations, the strongest mixing conditions are most useful in the case of invertible transformations.

It will be helpful to present a selection of simple examples, relative to which we will be able to explore ergodicity and the various notions of mixing. These examples and the lemmas necessary to show that they are measure-preserving transformations as claimed may be found in the books of Petersen [64], Rudolph [76] and Walters [92]. More details on these examples can also be found in the chapter on [► Ergodic Theory: Basic Examples and Constructions](#).

Example 1 (Rotation on the circle) Let $\alpha \in \mathbb{R}$. Let $R_\alpha: [0, 1) \rightarrow [0, 1)$ be defined by $R_\alpha(x) = x + \alpha \bmod 1$. It is straightforward to verify that R_α preserves the restriction of Lebesgue measure λ to $[0, 1)$ (it is sufficient to check that $\lambda(R_\alpha^{-1}(J)) = \lambda(J)$ for an interval J).

Example 2 (Doubling Map) Let $M_2: [0, 1) \rightarrow [0, 1)$ be defined by $M_2(x) = 2x \bmod 1$. Again, Lebesgue measure is invariant under M_2 (to see this, one observes that for an interval J , $M_2^{-1}(J)$ consists of two intervals, each of half the length of J). This may be generalized in the obvious way to a map M_k for any integer $k \geq 2$.

Example 3 (Interval Exchange Transformation) The class of interval exchange transformations was introduced by Sinai [85]. An interval exchange transformation is the map obtained by cutting the interval into a finite number of pieces and permuting them in such a way that the resulting map is invertible, and restricted to each interval is an order-preserving isometry.

More formally, one takes a sequence of positive lengths $\ell_1, \ell_2, \dots, \ell_k$ summing to 1 and a permutation π of $\{1, \dots, k\}$ and defines $a_i = \sum_{j < i} \ell_j$ and $b_i = \sum_{\pi(j) < \pi(i)} \ell_j$ (again with $b_0 = 0$). The interval exchange transformation defined by (ℓ_1, \dots, ℓ_k) and π is the map $T: [0, 1) \rightarrow [0, 1)$ defined by $T|_{[a_i, a_{i+1})}(x) = x + (b_i - a_i)$. It is straightforward to check that any such interval exchange transformation preserves Lebesgue measure on the unit interval.

Example 4 (Bernoulli Shift) Let A be a finite set and fix a vector $(p_i)_{i \in A}$ of positive numbers that sum to 1. Let $A^{\mathbb{N}}$ denote the set of sequences of the form $x_0 x_1 x_2 \dots$, where $x_n \in A$ for each $n \in \mathbb{N}$ and let $A^{\mathbb{Z}}$ denote the set of bi-infinite sequences of the form $\dots x_{-2} x_{-1} \cdot x_0 x_1 x_2 \dots$ (the \cdot is a placeholder that allows us to distinguish (for example) between the sequences $\dots 01010 \cdot 10101 \dots$ and $\dots 10101 \cdot 01010 \dots$).

We define a map (the *shift map*) S on $A^{\mathbb{N}}$ by $(S(x))_n = x_{n+1}$ and define S on $A^{\mathbb{Z}}$ by the same formula. Note that S is invertible as a transformation on $A^{\mathbb{Z}}$ but non-invertible as a transformation on $A^{\mathbb{N}}$.

We need to equip $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ with measures. This is done by defining the measure of a preferred class of sets, checking certain consistency conditions and appealing to the Kolmogorov extension theorem. Here the preferred sets are the *cylinder sets*. Given $m \leq n$ in the invertible case and a sequence $a_m \dots a_n$, we let $[a_m \dots a_n]_m^n$ denote $\{x \in A^{\mathbb{Z}}: x_m = a_m, \dots, x_n = a_n\}$ and define $\mu([a_m \dots a_n]_m^n) = p_{a_m} p_{a_{m+1}} \dots p_{a_n}$. This is then shown to uniquely define a measure μ on the σ -algebra of $A^{\mathbb{Z}}$ generated by the cylinder sets. It is immediate to see that for any cylinder set C , $\mu(S^{-1}C) = \mu(C)$, and it follows that S is a measure-preserving transformation of $(A^{\mathbb{Z}}, \mathcal{B}, \mu)$. The construction is exactly analogous in the non-invertible case. See the chapter on [► Measure Preserving Systems](#) or the books of Walters [92] or Rudolph [76] for more details of defining measures in these systems.

The class of Bernoulli shifts will play a distinguished role in what follows.

Example 5 (Markov Shift) The spaces $A^{\mathbb{N}}$ and $A^{\mathbb{Z}}$ are exactly as above, as is the shift map. All that changes is the measure.

To define a Markov shift, we need a stochastic matrix P (i.e. a matrix with non-negative entries whose rows sum to 1) with rows and columns indexed by A and a left eigenvector π for P with eigenvalue 1 with the property that the entries of π are non-negative and sum to 1. The existence of such an eigenvector is a consequence of the Perron–Frobenius theory of positive matrices. Provided that the matrix P is irreducible (for each a and a' in A , there is an $n > 0$ such that $P_{a,a'}^n > 0$), the eigenvector π is unique.

Given the pair (P, π) , one defines the measure of a cylinder set by $\mu([a_m \dots a_n]_m^n) = \pi_{a_m} P_{a_m a_{m+1}} \dots P_{a_{n-1} a_n}$ and extends μ as before to a probability measure on $A^{\mathbb{N}}$ or $A^{\mathbb{Z}}$.

Example 6 (Hard Sphere Gases and Billiards) We wish to model the behavior of a gas in a bounded region. We make the assumption that the gas consists of a large number N

of identical balls which move at constant velocity until two balls collide, whereupon they elastically swap momentum along the direction of contact. The phase space for this system is a region of \mathbb{R}^{6N} (with N 3-dimensional position vectors and N 3-dimensional velocity vectors). More abstractly, the system is equivalent to the motion of a single point particle in a region of $\mathbb{R}^M \times \mathbb{R}^M$ (with the first M -vector representing position and the second representing velocity). The system is constrained in that its position is required to lie in a bounded region S of \mathbb{R}^M with a piecewise smooth boundary. The system evolves by moving the position at a constant rate in the direction of the velocity vector until the point reaches ∂S , at which time the component of the velocity parallel to the normal to ∂S is reversed. This then defines a flow (i. e. a family of maps $(T_t)_{t \in \mathbb{R}}$ satisfying $T_{t+s} = T_t \circ T_s$) on the phase space. Since the magnitude of the velocity is conserved, it is convenient to restrict to flows with speed 1. This system is clearly the closest of the examples that we consider to the situation envisaged by Boltzmann. Perhaps not surprisingly, proofs of even the most basic properties for this system are much harder than those for the other examples that we consider.

We will need to make use of the concept of measure-theoretic isomorphism. Two measure-preserving transformations T of (X, \mathcal{B}, μ) and S of (Y, \mathcal{F}, ν) are *measure-theoretically isomorphic* (or just *isomorphic*) if there exist measurable maps $g: X \rightarrow Y$ and $h: Y \rightarrow X$ such that

1. $g \circ h$ and $h \circ g$ agree with the respective identity maps almost everywhere;
2. $\mu(g^{-1}F) = \nu(F)$ and $\nu(h^{-1}B) = \mu(B)$ for all $F \in \mathcal{F}$ and $B \in \mathcal{B}$; and
3. $S \circ g(x) = g \circ T(x)$ for μ -almost every x (or equivalently $T \circ h(y) = h \circ S(y)$ for ν -almost every y).

Measure-theoretic isomorphism is the basic notion of ‘sameness’ in ergodic theory. It is in some sense quite weak, so that systems may be isomorphic that feel very different (for example, as we discuss later, the time one map of a geodesic flow is isomorphic to a Bernoulli shift). For comparison, the notion of sameness in topological dynamical systems (topological conjugacy) is far stronger.

As an example of measure-theoretic isomorphism, it may be seen that the doubling map is isomorphic to the one-sided Bernoulli shift on $\{0, 1\}$ with $p_0 = p_1 = 1/2$ (the map g takes an $x \in [0, 1)$ to the sequence of 0’s and 1’s in its binary expansion (choosing the sequence ending with 0’s, for example, if x is of the form $p/2^n$) and the inverse map h takes a sequence of 0’s and 1’s to the point in $[0, 1)$ with that binary expansion.)

Given a measure-preserving transformation T of a probability space (X, \mathcal{B}, μ) , T is associated to an isometry of $L^2(X, \mathcal{B}, \mu)$ by $U_T(f) = f \circ T$. This operator is known as the *Koopman Operator*. In the case where T is invertible, the operator U_T is unitary. Two measure-preserving transformations T and S of (X, \mathcal{B}, μ) and (Y, \mathcal{F}, ν) are *spectrally isomorphic* if there is a Hilbert space isomorphism Θ from $L^2(X, \mathcal{B}, \mu)$ to $L^2(Y, \mathcal{F}, \nu)$ such that $\Theta \circ U_T = U_S \circ \Theta$. As we shall see below, spectral isomorphism is a strictly weaker property than measure-theoretic isomorphism.

Since in ergodic theory, measure-theoretic isomorphism is the basic notion of sameness, all properties that are used to describe measure-preserving systems are required to be invariant under measure-theoretic isomorphism (i. e. if two measure-preserving transformations are measure-theoretically isomorphic, the first has a given property if and only if the second does). On the other hand, we shall see that some mixing-type properties are invariant under spectral isomorphism, while others are not. If a property is invariant under spectral isomorphism, we say that it is a *spectral property*.

There are a number of mixing type properties that occur in the probability literature (α -mixing, β -mixing, ϕ -mixing, ψ -mixing etc.) (see Bradley’s survey [12] for a description of these conditions). Many of these are stronger than the Bernoulli property, and are therefore not preserved under measure-theoretic isomorphism. For this reason, these properties are not widely used in ergodic theory, although β -mixing turns out to be equivalent to the so-called *weak Bernoulli* property (which turns out to be stronger than the Bernoulli property that we discuss in this article – see Smorodinsky’s paper [87]) and α -mixing is equivalent to strong-mixing.

A basic construction (see the article on [Ergodic Theory: Basic Examples and Constructions](#)) that we shall require in what follows is the product of a pair of measure-preserving transformations: given transformations T of (X, \mathcal{B}, μ) and S of (Y, \mathcal{F}, ν) , we define the *product transformation* $T \times S: (X \times Y, \mathcal{B} \otimes \mathcal{F}, \mu \times \nu)$ by $(T \times S)(x, y) = (Tx, Sy)$.

One issue that we face on occasion is that it is sometimes convenient to deal with invertible measure-preserving transformations. It turns out that given a non-invertible measure-preserving transformation, there is a natural way to uniquely associate an invertible measure-preserving transformation sharing almost all of the ergodic properties of the original transformation. Specifically, given a non-invertible measure-preserving transformation T of (X, \mathcal{B}, μ) , one lets $\bar{X} = \{(x_0, x_1, \dots): x_n \in X \text{ and } T(x_n) = x_{n-1} \text{ for all } n\}$, $\bar{\mathcal{B}}$

be the σ -algebra generated by sets of the form $\bar{A}_n = \{\bar{x} \in \bar{X} : x_n \in A\}$, $\bar{\mu}(\bar{A}_n) = \mu(A)$ and $\bar{T}(x_0, x_1, \dots) = (T(x_0), x_0, x_1, \dots)$. The transformation \bar{T} of $(\bar{X}, \bar{\mathcal{B}}, \bar{\mu})$ is called the *natural extension* of the transformation T of (X, \mathcal{B}, μ) (see the chapter on [Ergodic Theory: Basic Examples and Constructions](#) for more details). In situations where one wants to use invertibility, it is often possible to pass to the natural extension, work there and then derive conclusions about the original non-invertible transformation.

Ergodicity

Given a measure-preserving transformation $T: X \rightarrow X$, if $T^{-1}A = A$, then $T^{-1}A^c = A^c$ also. This allows us to decompose the transformation X into two pieces A and A^c and study the transformation T separately on each. In fact the same situation holds if $T^{-1}A$ and A agree up to a set of measure 0. For this reason, we call a set A *invariant* if $\mu(T^{-1}A \Delta A) = 0$.

Returning to Boltzmann's ergodic hypothesis, existence of an invariant set of measure between 0 and 1 would be a bad situation as his essential idea was that the orbit of a single point would 'see' all of X , whereas if X were decomposed in this way, the most that a point in A could see would be all of A , and similarly the most that a point in A^c could see would be all of A^c .

A measure-preserving transformation will be called *ergodic* if it has no non-trivial decomposition of this form. More formally, let T be a measure-preserving transformation of a probability space (X, \mathcal{B}, μ) . The transformation T is said to be ergodic if for all invariant sets, either the set or its complement has measure 0.

Unlike the remaining concepts that we discuss in this article, this definition of ergodicity applies also to infinite measure-preserving transformations and even to certain non-measure-preserving transformations. See Aaronson's book [1] for more information.

The following lemma is often useful:

Lemma 1 *Let (X, \mathcal{B}, μ) be a probability space and let $T: X \rightarrow X$ be a measure-preserving transformation. Then T is ergodic if and only if the only measurable functions f satisfying $f \circ T = f$ (up to sets of measure 0) are constant almost everywhere.*

For the straightforward proof, we notice that if the condition in the lemma holds and A is an invariant set, then $\mathbf{1}_A \circ T = \mathbf{1}_A$ almost everywhere, so that $\mathbf{1}_A$ is an a.e. constant function and so A or A^c is of measure 0. Conversely, if f is an invariant function, we see that for each α , $\{x: f(x) < \alpha\}$ is an invariant set and hence of measure 0

or 1. It follows that f is constant almost everywhere. We remark for future use that it is sufficient to check that the bounded measurable invariant functions are constant.

The following corollary of the lemma shows that ergodicity is a spectral property.

Corollary 2 *Let T be a measure-preserving transformation of the probability space (X, \mathcal{B}, μ) . Then T is ergodic if and only if 1 is a simple eigenvalue of U_T .*

The ergodic theorems mentioned earlier due to von Neumann and Birkhoff are the following (see also the chapter on [Ergodic Theorems](#)).

Theorem 3 (von Neumann Mean Ergodic Theorem [91]) *Let T be a measure-preserving transformation of the probability space (X, \mathcal{B}, μ) . For $f \in L^2(X, \mathcal{B}, \mu)$, let $A_N f = 1/N(f + f \circ T + \dots + f \circ T^{N-1})$. Then for all $f \in L^2(X, \mathcal{B}, \mu)$, $A_N f$ converges in L^2 to an invariant function f^* .*

Theorem 4 (Birkhoff Pointwise Ergodic Theorem [6]) *Let T be a measure-preserving transformation of the probability space (X, \mathcal{B}, μ) . Let $f \in L^1(X, \mathcal{B}, \mu)$. Let $A_N f$ be as above. Then for μ -almost every $x \in X$, $(A_N f(x))$ is a convergent sequence.*

Of these two theorems, the pointwise ergodic theorem is the deeper result, and it is straightforward to deduce the mean ergodic theorem from the pointwise ergodic theorem. The mean ergodic theorem was reproved very concisely by Riesz [71] and it is this proof that is widely known now. Riesz's proof is reproduced in Parry's book [63]. There have been many different proofs given of the pointwise ergodic theorem. Notable amongst these are the argument due to Garsia [23] and a proof due to Katznelson and Weiss [40] based on work of Kamae [35], which appears in a simplified form in work of Keane and Petersen [42].

If the measure-preserving transformation T is ergodic, then by virtue of Lemma 1, the limit functions appearing in the ergodic theorems are constant. One sees that the constant is simply the integral of f with respect to μ , so that in this situation $A_N f(x)$ converges to $\int f d\mu$ in norm and pointwise almost everywhere, thereby providing a justification of Boltzmann's original claim: for ergodic measure-preserving transformations, *time averages agree with spatial averages*. In the case where T is not ergodic, it is also possible to identify the limit in the ergodic theorems: we have $f^* = \mathbb{E}(f|I)$, where I is the σ -algebra of T -invariant sets.

Note that the set on which the almost everywhere convergence in Birkhoff's theorem takes place depends on the L^1 function f that one is considering. Straightforward considerations show that there is no single full

measure set that works simultaneously for all L^1 functions. In the case where X is a compact metric space, it is well known that $C(X)$, the space of continuous functions on X with the uniform norm has a countable dense set, $(f_n)_{n \geq 1}$ say. If the invariant measure μ is ergodic, then for each n , there is a set B_n of measure 1 such that for all $x \in B_n$, $A_N f_n(x) \rightarrow \int f_n d\mu$. Letting $B = \bigcap_n B_n$, one obtains a full measure set such that for all n and all $x \in B$, $A_N f_n(x) \rightarrow \int f_n d\mu$. A simple approximation argument then shows that for all $x \in B$ and all $f \in C(X)$, $A_N f(x) \rightarrow \int f d\mu$. A point x with this property is said to be *generic* for μ . The observations above show that for an ergodic invariant measure μ , we have $\mu\{x: x \text{ is generic for } \mu\} = 1$.

If T is ergodic, but T^n is not ergodic for some n , then one can show that the space X splits up as A_1, \dots, A_d for some $d|n$ in such a way that $T(A_i) = A_{i+1}$ for $i < d$ and $T(A_d) = A_1$ with T^n acting ergodically on each A_i . The transformation T is *totally ergodic* if T^n is ergodic for all $n \in \mathbb{N}$. One can check that a non-invertible transformation T is ergodic if and only if its natural extension is ergodic.

The following lemma gives an alternative characterization of ergodicity, which in particular relates it to mixing.

Lemma 5 (Ergodicity as a Mixing Property) *Let T be a measure-preserving transformation of the probability space (X, \mathcal{B}, μ) . Then T is ergodic if and only if for all f and g in L^2 ,*

$$\frac{1}{N} \sum_{n=0}^N \langle f, g \circ T^n \rangle \rightarrow \langle f, 1 \rangle \langle 1, g \rangle.$$

In particular, if T is ergodic, then $(1/N) \sum_{n=0}^{N-1} \mu(A \cap T^{-n}B) \rightarrow \mu(A)\mu(B)$ for all measurable sets A and B .

Proof Suppose that T is ergodic. Then the left-hand side of the equality is equal to $\langle f, (1/N) \sum_{n=0}^{N-1} g \circ T^n \rangle$. The mean ergodic theorem shows that the second term converges in L^2 to the constant function with value $\int g d\mu = \langle g, 1 \rangle$, and the equality follows.

Conversely, if the equation holds for all f and g in L^2 , suppose that A is an invariant set. Let $f = g = \mathbf{1}_A$. Then since $g \circ T^n = \mathbf{1}_A$ for all n , the left-hand side is $\langle \mathbf{1}_A, \mathbf{1}_A \rangle = \mu(A)$. On the other hand, the right-hand side is $\mu(A)^2$, so that the equation yields $\mu(A) = \mu(A)^2$, and $\mu(A)$ is either 0 or 1 as required.

Taking $f = \mathbf{1}_A$ and $g = \mathbf{1}_B$ for measurable sets A and B gives the final statement. \square

We now examine the ergodicity of the examples presented above. Firstly, for the rotation of the circle, we claim that

the transformation is ergodic if and only if the ‘angle’ α is irrational. To see this, we argue as follows. If $\alpha = p/q$, then we see that $f(x) = e^{2\pi i q x}$ is a non-constant R_α -invariant function, and hence R_α is not ergodic. On the other hand, if α is irrational, suppose f is a bounded measurable invariant function. Since f is bounded, it is an L^2 function, and so f may be expressed in L^2 as a Fourier series: $f = \sum_{n \in \mathbb{Z}} c_n e_n$ where $e_n(x) = e^{2\pi i n x}$. We then see that $f \circ R_\alpha = \sum_{n \in \mathbb{Z}} e^{2\pi i n \alpha} c_n e_n$. In order for f to be equal in L^2 to $f \circ R_\alpha$, they must have the same Fourier coefficients, so that $c_n = e^{2\pi i n \alpha} c_n$ for each n . Since α is irrational, this forces $c_n = 0$ for all $n \neq 0$, so that f is constant as required.

The doubling map and the Bernoulli shift are both ergodic, although we defer proof of this for the time being, since they in fact have the strong-mixing property. A Markov chain with matrix P and vector π is ergodic if and only if for all i and j in A with $\pi_i > 0$ and $\pi_j > 0$, there exists an $n \geq 0$ with $P_{ij}^n > 0$. This follows from the ergodic theorem for Markov chains (which is derived from the Strong Law of Large Numbers) (see [18] for details). In particular, if the underlying Markov chain is irreducible, then the measure is ergodic.

In the case of interval exchange transformations, there is a simple necessary condition on the permutation for irreducibility, namely for $1 \leq j < k$, we do not have $\pi\{1, \dots, j\} = \{1, \dots, j\}$. Under this condition, Masur [49] and Veech [88] independently showed that for almost all values of the sequence of lengths $(\ell_i)_{1 \leq i \leq k}$, the interval exchange transformation is ergodic. (In fact they showed the stronger condition of *unique ergodicity*: that the transformation has no other invariant measure than Lebesgue measure. This implies that Lebesgue measure is ergodic, because if there were a non-trivial invariant set, then the restriction of Lebesgue measure to that set would be another invariant measure).

For the hard sphere systems, there are no results on ergodicity in full generality. Important special cases have been studied by Sinai [84], Sinai and Chernov [86], Krámli, Simányi and Szász [45], Simányi and Szász [81], Simányi [79,80] and Young [95].

Ergodic Decomposition

We already observed that if a transformation is not ergodic, then it may be decomposed into parts. Clearly if these parts are not ergodic, they may be further decomposed. It is natural to ask whether the transformation can be decomposed into ergodic parts, and if so what form does the decomposition take? In fact such a decomposition does exist, but rather than decompose the transforma-

tion, it is necessary to decompose the measure into ergodic pieces. This is known as ergodic decomposition.

The set of invariant measures for a measurable map T of a measurable space (X, \mathcal{B}) to itself forms a simplex. General functional analytic considerations (due to Choquet [14,15] – see also Phelps' account [66] of this theory) mean that it is possible to write any member of the simplex as an integral-convex combination of the extreme points. Further, the extreme points of the simplex may be identified as precisely the ergodic invariant measures for T . It follows that any invariant probability measure μ for T may be uniquely expressed in the form

$$\mu(A) = \int_{M_{\text{erg}}(X, T)} \nu(A) \, d m(\nu),$$

where $M_{\text{erg}}(X, T)$ denotes the set of ergodic T -invariant measures on X and m is a measure on $M_{\text{erg}}(X, T)$.

We will give a proof of this theorem in the special case of a continuous transformation of a compact space. Our proof is based on the Birkhoff ergodic theorem and the Riesz Representation Theorem identifying the dual space of the space of continuous functions on a compact space as the set of bounded signed measures on the space (see Rudin's book [75] for details). We include it here because this special case covers many cases that arise in practice, and because few of the standard ergodic theory references include a proof of ergodic decomposition. An exception to this is Rudolph's book [76] which gives a full proof in the case that X is a Lebesgue space. This is based on a detailed development of the theory of these spaces and builds measures using conditional expectations. Kalikow's notes [32] give a brief outline of a proof similar to that which follows. Oxtoby [62] also wrote a survey article containing much of the following (and much more besides).

Theorem 6 *Let X be a compact metric space, \mathcal{B} be the Borel σ -algebra, μ be an invariant Borel probability measure and T be a continuous measure-preserving transformation of (X, \mathcal{B}, μ) . Then for each $x \in X$, there exists an invariant Borel measure μ_x such that:*

1. For $f \in L^1(X, \mathcal{B}, \mu)$, $\int f \, d\mu = \int (\int f \, d\mu_x) \, d\mu(x)$;
2. Given $f \in L^1(X, \mathcal{B}, \mu)$, for μ -almost every $x \in X$, one has $A_N f(x) \rightarrow \int f \, d\mu_x$;
3. The measure μ_x is ergodic for μ -almost every $x \in X$.

Notice that conclusion (2) shows that μ_x can be understood as the distribution on the phase space “seen” if one starts the system in an initial condition of x . This interpretation of the measures μ_x corresponds closely with the ideas of Boltzmann and the Ehrenfests in the formulation

of the ergodic and quasi-ergodic hypotheses, which can be seen as demanding that μ_x is equal to μ for (almost) all x .

Proof The proof will be divided into 3 main steps: defining the measures μ_x , proving measurability with respect to x and proving ergodicity of the measures.

Step 1: Definition of μ_x

Given a function $f \in L^1(X, \mathcal{B}, \mu)$, Birkhoff's theorem states that for μ -almost every $x \in X$, $(A_N f(x))$ is convergent. It will be convenient to denote the limit by $\tilde{f}(x)$. Let f_1, f_2, \dots be a sequence of continuous functions that is dense in $C(X)$. For each k , there is a set B_k of x 's measure 1 for which $(A_n f_k(x))_{n=1}^\infty$ is a convergent sequence. Intersecting these gives a set B of full measure such that for $x \in B$, for each $k \geq 1$, $A_n f_k(x)$ is convergent. A simple approximation argument shows that for $x \in B$ and f an arbitrary continuous function, $A_n f(x)$ is convergent. Given $x \in B$, define a map $L_x: C(X) \rightarrow \mathbb{R}$ by $L_x(f) = \tilde{f}(x)$. This is a continuous linear functional on $C(X)$, and hence by the Riesz Representation Theorem there exists a Borel measure μ_x such that $\tilde{f}(x) = \int f \, d\mu_x$ for each $f \in C(X)$ and $x \in B$. Since $L_x(f) \geq 0$ when f is a non-negative function and $L_x(1) = 1$, the measure μ_x is a probability measure. Since $L_x(f \circ T) = L_x(f)$ for $f \in C(X)$, one can check that μ_x must be an invariant probability measure. For $x \notin B$, simply define $\mu_x = \mu$. Since B^c is a set of measure 0, this will not affect any of the statements that we are trying to prove.

Now for f continuous, we have $A_N f$ is a bounded sequence of functions with $A_N f(x)$ converging to $\int f \, d\mu_x$ almost everywhere and $\int A_N f \, d\mu = \int f \, d\mu$ since T is measure-preserving. It follows from the bounded convergence theorem that for $f \in C(X)$,

$$\int f \, d\mu = \int \left(\int f \, d\mu_x \right) d\mu(x). \quad (1)$$

Step 2: Measurability of $x \mapsto \mu_x(A)$

Lemma 7 *Let $C \in \mathcal{B}$ satisfy $\mu(C) = 0$. Then $\mu_x(C) = 0$ for μ -almost every $x \in X$.*

Proof Using regularity of Borel probability measures (see Rudin's book [75] for details), there exist open sets $U_1 \supset U_2 \supset \dots \supset C$ with $\mu(U_k) < 1/k$. There exist continuous functions $g_{k,m}$ with $(g_{k,m}(x))_{m=1}^\infty$ increasing to $\mathbf{1}_{U_k}$ everywhere (e.g. $g_{k,m}(x) = \min(1, m \cdot d(x, U_k^c))$). By (1), we have $\int (\int g_{k,m} \, d\mu_x) \, d\mu(x) < 1/k$ for all k, m . Note that $\int g_{k,m} \, d\mu_x = \lim_{n \rightarrow \infty} A_n g_{k,m}(x)$ is a measurable function of x , so that using the monotone convergence theorem (taking the limit in m), $x \mapsto \int \mathbf{1}_{U_k} \, d\mu_x = \mu_x(U_k)$ is measurable and $\int (\int \mathbf{1}_{U_k} \, d\mu_x) \, d\mu(x) \leq 1/k$. We now see that $x \mapsto \lim_{k \rightarrow \infty} \mu_x(U_k) = \mu_x(\bigcap U_k)$

is also measurable, and by monotone convergence we see $\int \mu_x(\cap U_k) d\mu(x) = 0$. It follows that $\mu_x(\cap U_k) = 0$ for μ -almost every x . Since $\cap U_k \supset C$, the lemma follows. \square

Given a set $A \in \mathcal{B}$, let f_k be a sequence of continuous functions (uniformly bounded by 1) satisfying $\|f_k - \mathbf{1}_A\|_{L^1(\mu)} < 2^{-n}$, so that in particular $f_k(x) \rightarrow \mathbf{1}_A(x)$ for μ -almost every x . For each k , $x \mapsto \int f_k d\mu_x = \lim_{n \rightarrow \infty} A_n f_k(x)$ is a measurable function. By Lemma 7, for μ -almost every x , $f_k \rightarrow \mathbf{1}_A$ μ_x -almost everywhere, so that by the bounded convergence theorem $\lim_{k \rightarrow \infty} \int f_k d\mu_x = \mu_x(A)$ for μ -almost every x . Since the limit of measurable functions is measurable, it follows that $x \mapsto \mu_x(A)$ is measurable for any measurable set $A \in \mathcal{B}$.

This allows us to define a measure ν by $\nu(A) = \int \mu_x(A) d\mu(x)$. For a bounded measurable function f , we have $\int f d\nu = \int (\int f d\mu_x) d\mu(x)$. Since this agrees with $\int f d\mu$ for continuous functions by (1), it follows that $\mu = \nu$. Conclusion (1) of the theorem now follows easily.

Given $f \in L^1(X)$, we let (f_k) be a sequence of continuous functions such that $\|f_k - f\|_{L^1(\mu)}$ is summable. This implies that $\|f_k - f\|_{L^1(\mu_x)}$ is summable for μ -almost every x and in particular, $\int f_k d\mu_x \rightarrow \int f d\mu_x$ for almost every x . On the other hand, by the remark following the statement of Birkhoff's theorem, we have $\tilde{f}_k = \mathbb{E}(f_k | \mathcal{I})$ so that $\|\tilde{f} - \tilde{f}_k\|_{L^1(\mu)}$ is summable and $\tilde{f}_k(x) \rightarrow \tilde{f}(x)$ for μ -almost every x . Combining these two statements, we see that for μ -almost every x , we have

$$\tilde{f}(x) = \lim_{k \rightarrow \infty} \tilde{f}_k(x) = \lim_{k \rightarrow \infty} \int f_k d\mu_x = \int f d\mu_x.$$

This establishes conclusion (2) of the theorem.

Step 3: Ergodicity of μ_x

We have shown how to disintegrate the invariant measure μ as an integral combination of μ_x 's, and we have interpreted the μ_x 's as describing the average behavior starting from x . It remains to show that the μ_x 's are ergodic measures.

Fix for now a continuous function f and a number $0 < \epsilon < 1$. Since $A_n f(x) \rightarrow \tilde{f}(x)$ μ -almost everywhere, there exists an N such that $\mu\{x: |A_N f(x) - \tilde{f}(x)| > \epsilon/2\} < \epsilon^3/8$.

We now claim the following:

$$\mu\{x: \mu_x\{y: |\tilde{f}(y) - \int f d\mu_x| > \epsilon\} > \epsilon\} < \epsilon. \quad (2)$$

To see this, note that $\{y: |\tilde{f}(y) - \int f d\mu_x| > \epsilon\} \subset \{y: |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} \cup \{y: |A_N f(y) - \tilde{f}(y)| > \epsilon/2\}$, so that if $\mu_x\{y: |\tilde{f}(y) - \int f d\mu_x| > \epsilon\} > \epsilon$, then either

$\mu_x\{y: |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} > \epsilon/2$ or $\mu_x\{y: |A_N f(y) - \tilde{f}(y)| > \epsilon/2\} > \epsilon/2$. We show that the set of x 's satisfying each condition is small.

Firstly, we have $\epsilon^3/8 > \mu\{y: |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} = \int \mu_x\{y: |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} d\mu(x)$, so that $\mu\{x: \mu_x\{y: |\tilde{f}(y) - A_N f(y)| > \epsilon/2\} > \epsilon/2\} < \epsilon^2/4 < \epsilon/2$.

For the second term, given $c \in \mathbb{R}$, let $F_c(x) = |A_N f(x) - c|$ and $G(x) = F_{\tilde{f}(x)}(x)$. Note that

$$\int F_{\tilde{f}(x)}(y) d\mu_x(y) = \lim_{n \rightarrow \infty} A_n F_{\tilde{f}(x)}(x) = \lim_{n \rightarrow \infty} A_n G(x)$$

(using the facts that $y \mapsto F_{\tilde{f}(x)}(y)$ is a continuous function and that since $\tilde{f}(x)$ is an invariant function, $F_{\tilde{f}(x)}(T^k x) = G(T^k x)$). Since $\int G(x) d\mu(x) < \epsilon^3/8$, it follows that $\int F_{\tilde{f}(x)}(y) d\mu_x(y) \leq \epsilon^2/4$ except on a set of x 's of measure less than $\epsilon/2$. Outside this bad set, we have $\mu_x\{y: |A_N f(y) - \tilde{f}(x)| > \epsilon/2\} < \epsilon/2$ so that $\mu\{x: \mu_x\{y: |A_N f(y) - \tilde{f}(x)| > \epsilon/2\} > \epsilon/2\} < \epsilon/2$ as required.

This establishes our claim (2) above. Since $\epsilon > 0$ is arbitrary, it follows that for each $f \in C(X)$, for μ -almost every x , μ_x -almost every y satisfies $\tilde{f}(y) = \int f d\mu_x$. As usual, taking a countable dense sequence (f_k) in $C(X)$, it is the case that for all k and μ -almost every x , $\tilde{f}_k(y) = \int f_k d\mu_x$ μ_x -almost everywhere. Let the set of x 's with this property be D . We claim that for $x \in D$, μ_x is ergodic. Suppose not. Then let $x \in D$ and let J be an invariant set of μ_x measure between δ and $1 - \delta$ for some $\delta > 0$. Then by density of $C(X)$ in $L^1(\mu_x)$, there exists an f_k with $\|f_k - \mathbf{1}_J\|_{L^1(\mu_x)} < \delta$. Since $\mathbf{1}_J$ is an invariant function, we have $\tilde{\mathbf{1}}_J = \mathbf{1}_J$. On the other hand, \tilde{f}_k is a constant function. It follows that $\|\tilde{f}_k - \tilde{\mathbf{1}}_J\|_{L^1(\mu_x)} \geq \delta > \|f_k - \mathbf{1}_J\|_{L^1(\mu_x)}$. This contradicts the identification of the limit as a conditional expectation and concludes the proof of the theorem. \square

Mixing

As mentioned above, ergodicity may be seen as an independence on average property. More specifically, one wants to know whether in some sense $\mu(A \cap T^{-n}B)$ converges to $\mu(A)\mu(B)$ as $n \rightarrow \infty$. Ergodicity is the property that there is convergence in the Césaro sense. Weak-mixing is the property that there is convergence in the strong Césaro sense. That is, a measure-preserving transformation T is *weak-mixing* if

$$\frac{1}{N} \sum_{n=0}^{N-1} |\mu(A \cap T^{-n}B) - \mu(A)\mu(B)| \rightarrow 0 \quad \text{as } N \rightarrow \infty.$$

In order for T to be *strong-mixing*, we require simply $\mu(A \cap T^{-N}B) \rightarrow \mu(A)\mu(B)$ as $N \rightarrow \infty$. It is clear that strong-mixing implies weak-mixing and weak-mixing implies ergodicity.

If T^d is not ergodic (so that $T^{-d}A = A$ for some A of measure strictly between 0 and 1), then $|\mu(T^{-nd}A \cap A) - \mu(A)^2| = \mu(A)(1 - \mu(A))$, so that T is not weak-mixing.

An alternative characterization of weak-mixing is as follows:

Lemma 8 *The measure-preserving transformation T is weak-mixing if and only if for every pair of measurable sets A and B , there exists a subset J of \mathbb{N} of density 1 (i. e. $\#(J \cap \{1, \dots, N\})/N \rightarrow 1$) such that*

$$\lim_{n \rightarrow \infty, n \notin J} \mu(A \cap T^{-n}B) = \mu(A)\mu(B). \quad (3)$$

By taking a countable family of measurable sets that are dense (with respect to the metric $d(A, B) = \mu(A \Delta B)$) and taking a suitable intersection of the corresponding J sets, one shows that for a given weak-mixing measure-preserving transformation, there is a single set $J \subset \mathbb{N}$ such that (3) holds for *all* measurable sets A and B (see Petersen [64] or Walters [92] for a proof).

We show that an irrational rotation of the circle is not weak-mixing as follows: let $\alpha \in \mathbb{R} \setminus \mathbb{Q}$ and let A be the interval $[\frac{1}{4}, \frac{3}{4})$. There is a positive proportion of n 's in the natural numbers (in fact proportion 1/3) with the property that $|T^n(\frac{1}{2}) - \frac{1}{2}| < \frac{1}{6}$. For these n 's $\mu(A \cap T^{-n}A) > \frac{1}{3}$, so that in particular $|\mu(A \cap T^{-n}A) - \mu(A)\mu(A)| > \frac{1}{12}$. Clearly this precludes the required convergence to 0 in the definition of weak-mixing, so that an irrational rotation is ergodic but not weak-mixing. Since $R_\alpha^n = R_{n\alpha}$, the earlier argument shows that R_α^n is ergodic, so that R_α is totally ergodic.

On the other hand, we show that any Bernoulli shift is strong-mixing. To see this, let A and B be arbitrary measurable sets. By standard measure-theoretic arguments, A and B may each be approximated arbitrarily closely by a finite union of cylinder sets. Since if A' and B' are finite unions of cylinder sets, we have that $\mu(A' \cap T^{-n}B')$ is equal to $\mu(A')\mu(B')$ for large n , it is easy to deduce that $\mu(A \cap T^{-n}B) \rightarrow \mu(A)\mu(B)$ as required. Since the doubling map is measure-theoretically isomorphic to a one-sided Bernoulli shift, it follows that the doubling map is also strong-mixing.

Similarly, if a Markov Chain is irreducible (i. e. for any states i and j , there exists an $n \geq 0$ such that $P_{ij}^n > 0$) and aperiodic (there is a state i such that $\gcd\{n: P_{ii}^n > 0\} = 1$), then given any pair of cylinder sets A' and B' we have by standard theorems of Markov chains

$\mu(A' \cap T^{-n}B') \rightarrow \mu(A')\mu(B')$. The same argument as above then shows that an aperiodic irreducible Markov Chain is strong-mixing. On the other hand, if a Markov chain is periodic ($d = \gcd\{n: P_{ii}^n > 0\} > 0$), then letting $A = B = \{x: x_0 = i\}$, we have that $\mu(A \cap T^{-n}B) = 0$ whenever $d \nmid n$. It follows T^d is not ergodic, so that T is not weak-mixing.

Both weak- and strong-mixing have formulations in terms of functions:

Lemma 9 *Let T be a measure-preserving transformation of the probability space (X, \mathcal{B}, μ) .*

1. *T is weak-mixing if and only if for every $f, g \in L^2$ one has*

$$\frac{1}{N} \sum_{n=0}^{N-1} |\langle f, g \circ T^n \rangle - \langle f, 1 \rangle \langle 1, g \rangle| \rightarrow 0 \text{ as } N \rightarrow \infty.$$

2. *T is strong-mixing if and only if for every $f, g \in L^2$, one has*

$$\langle f, g \circ T^N \rangle \rightarrow \langle f, 1 \rangle \langle 1, g \rangle \text{ as } N \rightarrow \infty.$$

Using this, one can see that both mixing conditions are spectral properties.

Lemma 10 *Weak- and strong-mixing are spectral properties.*

Proof Suppose S is a weak-mixing transformation of (Y, \mathcal{F}, ν) and the transformation T of (X, \mathcal{B}, μ) is spectrally isomorphic to S by the Hilbert space isomorphism Θ . Then for $f, g \in L^2(X, \mathcal{B}, \mu)$, $\langle f, g \circ T^n \rangle_X - \langle f, 1 \rangle_X \langle 1, g \rangle_X = \langle \Theta(f), \Theta(g) \circ S^n \rangle_Y - \langle \Theta(f), \Theta(1) \rangle_Y \langle \Theta(1), \Theta(g) \rangle_Y$. Since 1 is an eigenfunction of U_T with eigenvalue 1, $\Theta(1)$ is an eigenfunction of U_S with an eigenvalue 1, so since S is ergodic, $\Theta(1)$ must be a constant function. Since Θ preserves norms, $\Theta(1)$ must have a constant value of absolute value 1 and hence $\langle f, g \circ T^n \rangle_X - \langle f, 1 \rangle_X \langle 1, g \rangle_X = \langle \Theta(f), \Theta(g) \circ S^n \rangle_Y - \langle \Theta(f), 1 \rangle_Y \langle 1, \Theta(g) \rangle_Y$. It follows from Lemma 9 that T is weak-mixing.

A similar proof shows that strong-mixing is a spectral property. \square

Both weak- and strong-mixing properties are preserved by taking natural extensions.

Recent work of Avila and Forni [4] shows that for interval exchange transformations of $k \geq 3$ intervals with the underlying permutation satisfying the non-degeneracy condition above, almost all divisions of the interval (with respect to Lebesgue measure on the $k-1$ -dimensional simplex) lead to weak-mixing transformations. On the

other hand, work of Katok [36] shows that no interval exchange transformation is strong-mixing.

It is of interest to understand the behavior of the ‘typical’ measure-preserving transformation. There are a number of Baire category results addressing this. In order to state them, one needs a set of measure-preserving transformations and a topology on them. As mentioned earlier, it is effectively no restriction to assume that a transformation is a Lebesgue-measurable map on the unit interval preserving Lebesgue measure. The classical category results are then on the collection of invertible Lebesgue-measure preserving transformations of the unit interval. One topology on these is the ‘weak’ topology, where a sub-base is given by sets of the form $N(T, A, \epsilon) = \{S: \lambda(S(A)\Delta T(A)) < \epsilon\}$. With respect to this topology, Halmos [26] showed that a residual set (i. e. a dense G_δ set) of invertible measure-preserving transformations is weak-mixing (see also work of Alpern [3]), while Rokhlin [72] showed that the set of strong-mixing transformations is meagre (i. e. a nowhere dense F_σ set), allowing one to conclude that with respect to this topology, the typical transformation is weak- but not strong-mixing.

As often happens in these cases, even when a certain kind of behavior is typical, it may not be simple to exhibit concrete examples. In this case, a well-known example of a transformation that is weak-mixing but not strong-mixing was given by Chacon [13].

While on the face of it the formulation of weak-mixing is considerably less natural than that of strong-mixing, the notion of weak-mixing turns out to be extremely natural from a spectral point of view. Given a measure-preserving transformation T , let U_T be the Koopman operator described above. Since this operator is an isometry, any eigenvalue must lie on the unit circle. The constant function 1 is always an eigenfunction with eigenvalue 1. If T is ergodic and g and h are eigenfunctions of U_T with eigenvalue λ , then $g\bar{h}$ is an eigenfunction with eigenvalue 1, hence invariant, so that $g = Kh$ for some constant K . We see that for ergodic transformations, up to rescaling, there is at most one eigenfunction with any given eigenvalue.

If U_T has a non-constant eigenfunction f , then one has $|\langle U_T^n f, f \rangle| = \|f\|^2$ for each n , whereas by Cauchy–Schwarz, $|\langle f, 1 \rangle|^2 < \|f\|^2$. It follows that $|\langle U_T^n f, f \rangle - \langle f, 1 \rangle \langle 1, f \rangle| \geq c$ for some positive constant c , so that using Lemma 9, T is not weak-mixing.

Using the spectral theorem, the converse is shown to hold.

Theorem 11 *The measure-preserving transformation T is weak-mixing if and only if U_T has no non-constant eigenfunctions.*

Of course this also shows that weak-mixing is a spectral property. Equivalently, this says that the transformation T is weak-mixing if and only if the apart from the constant eigenfunction, the operator U_T has only continuous spectrum (that is, the operator has no other eigenfunctions). For a very nice and concise development of the part of spectral theory relevant to ergodic theory, the reader is referred to the Appendix in Parry’s book [63].

Using this theory, one can establish the following:

Theorem 12

1. T is weak-mixing if and only if $T \times T$ is ergodic;
2. If T and S are ergodic, then $T \times S$ is ergodic if and only if U_S and U_T have no common eigenvalues other than 1.

Proof The main factor in the proof is that the eigenvalues of $U_{T \times S}$ are precisely the set of $\alpha\beta$, where α is an eigenvalue of U_T and β is an eigenvalue of U_S . Further, the eigenfunctions of $U_{T \times S}$ with eigenvalue γ are spanned by eigenfunctions of the form $f \otimes g$, where f is an eigenfunction of U_T , g is an eigenfunction of U_S , and the product of the eigenvalues is γ .

Suppose that T is weak-mixing. Then the only eigenfunction is the constant function, so that the only eigenfunction of $U_{T \times T}$ is the constant function, proving that $T \times T$ is ergodic. Conversely, if U_T has an eigenvalue (so that $f \circ T = \alpha f$ for some non-constant f) then $f \otimes \bar{f}$ is a non-constant invariant function of $T \times T$ so that $T \times T$ is not ergodic.

For the second part, if U_S and U_T have a common eigenvalue other than 1 (say $f \circ T = \alpha f$ and $g \circ T = \alpha g$), then $f \otimes \bar{g}$ is a non-constant invariant function. Conversely, if $T \times S$ has a non-constant invariant function h , then h can be decomposed into functions of the form $f \otimes g$, where f and g are eigenfunctions of U_T and U_S respectively with eigenvalues α and β satisfying $\alpha\beta = 1$. Since the eigenvalues of S are closed under complex conjugation, we see that U_T and U_S have a common eigenvalue other than 1 as required. \square

For a measure-preserving transformation T , we let K be the subspace of L^2 spanned by the eigenfunctions of U_T . It is a remarkable fact that K may be identified as $L^2(X, \mathcal{B}', \mu)$ where \mathcal{B}' is a sub- σ -algebra of \mathcal{B} . The space K is called the *Kronecker factor* of T . The terminology comes from the fact that any sub- σ -algebra \mathcal{F} of \mathcal{B} gives rise to a factor mapping $\pi: (X, \mathcal{B}, \mu) \rightarrow (X, \mathcal{F}, \mu)$ with $\pi(x) = x$. By construction $L^2(X, \mathcal{B}', \mu)$ is the closed linear span of the eigenfunctions of T considered as a measure-preserving transformation of (X, \mathcal{B}', μ) . By the Discrete Spectrum Theorem of Halmos and von Neumann [27],

T acting on (X, \mathcal{B}', μ) is measure-theoretically isomorphic to a rotation on a compact group. This allows one to split $L^2(X, \mathcal{B}, \mu)$ as $L^2(X, \mathcal{B}', \mu) \oplus L^2_c(X, \mathcal{B}, \mu)$, where, as mentioned above the first part is the discrete spectrum part, spanned by eigenfunctions, and the second part is the continuous spectrum part, consisting of functions whose spectral measure is continuous. Since we have split L^2 into a discrete part and a continuous part, it is natural to ask whether the underlying transformation T can be split up in some way into a weak-mixing part and a discrete spectrum (compact group rotation) part, somewhat analogously to the ergodic decomposition. Unfortunately, there is no such decomposition available. However for some applications, for example to multiple recurrence (starting with the work of Furstenberg [20,21]), the decomposition of L^2 (possibly into more complicated parts) plays a crucial role (see the chapters on ► [Ergodic Theory: Recurrence](#) and ► [Ergodic Theory: Interactions with Combinatorics and Number Theory](#)).

For non-invertible measure-preserving transformations, the transformation is weak- or strong-mixing if and only if its natural extension has that property.

The understanding of weak-mixing in terms of the discrete part of the spectrum of the operator also extends to total ergodicity. T^n is ergodic if and only if T has no eigenvalues of the form $e^{2\pi i p/n}$ other than 1. From this it follows that an ergodic measure-preserving transformation T is totally ergodic if and only if it has no *rational spectrum* (i. e. no eigenvalues of the form $e^{2\pi i p/q}$ other than the simple eigenvalue 1).

An intermediate mixing condition between strong- and weak- mixing is that a measure-preserving transformation is *mild-mixing* if whenever $f \circ T^{n_i} \rightarrow f$ for an L^2 function f and a sequence $n_i \rightarrow \infty$, then f is a.e. constant. Clearly mild-mixing is a spectral property. If a transformation has an eigenfunction f , then it is straightforward to find a sequence n_i such that $f \circ T^{n_i} \rightarrow f$, so we see that mild-mixing implies weak-mixing. To see that strong-mixing implies mild-mixing, suppose that T is strong-mixing and that $f \circ T^{n_i} \rightarrow f$. Then we have $\int f \circ T^{n_i} \bar{f} \rightarrow \|f\|^2$. On the other hand, the strong mixing property implies that $\int f \circ T^{n_i} \bar{f} \rightarrow |\langle f, 1 \rangle|^2$. The equality of these implies that f is a.e. constant. Mild-mixing has a useful reformulation in terms of ergodicity of general (not necessarily probability) measure-preserving transformations: A transformation T is mild-mixing if and only if for every conservative ergodic measure-preserving transformation S , $T \times S$ is ergodic. See Furstenberg and Weiss' article [22] for further information on mild-mixing.

The strongest spectral property that we consider is that of having countable Lebesgue spectrum. While we

will avoid a detailed discussion of spectral theory in this article, this is a special case that can be described simply. Specifically, let T be an invertible measure-preserving transformation. Then T has countable Lebesgue spectrum if there is a sequence of functions f_1, f_2, \dots such that $\{1\} \cup \{U_T^n f_j: n \in \mathbb{Z}, j \in \mathbb{N}\}$ forms an orthonormal basis for $L^2(X)$.

To see that this property is stronger than strong-mixing, we simply observe that it implies that $\langle U_T^t U_T^n f_j, U_T^m f_k \rangle \rightarrow 0$ as $t \rightarrow \infty$. Then by approximating f and g by their expansions with respect to a finite part of the basis, we deduce that $\langle U_T^n f, g \rangle \rightarrow \langle f, 1 \rangle \langle 1, g \rangle$ as required. Since already strong-mixing is atypical from the topological point of view, it follows that countable Lebesgue spectrum has to be atypical. In fact, Yuzvinskii [96] showed that the typical invertible measure-preserving transformation has simple singular spectrum.

The property of countable Lebesgue spectrum is by definition a spectral property. Since it completely describes the transformation up to spectral isomorphism, there can be no stronger spectral properties. The remaining properties that we shall examine are invariant under measure-theoretic isomorphisms only.

An invertible measure-preserving transformation T of (X, \mathcal{B}, μ) is said to be K (for Kolmogorov) if there is a sub- σ -algebra \mathcal{F} of \mathcal{B} such that

1. $\bigcap_{n=1}^{\infty} T^{-n} \mathcal{F}$ is the trivial σ -algebra up to sets of measure 0 (i. e. the intersection consists only of null sets and sets of full measure).
2. $\bigvee_{n=1}^{\infty} T^n \mathcal{F} = \mathcal{B}$ (i. e. the smallest σ -algebra containing $T^n \mathcal{F}$ for all $n > 0$ is \mathcal{B}).

The K property has a useful reformulation in terms of entropy as follows: T is K if and only if for every non-trivial partition \mathcal{P} of X , the entropy of T with respect to the partition \mathcal{P} is positive: T has *completely positive entropy*. See the chapter on ► [Entropy in Ergodic Theory](#) for the relevant definitions. The equivalence of the K property and completely positive entropy was shown by Rokhlin and Sinai [74]. For a general transformation T , one can consider the collection of all subsets B of X such that with respect to the partition $\mathcal{P}_B = \{B, B^c\}$, $h(\mathcal{P}_B) = 0$. One can show that this is a σ -algebra. This σ -algebra is known as the *Pinsker σ -algebra*. The above reformulation allows us to say that a transformation is K if and only if it has a trivial Pinsker σ -algebra.

The K property implies countable Lebesgue spectrum (see Parry's book [63] for a proof). To see that K is not implied by countable Lebesgue spectrum, we point out that certain measure-preserving transformations derived from Gaussian systems (see for example the paper of Parry and

Newton [52]) have countable Lebesgue spectrum but zero entropy.

The fact that (two-sided) Bernoulli shifts have the K property follows from Kolmogorov's 0–1 law by taking $\mathcal{F} = \bigvee_{n=0}^{\infty} T^{-n}\mathcal{P}$, where \mathcal{P} is the partition into cylinder sets (see Williams's book [93] for details of the 0–1 law).

Although the K property is explicitly an invertible property, it has a non-invertible counterpart, namely exactness. A transformation T of (X, \mathcal{B}, μ) is *exact* if $\bigcap_{n=0}^{\infty} T^{-n}\mathcal{B}$ consists entirely of null sets and sets of measure 1. It is not hard to see that a non-invertible transformation is exact if and only if its natural extension is K.

The final and strongest property in our list is that of being measure-theoretically isomorphic to a Bernoulli shift. If T is measure-theoretically isomorphic to a Bernoulli shift, we say that T has the *Bernoulli property*. While in principle this could apply to both invertible and non-invertible transformations, in practice the definition applies to a large class of invertible transformations, but occurs comparatively seldom for non-invertible transformations. For this reason, we will restrict ourselves to a discussion of the Bernoulli property for invertible transformations (see however work of Hoffman and Rudolph [29] and Hecklen and Hoffman [28] for work on the one-sided Bernoulli property).

In the case of invertible Bernoulli shifts, Ornstein [53,58] developed in the early 1970s a powerful isomorphism theory, showing that two Bernoulli shifts are measure-theoretically isomorphic if and only if they have the same entropy. Entropy had already been identified as an invariant by Kolmogorov and Sinai [43,82], so this established that it was a complete invariant for Bernoulli shifts. Keane and Smorodinsky [41] gave a proof which showed that two Bernoulli shifts of the same entropy are isomorphic using a conjugating map that is continuous almost everywhere. With other authors, this theory was extended to show that the property of being isomorphic to a Bernoulli shift applied to a surprisingly large class of measure-preserving transformations (e.g. geodesic flows on manifolds of constant negative curvature (Ornstein and Weiss [60]), aperiodic irreducible Markov chains (Friedman and Ornstein [19]), toral automorphisms (Katznelson [39]) and more generally many Gibbs measures for hyperbolic dynamical systems (see the book of Bowen [11])).

Initially, it was conjectured that the properties of being K and Bernoulli were the same, but since then a number of measure-preserving transformations that are K but not Bernoulli have been identified. The earliest was due to Ornstein [55]. Ornstein and Shields [59] then provided an uncountable family of non-isomorphic K automorphisms.

Katok [37] gave an example of a smooth diffeomorphism that is K but not Bernoulli; and Kalikow [33] gave a very natural probabilistic example of a transformation that has this property (the T, T^{-1} process).

While in systems that one regularly encounters there is a correlation between positive entropy and the stronger mixing properties that we have discussed, these properties are logically independent (for example taking the product of a Bernoulli shift and the identity transformation gives a positive entropy transformation that fails to be ergodic; also, the zero entropy Gaussian systems with countable Lebesgue spectrum mentioned above have relatively strong mixing properties but zero entropy).

In many of the mixing criteria discussed above we have considered a pair of sets A and B and asked for asymptotic independence of A and B (so that for large n , A and $T^{-n}B$ become independent). It is natural to ask, given a finite collection of sets A_0, A_1, \dots, A_k , under what conditions $\mu(A_0 \cap T^{-n_1}A_1 \cap \dots \cap T^{-n_k}A_k)$ converges to $\prod_{j=0}^k \mu(A_j)$.

A measure-preserving transformation is said to be *mixing of order $k+1$* if for all measurable sets A_0, \dots, A_k ,

$$\lim_{n_1 \rightarrow \infty, n_{j+1} - n_j \rightarrow \infty} \mu(A_0 \cap T^{-n_1}A_1 \cap \dots \cap T^{-n_k}A_k) = \prod_{j=0}^k \mu(A_j).$$

An outstanding open question asked by Rokhlin [73] appearing already in Halmos' 1956 book [27] is to determine whether mixing (i.e. mixing of order 2) implies mixing of all orders. Kalikow [34] showed that mixing implies mixing of all orders for rank 1 transformations (existence of rank one mixing transformations having been previously established by Ornstein in [54]). Later Ryzhikov [77] used joining methods to establish the result for transformations with finite rank, and Host [30] also used joining methods to establish the result for measure-preserving transformations with singular spectrum, but the general question remains open.

It is not hard to show using martingale arguments that K automorphisms and hence all Bernoulli measure-preserving transformations are mixing of all orders.

For weak-mixing transformations, Furstenberg [21] has established the following *weak-mixing of all orders* statement: if a measure-preserving transformation T is weak-mixing, then given sets A_0, \dots, A_k , there is a subsequence J of the integers of density 0 such that

$$\lim_{n \rightarrow \infty, n \notin J} \mu(A_0 \cap T^{-n}A_1 \cap \dots \cap T^{-kn}A_k) = \prod_{i=0}^k \mu(A_i).$$

Bergelson [5] generalized this by showing that

$$\lim_{n \rightarrow \infty, n \notin J} \mu(A_0 \cap T^{-p_1(n)} A_1 \cap \cdots \cap T^{-p_k(n)} A_k) = \prod_{i=0}^k \mu(A_i)$$

whenever $p_1(n), \dots, p_k(n)$ are non-constant integer-valued polynomials such that $p_i(n) - p_j(n)$ is unbounded for $i \neq j$. The method of proof of both of these results was a Hilbert space version of the van der Corput inequality of analytic number theory. Furstenberg's proof played a key role in his ergodic proof [20] of Szemerédi's theorem on the existence of arbitrarily long arithmetic progressions in a subset of the integers of positive density (see the chapter on [Ergodic Theory: Interactions with Combinatorics and Number Theory](#) for more information about this direction of study).

The conclusions that one draws here are much weaker than the requirement for mixing of all orders. For mixing of all orders, it was required that provided the gaps between $0, n_1, \dots, n_k$ diverge to infinity, one achieves asymptotic independence, whereas for these weak-mixing results, the gaps are increasing along prescribed sequences with regular growth properties.

It is interesting to note that the analogous question of whether mixing implies mixing of all orders is known to fail in higher-dimensional actions. Here, rather than a \mathbb{Z} action, in which there is a single measure-preserving transformation (so that the integer n acts on a point $x \in X$ by mapping it to $T^n x$), one takes a \mathbb{Z}^d action. For such an action, one has d commuting transformations T_1, \dots, T_d and a vector (n_1, \dots, n_d) acts on a point x by sending it to $T_1^{n_1} \cdots T_d^{n_d} x$. Ledrappier [46] studied the following two-dimensional action. Let $X = \{x \in \{0, 1\}^{\mathbb{Z}^2} : x_v + x_{v+e_1} + x_{v+e_2} = 0 \pmod{2}\}$ and let $T_i(x)_v = x_{v+e_i}$. Since X is a compact Abelian group, it has a natural measure μ invariant under the group operations (the Haar measure). It is not hard to show that this system is mixing (i. e. given any measurable sets A and B , $\mu(A \cap T_1^{-n_1} T_2^{-n_2} B) \rightarrow \mu(A)\mu(B)$ as $\|(n_1, n_2)\| \rightarrow \infty$). Ledrappier showed that the system fails to be 3-mixing. Subsequently Masser [48] established necessary and sufficient conditions for similar higher-dimensional algebraic actions to be mixing of order k but not order $k + 1$ for any given k .

Hyperbolicity and Decay of Correlations

One class of systems in which the stronger mixing properties are often found is the class of smooth systems possessing uniform hyperbolicity (i. e. the tangent space to the manifold at each point splits into stable

and unstable subspaces $E^s(x)$ and $E^u(x)$ such that the $\|DT|_{E^s(x)}\| \leq a < 1$ for all x and $\|DT^{-1}|_{E^u(x)}\| \leq a$ and $DT(E^s(x)) = E^s(T(x))$ and $DT(E^u(x)) = E^u(T(x))$). In some cases similar conclusions are found in systems possessing non-uniform hyperbolicity. See Katok and Hasselblatt's book [38] for an overview of hyperbolic dynamical systems, as well as the chapter in this volume on [Smooth Ergodic Theory](#).

In the simple case of expanding piecewise continuous maps of the interval (that is, maps for which the absolute value of the derivative is uniformly bounded below by a constant greater than 1), it is known that if they are totally ergodic and topologically transitive (i. e. the forward images of any interval cover the entire interval), then provided that the map has sufficient smoothness (e. g. the map is C^1 and the derivative satisfies a certain additional summability condition), the map has a unique absolutely continuous invariant measure which is exact and whose natural extension is Bernoulli (see the paper of Góra [25] for results of this type proved under some of the mildest hypotheses). These results were originally established for maps that were twice continuously differentiable, and the hypotheses were progressively weakened, approaching, but never meeting, C^1 . Subsequent work of Quas [68,69] provided examples of C^1 expanding maps of the interval for which Lebesgue measure was invariant, but respectively not ergodic and not weak-mixing. Some of the key tools in controlling mixing in one-dimensional expanding maps that are absent in the C^1 case are bounded distortion estimates. Here, there is a constant $1 \leq C < \infty$ such that given any interval I on which some power T^n of T acts injectively and any sub-interval J of I , one has $1/C \leq (|T^n J|/|T^n I|)/(|J|/|I|) \leq C$. An early place in which bounded distortion estimates appear is the work of Rényi [70].

One important class of results for expanding maps establishes an exponential decay of correlations. Here, one starts with a pair of smooth functions f and g and one estimates $\int f \cdot g \circ T^n d\mu - \int f d\mu \int g d\mu$, where μ is an absolutely continuous invariant measure. If μ is mixing, we expect this to converge to 0. In fact though, in good cases this converges to 0 at an exponential rate for each pair of functions f and g belonging to a sufficiently smooth class. In this case, the measure-preserving transformation T is said to have exponential decay of correlations. See Liverani's article [47] for an introduction to a method of establishing this based on cones. Exponential decay of correlations implies in particular that the natural extension is Bernoulli.

Hu [31] has studied the situation of maps of the interval for which the derivative is bigger than 1 everywhere

except at a fixed point, where the local behavior is of the form $x \mapsto x + x^{1+\alpha}$ for $0 < \alpha < 1$. In this case, rather than exhibiting exponential decay of correlations, the map has polynomial decay of correlations with a rate depending on α .

In Young's survey [94], a variety of techniques are outlined for understanding the strong ergodic properties of non-uniformly hyperbolic diffeomorphisms. In her article [95], methods are introduced for studying many classes of non-uniformly hyperbolic systems by looking at suitably high powers of the map, for which the power has strong hyperbolic behavior. The article shows how to understand the ergodic behavior of these systems. These methods are applied (for example) to billiards, one-dimensional quadratic maps and Hénon maps.

Future Directions

Problem 1 (Mixing of all orders) Does mixing imply mixing of all orders? Can the results of Kalikow, Ryzhikov and Host be extended to larger classes of measure-preserving transformations? Thouvenot observed that it is sufficient to establish the result for measure-preserving transformations of entropy 0. This observation (whose proof is based on the Pinsker σ -algebra) was stated in Kalikow's paper [34] and is reproduced as Proposition 3.2 in recent work of de la Rue [16] on the mixing of all orders problem.

Problem 2 (Multiple weak-mixing) As mentioned above, Bergelson [5] showed that if T is a weak-mixing transformation, then there is a subset J of the integers of density 0 such that

$$\lim_{n \rightarrow \infty, n \notin J} \mu(A_0 \cap T^{-p_1(n)} A_1 \cap \dots \cap T^{-p_k(n)} A_k) = \prod_{i=0}^k \mu(A_i)$$

whenever $p_1(n), \dots, p_k(n)$ are non-constant integer-valued polynomials such that $p_i(n) - p_j(n)$ is unbounded for $i \neq j$. It is natural to ask what is the most general class of times that can replace the sequences $(p_1(n)), \dots, (p_k(n))$. In unpublished notes, Bergelson and Håland considered as times the values taken by a family of integer-valued generalized polynomials (those functions of an integer variable that can be obtained by the operations of addition, multiplication, addition of or multiplication by a real constant and taking integer parts (e.g. $g(n) = \lfloor \sqrt{2} \lfloor \pi n \rfloor + \lfloor \sqrt{3} n \rfloor^2 \rfloor$)). They conjectured necessary and sufficient conditions for the analogue of Bergelson's weak-mixing polynomial ergodic theorem to hold, and proved the conjecture in certain cases.

In a recent paper of McCutcheon and Quas [50], the analogous question was addressed in the case where T is a mild-mixing transformation.

Problem 3 (Pascal adic transformation) Vershik [89,90] introduced a family of transformations known as the adic transformations. The underlying spaces for these transformations are certain spaces of paths on infinite graphs, and the transformations act by taking a path to its lexicographic neighbor. Amongst the adic transformations, the so-called Pascal adic transformation (so-called because the underlying graph resembles Pascal's triangle) has been singled out for attention in work of Petersen and others [2,10,51,65]. In particular, it is unresolved whether this transformation is weak-mixing with respect to any of its ergodic measures. Weak-mixing has been shown by Petersen and Schmidt to follow from a number-theoretic condition on the binomial coefficients [2,10].

Problem 4 (Weak Pinsker Conjecture) Pinsker [67] conjectured that in a measure-preserving transformation with positive entropy, one could express the transformation as a product of a Bernoulli shift with a system with zero entropy. This conjecture (now known as the *Strong Pinsker Conjecture*) was shown to be false by Ornstein [56,57]. Shields and Thouvenot [78] showed that the collection of transformations that can be written as a product of a zero entropy transformation with a Bernoulli shift is closed in the so-called \bar{d} -metric that lies at the heart of Ornstein's theory.

It is, however, the case that if $T: X \rightarrow X$ has entropy $h > 0$, then for all $h' \leq h$, T has a factor S with entropy h' (this was originally proved by Sinai [83] and reproved using the Ornstein machinery by Ornstein and Weiss in [61]). The *Weak Pinsker Conjecture* states that if a measure-preserving transformation T has entropy $h > 0$, then for all $\epsilon > 0$, T may be expressed as a product of a Bernoulli shift and a measure-preserving transformation with entropy less than ϵ .

Bibliography

1. Aaronson J (1997) An Introduction to Infinite Ergodic Theory. American Mathematical Society, Providence
2. Adams TM, Petersen K (1998) Binomial coefficient multiples of irrationals. *Monatsh Math* 125:269–278
3. Alpern S (1976) New proofs that weak mixing is generic. *Invent Math* 32:263–278
4. Avila A, Forni G (2007) Weak-mixing for interval exchange transformations and translation flows. *Ann Math* 165:637–664
5. Bergelson V (1987) Weakly mixing pet. *Ergodic Theory Dynam Systems* 7:337–349
6. Birkhoff GD (1931) Proof of the ergodic theorem. *Proc Nat Acad Sci* 17:656–660

7. Birkhoff GD, Smith PA (1924) Structural analysis of surface transformations. *J Math* 7:345–379
8. Boltzmann L (1871) Einige allgemeine Sätze über Wärme-gleichgewicht. *Wiener Berichte* 63:679–711
9. Boltzmann L (1909) *Wissenschaftliche Abhandlungen*. Akademie der Wissenschaften, Berlin
10. Boshernitzan M, Berend D, Kolesnik G (2001) Irrational dilations of Pascal's triangle. *Mathematika* 48:159–168
11. Bowen R (1975) *Equilibrium States and the Ergodic Theory of Anosov Diffeomorphisms*. Springer, Berlin
12. Bradley RC (2005) Basic properties of strong mixing conditions. A survey and some open questions. *Probab Surv* 2:107–144
13. Chacon RV (1969) Weakly mixing transformations which are not strongly mixing. *Proc Amer Math Soc* 22:559–562
14. Choquet G (1956) Existence des représentations intégrales au moyen des points extrémaux dans les cônes convexes. *C R Acad Sci Paris* 243:699–702
15. Choquet G (1956) Unicité des représentations intégrales au moyen de points extrémaux dans les cônes convexes réticulés. *C R Acad Sci Paris* 243:555–557
16. de la Rue T (2006) 2-fold and 3-fold mixing: why 3-dot-type counterexamples are impossible in one dimension. *Bull Braz Math Soc (NS)* 37(4):503–521
17. Ehrenfest P, Ehrenfest T (1911) Begriffliche Grundlage der statistischen Auffassung in der Mechanik. No. 4. In: *Encyclopädie der mathematischen Wissenschaften*. Teubner, Leipzig
18. Feller W (1950) *An Introduction to Probability and its Applications*. Wiley, New York
19. Friedman NA, Ornstein DS (1970) On isomorphism of weak Bernoulli transformations. *Adv Math* 5:365–394
20. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J Analyse Math* 31:204–256
21. Furstenberg H (1981) *Recurrence in Ergodic Theory and Combinatorial Number Theory*. Princeton University Press, Princeton
22. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. In: *The Structure of Attractors in Dynamical Systems*. Proc Conf, North Dakota State Univ, Fargo, N.D., 1977. Springer, Berlin
23. Garsia AM (1965) A simple proof of E. Hopf's maximal ergodic theory. *J Math Mech* 14:381–382
24. Girsanov IV (1958) Spectra of dynamical systems generated by stationary Gaussian processes. *Dokl Akad Nauk SSSR* 119:851–853
25. Góra P (1994) Properties of invariant measures for piecewise expanding one-dimensional transformations with summable oscillations of derivative. *Ergodic Theory Dynam Syst* 14:475–492
26. Halmos PA (1944) In general a measure-preserving transformation is mixing. *Ann Math* 45:786–792
27. Halmos P (1956) *Lectures on Ergodic Theory*. Chelsea, New York
28. Hoffman C, Hecklen D (2002) Rational maps are d -adic bernoulli. *Ann Math* 156:103–114
29. Hoffman C, Rudolph DJ (2002) Uniform endomorphisms which are isomorphic to a Bernoulli shift. *Ann Math* 76:79–101
30. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. *Israel J Math* 76:289–298
31. Hu H (2004) Decay of correlations for piecewise smooth maps with indifferent fixed points. *Ergodic Theory Dynam Syst* 24:495–524
32. Kalikow S () Outline of ergodic theory. Notes freely available for download. See <http://www.math.uvic.ca/faculty/aquas/kalikow/kalikow.html>
33. Kalikow S (1982) T, T^{-1} transformation is not loosely Bernoulli. *Ann Math* 115:393–409
34. Kalikow S (1984) Twofold mixing implies threefold mixing for rank one transformations. *Ergodic Theory Dynam Syst* 2:237–259
35. Kamae T (1982) A simple proof of the ergodic theorem using non-standard analysis. *Israel J Math* 42:284–290
36. Katok A (1980) Interval exchange transformations and some special flows are not mixing. *Israel J Math* 35:301–310
37. Katok A (1980) Smooth non-Bernoulli K-automorphisms. *Invent Math* 61:291–299
38. Katok A, Hasselblatt B (1995) *Introduction to the Modern Theory of Dynamical Systems*. Cambridge, Cambridge
39. Katznelson Y (1971) Ergodic automorphisms of \mathbb{T}^n are Bernoulli shifts. *Israel J Math* 10:186–195
40. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. *Israel J Math* 42:291–296
41. Keane M, Smorodinsky M (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. *Ann Math* 109:397–406
42. Keane MS, Petersen KE (2006) Nearly simultaneous proofs of the ergodic theorem and maximal ergodic theorem. In: *Dynamics and Stochastics: Festschrift in Honor of M.S. Keane*. Institute of Mathematical Statistics, pp 248–251, Bethesda MD
43. Kolmogorov AN (1958) New metric invariant of transitive dynamical systems and endomorphisms of Lebesgue spaces. *Dokl Russ Acad Sci* 119:861–864
44. Koopman BO (1931) Hamiltonian systems and Hilbert space. *Proc Nat Acad Sci* 17:315–218
45. Krámlí A, Simányi N, Szász D (1991) The K-property of three billiard balls. *Ann Math* 133:37–72
46. Ledrappier F (1978) Un champ Markovien peut être d'entropie nulle et mélangeant. *C R Acad Sci Paris, Sér A-B* 287:561–563
47. Liverani C (2004) Decay of correlations. *Ann Math* 159:1275–1312
48. Masser DW (2004) Mixing and linear equations over groups in positive characteristic. *Israel J Math* 142:189–204
49. Masur H (1982) Interval exchange transformations and measured foliations. *Ann Math* 115:169–200
50. McCutcheon R, Quas A (2007) Generalized polynomials and mild mixing systems. *Canad J Math* (to appear)
51. Méla X, Petersen K (2005) Dynamical properties of the Pascal adic transformation. *Ergodic Theory Dynam Syst* 25:227–256
52. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. *Ann Math Statist* 37:1528–1533
53. Ornstein DS (1970) Bernoulli shifts with the same entropy are isomorphic. *Adv Math* 4:337–352
54. Ornstein DS (1972) On the root problem in ergodic theory. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Univ. California, Berkeley, Calif., 1970/1971), vol II: Probability theory. Univ. California Press, pp 347–356
55. Ornstein DS (1973) An example of a Kolmogorov automorphism that is not a Bernoulli shift. *Adv Math* 10:49–62
56. Ornstein DS (1973) A K-automorphism with no square root and Pinsker's conjecture. *Adv Math* 10:89–102

57. Ornstein DS (1973) A mixing transformation for which Pinsker's conjecture fails. *Adv Math* 10:103–123
58. Ornstein DS (1974) *Ergodic Theory, Randomness, and Dynamical Systems*. Yale University Press, Newhaven
59. Ornstein DS, Shields PC (1973) An uncountable family of K-automorphisms. *Adv Math* 10:89–102
60. Ornstein DS, Weiss B (1973) Geodesic flows are Bernoullian. *Israel J Math* 14:184–198
61. Ornstein DS, Weiss B (1975) Unilateral codings of Bernoulli systems. *Israel J Math* 21:159–166
62. Oxtoby JC (1952) Ergodic sets. *Bull Amer Math Soc* 58:116–136
63. Parry W (1981) *Topics in Ergodic Theory*. Cambridge, Cambridge
64. Petersen K (1983) *Ergodic Theory*. Cambridge, Cambridge
65. Petersen K, Schmidt K (1997) Symmetric Gibbs measures. *Trans Amer Math Soc* 349:2775–2811
66. Phelps R (1966) *Lectures on Choquet's Theorem*. Van Nostrand, New York
67. Pinsker MS (1960) Dynamical systems with completely positive or zero entropy. *Soviet Math Dokl* 1:937–938
68. Quas A (1996) A C^1 expanding map of the circle which is not weak-mixing. *Israel J Math* 93:359–372
69. Quas A (1996) Non-ergodicity for C^1 expanding maps and g -measures. *Ergodic Theory Dynam Systems* 16:531–543
70. Rényi A (1957) Representations for real numbers and their ergodic properties. *Acta Math Acad Sci Hungar* 8:477–493
71. Riesz F (1938) Some mean ergodic theorems. *J Lond Math Soc* 13:274–278
72. Rokhlin VA (1948) A 'general' measure-preserving transformation is not mixing. *Dokl Akad Nauk SSSR Ser Mat* 60:349–351
73. Rokhlin VA (1949) On endomorphisms of compact commutative groups. *Izvestiya Akad Nauk SSSR Ser Mat* 13:329–340
74. Rokhlin VA, Sinai Y (1961) Construction and properties of invariant measurable partitions. *Dokl Akad Nauk SSSR* 141:1038–1041
75. Rudin W (1966) *Real and Complex Analysis*. McGraw Hill, New York
76. Rudolph DJ (1990) *Fundamentals of Measurable Dynamics*. Oxford, Oxford University Press
77. Ryzhikov VV (1993) Joinings and multiple mixing of the actions of finite rank. *Funct Anal Appl* 27:128–140
78. Shields P, Thouvenot J-P (1975) Entropy zero \times Bernoulli processes are closed in the \bar{d} -metric. *Ann Probab* 3:732–736
79. Simányi N (2003) Proof of the Boltzmann–Sinai ergodic hypothesis for typical hard disk systems. *Invent Math* 154:123–178
80. Simányi N (2004) Proof of the ergodic hypothesis for typical hard ball systems. *Ann Henri Poincaré* 5:203–233
81. Simányi N, Szász D (1999) Hard ball systems are completely hyperbolic. *Ann Math* 149:35–96
82. Sinai YG (1959) On the notion of entropy of a dynamical system. *Dokl Russ Acad Sci* 124:768–771
83. Sinai YG (1964) On a weak isomorphism of transformations with invariant measure. *Mat Sb (NS)* 63:23–42
84. Sinai YG (1970) Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Uspehi Mat Nauk* 25:141–192
85. Sinai Y (1976) *Introduction to Ergodic Theory*. Princeton, Princeton, translation of the 1973 Russian original
86. Sinai YG, Chernov NI (1987) Ergodic properties of some systems of two-dimensional disks and three-dimensional balls. *Uspekhi Mat Nauk* 42:153–174, 256
87. Smorodinsky M (1971) A partition on a Bernoulli shift which is not weakly Bernoulli. *Math Systems Th* 5:201–203
88. Veech W (1982) Gauss measures for transformations on the space on interval exchange maps. *Ann Math* 115:201–242
89. Vershik A (1974) A description of invariant measures for actions of certain infinite-dimensional groups. *Soviet Math Dokl* 15:1396–1400
90. Vershik A (1981) Uniform algebraic approximation of shift and multiplication operators. *Soviet Math Dokl* 24:97–100
91. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. *Proc Nat Acad Sci USA* 18:70–82
92. Walters P (1982) *An Introduction to Ergodic Theory*. Springer, Berlin
93. Williams D (1991) *Probability with Martingales*. Cambridge, Cambridge
94. Young LS (1995) Ergodic theory of differentiable dynamical systems. In: *Real and Complex Dynamical Systems* (Hillerød, 1993). Kluwer, Dordrecht, pp 293–336
95. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. *Ann Math* 147:585–650
96. Yuzvinskii SA (1967) Metric automorphisms with a simple spectrum. *Soviet Math Dokl* 8:243–245

Ergodic Theorems

ANDRÉS DEL JUNCO

Department of Mathematics, University of Toronto,
Toronto, Canada

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Ergodic Theorems for Measure-Preserving Maps](#)

[Generalizations to Continuous Time](#)

[and Higher-Dimensional Time](#)

[Pointwise Ergodic Theorems for Operators](#)

[Subadditive and Multiplicative Ergodic Theorems](#)

[Entropy and the Shannon–McMillan–Breiman Theorem](#)

[Amenable Groups](#)

[Subsequence and Weighted Theorems](#)

[Ergodic Theorems and Multiple Recurrence](#)

[Rates of Convergence](#)

[Ergodic Theorems for Non-amenable Groups](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Dynamical system in its broadest sense, any set X , with a map $T: X \rightarrow X$. The classical example is: X is a set

whose points are the states of some physical system and the state x is succeeded by the state Tx after one unit of time.

Iteration repeated applications of the map T above to arrive at the state of the system after n units of time.

Orbit of x the forward images x, Tx, T^2x, \dots of $x \in X$ under iteration of T . When T is invertible one may consider the forward, backward or two-sided orbit of x .

Automorphism a dynamical system $T: X \rightarrow X$, where X is a measure space and T is an invertible map preserving measure.

Ergodic average if f is a function on X let $A_n f(x) = n^{-1} \sum_{i=0}^{n-1} f(T^i x)$; the average of the values of f over the first n points in the orbit of x .

Ergodic theorem an assertion that ergodic averages converge in some sense.

Mean ergodic theorem an assertion that ergodic averages converge with respect to some norm on a space of functions.

Pointwise ergodic theorem an assertion that ergodic averages $A_n f(x)$ converge for some or all $x \in X$, usually for a.e. x .

Stationary process a sequence (X_1, X_2, \dots) of random variables (real or complex-valued measurable functions) on a probability space whose joint distributions are invariant under shifting (X_1, X_2, \dots) to (X_2, X_3, \dots) .

Uniform distribution a sequence $\{x_n\}$ in $[0, 1]$ is uniformly distributed if for each interval $I \subset [0, 1]$, the time it spends in I is asymptotically proportional to the length of I .

Maximal inequality an inequality which allows one to bound the pointwise oscillation of a sequence of functions. An essential tool for proving pointwise ergodic theorems.

Operator any linear operator U on a vector space of functions on X , for example one arising from a dynamical system T by setting $Uf(x) = f(Tx)$. More generally any linear transformation on a real or complex vector space.

Positive contraction an operator T on a space of functions endowed with a norm $\|\cdot\|$ such that T maps positive functions to positive functions and $\|Tf\| \leq \|f\|$.

Definition of the Subject

Ergodic theorems are assertions about the long-term statistical behavior of a dynamical system. The subject arose out of Boltzmann's ergodic hypothesis which sought to equate the spatial average of a function over the set of

states in a physical system having a fixed energy with the time average of the function observed by starting with a particular state and following its evolution over a long time period.

Introduction

Suppose that (X, \mathcal{B}, μ) is a measure space and $T: (X, \mathcal{B}, \mu) \rightarrow (X, \mathcal{B}, \mu)$ is a measurable and measure-preserving transformation, that is $\mu(T^{-1}E) = \mu(E)$ for all $E \in \mathcal{B}$. One important motivation for studying such maps is that a Hamiltonian physical system (see the article by Petersen in this collection) gives rise to a one-parameter group $\{T_t : t \in \mathbb{R}\}$ of maps in the phase space of the system which preserve Lebesgue measure. The ergodic theorem of Birkhoff asserts that for $f \in L_1(\mu)$

$$\frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) \quad (1)$$

converges a.e. and that if T is *ergodic* (to be defined shortly) then the limit is $\int f d\mu$. This may be viewed as a justification for Boltzmann's ergodic hypothesis that "space averages equal time averages". See Zund [214] for some history of the ergodic hypothesis. For physicists, then, the problem is reduced to showing that a given physical system is ergodic, which can be very difficult. However ergodic systems arise in many natural ways in mathematics. One example is rotation $z \mapsto \lambda z$ of the unit circle if λ is a complex number of modulus one which is not a root of unity. Another is the shift transformation on a sequence of i.i.d. random variables, for example a coin-tossing sequence. Another is an automorphism of a compact Abelian group. Often a transformation possesses an invariant measure which is not obvious at first sight. Knowledge of such a measure can be a very useful tool. See Petersen's article for more examples.

If (X, \mathcal{B}, μ) is a probability space and T is ergodic then Birkhoff's ergodic theorem implies that if A is a measurable subset of X then for almost every x , the frequency with which x visits A is asymptotically equal to $\mu(A)$, a very satisfying justification of intuition. For example, applying this to the coin-tossing sequence one obtains the *strong law of large numbers* which asserts that almost every infinite sequence of coin tosses has tails occurring with asymptotic frequency $\frac{1}{2}$. One also obtains Borel's theorem on normal numbers which asserts that for almost all $x \in [0, 1]$ each digit $0, 1, 2, \dots, 9$ occurs with limiting frequency $\frac{1}{10}$. The so-called *continued fraction transformation* $x \mapsto x^{-1} \bmod 1$ on $(0, 1)$ has a finite invariant measure $\frac{dx}{1+x}$. (Throughout this article $x \bmod 1$ denotes the

fractional part of x .) Applying Birkhoff's theorem then gives precise information about the frequency of occurrence of any $n \in \mathbb{N}$ in the continued fraction expansion of x , for a.e. x . See for example Billingsley [34].

These are the classical roots of the subject of ergodic theorems. The subject has evolved from these simple origins into a vast field in its own right, quite independent of physics or probability theory. Nonetheless it still has close ties to both these areas and has also forged new links with many other areas of mathematics.

Our purpose here is to give a broad overview of the subject in a historical perspective. There are several excellent references, notably the books of Krengel [135] and Tempelman [194] which give a good picture of the state of the subject at the time they appeared. There has been tremendous progress since then. The time is ripe for a much more comprehensive survey of the field than is possible here.

Many topics are necessarily absent and many are only glimpsed. For example this article will not touch on random ergodic theorems. See the articles [75,143] for some references on this topic.

I thank Mustafa Akcoglu, Ulrich Krengel, Michael Lin, Dan Rudolph and particularly Joe Rosenblatt and Vitaly Bergelson for many helpful comments and suggestions. I would like to dedicate this article to Mustafa Akcoglu who has been such an important contributor to the development of ergodic theorems over the past 40 years. He has also played a vital role in my mathematical development as well as of many other mathematicians. He remains a source of inspiration to me, and a valued friend.

Ergodic Theorems for Measure-Preserving Maps

Suppose (X, \mathcal{B}, μ) is a measure space. A (measure-preserving) *endomorphism* of (X, \mathcal{B}, μ) is a measurable mapping $T: X \rightarrow X$ such that $\mu(T^{-1}E) = \mu(E)$ for any measurable subset $E \subset X$. If T has a measurable inverse then one says that it is an *automorphism*. In this article the unqualified terms “endomorphism” or “automorphism” will always mean measure-preserving. T is called *ergodic* if for all measurable E one has

$$T^{-1}E = E \Rightarrow \mu(E) = 0 \quad \text{or} \quad \mu(E^c) = 0.$$

A very general class of examples comes from the notion of a *stationary stochastic process* in probability theory. A stochastic process is a sequence of measurable functions f_1, f_2, \dots on a probability space (X, \mathcal{B}, μ) taking values in a measurable space (Y, \mathcal{C}) . The *distribution* of the process, a measure ν on $Y^{\mathbb{N}}$, is defined as the image of μ under the

map $(f_1, f_2, \dots): X \rightarrow Y^{\mathbb{N}}$. ν captures all the essential information about the process $\{f_i\}$. In effect one may view any process $\{f_i\}$ as a probability measure on $Y^{\mathbb{N}}$. The process is said to be *stationary* if ν is invariant under the left shift transformation S on $Y^{\mathbb{N}}$, $S(y)(i) = y(i+1)$, that is S is an endomorphism of the probability space $(Y^{\mathbb{N}}, \nu)$.

From a probabilistic point of view the most natural examples of stationary stochastic processes are independent identically distributed processes, namely the case when ν is a product measure $\lambda^{\mathbb{N}}$ for some measure λ on (Y, \mathcal{C}) . More generally one can consider a stationary Markov process defined by transition probabilities on the state space Y and an invariant probability on Y . See for example Chap. 7 in [50], also Sect. “Pointwise Ergodic Theorems for Operators” below.

The first and most fundamental result about endomorphisms is the celebrated recurrence theorem of Poincaré [173]).

Theorem 1 *Suppose μ is finite, $A \in \mathcal{B}$ and $\mu(A) > 0$. Then for a.e. $x \in A$ there is an $n > 0$ such that $T^n x \in A$, in fact there are infinitely many such n .*

It may be viewed as an ergodic theorem, in that it is a qualitative statement about how x behaves under iteration of T . For a proof, observe that if $E \subset A$ is the measurable set of points which never return to A then for each $n > 0$ the set $T^{-n}E$ is disjoint from E . Applying T^{-m} one gets also $T^{-(n+m)}E \cap T^{-m}E = \emptyset$. Thus $E, T^{-1}E, T^{-2}E, \dots$ is an infinite sequence of disjoint sets all having measure $\mu(E)$ so $\mu(E) = 0$ since $\mu(X) < \infty$. Much later Kac [120] formulated the following quantitative version of Poincaré's theorem.

Theorem 2 *Suppose that $\mu(X) = 1$, T is ergodic and let $r_A x$ denote the time of first return to A , that is $r_A(x)$ is the least $n > 0$ such that $T^n x \in A$. Then*

$$\frac{1}{\mu(A)} \int_A r_A d\mu = \frac{1}{\mu(A)}, \quad (2)$$

that is, the expected value of the return time to A is $\mu(A)^{-1}$.

Koopman [131] made the observation that associated to an automorphism T there is a unitary operator $U = U_T$ defined on the Hilbert space $L_2(\mu)$ by the formula $Uf = f \circ T$. This led von Neumann [151] to prove his mean ergodic theorem.

Theorem 3 *Suppose \mathcal{H} is a Hilbert space, U is a unitary operator on \mathcal{H} and let P denote the orthogonal projection on the subspace of U -invariant vectors. Then for any $x \in \mathcal{H}$ one has*

$$\left\| \frac{1}{n} \sum_{i=0}^{n-1} U^i x - Px \right\| \rightarrow 0 \quad \text{as } n \rightarrow \infty. \quad (3)$$

Von Neumann's theorem is usually quoted as above but to be historically accurate he dealt with unitary operators indexed by a continuous time parameter.

Inspired by von Neumann's theorem Birkhoff very soon proved his pointwise ergodic theorem [35]. In spite of the later publication by von Neumann his result did come first. See [29,214] for an interesting discussion of the history of the two theorems and of the interaction between Birkhoff and von Neumann.

Theorem 4 *Suppose (X, \mathcal{B}, μ) is a measure space and T is an endomorphism of (X, \mathcal{B}, μ) . Then for any $f \in L_1 = L_1(\mu)$ there is a T -invariant function $g \in L_1$ such that*

$$A_n f(x) = \frac{1}{n} \sum_{i=0}^{n-1} f(T^i x) \rightarrow g(x) \quad \text{a.e.} \quad (4)$$

Moreover if μ is finite then the convergence also holds with respect to the L_1 norm and one has $\int_E g \, d\mu = \int_E f \, d\mu$ for all T -invariant subsets E .

Again this formulation of Birkhoff's theorem is not historically accurate as he dealt with a smooth flow on a manifold. It was soon observed that the theorem, and its proof, remain valid for an abstract automorphism of a measure space although the realization that T need not be invertible seems to have taken a little longer.

The notation $A_n f = A_n(T)f$ as above will occur often in the sequel. Whenever T is an endomorphism one uses the notation $Tf = f \circ T$ and with this notation $A_n(T) = \frac{1}{n} \sum_{i=0}^{n-1} T^i$. When the scalars (\mathbb{R} or \mathbb{C}) for an L_1 space are not specified the notation should be understood as referring to either possibility. In most of the theorems in this article the complex case follows easily from the real and any indications about proofs will refer to the real case.

Although von Neumann originally used spectral theory to prove his result, there is a quick proof, attributed to Riesz by Hopf in his 1937 book [100], which uses only elementary properties of Hilbert space. Let I denote the (closed) subspace of U -invariant vectors and I' the (usually not closed) subspace of vectors of the form $f - Uf$. It is easy to check that any vector orthogonal to I' must be in I , whence the subspace $I + I'$ is dense in \mathcal{H} . For any vector of the form $x = y + y'$, $y \in I$, $y' \in I'$ it is clear that $A_n x = \frac{1}{n} \sum_{i=0}^{n-1} U^i x$ converges to $y = Px$, since $A_n y = y$ and if $y' = z - Uz$ then the telescoping sum $A_n y' = n^{-1}(z - U^n z)$ converges to 0. This establishes the desired convergence for $x \in I + I'$ and it is easy to extend it to the closure of $I + I'$ since the operators A_n are contractions of \mathcal{H} ($\|A_n\| \leq 1$). Lorch [147] used a soft argument in a similar spirit to extend von Neumann's theorem

from the case of a unitary operator on a Hilbert space to that of an arbitrary linear contraction on any reflexive Banach space. Sine [188] gave a necessary and sufficient condition for the strong convergence of the ergodic averages of a contraction on an arbitrary Banach space.

Birkhoff's theorem has the distinction of being one of the most reproved theorems of twentieth century mathematics. One approach to the pointwise convergence, parallel to the argument just seen, is to find a dense subspace E of L_1 so that the convergence holds for all $f \in E$ and then try to extend the convergence to all $f \in L_1$ by an approximation argument. The first step is not too hard. For simplicity assume that μ is a finite measure. As in the proof of von Neumann's theorem the subspace E spanned by the T -invariant L_1 functions together with functions of the form $g - Tg$, $g \in L_1$, is dense in $L_1(\mu)$. This can be seen by using the Hahn–Banach theorem and the duality of L_1 and L_∞ . (Here one needs finiteness of μ to know that $L_\infty \subset L_1$.) The pointwise convergence of $A_n f$ for invariant f is trivial and for $f = g - Tg$ it follows from telescoping of the sum and the fact that $n^{-1}T^n g \rightarrow 0$ a.e. This last can be shown by using the Borel–Cantelli lemma. The second step, extending pointwise convergence, as opposed to norm convergence, for f in a dense subspace to all f in L_1 is a delicate matter, requiring a *maximal inequality*.

Roughly speaking a maximal inequality is an inequality which bounds the pointwise oscillation of $A_n f$ in terms of the norm of f . The now standard maximal inequality in the context of Birkhoff's theorem is the following, due to Kakutani and Yosida [209]. Birkhoff's proof of his theorem includes a weaker version of this result. Let $S_n f = nA_n f$, the n th partial sum of the iterates of f .

Theorem 5 *Given any real $f \in L_1$ let $A = \bigcup_{n \geq 1} \{S_n f \geq 0\}$. Then*

$$\int_A f \, d\mu \geq 0. \quad (5)$$

Moreover if one sets $Mf = \sup_{n \geq 1} A_n f$ then for any $\alpha > 0$

$$\mu\{Mf > \alpha\} \leq \frac{1}{\alpha} \|f\|_1 \quad (6)$$

A distributional inequality such as (6) will be referred to as a *weak L_1 inequality*. Note that (6) follows easily from (5) by applying (5) to $f - \alpha$, at least in the case when μ is finite. With the maximal inequality in hand it is straightforward to complete the proof of Birkhoff's theorem. For a real-valued function f let

$$\text{Osc } f = \limsup A_n f - \liminf A_n f. \quad (7)$$

$\text{Osc } f = 0$ a.e. if and only if $A_n f$ converges a.e. (to a possibly infinite limit). One has $\limsup A_n f \leq Mf \leq M|f|$ and by symmetry $\liminf A_n f \geq M|f|$, so $\text{Osc } f \leq 2M|f|$. To establish the convergence of $A_n f$ for a real-valued $f \in L_1$ let $\epsilon > 0$ and write $f = g + h$ with $g \in E$ (the subspace where convergence has already been established), $h \in L^1$ and $\|h\| < \epsilon$. Then since $\text{Osc } g = 0$ one has $\text{Osc } f = \text{Osc } h$. Thus for any fixed $\alpha > 0$, using (6)

$$\begin{aligned} \mu\{\text{Osc } f > \alpha\} &= \mu\{\text{Osc } h > \alpha\} \leq \mu\left\{Mh > \frac{\alpha}{2}\right\} \\ &\leq \frac{2\|h\|_1}{\alpha} < \frac{2\epsilon}{\alpha}. \end{aligned} \quad (8)$$

Since $\epsilon > 0$ was arbitrary one concludes that $\mu\{\text{Osc } f > \alpha\} = 0$ and since $\alpha > 0$ is arbitrary it follows that $\mu\{\text{Osc } f > 0\} = 0$, establishing the a.e. convergence. Moreover a simple application of Fatou's lemma shows that the limiting function is in L_1 , hence finite a.e.

There are many proofs of (5). Two of particular interest are Garsia's [89], perhaps the shortest and most mysterious, and the proof via the filling scheme of Chacón and Ornstein [61], perhaps the most intuitive, which goes like this. Given a function $g \in L_1$ write $g^+ = \max(g, 0)$, $g^- = g^+ - g$ and let $Ug = Tg^+ - g^-$. Interpretation: the region between the graph of g^+ and the X -axis is a sheaf of vertical spaghetti sticks, the intervals $[0, g^+(x)]$, x in X , and $-g^-$ is a hole. Now move the spaghetti (horizontally) by T and then let it drop (vertically) into the hole leaving a new hole and a new sheaf which are the negative and positive parts of Ug . Now let $E' = \bigcup_{n \geq 1} \{U^n f \geq 0\}$, the set of points x at which the hole is eventually filled after finitely many iterations of U .

The key point is that $E = E'$. Indeed if $S_n f(x) \geq 0$ for some n , then the total linear height of sticks over $x, Tx, \dots, T^{n-1}x$ is greater than the total linear depth of holes at these points. The only way that spaghetti can escape from these points is by first filling the hole at x , which shows $x \in E'$. Similar thinking shows that if $x \in E'$ and the hole at x is filled for the first time at time n then $S_n f(x) \geq 0$, so $x \in E$, and that all the spaghetti that goes into the hole at x comes from points $T^i x$ which belong to E' . This shows that $E = E'$ and that the part of the hole lying beneath E is eventually filled by spaghetti coming from $E' = E$. Thus the amount of spaghetti over E is no less than the size of the hole under E , that is $\int_E f d\mu \geq 0$.

Most proofs of Birkhoff's theorem use a maximal inequality in some form but a few avoid it altogether, for example [126, 186]. It is also straightforward to deduce Birkhoff's theorem directly from a maximal inequality, as Birkhoff does, without first establishing convergence on a dense subspace. However the technique of proving

a pointwise convergence theorem by finding an appropriate dense subspace and a suitable maximal inequality has proved extremely useful, not only in ergodic theory.

Indeed, in some sense maximal inequalities are unavoidable: this is the content of the following principle proved already in 1926 by Banach. The principle has many formulations; the following one is a slight simplification of the one to be found in [135]. Suppose B is a Banach space, (X, \mathcal{B}, μ) is a finite measure space and let E denote the space of μ -equivalence classes of measurable real-valued functions on X . A linear map $T: B \rightarrow E$ is said to be *continuous in measure* if for each $\epsilon > 0$

$$\|x_n - x\| \rightarrow 0 \Rightarrow \mu\{|Tx_n - Tx| > \epsilon\} \rightarrow 0.$$

Suppose that T_n is a sequence of linear maps from B to E which are continuous in measure and let $Mx = \sup_n |T_n x|$. Of course if $T_n x$ converges a.e. to a finite limit then $Mx < \infty$ a.e.

Theorem 6 (Banach principle) *Suppose $Mx < \infty$ a.e. for each $x \in X$. Then there is a function $C(\lambda)$ such that $C(\lambda) \rightarrow 0$ as $\lambda \rightarrow \infty$ such that for all $x \in B$ and $\lambda > 0$ one has*

$$\mu\{Mx \geq \lambda\|x\|\} \leq C(\lambda). \quad (9)$$

Moreover the set of x for which $T_n x$ converges a.e. is closed in B .

The first chapter of Garsia's book [89] contains a nice introduction to the Banach principle.

It should be noted that, for general integrable f , Mf need not be in L^1 . However if $f \in L^p$ for $p > 1$ then Mf does belong to L^p and one has the estimate

$$\|Mf\|_p \leq \frac{p}{p-1} \|f\|_p. \quad (10)$$

This can be derived from (6) using also the (obvious) fact that $\|Mf\|_\infty \leq \|f\|_\infty$. Any such estimate on the norm of a maximal function will be called a *strong L_p inequality*. It also follows from (6) that if $\mu(X)$ is finite and $f \in L \log L$, that is $\int |f| \log^+ |f| d\mu < \infty$, then $Mf \in L^1$. In fact Ornstein [160] has shown that the converse of this last statement holds provided T is ergodic.

There is a special setting where one has uniform convergence in the ergodic theorem. Suppose T is a homeomorphism of a compact metric space X . By a theorem of Krylov and Bogoliouboff [138] there is at least one probability measure on the Borel σ -algebra of X which is invariant under T . T is said to be *uniquely ergodic* if there is only one Borel probability measure, say μ , invariant under T . It is easy to see that when this is the

case then T is an ergodic automorphism of (X, \mathcal{B}, μ) . As an example, if α is an irrational number then the rotation $z \mapsto e^{2\pi i \alpha} z$ is a uniquely ergodic transformation of the circle $\{|z| = 1\}$. Equivalently $x \mapsto x + \alpha \pmod{1}$ is a uniquely ergodic map on $[0, 1]$. A quick way to see this is to show that the Fourier co-efficients $\hat{\mu}(n)$ of any invariant probability μ are zero for $n \neq 0$. The Jewett–Krieger theorem (see Jewett [109] and Krieger [137]) guarantees that unique ergodicity is ubiquitous in the sense that any automorphism of a probability space is measure-theoretically isomorphic to a uniquely ergodic homeomorphism. The following important result is due to Oxtoby [165]).

Theorem 7 *If T is uniquely ergodic, μ is its unique invariant probability measure and $f \in C(X)$ then the ergodic averages $A_n(f)$ converge uniformly to $\int f d\mu$.*

This result can be proved along the same lines as the proof given above of von Neumann’s theorem. In a nutshell, one uses the fact that the dual of $C_{\mathbb{R}}(X)$ is the space of finite signed measures on X and the unique ergodicity to show that functions of the form $f - f \circ T$ together with the invariant functions (which are just constant functions) span a dense subspace of $C(X)$.

A sequence $\{x_n\}$ in the interval $[0, 1]$ is said to be *uniformly distributed* if $\frac{1}{n} \sum_{i=1}^n f(x_i) \rightarrow \int_0^1 f(x) dx$ for any $f \in C[0, 1]$ (equivalently for any Riemann integrable f or for any $f = 1_I$ where I is any subinterval of $[0, 1]$). As a simple application of Theorem 7 one obtains the linear case of the following result of Weyl [202].

Theorem 8 *If α is irrational and $p(x)$ is any non-constant polynomial with integer coefficients then the sequence $\{p(n)\alpha \pmod{1}\}$ is uniformly distributed.*

Furstenberg [84], see also [86], has shown that Weyl’s result, in full generality, can be deduced from the unique ergodicity of certain affine transformations of higher dimensional tori. For polynomials of degree $k > 1$ Weyl’s result is usually proved by inductively reducing to the case $k = 1$, using the following important lemma of van der Corput (see for example [139]).

Theorem 9 *Suppose that for each fixed $h > 0$ the sequence*

$$\{x_{n+h} - x_n \pmod{1}\}_n$$

is uniformly distributed. Then $\{x_n\}$ is uniformly distributed.

When μ is infinite and T is ergodic the limiting function in Birkhoff’s theorem is 0 a.e. In 1937 Hopf [100] proved a generalization of Birkhoff’s theorem which is more meaningful in the case of an infinite invariant measure.

It is a special case of a later theorem of Hurewicz [105], which we will discuss first.

Suppose that (X, \mathcal{B}, μ) is a σ -finite measure space. If ν is another σ -finite measure on \mathcal{B} write $\nu \ll \mu$ (ν is *absolutely continuous* relative to μ) if $\mu(E) = 0$ implies $\nu(E) = 0$ and one writes $\nu \sim \mu$ if $\nu \ll \mu$ and $\mu \ll \nu$. Consider a *non-singular automorphism* $\tau: X \rightarrow X$, meaning that τ is measurable with a measurable inverse and that $\mu(E) = 0$ if and only if $\mu(\tau E) = 0$. In other words $\nu = \mu \circ \tau \sim \mu$. By the Radon–Nikodym theorem there is a function $\rho \in L_1(\mu)$ such that $\rho > 0$ a.e. and $\nu(E) = \int_E \rho d\mu$ for all measurable E . In order to obtain an associated operator T on L_1 which is an (invertible) isometry one defines

$$Tf(x) = \rho(x)f(\tau x). \quad (11)$$

The dual operator on L_∞ is then given by $T^*g = g \circ \tau^{-1}$.

If ν is a σ -finite measure equivalent to μ which is invariant under τ then the ergodic theory of τ can be reduced to the measure-preserving case using σ . The interesting case is when there is no such ν . It was an open problem for some time whether there is always an equivalent invariant measure. In 1960 Ornstein [163] gave an example of a τ which does not have an equivalent invariant measure. It is curious that, with hindsight, such examples were already known in the fifties to people studying von Neumann algebras.

For $f \in L_1$ let $S_n f = \sum_{i=0}^n T^i f$. τ is said to be *conservative* if there is no set E with $\mu(E) > 0$ such that $\tau^{-i} E$, $i = 0, 1, \dots$ are pairwise disjoint, that is if the Poincaré recurrence theorem remains valid. For example, the shift on \mathbb{Z} is not conservative. Hurewicz [105] proved the following ratio ergodic theorem.

Theorem 10 *Suppose τ is conservative, $f, g \in L_1$ and $g(x) > 0$ a.e. Then $S_n f / S_n g$ converges a.e. to a τ -invariant limit h . If τ is ergodic then $h = \frac{\int f d\mu}{\int g d\mu}$.*

In the case when μ is τ -invariant one has $Tf = f \circ \tau$. If μ is invariant and finite, taking $g = 1$ one recovers Birkhoff’s theorem. If μ is invariant and σ -finite then Hurewicz’s theorem becomes the theorem of Hopf alluded to earlier.

Wiener and Wintner [204] proved the following variant of Birkhoff’s theorem.

Theorem 11 (Wiener–Wintner) *Suppose T is an automorphism of a probability space (X, \mathcal{B}, μ) . Then for any $f \in L_1$ there is a subset $X' \subset X$ of measure one such that for each $x \in X'$ and $\lambda \in \mathbb{C}$ of modulus 1 the sequence $\frac{1}{n} \sum_{i=0}^{n-1} \lambda^i T^i f(x)$ converges.*

It is an easy consequence of the ergodic theorem that one has a.e. convergence for a given f and λ but the point here

is that the set on which the convergence occurs is independent of λ .

Generalizations to Continuous Time and Higher-Dimensional Time

A (measure-preserving) flow is a one-parameter group $\{T_t, t \in \mathbb{R}\}$ of automorphisms of (X, \mathcal{B}, μ) , that is $T_{t+s} = T_t T_s$, such that $T_t x$ is measurable as a function of (t, x) . It will always be implicitly assumed that the map $(t, x) \mapsto T_t x$ from $\mathbb{R} \times X$ to X is measurable. Theorem 4 generalizes to flows by replacing sums with integrals and this generalization follows without difficulty from Theorem 4. (As already observed this observation reverses the historical record.) Theorem 4 may be viewed as a theorem about the “discrete flow” $\{T_n = T^n : n \in \mathbb{Z}\}$. Wiener was the first to generalize Birkhoff’s theorem to families of automorphisms $\{T_g\}$ indexed by groups more general than \mathbb{R} or \mathbb{Z} .

A measure-preserving flow is an *action* of \mathbb{R} while a single automorphism corresponds to an action of \mathbb{Z} . A (measure-preserving) action of a group G is a homomorphism $T: g \mapsto T_g$ from G into the group of automorphisms of a measure space (X, \mathcal{B}, μ) (satisfying the appropriate joint measurability condition in case G is not discrete). Suppose now that $G = \mathbb{R}^k$ or \mathbb{Z}^k and T is an action of G on (X, \mathcal{B}, μ) . In the case of \mathbb{Z}^k an action amounts to an arbitrary choice of commuting maps $T_1, \dots, T_k, T_i = T(e_i)$ where e_i is the standard basis of \mathbb{Z}^k . In the case of \mathbb{R}^k one must specify k commuting flows.

Let m denote counting measure on G in case $G = \mathbb{Z}^k$ and Lebesgue measure in case $G = \mathbb{R}^k$. For any subset E of G with $m(E) < \infty$ let

$$A_E f(x) = \frac{1}{m(E)} \int_E f(T_g x) dm(g). \quad (12)$$

One may then ask whether $A_E f$ converges to a limit, either in the mean or pointwise, as E varies through some sequence of sets which “grow large” or, in case $G = \mathbb{R}^k$, “shrink to 0”. The second case is referred to as a *local ergodic theorem*. In the case of ergodic theorems at infinity the continuous and discrete theories are rather similar and often the continuous analogue of a discrete result can be deduced from the discrete result.

In Wiener [203] proved the following result for actions of $G = \mathbb{R}^k$ and ergodic averages over Euclidean balls $B_r = \{x \in \mathbb{R}^k : \|x\|_2 \leq r\}$.

Theorem 12 Suppose T is an action of \mathbb{R}^k on (X, \mathcal{B}, μ) and $f \in L_1(\mu)$.

(a) For $f \in L_1 \lim_{r \rightarrow \infty} A_{B_r} f = g$ exists a.e. If μ is finite the convergence also holds with respect to the L_1 -norm,

g is T -invariant and $\int_I g d\mu = \int_I f d\mu$ for every T -invariant set I .

(b) $\lim_{r \rightarrow 0} A_{B_r} f = f$ a.e.

The local aspect of Wiener’s theorem is closely related to the Lebesgue differentiation theorem, see for example Proposition 3.5.4 of [132], which, in its simplest form, states that for $f \in L^1(\mathbb{R}^k, m)$ one has a.e. convergence of $\frac{1}{m(B_r)} \int_{B_r} f(x+t) dt$ to $f(x)$ as $r \rightarrow 0$. The local ergodic theorem implies Lebesgue’s theorem, simply by considering the action of \mathbb{R}^k on itself by translation. In fact the local ergodic theorem can also be deduced from Lebesgue’s theorem by a simple application of Fubini’s theorem (see for example [135], Chap. 1, Theorem 2.4 in the case $k = 1$).

The key point in Wiener’s proof is the following weak L_1 maximal inequality, similar to (6).

Theorem 13 Let $Mf = \sup_{r>0} |A_{B_r} f|$. Then one has

$$\mu\{Mf > \alpha\} \leq \frac{C}{\alpha} \|f\|_1, \quad (13)$$

where C is a constant depending only on the dimension d . (In fact one may take $C = 3^d$.)

In the case when T is the action of \mathbb{R}^k on itself by translation (13) is the well-known maximal inequality for the Hardy–Littlewood maximal function ([132], Lemma 3.5.3). Wiener proves (13) by way of the following *covering lemma*. If B is a ball in \mathbb{R}^k let B' denote the concentric ball with three times the radius.

Theorem 14 Suppose a compact subset K of \mathbb{R}^k is covered by a (finite) collection \mathcal{U} of open balls. Then there exist pairwise disjoint $B_i \in \mathcal{U}, i = 1, \dots, k$ such that $\bigcup_i B'_i$ covers K .

To find the B_i it suffices to let B_1 be the largest ball in \mathcal{U} , then B_2 the largest ball which does not intersect B_1 and in general B_n the largest ball which does not intersect $\bigcup_{i=1}^{n-1} B_i$. Then it is not hard to argue that the B'_i cover K .

In general, a covering lemma is, roughly speaking, an assertion that, given a cover \mathcal{U} of a set K in some space (often a group), one may find a subcollection $\mathcal{U}' \subset \mathcal{U}$ which covers a substantial part of K and is in some sense efficient in that $\sum_{U \in \mathcal{U}'} 1_U \leq C$, C some absolute constant. Covering lemmas play an important role in the proofs of many maximal inequalities. See Sect. 3.5 of [132] for a discussion of several of the best-known classical covering lemmas.

As Wiener was likely aware, the same kind of covering argument easily leads to maximal inequalities and ergodic theorems for averages over “sufficiently regular” sets. For example, in the case of $G = \mathbb{Z}^k$ use the standard total order $<$ on G and for $n \in \mathbb{N}^k$ let $S_n = \{m \in \mathbb{Z}^k : 0 < m <$

$n\}$. Let $e(n)$ denote the maximum value of n_i/n_j . Then one has the following result.

Theorem 15 For any $e > 0$ and any integrable f

$$\lim_{n \rightarrow \infty, e(n) < e} A_{S_n} f$$

exists a.e. and the limit is characterized by the same properties as in Theorem 12. Moreover there is a maximal inequality

$$\mu \left\{ \sup_{e(n) < e} |A_{S_n} f| > \alpha \right\} \leq \frac{C}{\alpha} \|f\|_1, \quad (14)$$

where C is a constant depending only on e and the dimension k .

If one does not impose a bound on $e(n)$ the a.e. convergence of $A_{S_n} f$ may fail for $f \in L_1$, (although it does hold for any increasing sequence of n 's by Theorem 30 below). Nonetheless Dunford [72] and independently Zygmund [215] showed that one does have unrestricted convergence if $f \in L_p$ for some $p > 1$, provided μ is finite. Let T_1, \dots, T_k be any k (possibly non-commuting!) automorphisms of a finite measure space (X, \mathcal{B}, μ) . For $n = (n_1, n_2, \dots, n_k)$ let $T_n = T_1^{n_1} \dots T_k^{n_k}$. (This is not an action of \mathbb{Z}^k unless the T_i commute with each other.) As before when $F \subset \mathbb{Z}^k$ is a finite subset write $A_F f = \frac{1}{|F|} \sum_{n \in F} T_n f$. Finally let $P_j f = \lim_n A_n(T_j) f$.

Theorem 16 For $f \in L_p$

$$\lim_{n \rightarrow \infty} A_{S_n} f = P_1 \dots P_k f \quad (15)$$

both a.e. and in L_p .

The proof of Theorem 16 uses repeated applications of Birkhoff's theorem and hinges on (10).

Somewhat surprisingly Hurewicz's theorem was not generalized to higher dimensions until the very recent work of Feldman [79]. In fact the theorem fails if one considers averages over the cube $[0, n-1]^d$ in \mathbb{Z}^d . However Feldman was able to prove a suitably formulated generalization of Hurewicz's theorem for averages over symmetric cubes.

For $f \in L^1(\mathbb{R})$ the classical Hilbert transform is defined for a.e. t by

$$Hf(t) = \frac{1}{\pi} \lim_{\epsilon \rightarrow 0^+} \int_{|s| > \epsilon} \frac{f(t-s)}{s} ds \quad (16)$$

Let $H^* f(t) = \frac{1}{\pi} \sup_{\epsilon > 0} \left| \int_{|s| > \epsilon} \frac{f(t-s)}{s} ds \right|$, the corresponding maximal function. The proof that the limit (16) exists a.e. is based on the maximal inequality

$$m\{Hf > \lambda\} \leq \frac{C}{\lambda} \|f\|_1, \quad (17)$$

where C is an absolute constant. See Loomis [146] (1946) for a proof of (17) using real-variable methods. Cotlar [65] proved the existence of the following ergodic analogue of the Hilbert transform (actually for the n -dimensional Hilbert transform).

Theorem 17 Suppose $\{T_t\}$ is a measure-preserving flow on a probability space (X, \mathcal{B}, μ) and $f \in L^1$. Then

$$\lim_{\epsilon \rightarrow 0} \int_{\epsilon < t < \epsilon^{-1}} \frac{f(T_t x)}{t} dt \quad (18)$$

exists for a.a. x .

Motivated by Cotlar's result Calderón [54] proved a general transfer principle which allows one to transfer maximal inequalities and convergence theorems for functions of a real or integer variable to the ergodic setting. Although stated for functions on \mathbb{R} it applies equally well to \mathbb{R}^k or \mathbb{Z}^k . This principle subsumes Birkhoff's theorem, Wiener's theorem and Cotlar's result. For simplicity only a special case of Calderón's result, in the discrete case, will be stated here.

Let T be an automorphism of a probability space (X, \mathcal{B}, μ) , let m denote counting measure on \mathbb{Z} and suppose σ is a probability measure on \mathbb{Z} . For $g \in L_1(m)$ define $\sigma(g)(n) = \sum_{i \in \mathbb{Z}} \sigma(i) g(n+i)$. For $f \in L_1(\mu)$ let $\sigma(T)f = \sum_{i \in \mathbb{Z}} \sigma(i) T^i f$, which is easily seen to converge a.e. and in L_1 -norm. Given a fixed sequence σ_n of probabilities define $Mg = \sup_n \sigma_n g$ and $M(T)f = \sup_n \sigma_n(T)f$.

Theorem 18 (Calderón) Suppose there is a constant C such that

$$m\{Mg > \lambda\} < \frac{C}{\lambda} \|g\|_1 \quad \text{for all } g \in L_1(m). \quad (19)$$

Then one also has

$$\mu\{M(T)f > \lambda\} < \frac{C}{\lambda} \|f\|_1 \quad \text{for all } T \quad \text{and } g \in L_1(\mu). \quad (20)$$

In other words, in order to prove a maximal inequality for general T it suffices to prove it in case T is the shift map on the integers. The idea of transference could already be seen in Wiener's proof of the \mathbb{R}^n ergodic theorem. Transference principles in various forms of have become an important tool in the study of ergodic theorems. See Bellow [18] for a very readable overview.

Pointwise Ergodic Theorems for Operators

Early in the history of ergodic theorems there were attempts to generalize the ergodic theorem to more general

linear operators on L_p spaces, that is, operators which do not arise by composition with a mapping of X . In the case $p = 1$ the main motivation for this comes from the theory of Markov processes.

If (X, \mathcal{B}) is a measurable space a *sub-stochastic kernel* on X is a non-negative function P on $X \times \mathcal{B}$ such that

- (a) for each $x \in X$ $P(x, \cdot) = P_x$ is a measure on \mathcal{B} such that $P_x(X) \leq 1$ and
- (b) $P(\cdot, A)$ is a measurable function for each $A \in \mathcal{B}$.

It is most intuitive to think about the *stochastic* case, namely when each P_x is a probability measure. One then views $P(x, A)$ as the probability that the point x moves into the set A in one unit of time, so one has stochastic dynamics as opposed to deterministic dynamics, namely the case when $P_x = \delta_{Tx}$ for a map T . In this case the measures P_x are called *transition probabilities*.

If μ is a σ -finite measure on X one may define the measure $P\mu = \int P_x d\mu(x)$. $P\nu$ is also meaningful if ν is a finite signed measure. The case when $P\mu = \mu$ is the stochastic analogue of measure-preserving dynamics and the case when $P\mu \ll \mu$ is the analogue of non-singular dynamics. It is easy to see that given any σ -finite measure λ there is always a μ such that $\lambda \ll \mu$ and $P\mu \ll \mu$. Let \tilde{L}_1 denote the space of finite signed measures ν such that $\nu \ll \mu$, which is identified with $L_1 = L_1(\mu, \mathbb{R})$ via the Radon–Nikodym theorem. If $P\mu \ll \mu$ then P maps $\tilde{L}_1(\mu)$ into itself so the restriction of P is an operator T on $L_1(\mu)$. T is a positive contraction, that is $\|T\| \leq 1$ and T maps non-negative functions to non-negative functions. As proved in, for example, [153], every positive contraction arises in this way from a substochastic kernel under the assumption that (X, \mathcal{B}) is standard. This simply means that there is some complete metric on X for which \mathcal{B} is the σ -algebra of Borel sets. Virtually all measurable spaces encountered in analysis are standard so this should be viewed as a technicality only. See [153], [80] and [135] for more about the relation between kernels and positive contractions.

The case when X is finite, P is stochastic and μ is a probability measure is classical in probability theory. P and μ determine a probability measure ν on $X^{\mathbb{N}}$ characterized by its values on cylinder sets, namely for all $x_1, \dots, x_n \in X$

$$\begin{aligned} \nu\{x \in X^{\mathbb{N}} : x(i) = x_i, 1 \leq i \leq n-1\} \\ = \mu(x_1) \prod_{i=1}^{n-1} P_{x_i}(x_{i+1}). \end{aligned} \quad (21)$$

The co-ordinate functions $X_i(x) = x(i)$ on the space $X^{\mathbb{N}}$ endowed with the probability ν form a *Markov process*,

which will be stationary if and only if μ is P -invariant. For a general X the analogous construction is possible provided X is standard.

Hopf [101] initiated the systematic study of positive L_1 contractions and proved the following ergodic theorem.

Theorem 19 Suppose (X, \mathcal{B}, μ) is a probability space and T is a positive contraction on $L^1(\mu)$ satisfying $T1 = 1$ and $T^*1 = 1$. Then $\lim_{n \rightarrow \infty} n^{-1} \sum_{k=0}^{n-1} T^k f$ exists a.e.

The importance of Hopf's article lies less in this convergence result than in the methods he developed. He proved that the maximal inequality (5) generalizes to all positive L_1 contractions T and used this to obtain the decomposition of X into its conservative and dissipative parts C and D , characterized by the fact that for any $p \in L_1$ such that $p > 0$ a.e. $\sum_{k=0}^{\infty} T^k p$ is infinite a.e. on C and finite a.e. on D . These results are the cornerstone of much of the subsequent work on L_1 contractions.

Theorem 19 contains Birkhoff's theorem for an automorphism τ of (X, \mathcal{B}, μ) , simply by defining $Tf = f \circ \tau$. In fact one can also deduce Theorem 19 from Birkhoff's theorem (if one assumes only that X is standard). Indeed the hypotheses of Hopf's theorem imply that the kernel P associated to T is stochastic ($T^*1 = 1$) and that μ is P -invariant ($T1 = 1$). Hopf's theorem then follows by applying Birkhoff's theorem to the shift on the stationary Markov process associated to P and μ . In fact Kakutani [123] (see also Doob [71]) had already made essentially the same observation, except that his result assumes the stationary Markov process to be already given.

In 1955 Dunford and Schwartz [73] made essential use of Hopf's work to prove the following result.

Theorem 20 Suppose $\mu(X) = 1$ and T is a (not necessarily positive) contraction with respect to both the L_1 and L_{∞} norms. Then the conclusion of Hopf's theorem remains valid.

Note that the assumption that T contracts the L_{∞} -norm is meaningful, as $L_{\infty} \subset L_1$. The proof of the result is reduced to the positive case by defining a positive contraction $|T|$ analogous to the total variation of a complex measure.

Then in 1960 Chacón and Ornstein [61] proved a definitive ratio ergodic theorem for all positive contractions of L_1 which generalizes both Hopf's theorem and the Hurewicz theorem.

Theorem 21 Suppose T is a positive contraction of $L_1(\mu)$, where μ is σ -finite, $f, g \in L_1$ and $g \geq 0$. Then

$$\frac{\sum_{i=0}^n T^i f}{\sum_{i=0}^n T^i g} \quad (22)$$

converges a.e. on the set $\{\sum_{i=0}^{\infty} T^i g > 0\}$.

In 1963 Chacón [58] proved a very general theorem for non-positive operators which includes the Chacón–Ornstein theorem as well as the Dunford–Schwartz theorem.

Theorem 22 Suppose T is a contraction of L_1 and $p_n \geq 0$ is a sequence of measurable functions with the property that

$$g \in L_1, |g| \leq p_n \Rightarrow |Tg| \leq p_{n+1}. \quad (23)$$

Then

$$\frac{\sum_{i=0}^n T^i f}{\sum_{i=0}^n p_i} \quad (24)$$

converges a.e. to a finite limit on the set $\{\sum_{i=0}^{\infty} p_i > 0\}$.

If T is an L_1, L_{∞} -contraction and $p_n = 1$ for all n then the hypotheses of this theorem are satisfied, so Theorem 22 reduces to the result of Dunford and Schwartz. See [59] for a concise overview of all of the above theorems in this section and the relations between them.

The identification of the limit in the Chacón–Ornstein theorem on the conservative part C of X is a difficult problem. It was solved by Neveu [152] in case $C = X$, and in general by Chacón [57]. Chacón [60] has shown that there is a non-singular automorphism τ of (X, \mathcal{B}, μ) such that for the associated invertible isometry T of $L_1(\mu)$ given by (11) there is an $f \in L_1(\mu)$ such that $\limsup A_n f = \infty$ and $\liminf A_n f = -\infty$.

In 1975 Akcoglu [3] solved a major open problem when he proved the following celebrated theorem.

Theorem 23 Suppose $T: L_p \mapsto L_p$ is a positive contraction. Then $A_n f = \frac{1}{n} \sum_{i=0}^{n-1} T^i f$ converges a.e. Moreover one has the strong L_p inequality

$$\left\| \sup_n |A_n f| \right\|_p \leq \frac{p}{p-1} \|f\|_p. \quad (25)$$

As usual, the maximal inequality (25) is the key to the convergence. Note that it is identical in form to (10) the classical strong L_p inequality for automorphisms. (25) was proved by A. Ionesco–Tulcea (now Bellow) [106] in the case of positive invertible isometries of L_p . It is a result of Banach [16], see also [140], that in this case T arises from a non-singular automorphism τ of (X, \mathcal{B}, μ) in the form $Tf = \rho^{1/p} f \circ \tau$. By a series of reductions Bellow was able to show that in this case (25) can be deduced from (10).

Akcoglu's brilliant idea was to consider a *dilation* S of T which is a positive invertible isometry on a larger L_p space $\tilde{L}_p = L_p(Y, \mathcal{C}, \nu)$. What this means is that there is a positive isometric injection $D: L_p \rightarrow \tilde{L}_p$ and a positive projection P on \tilde{L}_p whose range is $D(L_p)$ such that

$DT^n = PS^n D$ for all $n \geq 0$. Given the existence of such an S it is not hard to deduce (25) for T from (25) for S . In fact Akcoglu constructs a dilation only in the case when L_p is finite dimensional and shows how to reduce the proof of (25) to this case. In the finite dimensional case the construction is very concrete and P is a conditional expectation operator. Later Akcoglu and Kopp [7] gave a construction in the general case. It is noteworthy that the proof of Akcoglu's theorem consists ultimately of a long string of reductions to the classical strong L_p inequality (10), which in turn is a consequence of (5).

Subadditive and Multiplicative Ergodic Theorems

Consider a family $\{X_{n,m}\}$ of real-valued random variables on a probability space indexed by the set of pairs $(n, m) \in \mathbb{Z}^2$ such that $0 \leq n < m$. $\{X_{(n,m)}\}$ is called a (stationary) *subadditive process* if

- (a) the joint distribution of $\{X_{n,m}\}$ is the same as that of $\{X_{n+1,m+1}\}$
- (b) $X_{n,m} \leq X_{n,l} + X_{l,m}$ whenever $n < l < m$.

Denoting the index set by $\{n < m\} \subset \mathbb{Z}^2$ the distribution of the process is a measure μ on $\mathbb{R}^{\{n < m\}}$ which is invariant under the shift $Tx(n, m) = x(n+1, m+1)$. Thus there is no loss of generality in assuming that there is an underlying endomorphism T such that $X_{n,m} \circ T = X_{n+1,m+1}$. In 1968 Kingman [129] proved the following generalization of Birkhoff's theorem. $\gamma = \inf \frac{1}{n} \int X_{0,n} d\mu$ is called the *time constant* of the process.

Theorem 24 If the $X_{n,m}$ are integrable and $\gamma > -\infty$ then $\frac{1}{n} X_{0,n}$ converges a.e. and in L_1 -norm to a \mathbb{T} -invariant limit $\tilde{X} \in L_1(\mu)$ satisfying $\int \tilde{X} d\mu = \gamma$.

It is easy to deduce from the above that if one assumes only that $X_{0,1}^+$ is integrable then $\frac{1}{n} X_{0,n}$ still converges a.e. to a T -invariant limit \tilde{X} taking values in $[-\infty, \infty)$.

Subadditive processes first arose in the work of Hammersley and Welsh [99] on percolation theory. Here is an example. Let G be the graph with vertex set \mathbb{Z}^2 and with edges joining every pair of nearest neighbors. Let E denote the edge set and let $\{T_e: e \in E\}$ be non-negative integrable i.i.d. random variables. To each finite path P in G associate the “travel time” $T(P) = \sum_{e \in E} T_e$. For integers $m > n \geq 0$ let $X_{n,m}$ be the infimum of $T(P)$ over all paths P joining $(0, n)$ to $(0, m)$. This is a subadditive process with $0 \leq \gamma < \int T_e d\mu$ and it is not hard to see that the underlying endomorphism is ergodic. Thus Kingman's theorem yields the result that $\frac{1}{n} X_{0,n} \rightarrow \gamma$ a.e.

Suppose now that T is an ergodic automorphism of a probability space (X, \mathcal{B}, μ) and P is a function on X

taking values in the space of $d \times d$ real matrices. Define $P_{n,m} = P(T^{m-1}x)P(T^{m-2}x) \dots P(T^n x)$ and let $P_{n,m}(i, j)$ denote the i, j entry of $P_{n,m}$. Then $X_{n,m} = \log(\|P_{n,m}\|)$ (use any matrix norm) is a subadditive process so one obtains the first part of the following result of Furstenberg and Kesten [87] (1960) originally proved by more elaborate methods. The second part can also be deduced from the subadditive theorem with a little more work. See Kingman [130] for details and for some other applications of subadditive processes.

Theorem 25

- (a) Suppose $\int \log^+(|P|)d\mu < \infty$. Then $\|P_{0,n}\|^{1/n}$ converges a.e. to a finite limit.
- (b) Suppose that for each i, j $P(i, j)$ is a strictly positive function such that $\log P(i, j)$ is integrable. Then the limit $p = \lim(P_{0,n}(i, j))^{\frac{1}{n}}$ exists a.e. and is independent of i and j .

Partial results generalizing Kingman's theorem to the multiparameter case were obtained by Smythe [189] and Nguyen [157]. In 1981 Akcoglu and Krengel [8] obtained a definitive multi-parameter subadditive theorem. They consider an action $\{T_m\}$ of the semigroup $G = \mathbb{Z}_{\geq 0}^d$ by endomorphisms of a measure space (X, \mathcal{B}, μ) . Using the standard total ordering $<$ of G an interval in G is any set of the form $\{k \in G : m < k < n\}$ for any $m < n \in G$. Let \mathcal{I} denote the set of non-empty intervals. Reversing the direction of the inequality, they define a superadditive process as a collection of integrable functions $F_I, I \in \mathcal{I}$, such that

- (a) $F_I \circ T_m = F_{I+m}$,
- (b) $F_I \geq F_{I_1} + \dots + F_{I_k}$ whenever I is the disjoint union of I_1, \dots, I_k and
- (c) $\gamma = \sup_{I \in \mathcal{I}} |I|^{-1} \int F_I d\mu < \infty$.

A sequence $\{I_n\}$ of sets in \mathcal{I} is called *regular* if there is an increasing sequence I'_n such that $I_n \subset I'_n$ and $|I'_n| \leq C|I_n|$ for some constant C .

Theorem 26 (Akcoglu–Krengel) Suppose F_I is a superadditive process and $\{I_n\}$ is regular. Then $\frac{1}{|I_n|} \int F_{I_n} d\mu$ converges a.e.

$\{F_I\}$ is additive if the inequality in (b) is replaced by equality. In this case $F_I = \sum_{n \in I} f \circ T_n$ where f is an integrable function. Thus in the additive case the Akcoglu–Krengel result is a theorem about ordinary multi-dimensional ergodic averages, which is in fact a special case of an earlier result of Tempelman [196] (see Sect. “Amenable Groups” below).

Kingman's proof of Theorem 24 hinged on the existence of a certain (typically non-unique) decomposition

for subadditive processes. Akcoglu and Krengel's proof of the multi-parameter result does not depend on a Kingman-type decomposition, in fact they show that there is no such decomposition in general. They prove a weak maximal inequality

$$\mu\{\sup |I_n|^{-1} F_{I_n} > \lambda\} < \frac{C}{\lambda} \gamma, \quad (26)$$

where C is a constant depending only on the dimension, and show that this is sufficient to prove their result. In the case $d = 1$ the Akcoglu–Krengel argument provides a new and more natural proof of Kingman's theorem, similar in spirit to Wiener's arguments.

Akcoglu and Sucheston [9] have proved a ratio ergodic theorem for subadditive processes with respect to a positive L_1 contraction, generalizing both the Chacón–Ornstein theorem and Kingman's theorem.

In 1968 Oseledec [164] proved his celebrated multiplicative ergodic theorem, which gives very precise information about the random matrix products studied by Furstenberg and Kesten. His theorem is an important tool for the study of Lyapunov exponents in differentiable dynamics, see notably Pesin [169]. If A is a $d \times d$ matrix let $\|A\| = \sup\{\|Ax\| : \|x\| = 1\}$ where $\|x\|$ is the Euclidean norm on \mathbb{R}^n .

Theorem 27 Suppose T is an endomorphism of the probability space (X, \mathcal{B}, μ) . Suppose P is a measurable function on X whose values are $d \times d$ real matrices such that $\int \log^+ \|P\| d\mu < \infty$ and let $P_n(x) = P(T^{n-1}x)P(T^{n-2}x) \dots P(x)$. Then there is a T -invariant subset X' of X with measure 1 such that for $x \in X'$ the following hold.

- (a) $\lim_{n \rightarrow \infty} (P_n^*(x)P_n(x))^{\frac{1}{2n}} = A(x)$ exists.
- (b) Let $0 \leq \exp \lambda_1(x) \leq \exp \lambda_2(x) \leq \dots \leq \exp \lambda_r(x)$ be the distinct eigenvalues of $A(x)$ ($r = r(x)$ may depend on x and λ_1 may be $-\infty$) with multiplicities $m_1(x), \dots, m_r(x)$. Let $E_i(x)$ be the eigenspace corresponding to $\exp(\lambda_i(x))$ and set

$$F_i(x) = E_1(x) + \dots + E_i(x).$$

Then for each $u \in F_i(x) \setminus F_{i-1}(x)$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log \|P_n(x)u\| = \lambda_i(x).$$

- (c) The functions m_i and λ_i are T -invariant.
- (d) If T is ergodic, $\det P(x) = 1$ a.e. and

$$\limsup_n \frac{1}{n} \int \log \|P_n\| d\mu > 0$$

then the λ_i are constants, $\lambda_1 < 0$ and $\lambda_r > 0$.

Raghuathan [174] gave a much shorter proof of Osledec's theorem, valid for matrices with entries in a locally compact normed field. He showed that it could be reduced to the Furstenberg–Kesten theorem by considering the exterior powers of P . Ruelle [180] extended Osledec's theorem to the case where P takes values in the set of bounded operators on a Hilbert space. Walters [199] has given a proof (under slightly stronger hypotheses) which avoids the matrix calculations and tools from multilinear algebra used in other proofs.

Entropy and the Shannon–McMillan–Breiman Theorem

The notion of entropy was introduced by Shannon in his landmark work [185] which laid the foundations for a mathematical theory of information. Suppose (X, \mathcal{B}, μ) is a probability space, P is a finite measurable partition of X and T is an automorphism of (X, \mathcal{B}, μ) . $P(x)$ denotes the atom of P containing x . The *entropy* of P is

$$\begin{aligned} h(P) &= - \sum_{p \in P} \mu(p) \log(\mu(p)) \\ &= - \int \log(\mu(P(x))) d\mu(x) \geq 0. \end{aligned} \quad (27)$$

$-\log(\mu(A))$ may be viewed as a quantitative measure of the amount of information contained in the statement that a randomly chosen $x \in X$ happens to belong to A . So $h(P)$ is the expected information if one is about to observe which atom of P a randomly chosen point falls in. See Billingsley [34] for more motivation of this concept. See also the article in this collection on entropy by J. King or any introductory book on ergodic theory, e.g. Petersen [171].

If P and Q are partitions $P \vee Q$ denotes the common refinement which consists of all sets $p \cap q$, $p \in P$, $q \in Q$. It is intuitive and not hard to show that $h(P \vee Q) \leq h(P) + h(Q)$. Now let $P_0^n = \bigvee_{i=0}^{n-1} T^{-i}P$ and $h_n = h(P_0^n)$. The subadditivity of entropy implies that $h_{n+m} \leq h_n + h_m$, so by a well-known elementary lemma the limit

$$h(P, T) = \lim_n \frac{h_n}{n} = \inf_n \frac{h_n}{n} \geq 0 \quad (28)$$

exists.

If one thinks of $P(x)$ as a measurement performed on the space X and Tx as the state succeeding x after one second has elapsed then $h(P, T)$ is the expected information per second obtained by repeating the experiment every second for a very long time. See [177] for an alternative and very useful approach to $h(P, T)$ via *name-counting*.

The following result, which is known as the Shannon–McMillan–Breiman theorem, has proved to be of fundamental importance in ergodic theory, notably, for example, in the proof of Ornstein's celebrated isomorphism theorem for Bernoulli shifts [159].

Theorem 28 *If T is ergodic then*

$$\lim_{n \rightarrow \infty} -\frac{1}{n} \log(\mu(P_0^{n-1}(x))) = h(P, T) \quad (29)$$

a.e. and in L_1 -norm.

In other words, the actual information obtained per second by observing x over time converges to the constant $h(P, T)$, namely the limiting expected information per second. Shannon [185] formulated Theorem 28 and proved convergence in probability. McMillan [149] proved L_1 convergence and Breiman [49] obtained the a.e. convergence.

The original proofs of a.e. convergence used the martingale convergence theorem, were not very intuitive and did not generalize to \mathbb{Z}^n -actions, where the martingale theorem is not available. Ornstein and Weiss [161] (1983) found a beautiful and more natural argument which bypasses the martingale theorem and allows generalization to a class of groups which includes \mathbb{Z}^n .

Amenable Groups

Let G be any countable group and $T = \{T_g\}$ an action of G by automorphisms of a probability space. Suppose σ is a complex measure on G , that is, $\{\sigma(g)\}_{g \in G}$ is an absolutely summable sequence. Let T_g act on functions via $T_g f = f \circ T_g$, so T_g is an isometry of L_p for every $1 \leq p \leq \infty$. Let $\sigma(T) = \sum_{g \in G} \sigma(g) T_g$. A very general framework for formulating ergodic theorems is to consider a sequence $\{\sigma_n\}$ and ask whether $\sigma_n(T)f$ converges, a.e. or in mean, for f in some L_p space. When $\sigma_n(T)f$ converges for all actions T and all f in L_p in p -norm or a.e. then one says that σ_n is *mean* or *pointwise good* in L_p . When the σ_n are probability measures it is natural to call such results weighted ergodic theorems and this terminology is retained for complex σ as well.

Birkhoff's theorem says that if $G = \mathbb{Z}$ and σ_n is the normalized counting measure on $\{0, 1, \dots, n-1\}$ then $\{\sigma_n\}$ is pointwise good in L_1 . This section will be concerned only with sequences $\{\sigma_n\}$ such that σ_n is normalized counting measure on a finite subset $F_n \subset G$ so one speaks of mean or pointwise good sequences $\{F_n\}$. A natural condition to require of $\{F_n\}$, which will ensure that the limiting function is invariant, is that it be asymptoti-

cally (left) invariant in the sense that

$$\frac{|gF_n \Delta F_n|}{|F_n|} \rightarrow 0 \quad \forall g \in G. \quad (30)$$

Such a sequence is called a *Følner sequence* and a group G is *amenable* if it has a Følner sequence. As in most of this article G is restricted to be a discrete countable group for simplicity but most of the results to be seen actually hold for a general locally compact group.

Amenability of G is equivalent to the existence of a finitely additive left invariant probability measure on G . It is not hard to see that any Abelian, and more generally any solvable, group is amenable. On the other hand the free group F_2 on two generators is not amenable. See Patterson [167] for more information on amenable groups. The Følner property by itself is enough to give a mean ergodic theorem.

Theorem 29 *Any Følner sequence is mean good in L_p for $1 \leq p < \infty$.*

The proof of this result is rather similar to the proof of Theorem 3. In fact Theorem 29 is only a special case of quite general results concerning amenable semi-groups acting on abstract Banach spaces. See the book of Patterson [167] for more on this.

Turning to pointwise theorems, the Følner condition alone does not yield a pointwise theorem, even when $G = \mathbb{Z}$ and the F_n are intervals. For example Akcoglu and del Junco [6] have shown that when $G = \mathbb{Z}$ and $F_n = [n, n + \sqrt{n}] \cap \mathbb{Z}$ the pointwise ergodic theorem fails for any aperiodic T and for some characteristic function f . See also del Junco and Rosenblatt [119].

The following pointwise result of Tempelman [196] is often quoted. A Følner sequence $\{F_n\}$ is called *regular* if there is a constant C such that $|F_n^{-1}F_n| \leq C|F_n|$ and there is an increasing sequence F'_n such that $F_n \subset F'_n$ and $|F'_n| \leq C|F_n|$.

Theorem 30 *Any regular Følner sequence is pointwise good in L_1 .*

In case the F_n are intervals in \mathbb{Z}^n this result can be proved by a variant of Wiener's covering argument and in the general case by an abstraction thereof. The condition $|F_n^{-1}F_n| \leq C|F_n|$ captures the property of rectangles which is needed for the covering argument. Emerson [78] independently proved a very similar result.

The work on ergodic theorems for abstract locally compact groups was pioneered by Calderón [53] who built on Wiener's methods. The main result in this paper is somewhat technical but it already contains the germ of Tempelman's theorem. Other ergodic theorems

for amenable groups, whose main interest lies in the case of continuous groups, include Tempelman [195], Renaud [175], Greenleaf [93] and Greenleaf and Emerson [94]. The discrete versions of these results are all rather close to Tempelman's theorem.

Among pointwise theorems for discrete groups Tempelman's result was essentially the best available for a long time. It was not known whether every amenable group had a Følner sequence which is pointwise good for some L_p . In 1988 Shulman [187] introduced the notion of a *tempered Følner sequence* $\{F_n\}$, namely one for which

$$\left| \bigcup_{i < n} F_i^{-1} F_n \right| < C|F_n|, \quad (31)$$

for some constant C . The advantage of the tempered condition is that any Følner sequence has a tempered subsequence, and in particular any amenable group has a tempered Følner sequence.

Shulman proved a maximal inequality in L_2 for such F_n which implies that $\{F_n\}$ is pointwise good in L_2 . An account of this work may be found in Section 5.6 of Tempelman's book [194].

Lindenstrauss [145] was able to extend the result to L_1 .

Theorem 31 *Any tempered Følner sequence is pointwise good in L_1 .*

The key new idea in his proof is to use a probabilistic argument to establish a covering lemma. In the discrete case Ornstein and Weiss [201] have given a non-probabilistic proof of Lindenstrauss's covering lemma. Lindenstrauss also generalizes the a.e. convergence in the Shannon–McMillan–Breiman theorem to this setting. L_1 convergence was already established by Kieffer [128] in 1975.

Subsequence and Weighted Theorems

In this section G and T are as in the previous section and σ_n is a sequence of complex measures on G .

This section will be concerned with conditions on σ_n that ensure that it is mean or pointwise good. For the most part G will be \mathbb{Z} .

Hopf's ergodic theorem gives a class of examples for free. Choose any probability measure λ on G , define the operator $T_\lambda = \lambda(T)$ and observe that $(T_\lambda)^n = T_{\lambda^{*n}}$, where λ^{*n} denotes the convolution power. Since T_λ satisfies the hypotheses of Hopf's theorem it follows that the sequence $\sigma_n = \frac{1}{n} \sum_{i=0}^{n-1} \lambda^{*i}$ is pointwise good in L_1 .

Another sort of example is given by choosing a sequence g_n in G and letting $\sigma_n = \frac{1}{n} \sum_{i=0}^{n-1} \delta_{g_n}$. If one has convergence for such a sequence one speaks of a *subsequence ergodic theorem*. This section will mainly focus on

the case $G = \mathbb{Z}$ and sequences which are the increasing enumeration of a subset $S \subset \mathbb{N}$. Given $S \subset \mathbb{N}$ write $\sigma_{S,n}$ for the corresponding probabilities σ_n and say S is good if $\sigma_{S,n}$ is good.

Perhaps the first subsequence ergodic theorem is due to Blum and Hanson [38] who proved that an automorphism T of a probability space is *strongly mixing* if and only if $\frac{1}{n} \sum_{i=0}^{n-1} T^{m_i} f$ converges in L_2 norm for every $f \in L_2$ and every increasing $\{m_i\}$. Strong mixing means that $\mu(T^{-n}A \cap B) \rightarrow \mu(A)\mu(B)$. In 1969 Brunel and Keane [51] proved the first pointwise subsequence ergodic theorem

Theorem 32 *Suppose that T is a translation on a compact Abelian group G with Haar measure λ , $g \in G$, and $E \subset G$ is any Borel set with $\lambda(E) > 0$ and $\lambda(\partial E) = 0$. Let $S = \{i > 0 : T^i g \in E\}$. Then S is pointwise good in L_1 .*

Krengel [133] constructed the first example of a sequence $S \subset \mathbb{N}$ which is pointwise universally bad, in the strong sense that for any aperiodic T the a.e. convergence of $\sigma_{S,n}(T)f$ fails for some characteristic function f . Bellow [17] proved that any lacunary sequence (meaning $a_{n+1} > ca_n$ for some $c > 1$) is pointwise universally bad in L_1 .

Later Akcoglu et al. [1] were able to show that for lacunary sequences $\{\sigma_{S,n}\}$ is even *strongly sweeping out*. A sequence $\{\sigma_n\}$ of probability measures on \mathbb{Z} is said to be strongly sweeping out if for any ergodic T and for all $\delta > 0$ there is a characteristic function f with $\int f d\mu < \delta$ such that $\limsup \sigma_n(T)f = 1$ a.e. It is not difficult to show that if $\{\sigma_n\}$ is strongly sweeping out then there are characteristic functions f such that $\liminf \sigma_n(T)f = 0$ and $\limsup \sigma_n(T)f = 1$. Thus for lacunary sequences the ergodic theorem fails in the worst possible way.

Bellow and Losert [21] gave the first example of a sequence $S \subset \mathbb{Z}$ of density 0 which is universally good for pointwise convergence, answering a question posed by Furstenberg. They construct an S which is pointwise good in L_1 . This paper also contains a good overview of the progress on weighted and subsequence ergodic theorems at that time.

Weyl's theorem on uniform distribution (Theorem 9) suggests the possibility of an ergodic theorem for the sequence $\{n^2\}$. It is not hard to see that $\{n^2\}$ is mean good in L_2 . In fact the spectral theorem and the dominated convergence theorem show that it is enough to prove that the L_∞ -bounded sequence of functions $\frac{1}{n} \sum_{i=0}^{n-1} z^{n^2}$ on the unit circle converges at each point z of the unit circle. When z is not a root of unity the sequence converges to 0 by Weyl's result and when z is a root of unity the convergence is trivial because $\{z^{n^2}\}$ is periodic. In 1987 Bourgain [39,43]

proved his celebrated pointwise ergodic theorem for polynomial subsequences.

Theorem 33 *If p is any polynomial with rational coefficients taking integer values on the integers then $S = \{p(n)\}$ is pointwise good in L_2 .*

The first step in Bourgain's argument is to reduce the problem of proving a maximal inequality to the case of the shift map on the integers, via Calderón's transfer principle. Then the problem is transferred to the circle by using Fourier transforms. At this point the problem becomes a very delicate question about exponential sums and a whole arsenal of tools is brought to bear. See Rosenblatt and Wierdl [176] and Quas and Wierdl [28] (Appendix B) for nice expositions of Bourgain's methods.

Bourgain subsequently improved this to all L_p , $p > 1$ and also extended it to sequences $\{[q(n)]\}$ where now q is an arbitrary real polynomial and $[\cdot]$ denotes the greatest integer function. He also announced that his methods can be used to show that the sequence of primes is pointwise good in L_p for any $p > \frac{1+\sqrt{3}}{2}$. Wierdl [205] (1988) soon extended the result for primes to all $p > 1$.

Theorem 34 *The primes are pointwise good in L_p for $p > 1$.*

It has remained a major open question for quite some time whether any of these results hold for $p = 1$. In 2005 there appeared a preprint of Mauldin and Buczolic [148], which remains unpublished, showing that polynomial sequences are L_1 -universally bad.

Another major result of Bourgain's is the so-called return times theorem [44]. A simplification of Bourgain's original proof was published jointly with Furstenberg, Katznelson and Ornstein as an appendix to an article [47] of Bourgain. To state it let us agree to say that a sequence of complex numbers $\{a(n)\}_{n \geq 0}$ has property P if the sequence of complex measures $\sigma_n = \frac{1}{n} \sum_{i=0}^{n-1} a(i)\delta_i$ has property P , where δ_i denotes the point mass at i .

Theorem 35 (Bourgain) *Suppose T is an automorphism of a probability space (X, \mathcal{B}, μ) , $1 \leq p, q \leq \infty$ are conjugate exponents and $f \in L_p(\mu)$. Then for almost all x the sequence $\{f(T^n x)\}$ is pointwise good in L_q .*

Applying this to characteristic functions $f = 1_E$ one sees that the return time sequence $\{i > 0 : S^i x \in E\}$ is good for pointwise convergence in L_1 . Theorem 32 is a very special case. It is also easy to see that Theorem 35 contains the Wiener–Wintner theorem.

In 1998 Rudolph [179] proved a far-reaching generalization of the return times theorem using the technique of *joinings*. For an introduction to joinings see the article

by de la Rue in this collection and also Thouvenot [198], Glasner [90] (2003) and Rudolph's book [177]. Rudolph's result concerns the convergence of multiple averages

$$\frac{1}{N} \sum_{n=0}^{N-1} \prod_{j=1}^k f_j(T_j^n x) \quad (32)$$

where each T_j is an automorphism of a probability space $(X_j, \mathcal{B}_j, \mu_j)$ and the f_j are L_∞ functions. The point is that the convergence occurs whenever each $x_j \in X'_j$, sets of measure one which may be chosen sequentially for $j = 1, \dots, k$ without knowing what T_i or f_i are for any $i > j$. He actually proves something stronger, namely he identifies an intrinsic property of a sequence $\{a_i\}$, which he calls *fully generic*, such that the following hold.

- (a) The constant sequence $\{1\}$ is fully generic.
- (b) If $\{a_i\}$ is fully generic then for any T and $f \in L_\infty$ the sequence $a_i f(T^i x)$ is fully generic for almost all x .
- (c) Fully generic implies pointwise good in L_∞ .

The definition of fully generic will not be quoted here as it is somewhat technical.

For a proof of the basic return times theorem using joinings see Rudolph [178]. Assani, Lesigne and Rudolph [13] took a first step towards the multiple theorem, a Wiener–Wintner version of the return times theorem. Also Assani [11] independently gave a proof of Rudolph's result in the case when all the T_j are weakly mixing.

Ornstein and Weiss [162] have proved the following version of the return times theorem for abstract discrete groups. As with \mathbb{Z} , let us say that a sequence $\{a_g\}_{g \in G}$ of complex numbers has property P for $\{F_n\}$ if the sequence $\sigma_n = \frac{1}{|F_n|} \sum_{g \in F_n} a(g) \delta_g$ of complex measures has property P .

Theorem 36 *Suppose that the increasing Følner sequence $\{F_n\}$ satisfies the Tempelman condition $\sup_n |F_n^{-1} F_n| / |F_n| < \infty$ and $\bigcup F_n = G$. If $b \in L_\infty$ then for a.a. x the sequence $\{b(T_g x)\}$ is pointwise good in L_1 for $\{F_n\}$.*

Recently Demeter, Lacey, Tao and Thiele [67] have proved that the return times theorem remains valid for any $1 < p \leq \infty$ and $q \geq 2$. On the other hand Assani, Buczolich and Mauldin [14] (2005) showed that it fails for $p = q = 1$.

Bellow, Jones and Rosenblatt have a series of papers [22, 23, 24, 25] studying general weighted averages associated to a sequence σ_n of probability measures on \mathbb{Z} , and, in some cases, more general groups. The following are a few of their results. [23] is concerned with \mathbb{Z} -actions and *moving block averages* given by $\sigma_n = m_{I_n}$, where the

I_n are finite intervals and m_I denotes normalized counting measure on I . They resolve the problem completely, obtaining a checkable necessary and sufficient condition for such a sequence to be pointwise good in L_1 .

[24] gives sufficient conditions on a sequence σ_n for it to be pointwise good in L_p , $p > 1$, via properties of the Fourier transforms $\hat{\sigma}_n$. A particular consequence is that if $\lim_{n \rightarrow \infty} \sum_{k \in \mathbb{Z}} |\sigma_n(k) - \sigma_n(k-1)| = 0$ then $\{\sigma_n\}$ has a subsequence which is pointwise good in L_p , $p > 1$. In [25] they obtain convergence results for sequences $\sigma_n = \sigma^n$, the convolution powers of a probability measure σ . A consequence of one of their main results is that if the expectation $\sum_{k \in \mathbb{Z}} k \sigma(k)$ is zero, the second moment $\sum_{k \in \mathbb{Z}} k^2 \sigma(k)$ is finite and σ is *aperiodic* (its support is not contained in any proper coset in \mathbb{Z}) then σ^n is pointwise good in L_p for $p > 1$.

Bellow and Calderón [19] later showed that this last result is valid also for $p = 1$. This is a consequence of the following sufficient condition for a sequence T to satisfy a weak L_1 inequality. Given an automorphism of a probability space (X, \mathcal{B}, μ) let $Mf = \sup |\sigma_n(T)f(x)|$ be the maximal operator associated to $\{\sigma_n\}$.

Theorem 37 (Bellow and Calderón) *Suppose there is an $\alpha \in (0, 1]$ and $C > 0$ such that for each $n > 1$ one has*

$$|\sigma_n(x+y) - \sigma_n(x)| \leq C \frac{|y|^\alpha}{|x|^{1+\alpha}} \quad \text{for all } x, y \in \mathbb{Z}$$

such that $0 < 2|y| \leq |x|$

Then there is a constant D such that

$$\mu\{Mf > \lambda\} \leq \frac{D}{\lambda} \|f\|_1 \quad \text{for all } T, \quad f \in L_1(\mu)$$

and $\lambda > 0$.

Ergodic Theorems and Multiple Recurrence

Suppose $S \subset \mathbb{N}$. The *upper density* of S is

$$\bar{d}(S) = \limsup_n \frac{|S \cap [1, n]|}{n}. \quad (33)$$

and the *density* $d(S)$ is the limit of the same quantity, if it exists. In 1975 Szemerédi [190] proved the following celebrated theorem, answering an old question of Erdős and Turán.

Theorem 38 *Any subset of \mathbb{N} with positive upper density contains an arithmetic progression of length k for each $k \geq 1$.*

This result has a distinctly ergodic-theoretic flavor. Letting T denote the shift map on \mathbb{Z} , it says that for each k there

is an n such that $S' = \bigcap_{i=1}^k T^{-in} S$ is non-empty. In fact the result gives more: there is an n for which $\bar{d}(S') > 0$. In this light Szemerédi's theorem becomes a multiple recurrence theorem for the shift map on \mathbb{N} , equipped with the invariant “measure-like” quantity \bar{d} . Of course \bar{d} is not even finitely additive so it is not a measure. d , however, is at least finitely additive, when defined, and $d(\mathbb{N}) = 1$.

This point of view suggests the following multiple recurrence theorem.

Theorem 39 *Suppose T is an automorphism of a probability space (X, \mathcal{B}, μ) , $\mu(B) > 0$ and $k \geq 1$. Then there is an $n > 0$ such that $\mu(\bigcap_{i=1}^k T^{in} B) > 0$.*

In 1977, Furstenberg [85] proved the following ergodic theorem which implies the multiple recurrence theorem. He also established a general *correspondence principle* which puts the shaky analogy between the multiple recurrence theorem and Szemerédi's theorem on a firm footing and allows each to be deduced from the other. Thus he obtained an ergodic theoretic proof of Szemerédi's combinatorial result.

Theorem 40 *Suppose T is an automorphism of a probability space (X, \mathcal{B}, μ) , $f \in L_\infty$, $f \geq 0$, $\int f d\mu > 0$ and $k \geq 1$. Then*

$$\liminf_N \frac{1}{N} \sum_{n=0}^{N-1} \int \prod_{i=1}^k T^{in} f d\mu > 0. \quad (34)$$

Furstenberg's result opened the door to the study of so-called ergodic Ramsey theory which has yielded a vast array of deep results in combinatorics, many of which have no non-ergodic proof as yet. The focus of this article is not on this direction but the reader is referred to Furstenberg's book [86] for an excellent introduction and to Bergelson [27,28] for surveys of later developments. There is also the article by Frantzikinakis and McCutcheon in this collection.

Furstenberg's proof relies on a deep structure theorem for a general automorphism which was also developed independently by Zimmer [213], [212] in a more general context. A *factor* of T is any sub- σ -algebra $\mathcal{F} \subset \mathcal{B}$ such that $T(\mathcal{F}) = \mathcal{F}$. (It is more accurate to think of the factor as the action of T on the measure space $(X, \mathcal{F}, \mu|_{\mathcal{F}})$.) The structure theorem asserts that there is a transfinite increasing sequence of factors $\{\mathcal{F}_\alpha\}$ of T such that the following conditions hold.

- (a) $\mathcal{F}_{\alpha+1}$ is compact relative to \mathcal{F}_α .
- (b) $\mathcal{F}_\alpha = \bigvee_{\beta < \alpha} \mathcal{F}_\beta$ whenever α is a limit ordinal.
- (c) \mathcal{B} is weakly mixing relative to $\bigvee \mathcal{F}_\alpha$.

(\bigvee denotes the σ -algebra generated by a collection of σ -algebras.) $\bigvee_\alpha \mathcal{F}_\alpha$ is called the *maximal distal factor* of T and if $\bigvee \mathcal{F}_\alpha = \mathcal{B}$ then T is called *distal*. The definitions of the relative properties in (a) and (c) above are somewhat technical so only their absolute versions (i.e. relative to the trivial σ -algebra $\{\emptyset, X\}$) will be described here.

T is said to be compact if for each $f \in L_2$ the orbit $\{T^i f : i \in \mathbb{Z}\}$ is pre-compact in the norm topology of L_2 . This turns out to be equivalent to the statement that T is a translation on a compact Abelian group endowed with its Haar measure. The property of weak mixing is a fundamental notion in ergodic theory which has many equivalent definitions. The most appropriate for our purposes is that T is weakly mixing if it has no compact factors. This turns out to be equivalent to the ergodicity of $T \times T$ acting on the product measure space $(X, \mathcal{B}, \mu) \times (X, \mathcal{B}, \mu)$.

The verification of 37 in the case of compact T is rather easy. In this case it is not hard to prove that for any $f \in L_\infty$ and $\epsilon > 0$ the set $\{n \in \mathbb{Z} : \|T^n f - f\|_2 < \epsilon\}$ has bounded gaps and (34) follows easily. In the case when T is weakly mixing (34) is a consequence of the following theorem which Furstenberg proves in [85] (as a warm-up for its much harder relative version).

Theorem 41 *If T is weakly mixing and f_1, f_2, \dots, f_k are L_∞ functions then*

$$\lim_N \frac{1}{N} \sum_{n=0}^{N-1} \int \prod_{i=1}^k T^{in} f_i d\mu = \prod_{i=1}^k \int f_i \quad (35)$$

Later Bergelson [26] showed that the result can be obtained easily by an induction argument using the following Hilbert space generalization of van der Corput's lemma.

Theorem 42 (Bergelson) *Suppose $\{x_n\}$ is a bounded sequence of vectors in Hilbert space such that for each $h > 0$ one has $\frac{1}{N} \sum_{n=0}^{N-1} \langle x_{n+h}, x_n \rangle \rightarrow 0$ as $N \rightarrow \infty$. Then $\|\frac{1}{N} \sum_{n=0}^{N-1} x_n\| \rightarrow 0$.*

Ryzhikov has also given a beautiful short proof of Theorem 41 using joinings ([182]). Bergelson's van der Corput lemma and variants of it have been a key tool in subsequent developments in ergodic Ramsey theory and in the convergence results to be discussed in this section. Bergelson [26] used it to prove the following mean ergodic theorem for weakly mixing automorphisms.

Theorem 43 *Suppose T is weakly mixing, f_1, \dots, f_k are L_∞ functions and p_1, \dots, p_k are polynomials with rational coefficients taking integer values on the integers such that no*

$p_i - p_j$ is constant for $i \neq j$. Then

$$\lim_{N \rightarrow \infty} \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^k T^{p_i(n)} f_i - \prod_{i=1}^k \int f_i d\mu \right\| = 0. \quad (36)$$

Theorems 40 and 41 immediately raise the question of convergence of the multiple averages $\frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^k T^{in} f_i$ for a general T . Several authors obtained partial results on the question of mean convergence. It was finally resolved only recently by Host and Kra [104], who proved the following landmark theorem.

Theorem 44 Suppose $f_1, f_2, \dots, f_k \in L_\infty$. Then there is a $g \in L_\infty$ such that

$$\lim \left\| \frac{1}{N} \sum_{n=0}^{N-1} \prod_{i=1}^k T^{in} f_i - g \right\|_2 = 0. \quad (37)$$

Independently and somewhat later Ziegler [211] obtained the same result by somewhat different methods. Furstenberg had already established Theorem 44 for $k = 2$ in [85]. It was proved for $k = 3$ in the case of a totally ergodic T by Conze and Lesigne [63] and in general by Host and Kra [102]. It can also be obtained using the methods developed by Furstenberg and Weiss [88]. In this paper Furstenberg and Weiss proved a result for polynomial powers of T in the case $k = 2$. They also formalized the key notion of a *characteristic factor*. A factor C of T is said to be characteristic for the averages (37) if, roughly speaking, the L_2 limiting behavior of the averages is unchanged when any one of the f_i 's is replaced by its conditional expectation on C . This means that the question of convergence of these averages may be reduced to the case when f_i are all C -measurable. So the problem is to find the right (smallest) characteristic factor and prove convergence for that factor.

The importance of characteristic factors was already apparent in Furstenberg's original paper [85], where he showed that the maximal distal factor is characteristic for the averages (37). In fact he showed that for a given k a k -step distal factor is characteristic. (An automorphism is k -step distal if it is the top rung in a k -step ladder of factors as in the Furstenberg–Zimmer structure-theorem.) It turns out, though, that the right characteristic factor for (37) is considerably smaller. In their seminal paper [63] Conze and Lesigne identified the characteristic factor for $k = 3$, now called the Conze–Lesigné factor. As shown in [104], and [211], the characteristic factor for a general k is (isomorphic to) an inverse limit of k -step nilflows. A k -step nilflow is a compact homogeneous space N/Γ of a k -step nilpotent Lie group N , endowed with its

unique left-invariant probability measure, on which T acts via left translation by an element of N . Ergodic properties of nilflows have been studied for some time in ergodic theory, for example in Parry [166]. In this way the problem of L_2 -convergence of (37) is reduced to the case when T is a nilflow. In this case one has more: the averages converge pointwise by a result of Leibman [141] (See also Ziegler [210]).

There have already been a good many generalizations of (37). Host and Kra [103], Frantzikinakis and Kra [82, 83], and Leibman [141] have proved results which replace linear powers of T by polynomial powers. In increasing degrees of generality Conze and Lesigne [63], Frantzikinakis and Kra [81] and Tao [192] have obtained results which replace the maps T, T^2, \dots, T^k in (37) by commuting maps T_1, \dots, T_k . Bergelson and Leibman [30, 31] have obtained results, both positive and negative, in the case of two non-commuting maps.

In the direction of pointwise convergence the only general result is the following theorem of Bourgain [48] which asserts pointwise convergence in the case $k = 2$.

Theorem 45 Suppose S and T are powers of a single automorphism R and $f, g \in L_\infty$. Then $\frac{1}{N} \sum_{n=1}^N f(T^n x)g(S^n x)$ converges a.e.

When T is a K -automorphism Derrien and Lesigne [68] have proved that the averages (35) converge pointwise to the product of the integrals, even with polynomial powers of T replacing the linear powers.

Gowers [91] has given a new proof of Szemerédi's theorem by purely finite methods using only harmonic analysis on \mathbb{Z}_n . His results give better quantitative estimates in the statement of finite versions of Szemerédi's theorem. Although his proof contains no ergodic theory it is to some extent guided by Furstenberg's approach.

This section would be incomplete without mentioning the spectacular recent result of Green and Tao [92] on primes in arithmetic progression and the subsequent extensions of the Green–Tao theorem due to Tao [191] and Tao and Ziegler [193].

Rates of Convergence

There are many results which say in various ways that, in general, there are no estimates for the rate of convergence of the averages $A_n f$ in Birkhoff's theorem. For example there is the following result of Krengel [134].

Theorem 46 Suppose $\lim_{n \rightarrow \infty} c_n = \infty$ and T is any ergodic automorphism of a probability space. Then there is a bounded measurable f with $\int f d\mu = 0$ such that $\limsup c_n A_n f = \infty$ a.e.

See Part 1 of Derriennic [69] for a selection of other results in this direction. In spite of these negative results one can obtain quantitative estimates by reformulating the ergodic theorem in various ways. Bishop [37] proved the following result which is purely finite and constructive in nature and evidently implies the a.e. convergence in Birkhoff's theorem. If $y = (y_1, \dots, y_n)$ is a finite sequence of real numbers and $a < b$, an *upcrossing* of y over $[a, b]$ is a minimal integer interval $[k, l] \subset [1, n]$ satisfying $y_k < a$ and $y_l > b$.

Theorem 47 Suppose T is an automorphism of the probability space (X, \mathcal{B}, μ) . Let $U(n, a, b, f, x)$ be the number of upcrossings of the sequence $A_0 f(x), \dots, A_n f(x)$ over $[a, b]$. Then for every n

$$\mu\{x : U(n, a, b, f, x) > k\} \leq \frac{\|f\|_1}{(b-a)k}. \quad (38)$$

Ivanov [107] has obtained the following stronger upcrossing inequality for an arbitrary positive measurable f , which also implies Birkhoff's theorem.

Theorem 48 For any positive measurable f and $0 < a < b$

$$\mu\{x : U(n, a, b, f, x) > k\} \leq \left(\frac{a}{b}\right)^k. \quad (39)$$

Note the exponential decay and the remarkable fact that the estimate does not depend on f . Ivanov has also obtained the following result (Theorem 23 in [121]) about *fluctuations* of $A_n f$. An ϵ -*fluctuation* of a real sequence $y = (y_1 \dots y_n)$ is a minimal integer interval $[k, l]$ satisfying $|y_k - y_l| \geq \epsilon$. If $f \in L_1(\mathbb{R})$ let $F(\epsilon, f, x)$ be the number of ϵ -fluctuations of the sequence $\{A_n f(x)\}_{n=0}^\infty$.

Theorem 49

$$\mu\{x : F(\epsilon, f, x) \geq k\} \leq C \frac{(\log k)^{\frac{1}{2}}}{k} \quad (40)$$

where C is a constant depending only on $\|f\|_1/\epsilon$. If $f \in L_\infty$ then

$$\mu\{x : F(\epsilon, f, x) \geq k\} \leq A e^{-Bk}, \quad (41)$$

where A and B are constants depending only on $\|f\|_\infty/\epsilon$.

See Kachurovskii's survey [121] of results on rates of convergence in the ergodic theorem and also the article of Jones et al. [115] for more results on oscillation type inequalities.

Under special assumptions on f and T it is possible to give more precise results on speed of convergence. If the

sequence $T^i f$ is independent then there is a vast literature in probability theory giving very precise results, for example the central limit theorem and the law of the iterated logarithm (see, for example, [50]). See the surveys of Derriennic [69] (2006) and of Merlevède, Peligrad and Utev [150] for results on the central limit theorem for dynamical systems.

Ergodic Theorems for Non-amenable Groups

Guivarch [95] was the first to prove an ergodic theorem for a general pair of non-commuting unitary operators. Much work has been done in the last 15 years on mean and pointwise theorems for non-amenable groups. See the extensive survey by Nevo [155]. Here is just one result of Nevo [154] as a sample. Suppose G is a discrete group and $\{\sigma_n\}$ is a sequence of probability measures on G . Say that $\{\sigma_n\}$ is *mean ergodic* if for any unitary representation π of G on a Hilbert space \mathcal{H} and any $x \in \mathcal{H}$ one has $\sum_{g \in G} \sigma_n(g) \pi(g)x$ converges in norm to the projection of x on the subspace of π -invariant vectors. Say that $\{\sigma_n\}$ is *pointwise ergodic* if for any measure preserving action T of G on a probability space and $f \in L_2$ one has $\sum_{g \in G} \sigma_n(g) T_g f$ converges a.e. to the projection of f on the subspace of T -invariant functions.

Theorem 50 Let G be the free group on k generators, $k \geq 1$. Let τ_n be the normalized counting measure on the set of elements whose word length (in terms of the generators and their inverses) is n . Let $\sigma_n = (\tau_n + \tau_{n+1})/2$ and $\sigma'_n = \frac{1}{n} \sum_{i=1}^n \sigma_i$. Then σ_n and σ'_n are mean and pointwise ergodic but τ_n is not mean ergodic.

Future Directions

This section will be devoted to a few open problems and some questions. Many of these have been suggested by my colleagues acknowledged in the introduction.

The topic of convergence of Furstenberg's multiple averages has seen some remarkable achievements in the last few years and will likely continue to be vital for some time to come. The question of pointwise convergence of multiple averages is completely open, beyond Bourgain's result (Theorem 45). Even extending Bourgain's result to three different powers of R or to any two commuting automorphisms S and T would be a very significant achievement. Another natural question is whether one has pointwise convergence of the averages of the sequence $f(T^n(x))g(T^{n^2}(x))$, which are mean convergent by the result of Furstenberg and Weiss [88] (which is now subsumed in much more general results).

A long-standing open problem relating to Akcoglu's ergodic theorem (Theorem 23) for positive contractions of L_p is whether it can be extended to power-bounded operators T (this means that $\|T\|^n$ is bounded). It is also an open question whether it extends to non-positive contractions, excepting the case $p = 2$ where Burkholder [52] has shown that it fails.

There are many natural questions in the area of subsequence and weighted ergodic theorems. For example, which of Bourgain's several pointwise theorems can be extended to L_1 ? Are there other natural subsequences of an arithmetic character which have density 0 and for which an ergodic theorem is valid, either mean or pointwise, and in what L_p spaces might such theorems be valid? Are there higher dimensional analogues?

Since lacunary sequences are bad, to have any sort of pointwise theorem $l_n = \log \frac{a_n}{a_{n+1}}$ must get close to 0 and for simplicity let us assume that $\lim l_n = 0$. How fast must the convergence be in order to get an ergodic theorem? Jones and Wierdl [113] have shown that if $l_n > (\log n)^{-\frac{1}{2} + \epsilon}$ then the pointwise ergodic theorem fails in L_2 while Jones, Lacey and Wierdl [116] have shown that an only slightly faster rate permits a sequence which is pointwise good in L_2 . How well or badly does the ergodic theorem succeed or fail depending on the rate of convergence of l_n ? In particular is there a (slow) rate which still guarantees strong sweeping out? [116] contains some interesting conjectures in this direction.

There are also interesting questions concerning the mean and pointwise ergodic theorems for subsequences which are chosen randomly in some sense. See Bourgain [42] and [116] for some results in this direction. Again [116] contains some interesting conjectures along these lines.

In a recent paper [32] Bergelson and Leibman prove some very interesting and surprising results about the distribution of *generalized polynomials*. A generalized polynomial is any function which can be built starting with polynomials in $\mathbb{R}[x]$ using the operations of addition, multiplication and taking the greatest integer. As a consequence they derive a generalization of von Neumann's mean ergodic theorem to averages along generalized polynomial sequences. The following is a special case.

Theorem 51 *Suppose p is a generalized polynomial taking integer values on the integers and U is a unitary operator on \mathcal{H} . Then $\frac{1}{n} \sum_{i=0}^{n-1} U^{p(i)} x$ is norm convergent for all $x \in \mathcal{H}$.*

This begs the question: does one have pointwise convergence? If so, this would be a far-reaching generalization of Bourgain's polynomial ergodic theorem.

There are also lots of questions concerning the nature of Følner sequences $\{F_n\}$ in an amenable group which give a pointwise theorem. For example Lindenstrauss [145] has shown that in the *lamplighter* group, a semi-direct product of \mathbb{Z} with $\bigoplus_{i \in \mathbb{Z}} \mathbb{Z}/2\mathbb{Z}$ on which Z acts by the shift, there is no sequence satisfying Tempelman's condition and that any $\{F_n\}$ satisfying the Shulman condition must grow super-exponentially. So, it is natural to ask for slower rates of growth. In particular, in any amenable group is there always a sequence $\{F_n\}$ which is pointwise good and grows at most exponentially? Can one do better either in general or in particular groups?

Lindenstrauss's theorem at least guarantees the existence of Følner sequences which are pointwise good in L_1 but in particular groups there are often natural sequences which one hopes might be good. For example in $\bigoplus_{i=1}^{\infty} \mathbb{Z}$ one may take F_n to be a cube based at 0 of sidelength l_n and dimension d_n (that is, all but the first d_n co-ordinates are zero), where both sequences increase to ∞ . What conditions on l_n and d_n will give a good sequence? Note that no such sequence is regular in Tempelman's sense. If $d_n = n$ then $\{l_n\}$ must be superexponential to ensure Shulman's condition. Can one do better? What about $l_n = d_n = n$?

Bibliography

1. Akcoglu M, Bellow A, Jones RL, Losert V, Reinhold-Larsson K, Wierdl M (1996) The strong sweeping out property for lacunary sequences, Riemann sums, convolution powers, and related matters. *Ergodic Theory Dynam Systems* 16(2):207–253
2. Akcoglu M, Jones RL, Rosenblatt JM (2000) The worst sums in ergodic theory. *Michigan Math J* 47(2):265–285
3. Akcoglu MA (1975) A pointwise ergodic theorem in L_p -spaces. *Canad J Math* 27(5):1075–1082
4. Akcoglu MA, Chacon RV (1965) A convexity theorem for positive operators. *Z Wahrsch Verw Gebiete* 3:328–332 (1965)
5. Akcoglu MA, Chacon RV (1970) A local ratio theorem. *Canad J Math* 22:545–552
6. Akcoglu MA, del Junco A (1975) Convergence of averages of point transformations. *Proc Amer Math Soc* 49:265–266
7. Akcoglu MA, Kopp PE (1977) Construction of dilations of positive L_p -contractions. *Math Z* 155(2):119–127
8. Akcoglu MA, Krengel U (1981) Ergodic theorems for superadditive processes. *J Reine Angew Math* 323:53–67
9. Akcoglu MA, Sucheston L (1978) A ratio ergodic theorem for superadditive processes. *Z Wahrsch Verw Gebiete* 44(4):269–278
10. Alaoglu L, Birkhoff G (1939) General ergodic theorems. *Proc Nat Acad Sci USA* 25:628–630
11. Assani I (2000) Multiple return times theorems for weakly mixing systems. *Ann Inst H Poincaré Probab Statist* 36(2):153–165
12. Assani I (2003) Wiener–Wintner ergodic theorems. World Scientific Publishing Co. Inc., River Edge, NJ
13. Assani I, Lesigne E, Rudolph D (1995) Wiener–Wintner return-times ergodic theorem. *Israel J Math* 92(1–3):375–395

14. Assani I, Buczolich Z, Mauldin RD (2005) An L^1 counting problem in ergodic theory. *J Anal Math* 95:221–241
15. Auslander L, Green L, Hahn F (1963) Flows on homogeneous spaces. With the assistance of Markus L, Massey W and an appendix by Greenberg L *Annals of Mathematics Studies*, No 53. Princeton University Press, Princeton, NJ
16. Banach S (1993) *Théorie des opérations linéaires*. Éditions Jacques Gabay, Sceaux, reprint of the 1932 original
17. Bellow A (1983) On “bad universal” sequences in ergodic theory II. In: Belley JM, Dubois J, Morales P (eds) *Measure theory and its applications. Lecture Notes in Math*, vol 1033. Springer, Berlin, pp 74–78
18. Bellow A (1999) Transference principles in ergodic theory. In: Christ M, Kenig CE, Sadowsky C (eds) *Harmonic analysis and partial differential equations, Chicago Lectures in Math*. Univ Chicago Press, Chicago, pp 27–39
19. Bellow A, Calderón A (1999) A weak-type inequality for convolution products. In: Christ M, Kenig CE, Sadowsky C (eds) *Harmonic analysis and partial differential equations, Chicago Lectures in Math*. Univ Chicago Press, Chicago, pp 41–48
20. Bellow A, Jones R (eds) (1991) *Almost everywhere convergence, II*. Academic Press, Boston
21. Bellow A, Losert V (1985) The weighted pointwise ergodic theorem and the individual ergodic theorem along subsequences. *Trans Amer Math Soc* 288(1):307–345
22. Bellow A, Jones R, Rosenblatt J (1989) Almost everywhere convergence of powers. In: Edgar GA, Sucheston L (eds) *Almost everywhere convergence*. Academic, Boston, pp 99–120
23. Bellow A, Jones R, Rosenblatt J (1990) Convergence for moving averages. *Ergodic Theory Dynam Systems* 10(1):43–62
24. Bellow A, Jones RL, Rosenblatt J (1992) Almost everywhere convergence of weighted averages. *Math Ann* 293(3):399–426
25. Bellow A, Jones R, Rosenblatt J (1994) Almost everywhere convergence of convolution powers. *Ergodic Theory Dynam Systems* 14(3):415–432
26. Bergelson V (1987) Weakly mixing PET. *Ergodic Theory Dynam Systems* 7(3):337–349
27. Bergelson V (1996) Ergodic Ramsey theory—an update. In: Pollicott M, Schmidt K (eds) *Ergodic theory of \mathbb{Z}^d actions*. London Math Soc Lecture Note Ser, vol 228. Cambridge Univ Press, Cambridge, pp 1–61
28. Bergelson V (2006) Combinatorial and Diophantine applications of ergodic theory. In: Hasselblatt B, Katok A (eds) *Handbook of dynamical systems*, vol 1B, Appendix A by Leibman A, Appendix B by Quas A, Wierdl M. Elsevier, Amsterdam, pp 745–869
29. Bergelson V (2007) Some historical remarks and modern questions around the ergodic theorem. *Internat Math Nachrichten* 205:1–10
30. Bergelson V, Leibman A (2002) A nilpotent Roth theorem. *Invent Math* 147(2):429–470
31. Bergelson V, Leibman A (2004) Failure of the Roth theorem for solvable groups of exponential growth. *Ergodic Theory Dynam Systems* 24(1):45–53
32. Bergelson V, Leibman A (2007) Distribution of values of bounded generalized polynomials. *Acta Math* 198(2):155–230
33. Berkson E, Bourgain J, Gillespie TA (1991) On the almost everywhere convergence of ergodic averages for power-bounded operators on L^p -subspaces. *Integral Equ Operator Theory* 14(5):678–715
34. Billingsley P (1965) *Ergodic theory and information*. Wiley, New York
35. Birkhoff GD (1931) Proof of the ergodic theorem. *Proc Nat Acad Sci USA* 17:656–660
36. Bishop E (1966) An upcrossing inequality with applications. *Michigan Math J* 13:1–13
37. Bishop E (1967/1968) A constructive ergodic theorem. *J Math Mech* 17:631–639
38. Blum JR, Hanson DL (1960) On the mean ergodic theorem for subsequences. *Bull Amer Math Soc* 66:308–311
39. Bourgain J (1987) On pointwise ergodic theorems for arithmetic sets. *C R Acad Sci Paris Sér I Math* 305(10):397–402
40. Bourgain J (1988) Almost sure convergence and bounded entropy. *Israel J Math* 63(1):79–97
41. Bourgain J (1988) An approach to pointwise ergodic theorems. In: Lindenstrauss J, Milman VD (eds) *Geometric aspects of functional analysis (1986/87)*. Lecture Notes in Math, vol 1317. Springer, Berlin, pp 204–223
42. Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the integers. *Israel J Math* 61(1):39–72
43. Bourgain J (1988) On the pointwise ergodic theorem on L^p for arithmetic sets. *Israel J Math* 61(1):73–84
44. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. *C R Acad Sci Paris Sér I Math* 306(12):483–485
45. Bourgain J (1988) Temps de retour pour les systèmes dynamiques. *C R Acad Sci Paris Sér I Math* 306(12):483–485
46. Bourgain J (1989) Almost sure convergence in ergodic theory. In: Edgar GA, Sucheston L (eds) *Almost everywhere convergence*. Academic, Boston, pp 145–151
47. Bourgain J (1989) Pointwise ergodic theorems for arithmetic sets. *Inst Hautes Études Sci Publ Math* (69):5–45, with an appendix by the author, Furstenberg H, Katznelson Y, Ornstein DS
48. Bourgain J (1990) Double recurrence and almost sure convergence. *J Reine Angew Math* 404:140–161
49. Breiman L (1957) The individual ergodic theorem of information theory. *Ann Math Statist* 28:809–811
50. Breiman L (1968) *Probability*. Addison-Wesley, Reading
51. Brunel A, Keane M (1969) Ergodic theorems for operator sequences. *Z Wahrsch Verw Gebiete* 12:231–240
52. Burkholder DL (1962) Semi-Gaussian subspaces. *Trans Amer Math Soc* 104:123–131
53. Calderon AP (1953) A general ergodic theorem. *Ann Math* (2) 58:182–191
54. Calderón AP (1968) Ergodic theory and translation-invariant operators. *Proc Natl Acad Sci USA* 59:349–353
55. Calderón AP (1968) Ergodic theory and translation-invariant operators. *Proc Natl Acad Sci USA* 59:349–353
56. Calderon AP, Zygmund A (1952) On the existence of certain singular integrals. *Acta Math* 88:85–139
57. Chacon RV (1962) Identification of the limit of operator averages. *J Math Mech* 11:961–968
58. Chacon RV (1963) Convergence of operator averages. In: Wright FB (ed) *Ergodic theory*. Academic, New York, pp 89–120
59. Chacon RV (1963) Linear operators in L_1 . In: Wright FB (ed) *Ergodic theory*. Academic, New York, pp 75–87
60. Chacon RV (1964) A class of linear transformations. *Proc Amer Math Soc* 15:560–564

61. Chacon RV, Ornstein DS (1960) A general ergodic theorem. *Illinois J Math* 4:153–160
62. Conze JP, Lesigne E (1984) Théorèmes ergodiques pour des mesures diagonales. *Bull Soc Math France* 112(2):143–175
63. Conze JP, Lesigne E (1988) Sur un théorème ergodique pour des mesures diagonales. In: *Probabilités*, Publ Inst Rech Math Rennes, vol 1987. Univ Rennes I, Rennes, pp 1–31
64. Cotlar M (1955) On ergodic theorems. *Math Notae* 14:85–119 (1956)
65. Cotlar M (1955) A unified theory of Hilbert transforms and ergodic theorems. *Rev Mat Cuyana* 1:105–167 (1956)
66. Day M (1942) Ergodic theorems for Abelian semigroups. *Trans Amer Math Soc* 51:399–412
67. Demeter C, Lacey M, Tao T, Thiele C (2008) Breaking the duality in the return times theorem. *Duke Math J* 143(2):281–355
68. Derrien JM, Lesigne E (1996) Un théorème ergodique polynomial ponctuel pour les endomorphismes exacts et les K-systèmes. *Ann Inst H Poincaré Probab Statist* 32(6):765–778
69. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the “central limit theorem”. *Discrete Contin Dyn Syst* 15(1):143–158
70. Derriennic Y (2006) Some aspects of recent works on limit theorems in ergodic theory with special emphasis on the “central limit theorem”. *Discrete Contin Dyn Syst* 15(1):143–158
71. Doob JL (1938) Stochastic processes with an integral-valued parameter. *Trans Amer Math Soc* 44(1):87–150
72. Dunford N (1951) An individual ergodic theorem for non-commutative transformations. *Acta Sci Math Szeged* 14:1–4
73. Dunford N, Schwartz J (1955) Convergence almost everywhere of operator averages. *Proc Natl Acad Sci USA* 41:229–231
74. Dunford N, Schwartz JT (1988) *Linear operators, Part I*. Wiley Classics Library, Wiley, New York. General theory, with the assistance of Bade WG, Bartle RG, Reprint of the 1958 original, A Wiley–Interscience Publication
75. Durand S, Schneider D (2003) Random ergodic theorems and regularizing random weights. *Ergodic Theory Dynam Systems* 23(4):1059–1092
76. Eberlein WF (1949) Abstract ergodic theorems and weak almost periodic functions. *Trans Amer Math Soc* 67:217–240
77. Edgar G, Sucheston L (eds) (1989) *Almost everywhere convergence*. Academic Press, Boston
78. Emerson WR (1974) The pointwise ergodic theorem for amenable groups. *Amer J Math* 96:472–487
79. Feldman J (2007) A ratio ergodic theorem for commuting, conservative, invertible transformations with quasi-invariant measure summed over symmetric hypercubes. *Ergodic Theory Dynam Systems* 27(4):1135–1142
80. Foguel SR (1969) The ergodic theory of Markov processes. In: *Van Nostrand Mathematical Studies*, No 21. Van Nostrand Reinhold, New York
81. Frantzikinakis N, Kra B (2005) Convergence of multiple ergodic averages for some commuting transformations. *Ergodic Theory Dynam Systems* 25(3):799–809
82. Frantzikinakis N, Kra B (2005) Polynomial averages converge to the product of integrals. *Israel J Math* 148:267–276
83. Frantzikinakis N, Kra B (2006) Ergodic averages for independent polynomials and applications. *J London Math Soc* (2) 74(1):131–142
84. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math Systems Theory* 1:1–49
85. Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J Analyse Math* 31:204–256
86. Furstenberg H (1981) *Recurrence in ergodic theory and combinatorial number theory*. Princeton University Press, Princeton, NJ, m B Porter Lectures
87. Furstenberg H, Kesten H (1960) Products of random matrices. *Ann Math Statist* 31:457–469
88. Furstenberg H, Weiss B (1996) A mean ergodic theorem for $(1/N) \sum_{n=1}^N f(T^n x)g(T^{n^2} x)$. In: Bergelson V, March P, Rosenblatt J (eds) *Convergence in ergodic theory and probability*. Ohio State Univ Math Res Inst Publ, vol 5. de Gruyter, Berlin, pp 193–227
89. Garsia AM (1970) Topics in almost everywhere convergence. In: *Lectures in Advanced Mathematics*, vol 4. Markham, Chicago
90. Glasner E (2003) *Ergodic theory via joinings*, Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence
91. Gowers WT (2001) A new proof of Szemerédi’s theorem. *Geom Funct Anal* 11(3):465–588
92. Green B, Tao T (2004) The primes contain arbitrarily large arithmetic progressions. <http://arxiv.org/abs/mathNT/0404188>
93. Greenleaf FP (1973) Ergodic theorems and the construction of summing sequences in amenable locally compact groups. *Comm Pure Appl Math* 26:29–46
94. Greenleaf FP, Emerson WR (1974) Group structure and the pointwise ergodic theorem for connected amenable groups. *Adv Math* 14:153–172
95. Guivarc’h Y (1969) Généralisation d’un théorème de von Neumann. *C R Acad Sci Paris Sér A-B* 268:A1020–A1023
96. Halmos PR (1946) An ergodic theorem. *Proc Natl Acad Sci USA* 32:156–161
97. Halmos PR (1949) A nonhomogeneous ergodic theorem. *Trans Amer Math Soc* 66:284–288
98. Halmos PR (1960) *Lectures on ergodic theory*. Chelsea, New York
99. Hammersley JM, Welsh DJA (1965) First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory. In: *Proc Internat Res Semin. Statist Lab, University of California, Berkeley*. Springer, New York, pp 61–110
100. Hopf E (1937) Ergodentheorie. In: *Ergebnisse der Mathematik und ihrer Grenzgebiete*, vol 5. Springer, Berlin
101. Hopf E (1954) The general temporally discrete Markoff process. *J Rational Mech Anal* 3:13–45
102. Host B, Kra B (2001) Convergence of Conze–Lesigne averages. *Ergodic Theory Dynam Systems* 21(2):493–509
103. Host B, Kra B (2005) Convergence of polynomial ergodic averages. *Israel J Math* 149:1–19
104. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. *Ann Math* (2) 161(1):397–488
105. Hurewicz W (1944) Ergodic theorem without invariant measure. *Ann Math* (2) 45:192–206
106. Tulcea AI (1964) Ergodic properties of positive isometries. *Bull AMS* 70:366–371

107. Ivanov VV (1996) Geometric properties of monotone functions and the probabilities of random oscillations. *Sibirsk Mat Zh* 37(1):117–15
108. Ivanov VV (1996) Oscillations of averages in the ergodic theorem. *Dokl Akad Nauk* 347(6):736–738
109. Jewett RI (1969/1970) The prevalence of uniquely ergodic systems. *J Math Mech* 19:717–729
110. Jones R, Rosenblatt J, Tempelman A (1994) Ergodic theorems for convolutions of a measure on a group. *Illinois J Math* 38(4):521–553
111. Jones RL (1987) Necessary and sufficient conditions for a maximal ergodic theorem along subsequences. *Ergodic Theory Dynam Systems* 7(2):203–210
112. Jones RL (1993) Ergodic averages on spheres. *J Anal Math* 61:29–45
113. Jones RL, Wierdl M (1994) Convergence and divergence of ergodic averages. *Ergodic Theory Dynam Systems* 14(3):515–535
114. Jones RL, Olsen J, Wierdl M (1992) Subsequence ergodic theorems for L^p contractions. *Trans Amer Math Soc* 331(2):837–850
115. Jones RL, Kaufman R, Rosenblatt JM, Wierdl M (1998) Oscillation in ergodic theory. *Ergodic Theory Dynam Systems* 18(4):889–935
116. Jones RL, Lacey M, Wierdl M (1999) Integer sequences with big gaps and the pointwise ergodic theorem. *Ergodic Theory Dynam Systems* 19(5):1295–1308
117. Jones RL, Rosenblatt JM, Wierdl M (2001) Oscillation inequalities for rectangles. *Proc Amer Math Soc* 129(5):1349–1358 (electronic)
118. Jones RL, Rosenblatt JM, Wierdl M (2003) Oscillation in ergodic theory: higher dimensional results. *Israel J Math* 135:1–27
119. del Junco A, Rosenblatt J (1979) Counterexamples in ergodic theory and number theory. *Math Ann* 245(3):185–197
120. Kac M (1947) On the notion of recurrence in discrete stochastic processes. *Bull Amer Math Soc* 53:1002–1010
121. Kachurovskii AG (1996) Rates of convergence in ergodic theorems. *Uspekhi Mat Nauk* 51(4(310)):73–124
122. Kachurovskii AG (1996) Spectral measures and convergence rates in the ergodic theorem. *Dokl Akad Nauk* 347(5):593–596
123. Kakutani S (1940) Ergodic theorems and the Markoff process with a stable distribution. *Proc Imp Acad Tokyo* 16:49–54
124. Kalikow S, Weiss B (1999) Fluctuations of ergodic averages. In: *Proceedings of the Conference on Probability, Ergodic Theory, and Analysis*, vol 43. pp 480–488
125. Kamae T (1982) A simple proof of the ergodic theorem using nonstandard analysis. *Israel J Math* 42(4):284–290
126. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, *Encyclopedia of Mathematics and its Applications*, vol 54. Cambridge University Press, Cambridge, with a supplementary chapter by Katok A and Mendoza L
127. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. *Israel J Math* 42(4):291–296
128. Kieffer JC (1975) A generalized Shannon–McMillan theorem for the action of an amenable group on a probability space. *Ann Probability* 3(6):1031–1037
129. Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. *J Roy Statist Soc Ser B* 30:499–510
130. Kingman JFC (1976) Subadditive processes. In: *École d'Été de Probabilités de Saint-Flour, V–1975. Lecture Notes in Math*, vol 539. Springer, Berlin, pp 167–223
131. Koopman B (1931) Hamiltonian systems and transformations in hilbert spaces. *Proc Natl Acad Sci USA* 17:315–318
132. Krantz SG, Parks HR (1999) The geometry of domains in space. Birkhäuser Advanced Texts: Basler Lehrbücher. Birkhäuser Boston Inc, Boston, MA
133. Krengel U (1971) On the individual ergodic theorem for subsequences. *Ann Math Statist* 42:1091–1095
134. Krengel U (1978/79) On the speed of convergence in the ergodic theorem. *Monatsh Math* 86(1):3–6
135. Krengel U (1985) Ergodic theorems. In: *de Gruyter studies in mathematics*, vol 6. de Gruyter, Berlin, with a supplement by Antoine Brunel
136. Krengel U, Lin M, Wittmann R (1990) A limit theorem for order preserving nonexpansive operators in L_1 . *Israel J Math* 71(2):181–191
137. Krieger W (1972) On unique ergodicity. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol II. Probability Theory, University of California Press, Berkeley, pp 327–346
138. Kryloff N, Bogoliouboff N (1937) La théorie générale de la mesure dans son application à l'étude des systèmes dynamiques de la mécanique non linéaire. *Ann Math* (2) 38(1):65–113
139. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, New York, Pure and Applied Mathematics
140. Lamperti J (1958) On the isometries of certain function-spaces. *Pacific J Math* 8:459–466
141. Leibman A (2005) Convergence of multiple ergodic averages along polynomials of several variables. *Israel J Math* 146:303–315
142. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial actions of \mathbb{Z}^d by translations on a nilmanifold. *Ergodic Theory Dynam Systems* 25(1):215–225
143. Lemańczyk M, Lesigne E, Parreau F, Volný D, Wierdl M
144. Lesigne E (1989) Théorèmes ergodiques pour une translation sur un nilvariété. *Ergodic Theory Dynam Systems* 9(1):115–126
145. Lindenstrauss E (1999) Pointwise theorems for amenable groups. *Electron Res Announc Amer Math Soc* 5:82–90
146. Loomis LH (1946) A note on the Hilbert transform. *Bull Amer Math Soc* 52:1082–1086
147. Lorch ER (1939) Means of iterated transformations in reflexive vector spaces. *Bull Amer Math Soc* 45:945–947
148. Mauldin D, Buczolich Z (2005) Concepts behind divergent ergodic averages along the squares. In: *Assani I (ed) Ergodic theory and related fields. Contemp Math*, vol 430. Amer Math Soc, Providence, pp 41–56
149. McMillan B (1953) The basic theorems of information theory. *Ann Math Statistics* 24:196–219
150. Merlevède F, Peligrad M, Utev S (2006) Recent advances in invariance principles for stationary sequences. *Probab Surv* 3:1–36
151. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. *Proc Natl Acad Sci USA* 18:70–82
152. Neveu J (1961) Sur le théorème ergodique ponctuel. *C R Acad Sci Paris* 252:1554–1556
153. Neveu J (1965) Mathematical foundations of the calculus of probability. Translated by Amiel Feinstein, Holden-Day, San Francisco

154. Nevo A (1994) Harmonic analysis and pointwise ergodic theorems for noncommuting transformations. *J Amer Math Soc* 7(4):875–902
155. Nevo A (2006) Pointwise ergodic theorems for actions of groups. In: Hasselblatt B, Katok A (eds) *Handbook of dynamical systems* vol 1B. Elsevier, Amsterdam, pp 871–982
156. Nevo A, Stein EM (1994) A generalization of Birkhoff's pointwise ergodic theorem. *Acta Math* 173(1):135–154
157. Nguyen XX (1979) Ergodic theorems for subadditive spatial processes. *Z Wahrsch Verw Gebiete* 48(2):159–176
158. Orey S (1971) Lecture notes on limit theorems for Markov chain transition probabilities. In: Van Nostrand Reinhold Mathematical Studies, no 34. Van Nostrand Reinhold Co, London
159. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. *Adv Math* 4:337–352 (1970)
160. Ornstein D (1971) A remark on the Birkhoff ergodic theorem. *Illinois J Math* 15:77–79
161. Ornstein D, Weiss B (1983) The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Israel J Math* 44(1):53–60
162. Ornstein D, Weiss B (1992) Subsequence ergodic theorems for amenable groups. *Israel J Math* 79(1):113–127
163. Ornstein DS (1960) On invariant measures. *Bull Amer Math Soc* 66:297–300
164. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Ljapunov, exponents of dynamical systems. *Trudy Moskov Mat Obšč* 19:179–210
165. Oxtoby JC (1952) Ergodic sets. *Bull Amer Math Soc* 58:116–136
166. Parry W (1969) Ergodic properties of affine transformations and flows on nilmanifolds. *Amer J Math* 91:757–771
167. Paterson A (1988) Amenability, *Mathematical Surveys and Monographs*, vol 29. American Mathematical Society, Providence, RI
168. Peck JEL (1951) An ergodic theorem for a noncommutative semigroup of linear operators. *Proc Amer Math Soc* 2:414–421
169. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. *Uspehi Mat Nauk* 32(4(196)):55–112,287
170. Petersen K (1983) Another proof of the existence of the ergodic Hilbert transform. *Proc Amer Math Soc* 88(1):39–43
171. Petersen K (1989) *Ergodic theory*, Cambridge Studies in Advanced Mathematics, vol 2. Cambridge University Press, Cambridge
172. Pitt HR (1942) Some generalizations of the ergodic theorem. *Proc Cambridge Philos Soc* 38:325–343
173. Poincaré H (1987) *Les méthodes nouvelles de la mécanique céleste*. Tome I, II, III. Les Grands Classiques Gauthier–Villars. Librairie Scientifique et Technique Albert Blanchard, Paris
174. Raghuathan MS (1979) A proof of Oseledec's multiplicative ergodic theorem. *Israel J Math* 32(4):356–362
175. Renaud PF (1971) General ergodic theorems for locally compact groups. *Amer J Math* 93:52–64
176. Rosenblatt JM, Wierdl M (1995) Pointwise ergodic theorems via harmonic analysis. In: Peterson KE, Salama IA (eds) *Ergodic theory and its connections with harmonic analysis*, London Math Soc Lecture Note Ser, vol 205. Cambridge University Press, Cambridge, pp 3–151
177. Rudolph DJ (1990) Fundamentals of measurable dynamics. In: *Fundamentals of measurable dynamics: Ergodic theory on Lebesgue spaces*. Oxford Science Publications, The Clarendon Press, Oxford University Press, New York
178. Rudolph DJ (1994) A joinings proof of Bourgain's return time theorem. *Ergodic Theory Dynam Systems* 14(1):197–203
179. Rudolph DJ (1998) Fully generic sequences and a multiple-term return-times theorem. *Invent Math* 131(1):199–228
180. Ruelle D (1982) Characteristic exponents and invariant manifolds in Hilbert space. *Ann Math* (2) 115(2):243–290
181. Ryll–Nardzewski C (1951) On the ergodic theorems, II. Ergodic theory of continued fractions. *Studia Math* 12:74–79
182. Ryzhikov V (1994) Joinings, intertwining operators, factors and mixing properties of dynamical systems. *Russian Acad Sci Izv Math* 42:91–114
183. Shah NA (1998) Invariant measures and orbit closures on homogeneous spaces for actions of subgroups generated by unipotent elements. In: Dani SG (ed) *Lie groups and ergodic theory*. Tata Inst Fund Res Stud Math, vol 14. Tata Inst Fund Res, Bombay, pp 229–271
184. Shalom Y (1998) Random ergodic theorems, invariant means and unitary representation. In: Dani SG (ed) *Lie groups and ergodic theory*. Tata Inst Fund Res Stud Math, vol 14. Tata Inst Fund Res, Bombay, pp 273–314
185. Shannon CE (1948) A mathematical theory of communication. *Bell System Tech J* 27:379–423, 623–656
186. Shields PC (1987) The ergodic and entropy theorems revisited. *IEEE Trans Inform Theory* 33(2):263–266
187. Shulman A (1988) Maximal ergodic theorems on groups. *Dep Lit NIINTI No.* 2184
188. Sine R (1970) A mean ergodic theorem. *Proc Amer Math Soc* 24:438–439
189. Smythe RT (1976) Multiparameter subadditive processes. *Ann Probability* 4(5):772–782
190. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. *Acta Arith* 27:199–245
191. Tao T (2005) The gaussian primes contain arbitrarily shaped constellations. <http://arxiv.org/abs/math/0501314>
192. Tao T (2008) Norm convergence of multiple ergodic averages for commuting transformations. *Ergod Theory Dyn Syst* 28(2):657–688
193. Tao T, Ziegler T (2006) The primes contain arbitrarily long polynomial progressions. <http://frontmath.ucdavis.edu/06105050>
194. Tempelman A (1992) *Ergodic theorems for group actions, Mathematics and its Applications*, vol 78. Kluwer, Dordrecht, informational and thermodynamical aspects, Translated and revised from the 1986 Russian original
195. Tempel'man AA (1967) Ergodic theorems for general dynamical systems. *Dokl Akad Nauk SSSR* 176:790–793
196. Tempel'man AA (1972) Ergodic theorems for general dynamical systems. *Trudy Moskov Mat Obšč* 26:95–132
197. Tempel'man AA (1972) A generalization of a certain ergodic theorem of Hopf. *Teor Veroyatnost i Primenen* 17:380–383
198. Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: Peterson KE, Salama IA (eds) *Ergodic theory and its connections with harmonic analysis*, London Math Soc Lecture Note Ser, vol 205. Cambridge University Press, Cambridge, pp 207–235

199. Walters P (1993) A dynamical proof of the multiplicative ergodic theorem. *Trans Amer Math Soc* 335(1):245–257
200. Weber M (1998) Entropie métrique et convergence presque partout, *Travaux en Cours*, vol 58. Hermann, Paris
201. Weiss B (2003) Actions of amenable groups. In: Bezuglyi S, Kolyada S (eds) *Topics in dynamics and ergodic theory*, London Math Soc Lecture Note Ser, vol 310. Cambridge University Press, Cambridge, pp 226–262
202. Weyl H (1916) Über die Gleichverteilung von Zahlen mod Eins. *Math Ann* 77(3):313–352
203. Wiener N (1939) The ergodic theorem. *Duke Math J* 5(1):1–18
204. Wiener N, Wintner A (1941) Harmonic analysis and ergodic theory. *Amer J Math* 63:415–426
205. Wierdl M (1988) Pointwise ergodic theorem along the prime numbers. *Israel J Math* 64(3):315–336 (1989)
206. Wittmann R (1995) Almost everywhere convergence of ergodic averages of nonlinear operators. *J Funct Anal* 127(2):326–362
207. Yosida K (1940) An abstract treatment of the individual ergodic theorem. *Proc Imp Acad Tokyo* 16:280–284
208. Yosida K (1940) Ergodic theorems of Birkhoff–Khintchine's type. *Jap J Math* 17:31–36
209. Yosida K, Kakutani S (1939) Birkhoff's ergodic theorem and the maximal ergodic theorem. *Proc Imp Acad, Tokyo* 15:165–168
210. Ziegler T (2005) A non-conventional ergodic theorem for a nilsystem. *Ergodic Theory Dynam Systems* 25(4):1357–1370
211. Ziegler T (2007) Universal characteristic factors and Furstenberg averages. *J Amer Math Soc* 20(1):53–97 (electronic)
212. Zimmer RJ (1976) Ergodic actions with generalized discrete spectrum. *Illinois J Math* 20(4):555–588
213. Zimmer RJ (1976) Extensions of ergodic group actions. *Illinois J Math* 20(3):373–409
214. Zund JD (2002) George David Birkhoff and John von Neumann: a question of priority and the ergodic theorems, 1931–1932. *Historia Math* 29(2):138–156
215. Zygmund A (1951) An individual ergodic theorem for non-commutative transformations. *Acta Sci Math Szeged* 14:103–110

Ergodic Theory: Basic Examples and Constructions

MATTHEW NICOL¹, KARL PETERSEN²

¹ Department of Mathematics,
University of Houston, Houston, USA

² Department of Mathematics,
University of North Carolina, Chapel Hill, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Examples
Constructions

Future Directions

Bibliography

Glossary

- A transformation T of a measure space (X, \mathcal{B}, μ) is *measure-preserving* if $\mu(T^{-1}A) = \mu(A)$ for all measurable $A \in \mathcal{B}$.
- A measure-preserving transformation (X, \mathcal{B}, μ, T) is *ergodic* if $T^{-1}(A) = A \pmod{\mu}$ implies $\mu(A) = 0$ or $\mu(A^c) = 0$ for each measurable set $A \in \mathcal{B}$.
- A measure-preserving transformation (X, \mathcal{B}, μ, T) of a probability space is *weak-mixing* if $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} |\mu(T^{-i}A \cap B) - \mu(A)\mu(B)| = 0$ for all measurable sets $A, B \in \mathcal{B}$.

A measure-preserving transformation (X, \mathcal{B}, μ, T) of a probability space is *strong-mixing* if $\lim_{n \rightarrow \infty} \mu(T^{-n}A \cap B) = \mu(A)\mu(B)$ for all measurable sets $A, B \in \mathcal{B}$.

- A continuous transformation T of a compact metric space X is *uniquely ergodic* if there is only one T -invariant Borel probability measure on X . A continuous transformation of a topological space X is *topologically mixing* for any two open sets $U, V \subset X$ there exists $N > 0$ such that $T^{-n}(U) \cap V \neq \emptyset$, for each $n \geq N$.
- Suppose (X, \mathcal{B}, μ) is a probability space. A *finite partition* \mathcal{P} of X is a finite collection of disjoint (mod μ , i.e., up to sets of measure 0) measurable sets $\{P_1, \dots, P_n\}$ such that $X = \cup P_i \pmod{\mu}$. The *entropy* of \mathcal{P} with respect to μ is $H(\mathcal{P}) = -\sum_i \mu(P_i) \ln \mu(P_i)$ (other bases are sometimes used for the logarithm).
- The *metric (or measure-theoretic) entropy* of T with respect to \mathcal{P} is $h_\mu(T, \mathcal{P}) = \lim_{n \rightarrow \infty} \frac{1}{n} H(\mathcal{P} \vee \dots \vee T^{-n+1}(\mathcal{P}))$, where $\mathcal{P} \vee \dots \vee T^{-n+1}(\mathcal{P})$ is the partition of X into sets of points with the same coding with respect to \mathcal{P} under T^i , $i = 0, \dots, n-1$. That is x, y are in the same set of the partition $\mathcal{P} \vee \dots \vee T^{-n+1}(\mathcal{P})$ if and only if $T^i(x)$ and $T^i(y)$ lie in the same set of the partition \mathcal{P} for $i = 0, \dots, n-1$.
- The *metric entropy* $h_\mu(T)$ of (X, \mathcal{B}, μ, T) is the supremum of $h_\mu(T, \mathcal{P})$ over all finite measurable partitions \mathcal{P} .
- If T is a continuous transformation of a compact metric space X , then the *topological entropy* of T is the supremum of the metric entropies $h_\mu(T)$, where the supremum is taken over all T -invariant Borel probability measures.
- A system (X, \mathcal{B}, μ, T) is *loosely Bernoulli* if it is isomorphic to the first-return system to a subset of positive measure of an irrational rotation or a (positive or infinite entropy) Bernoulli system.

- Two systems are *spectrally isomorphic* if the unitary operators that they induce on their L^2 spaces are unitarily equivalent.
- A *smooth dynamical system* consists of a differentiable manifold M and a differentiable map $f: M \rightarrow M$. The degree of differentiability may be specified.
- Two submanifolds S_1, S_2 of a manifold M *intersect transversely* at $p \in M$ if $T_p(S_1) + T_p(S_2) = T_p(M)$.
- An $(\epsilon-)$ small C^r *perturbation* of a C^r map f of a manifold M is a map g such that $d_{C^r}(f, g) < \epsilon$ i.e. the distance between f and g is less than ϵ in the C^r topology.
- A map T of an interval $I = [a, b]$ is *piecewise smooth* (C^k for $k \geq 1$) if there is a finite set of points $a = x_1 < x_2 < \dots < x_n = b$ such that $T|(x_i, x_{i+1})$ is C^k for each i . The degree of differentiability may be specified.
- A measure μ on a measure space (X, \mathcal{B}) is *absolutely continuous* with respect to a measure ν on (X, \mathcal{B}) if $\nu(A) = 0$ implies $\mu(A) = 0$ for all measurable $A \in \mathcal{B}$.
- A Borel measure μ on a Riemannian manifold M is *absolutely continuous* if it is absolutely continuous with respect to the Riemannian volume on M .
- A measure μ on a measure space (X, \mathcal{B}) is *equivalent* to a measure ν on (X, \mathcal{B}) if μ is absolutely continuous with respect to ν and ν is absolutely continuous with respect to μ .

Definition of the Subject

Measure-preserving systems are a common model of processes which evolve in time and for which the rules governing the time evolution don't change. For example, in Newtonian mechanics the planets in a solar system undergo motion according to Newton's laws of motion: the planets move but the underlying rule governing the planets' motion remains constant. The model adopted here is to consider the time-evolution as a transformation (either a map in discrete time or a flow in continuous time) on a probability space or more generally a measure space. This is the setting of the subject called ergodic theory. Applications of this point of view include the areas of statistical physics, classical mechanics, number theory, population dynamics, statistics, information theory and economics. The purpose of this chapter is to present a flavor of the diverse range of examples of measure-preserving transformations which have played a role in the development and application of ergodic theory and smooth dynamical systems theory. We also present common constructions involving measure-preserving systems. Such constructions may be considered a way of putting 'building-block' dynamical systems to-

gether to construct examples or decomposing a complicated system into simple 'building-blocks' to understand it better.

Introduction

In this chapter we collect a brief list of some important examples of measure-preserving dynamical systems, which we denote typically by (X, \mathcal{B}, μ, T) or (T, X, \mathcal{B}, μ) or slight variations. These examples have played a formative role in the development of dynamical systems theory, either because they occur often in applications in one guise or another or because they have been useful simple models to understand certain features of dynamical systems. There is a fundamental difference in the dynamical properties of those systems which display hyperbolicity: roughly speaking there is some exponential divergence of nearby orbits under iteration of the transformation. In differentiable systems this is associated with the derivative of the transformation possessing eigenvalues of modulus greater than one on a 'dynamically significant' subset of phase space. Hyperbolicity leads to complex dynamical behavior such as positive topological entropy, exponential divergence of nearby orbits ("sensitivity to initial conditions") often coexisting with a dense set of periodic orbits. If ϕ, ψ are sufficiently regular functions on the phase space X of a hyperbolic measure-preserving transformation (T, X, μ) , then typically we have fast decay of correlations in the sense that

$$\left| \int_X \phi(T^n x) \psi(x) d\mu - \int \phi d\mu \int \psi d\mu \right| \leq Ca(n)$$

where $a(n) \rightarrow 0$. If $a(n) \rightarrow 0$ at an exponential rate we say that the system has exponential decay of correlations. A theme in dynamical systems is that the time series formed by sufficiently regular observations on systems with some degree of hyperbolicity often behave statistically like independent identically distributed random variables.

At this point it is appropriate to point out two pervasive differences between the usual probabilistic setting of a stationary stochastic process $\{X_n\}$ and the (smooth) dynamical systems setting of a time series of observations on a measure-preserving system $\{\phi \circ T^n\}$. The most crucial is that for deterministic dynamical systems the time series is usually not an independent process, which is a common assumption in the strictly probabilistic setting. Even if some weak-mixing is assumed in the probabilistic setting it is usually a mixing condition on the σ -algebras $\mathcal{F}_n = \sigma(X_1, \dots, X_n)$ generated by successive random variables, a condition which is not natural (and usually very difficult to check) for dynamical systems. Mix-

ing conditions on dynamical systems are given more naturally in terms of the mixing of the sets of the σ -algebra \mathcal{B} of the probability space (X, \mathcal{B}, μ) under the action of T and not by mixing properties of the σ -algebras generated by the random variables $\{\phi \circ T^n\}$. The other difference is that in the probabilistic setting, although $\{X_n\}$ satisfy moment conditions, usually no regularity properties, such as the Hölder property or smoothness, are assumed. In contrast in dynamical systems theory the transformation T is often a smooth or piecewise smooth transformation of a Riemannian manifold X and the observation $\phi: X \rightarrow \mathbb{R}$ is often assumed continuous or Hölder. The regularity of the observation ϕ turns out to play a crucial role in proving properties such as rates of decay of correlation, central limit theorems and so on.

An example of a hyperbolic transformation is an expanding map of the unit interval $T(x) = (2x)$ (where (x) is x modulo the integers). Here the derivative has modulus 2 at all points in phase space. This map preserves Lebesgue measure, has positive topological entropy, Lebesgue almost every point x has a dense orbit and periodic points for the map are dense in $[0, 1)$.

Non-hyperbolic systems are of course also an important class of examples, and in contrast to hyperbolic systems they tend to model systems of ‘low complexity’, for example systems displaying quasiperiodic behavior. The simplest non-trivial example is perhaps an irrational rotation of the unit interval $[0, 1)$ given by a map $T(x) = (x + \alpha)$, $\alpha \in \mathbb{R} \setminus \mathbb{Q}$. T preserves Lebesgue measure, every point has a dense orbit (there are no periodic orbits), yet the topological entropy is zero and nearby points stay the same distance from each other under iteration under T .

There is a natural notion of equivalence for measure-preserving systems. We say that measure-preserving systems (T, X, \mathcal{B}, μ) and (S, Y, \mathcal{C}, ν) are *isomorphic* if (possibly after deleting sets of measure 0 from X and Y) there is a one-to-one onto measurable map $\phi: X \rightarrow Y$ with measurable inverse ϕ^{-1} such that $\phi \circ T = S \circ \phi$ μ a.e. and $\mu(\phi^{-1}(A)) = \nu(A)$ for all $A \in \mathcal{C}$. If X, Y are compact topological spaces we say that T is *topologically conjugate* to S if there exists a homeomorphism $\phi: X \rightarrow Y$ such that $\phi \circ T = S \circ \phi$. In this case we call ϕ a *conjugacy*. If ϕ is C^r for some $r \geq 1$ we will call ϕ a *C^r -conjugacy* and similarly for other degrees of regularity.

We will consider $X = [0, 1)(\text{mod } 1)$ as a representation of the unit circle $S^1 = \{z \in \mathbb{C} : |z| = 1\}$ (under the map $x \rightarrow e^{2\pi i x}$) and similarly represent the k -dimensional torus $T^k = S^1 \times \cdots \times S^1$ (k -times). If the σ -algebra is clear from the context we will write (T, X, μ) instead of (T, X, \mathcal{B}, μ) when denoting a measure-preserving system.

Examples

Rigid Rotation of a Compact Group

If G is a compact group equipped with Haar measure and $a \in G$, then the transformation $T(x) = ax$ preserves Haar measure and is called a *rigid rotation* of G . If G is abelian and the transformation is ergodic (in this setting transitivity implies ergodicity), then the transformation is uniquely ergodic. Such systems always have zero topological entropy.

The simplest example of such a system is a circle rotation. Take $X = [0, 1)(\text{mod } 1)$, with

$$T(x) = (x + \alpha) \quad \text{where } \alpha \in \mathbb{R}.$$

Then T preserves Lebesgue (Haar) measure and is ergodic (in fact uniquely ergodic) if and only if α is irrational. Similarly, the map

$$T(x_1, \dots, x_k) = (x_1 + \alpha_1, \dots, x_k + \alpha_k),$$

where $\alpha_1, \dots, \alpha_k \in \mathbb{R}$,

preserves k -dimensional Lebesgue (Haar) measure and is ergodic (uniquely ergodic) if and only if there are no integers m_1, \dots, m_k , not all 0, which satisfy $m_1\alpha_1 + \cdots + m_k\alpha_k \in \mathbb{Z}$.

Adding Machines

Let $\{k_i\}_{i \in \mathbb{N}}$ be a sequence of integers with $k_i \geq 2$. Equip each cyclic group \mathbb{Z}_{k_i} with the discrete topology and form the product space $\Sigma = \prod_{i=1}^{\infty} \mathbb{Z}_{k_i}$ equipped with the product topology. An *adding machine* corresponding to the sequence $\{k_i\}_{i \in \mathbb{N}}$ is the topological space $\Sigma = \prod_{i=1}^{\infty} \mathbb{Z}_{k_i}$ together with the map

$$\sigma: \Sigma \rightarrow \Sigma$$

defined by

$$\sigma(k_1 - 1, k_2 - 1, \dots) = (0, 0, \dots)$$

if each entry in the \mathbb{Z}_{k_i} component is $k_i - 1$,

while

$$\sigma(k_1 - 1, k_2 - 1, \dots, k_n - 1, x_1, x_2, \dots) \\ = (0, 0, \dots, 0, \overbrace{x_1 + 1, x_2, x_3, \dots}^{n \text{ times}})$$

when $x_1 \neq k_{n+1} - 1$.

The map σ may be thought of as “add one and carry” and also as mapping each point to its successor in a certain order. See Sect. “[Adic Transformations](#)” for generalizations. If each $k_i = 2$ then the system is called the *dyadic*

(or von Neumann-Kakutani) adding machine or 2-odometer. Adding machines give examples of naturally occurring minimal systems of low orbit complexity in the sense that the topological entropy of an adding machine is zero. In fact if f is a continuous map of an interval with zero topological entropy and S is a closed, topologically transitive invariant set without periodic orbits, then the restriction of f to S is topologically conjugate to the dyadic adding machine (Theorem 11.3.13 in [46]).

We say a non-empty set Λ is an *attractor* for a map T if there is an open set U containing Λ such that $\Lambda = \bigcap_{n \geq 0} T^n(U)$ (other definitions are found in the literature). The dyadic adding machine is topologically conjugate to the Feigenbaum attractor at the limit point of period doubling bifurcations (see Sect. “Unimodal Maps”). Furthermore, attractors for continuous unimodal maps of the interval are either periodic orbits, transitive cycles of intervals, or Cantor sets on which the dynamics is topologically conjugate to an adding machine [32].

Interval Exchange Maps

A map $T: [0, 1] \rightarrow [0, 1]$ is an *interval exchange transformation* if it is defined in the following way. Suppose that π is a permutation of $\{1, \dots, n\}$ and $l_i > 0$, $i = 1, \dots, n$, is a sequence of subintervals of I (open or closed) with $\sum_i l_i = 1$. Define t_i by $l_i = t_i - t_{i-1}$ with $t_0 = 0$. Suppose also that σ is an n -vector with entries ± 1 . T is defined by sending the interval $t_{i-1} \leq x < t_i$ of length l_i to the interval

$$\sum_{\pi(j) < \pi(i)} l_{\pi(j)} \leq x < \sum_{\pi(j) > \pi(i)} l_{\pi(j)}$$

with orientation preserved if the i th entry of σ is $+1$ and orientation reversed if the i th entry of σ is -1 . Thus on each interval l_i , T has the form $T(x) = \sigma_i x + a_i$, where σ_i is ± 1 . If $\sigma_i = 1$ for each i , the transformation is called *orientation preserving*.

The transformation T has finitely many discontinuities (at the endpoints of each l_i), and modulo this set of discontinuities is smooth. T is also invertible (neglecting the finite set of discontinuities) and preserves Lebesgue measure. These maps have zero topological entropy and arise naturally in studies of polygonal billiards and more generally area-preserving flows. There are examples of minimal but non-ergodic interval exchange maps [56, 69].

Full Shifts and Shifts of Finite Type

Given a finite set (or alphabet) $A = \{0, \dots, d-1\}$, take $X = \Omega^+(A) = A^{\mathbb{N}}$ (or $X = A^{\mathbb{Z}}$) the sets of one-sided

(two-sided) sequences, respectively, with entries from A . For example sequences in $A^{\mathbb{N}}$ have the form $x = \cdot x_0 x_1 \dots x_n \dots$. A *cylinder set* $C(y_{n_1}, \dots, y_{n_k})$, $y_{n_i} \in A$, of length k is a subset of X defined by fixing k entries; for example,

$$C(y_{n_1}, \dots, y_{n_k}) = \{x: x_{n_1} = y_{n_1}, \dots, x_{n_k} = y_{n_k}\}.$$

We define the set A^k to consist of all cylinders $C(y_1, \dots, y_k)$ determined by fixing the first k entries, i. e. an element of A^k is specified by fixing the first k entries of a sequence $\cdot x_0 \dots x_k$ by requiring $x_i = y_i$, $i = 0, \dots, k$.

Let $p = (p_0, \dots, p_{d-1})$ be a probability vector: all $p_i \geq 0$ and $\sum_{i=0}^{d-1} p_i = 1$. For any cylinder $B = C(b_1, \dots, b_k) \in A^k$, define

$$g_k(B) = p_{b_1} \dots p_{b_k}. \quad (1)$$

It can be shown that these functions on A^k extend to a shift-invariant measure μ_p on $A^{\mathbb{N}}$ (or $A^{\mathbb{Z}}$) called product measure. (See the article on [Measure Preserving Systems](#).) The space $A^{\mathbb{N}}$ or $A^{\mathbb{Z}}$ may be given a metric by defining

$$d(x, y) = \begin{cases} 1 & \text{if } x_0 \neq y_0; \\ \frac{1}{2^{|n|}} & \text{if } x_n \neq y_n \text{ and } x_i = y_i \text{ for } |i| < n. \end{cases}$$

The *shift* $\sigma(\cdot x_0 x_1 \dots x_n \dots) = \cdot x_1 x_2 \dots x_n \dots$ is ergodic with respect to μ_p . The measure-preserving system $(\Omega, \mathcal{B}, \mu, \sigma)$ (with \mathcal{B} the σ -algebra of Borel subsets of $\Omega(A)$, or its completion), is denoted by $\mathcal{B}(p)$ and is called the *Bernoulli shift* determined by p . This system models an infinite number of independent repetitions of an experiment with finitely many outcomes, the i th of which has probability p_i on each trial.

These systems are mixing of all orders (i. e. σ^n is mixing for all $n \geq 1$) and have countable Lebesgue spectrum (hence are all spectrally isomorphic). Kolmogorov and Sinai showed that two of them cannot be isomorphic unless they have the same entropy; Ornstein [81] showed the converse. $\mathcal{B}(1/2, 1/2)$ is isomorphic to the Lebesgue-measure-preserving transformation $x \rightarrow 2x \bmod 1$ on $[0, 1]$; similarly, $\mathcal{B}(1/3, 1/3, 1/3)$ is isomorphic to $x \rightarrow 3x \bmod 1$. Furstenberg asked whether the only nonatomic measure invariant for both $x \rightarrow 2x \bmod 1$ and $x \rightarrow 3x \bmod 1$ on $[0, 1]$ is Lebesgue measure. Lyons [68] showed that if one of the actions is K , then the measure must be Lebesgue, and Rudolph [100] showed the same thing under the weaker hypothesis that one of the actions has positive entropy. For further work on this question, see [50, 87].

This construction can be generalized to model one-step finite-state Markov stochastic processes as dynamical systems. Again let $A = \{0, \dots, d-1\}$, and let $p = (p_0, \dots,$

p_{d-1}) be a probability vector. Let P be a $d \times d$ stochastic matrix with rows and columns indexed by A . This means that all entries of P are nonnegative, and the sum of the entries in each row is 1. We regard P as giving the transition probabilities between pairs of elements of A . Now we define for any cylinder $B = C(b_1, \dots, b_k) \in A^k$

$$\mu_{P,P}(B) = p_{b_1} P_{b_1 b_2} P_{b_2 b_3} \dots P_{b_{k-1} b_k}. \quad (2)$$

It can be shown that $\mu_{P,P}$ extends to a measure on the Borel σ -algebra of $\Omega^+(A)$, and its completion. (See the article on [► Measure Preserving Systems](#).) The resulting stochastic process is a (one-step, finite-state) Markov process. If p and P also satisfy

$$pP = p, \quad (3)$$

then the Markov process is stationary. In this case we call the (one or two-sided) measure-preserving system the Markov shift determined by p and P .

Aperiodic and irreducible Markov chains (those for which a power of the transition matrix P has all entries positive) are strongly mixing, in fact are isomorphic to Bernoulli shifts (usually by means of a complicated measure-preserving recoding).

More generally we say a set $\Lambda \subset A^{\mathbb{Z}}$ is a *subshift* if it is compact and invariant under σ . A subshift Λ is said to be of *finite type* (SFT) if there exists an $d \times d$ matrix $M = (a_{ij})$ such that all entries are 0 or 1 and $x \in \Lambda$ if and only if $a_{x_i x_{i+1}} = 1$ for all $i \in \mathbb{Z}$. Shifts of finite type are also called *topological Markov chains*. There are many invariant measures for a non-trivial shift of finite type. For example the orbit of each periodic point is the support of an invariant measure. An important role in the theory, derived from motivations of statistical mechanics, is played by *equilibrium measures* (or *equilibrium states*) for continuous functions $\phi: \Lambda \rightarrow \mathbb{R}$, i. e. those measures μ which maximize $\{h_\sigma(\mu) + \int_\Lambda \phi d\mu\}$ over all shift-invariant probability measures, where $h_\sigma(\mu)$ is the measure-theoretic entropy of σ with respect to μ . The study of full shifts or shifts of finite type has played a prominent role in the development of the hyperbolic theory of dynamical systems as physical systems with ‘chaotic’ dynamics ‘typically’ possess an invariant set with induced dynamics topologically conjugate to a shift of finite type (see the discussion by Smale in p. 147 in [108]). Dynamical systems in which there are transverse homoclinic connections are a common example (Theorem 5.3.5 in [45]). Furthermore in certain settings positive metric entropy implies the existence of shifts of finite type. One result along these lines is a theorem of Katok [53]. Let $h_{\text{top}}(f)$ denote the topological entropy of a map f and $h_\mu(f)$ denote metric entropy with respect to an invariant measure μ .

Theorem 1 (Katok) Suppose $T: M \rightarrow M$ is a $C^{1+\epsilon}$ diffeomorphism of a closed manifold and μ is an invariant measure with positive metric entropy (i. e. $h_\mu(T) > 0$). Then for any $0 < \epsilon < h_\mu(T)$ there exists an invariant set Λ topologically conjugate to a transitive shift of finite type with $h_{\text{top}}(T|_\Lambda) > h_\mu(T) - \epsilon$.

More Examples of Subshifts

We consider some further examples of systems that are given by the shift transformation on a subset of the set of (usually doubly-infinite) sequences on a finite alphabet, usually $\{0, 1\}$. Associated with each subshift is its *language*, the set of all finite blocks seen in all sequences in the subshift. These languages are *extractive* (or *factorial*) (every subword of a word in the language is also in the language) and *insertive* (or *extendable*) (every word in the language extends on both sides to longer words in the language). In fact these two properties characterize the languages (subsets of the set of finite-length words on an alphabet) associated with subshifts.

Prouhet–Thue–Morse An interesting (and often rediscovered) element of $\{0, 1\}^{\mathbb{Z}^+}$ is produced as follows. Start with 0 and at each stage write down the opposite ($0' = 1, 1' = 0$) or mirror image of what is available so far. Or, repeatedly apply the *substitution* $0 \rightarrow 01, 1 \rightarrow 10$.

0
0 1
0 1 10
0 1 10 0110
⋮

The n th entry is the sum, mod 2, of the digits in the dyadic expansion of n . Using Keane’s *block multiplication* [55] according to which if B is a block, $B \times 0 = B, B \times 1 = B'$, and $B \times (\omega_1 \dots \omega_n) = (B \times \omega_1) \dots (B \times \omega_n)$, we may also obtain this sequence as

$$0 \times 01 \times 01 \times 01 \times \dots$$

The orbit closure of this sequence is uniquely ergodic (there is a unique shift-invariant Borel probability measure, which is then necessarily ergodic). It is isomorphic to a skew product (see Sect. “[Skew Products](#)”) over the von Neumann-Kakutani adding machine, or odometer (see Sect. “[Adding Machines](#)”). *Generalized Morse systems*, that is, orbit closures of sequences like $0 \times 001 \times 001 \times 001 \times \dots$, are also isomorphic to skew products over compact group rotations.

Chacon System This is the orbit closure of the sequence generated by the substitution $0 \rightarrow 0010, 1 \rightarrow 1$. It is uniquely ergodic and is one of the first systems shown to be weakly mixing but not strongly mixing. It is *prime* (has no nontrivial factors) [30], and in fact has *minimal self joinings* [31]. It also has a nice description by means of cutting up the unit interval and stacking the pieces, using spacers (see Sect. “Cutting and Stacking”). This system has singular spectrum. It is not known whether or not its Cartesian square is loosely Bernoulli.

Sturmian Systems Take the orbit closure of the sequence $\omega_n = \chi_{[1-\alpha, 1)}(n\alpha)$, where α is irrational. This is a uniquely ergodic system that is isomorphic to rotation by α on the unit interval. These systems have *minimal complexity* in the sense that the number of n -blocks grows as slowly as possible $(n + 1)$ [29].

Toeplitz Systems A bi-infinite sequence (x_i) is a Toeplitz sequence if the set of integers can be decomposed into arithmetic progressions such that each x_i is constant on each arithmetic progression. A shift space X is a Toeplitz shift if it is the closure of the orbit of a Toeplitz sequence. It is possible to construct Toeplitz shifts which are uniquely ergodic and isomorphic to a rotation on a compact abelian group [34].

Sofic Systems These are images of SFT's under continuous factor maps (finite codes, or block maps). They correspond to *regular languages*—languages whose words are recognizable by finite automata. These are the same as the languages defined by *regular expressions*—finite expressions built up from \emptyset (empty set), ϵ (empty word), $+$ (union of two languages), \cdot (all concatenations of words from two languages), and $*$ (all finite concatenations of elements). They also have the characteristic property that the family of all *follower sets* of all blocks seen in the system is a finite family; similarly for *predecessor sets*. These are also generated by *phase-structure grammars* which are *linear*, in the sense that every production is either of the form $A \rightarrow Bw$ or $A \rightarrow w$, where A and B are variables and w is a string of terminals (symbols in the alphabet of the language).

(A *phase-structure grammar* consists of alphabets V of *variables* and A of *terminals*, a set of *productions*, which is finite set of pairs of words (α, w) , usually written $\alpha \rightarrow w$, of words on $V \cup A$, and a *start symbol* S . The associated language consists of all words on the alphabet A of terminals which can be made by starting with S and applying a finite sequence of productions.)

Sofic systems typically support many invariant measures (for example they have many periodic points) but topologically transitive ones (those with a dense orbit) have a unique measures of maximal entropy (see [65]).

Context-free Systems These are generated by phase-structure grammars in which all productions are of the form $A \rightarrow w$, where A is a variable and w is a string of variables and terminals.

Coded Systems These are systems all of whose blocks are concatenations of some (finite or infinite) list of blocks. These are the same as the closures of increasing sequences of SFT's [60]. Alternatively, they are the closures of the images under finite edge-labelings of irreducible countable-state topological Markov chains. They need not be context-free. Squarefree languages are not coded, in fact do not contain any coded systems of positive entropy. See [13, 14, 15].

Smooth Expanding Interval Maps

Take $X = [0, 1)(\text{mod } 1)$, $m \in \mathbb{N}$, $m > 1$ and define

$$T(x) = (mx).$$

Then T preserves Lebesgue measure μ (recall that T preserves μ if $\mu(T^{-1}A) = \mu(A)$ for all $A \in \mathcal{B}$). Furthermore it can be shown that T is ergodic.

This simple map exemplifies many of the characteristics of systems with some degree of hyperbolicity. It is isomorphic to a Bernoulli shift. The map has positive topological entropy and exponential divergence of nearby orbits, and Hölder functions have exponential decay of correlations and satisfy the central limit theorem and other strong statistical properties [20].

If $m = 2$ the system is isomorphic to a model of tossing a fair coin, which is a common example of randomness. To see this let $\mathcal{P} = \{P_0 = [0, 1/2), P_1 = [1/2, 1]\}$ be a partition of $[0, 1]$ into two subintervals. We code the orbit under T of any point $x \in [0, 1)$ by 0's and 1's by letting $x_k = i$ if $T^k x \in P_i$, $k = 0, 1, 2, \dots$. The map $\phi: X \rightarrow \{0, 1\}^{\mathbb{N}}$ which associates a point x to its *itinerary* in this way is a measure-preserving map from (X, μ) to $\{0, 1\}^{\mathbb{N}}$ equipped with the Bernoulli measure from $p_0 = p_1 = \frac{1}{2}$. The map ϕ satisfies $\phi \circ T = \sigma \circ \phi$, μ a.e. and is invertible a.e., hence is an isomorphism. Furthermore, reading the binary expansion of x is equivalent to following the orbit of x under T and noting which element of the partition \mathcal{P} is entered at each time. Borel's theorem on normal numbers (base m) may be seen as a special case of the Birkhoff Ergodic Theorem in this setting.

Piecewise C^2 Expanding Maps The main statistical features of the examples in Sect. “Smooth Expanding Interval Maps” generalize to a broader class of expanding maps of the interval. For example:

Let $X = [0, 1]$ and let $\mathcal{P} = \{I_1, \dots, I_n\}$ ($n \geq 2$) be a partition of X into intervals (closed, half-open or open) such that $I_i \cap I_j = \emptyset$ if $i \neq j$. Let I_i° denote the interior of I_i . Suppose $T: X \rightarrow X$ satisfies:

- (a) For each $i = 1, \dots, n$, $T|_{I_i}$ has a C^2 extension to the closure \bar{I}_i of I_i and $|T'(x)| \geq \alpha > 1$ for all $x \in I_i^\circ$.
- (b) $T(I_j) = \cup_{i \in P_j} I_i$ Lebesgue a.e. for some non-empty subset $P_j \subset \{1, \dots, n\}$.
- (c) For each I_j there exists n_j such that $T^{n_j}(I_j) = [0, 1]$ Lebesgue a.e.

Then T has an invariant measure μ which is absolutely continuous with respect to Lebesgue measure m , and there exists $C > 0$ such that $\frac{1}{C} \leq \frac{d\mu}{dm} \leq C$. Furthermore T is ergodic with respect to μ and displays the same statistical properties listed above for the C^2 expanding maps [20]. (See the “Folklore Theorem” in the article on ► [Measure Preserving Systems](#).)

More Interval Maps

Continued Fraction Map This is the map $T: [0, 1] \rightarrow [0, 1]$ given by $Tx = 1/x \bmod 1$, and it corresponds to the shift $[0; a_1, a_2, \dots] \rightarrow [0; a_2, a_3, \dots]$ on the continued fraction expansions of points in the unit interval (a map on $\mathbb{N}^{\mathbb{N}}$). It preserves a unique finite measure equivalent to Lebesgue measure, the *Gauss measure* $dx/(\log 2)(1+x)$. It is Bernoulli with entropy $\pi^2/6 \log 2$ (in fact the natural partition into intervals is a weak Bernoulli generator, for the definition and details see [91]). By using the Ergodic Theorem, Khintchine and Lévy showed that

$$(a_1 \cdots a_n)^{1/n} \rightarrow \prod_{k=1}^{\infty} \left[1 + \frac{1}{k^2 + 2k} \right]^{\frac{\log k}{\log 2}} \text{ a.e. as } n \rightarrow \infty;$$

$$\text{if } [0; a_1, \dots, a_n] = \frac{p_n}{q_n}, \text{ then } \frac{1}{n} \log q_n \rightarrow \frac{\pi^2}{12 \log 2} \text{ a.e.};$$

$$\frac{1}{n} \log \left| x - \frac{p_n(x)}{q_n(x)} \right| \rightarrow \frac{\pi^2}{6 \log 2} \text{ a.e.};$$

and if m is Lebesgue measure (or any equivalent measure) and μ is Gauss measure, then for each interval I , $m(T^{-n}I) \rightarrow \mu(I)$, in fact exponentially fast, with a best constant 0.30366... see [10,75].

The Farey Map This is the map $U: [0, 1] \rightarrow [0, 1]$ given by $Ux = x/(1-x)$ if $0 \leq x \leq 1/2$, $Ux = (1-x)/x$

if $1/2 \leq x \leq 1$. It is ergodic for the σ -finite infinite measure dx/x (Rényi and Parry). It is also ergodic for the *Minkowski measure* d , which is a measure of maximal entropy. This map corresponds to the shift on the *Farey tree* of rational numbers which provide the *intermediate convergence* (best one-sided) as well as the continued fraction (best two-sided) rational approximations to irrational numbers. See [61,62].

f -Expansions Generalizing the continued fraction map, let $f: [0, 1] \rightarrow [0, 1]$ and let $\{I_n\}$ be a finite or infinite partition of $[0, 1]$ into subintervals. We study the map f by coding itineraries with respect to the partition $\{I_n\}$. For many examples, absolutely continuous (with respect to Lebesgue measure) invariant measures can be found and their dynamical properties determined. See [104].

β -Shifts This is the special case of f -expansions when $f(x) = \beta x \bmod 1$ for some fixed $\beta > 1$. This map of the interval is called the β -transformation. With a proper choice of partition, it is represented by the shift on a certain subshift of the set of all sequences on the alphabet $D = \{0, 1, \dots, \lfloor \beta \rfloor\}$, called the β -shift. A point x is expanded as an infinite series in negative powers of β with coefficients from this set; $d_\beta(x)_n = \lfloor \beta f^n(x) \rfloor$. (By convention terminating expansions are replaced by eventually periodic ones.) A one-sided sequence on the alphabet D is in the β -shift if and only if all of its shifts are lexicographically less than or equal to the expansion $d_\beta(1)$ of 1 base β . A one-sided sequence on the alphabet D is the valid expansion of 1 for some β if and only if it lexicographically dominates all its shifts. These were first studied by Bissinger [11], Everett [35], Rényi [93] and Parry [84,85]; there are good summaries by Bertrand-Mathis [9] and Blanchard [12].

For $\beta = \frac{1+\sqrt{5}}{2}$, $d_\beta(1) = 10101010\dots$.

For $\beta = \frac{3}{2}$, $d_\beta(1) = 101000001\dots$ (not eventually periodic).

Every β -shift is coded.

The topological entropy of a β -shift is $\log \beta$. There is a unique measure of maximal entropy $\log \beta$.

A β -shift is a shift of finite type if and only if the β -expansion of 1 is finite. It is sofic if and only if the expansion of 1 is eventually periodic. If β is a Pisot–Vijayaragavhan number (algebraic integer all of whose conjugates have modulus less than 1), then the β -shift is sofic. If the β -shift is sofic, then β is a Perron number (algebraic integer of maximum modulus among its conjugates).

Theorem 2 (Parry [86]) Every strongly transitive (for every nonempty open set U , $\cup_{n \geq 0} T^n U = X$) piecewise mono-

tonic map on $[0, 1]$ is topologically conjugate to a β -transformation.

Gaussian Systems

Consider a real-valued stationary process $\{f_k: -\infty < k < \infty\}$ on a probability space (Ω, \mathcal{F}, P) . The process (and the associated measure-preserving system consisting of the shift and a shift-invariant measure on $\mathbb{R}^{\mathbb{Z}}$) is called *Gaussian* if for each $d \geq 1$, any d of the f_k form an \mathbb{R}^d -valued Gaussian random variable on Ω : this means that with $E(f_k) = m$ for all k and

$$A_{ij} = \int_{\Omega} (f_{k_i} - m)(f_{k_j} - m) dP = C(k_i - k_j) \quad \text{for } i, j = 1, \dots, d,$$

where $C(\cdot)$ is a function, for each Borel set $B \subset \mathbb{R}$,

$$\begin{aligned} P\{\omega: (f_{k_1}(\omega), \dots, f_{k_d}(\omega)) \in B\} \\ = \frac{1}{2\pi^{d/2} \sqrt{\det A}} \int_B \exp \left[-\frac{1}{2} (x - (m, \dots, m))^{\text{tr}} \right. \\ \left. \cdot A^{-1} (x - (m, \dots, m)) \right] dx_1 \cdots dx_d. \end{aligned}$$

where A is a matrix with entries (A_{ij}) . The function $C(k)$ is positive semidefinite and hence has an associated measure σ on $[0, 2\pi)$ such that

$$C(k) = \int_0^{2\pi} e^{ikt} d\sigma(t).$$

Theorem 3 (de la Rue [33]) *The Gaussian system is ergodic if and only if the “spectral measure” σ is continuous (i. e., nonatomic), in which case it is also weakly mixing. It is mixing if and only if $C(k) \rightarrow 0$ as $|k| \rightarrow \infty$. If σ is singular with respect to Lebesgue measure, then the entropy is 0; otherwise the entropy is infinite.*

For more details see [28].

Hamiltonian Systems

(This paragraph is from the article on [► Measure Preserving Systems](#).) Many systems that model physical situations can be studied by means of Hamilton’s equations. The state of the entire system at any time is specified by a vector $(q, p) \in \mathbb{R}^{2n}$, the *phase space*, with q listing the coordinates of the positions of all of the particles, and p listing the coordinates of their momenta. We assume there is a time-independent *Hamiltonian function* $H(q, p)$ such that the time development of the system satisfies *Hamilton’s equations*:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \quad \frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \quad i = 1, \dots, n. \quad (4)$$

Often in applications the Hamiltonian function is the sum of kinetic and potential energy:

$$H(q, p) = K(p) + U(q). \quad (5)$$

Solving these equations with initial state (q, p) for the system produces a flow $(q, p) \rightarrow T_t(q, p)$ in phase space which moves (q, p) to its position $T_t(q, p)t$ units of time later. According to *Liouville’s formula* Theorem 3.2 in [69], this flow preserves Lebesgue measure on \mathbb{R}^{2n} . Calculating dH/dt by means of the Chain Rule

$$\frac{dH}{dt} = \sum_i \left(\frac{\partial H}{\partial p_i} \frac{dp_i}{dt} + \frac{\partial H}{\partial q_i} \frac{dq_i}{dt} \right)$$

and using Hamilton’s equations shows that H is constant on orbits of the flow, and thus each set of constant energy $X(H_0) = \{(q, p): H(q, p) = H_0\}$ is an invariant set. There is a natural invariant measure on a constant energy set $X(H_0)$ for the restricted flow, namely the measure given by rescaling the volume element dS on $X(H_0)$ by the factor $1/||\nabla H||$.

Billiard Systems These form an important class of examples in ergodic theory and dynamical systems, motivated by natural questions in physics, particularly the behavior of gas models. Consider the motion of a particle inside a bounded region D in \mathbb{R}^d with piecewise smooth (C^1 at least) boundaries. In the case of planar billiards we have $d = 2$. The particle moves in a straight line with constant speed until it hits the boundary, at which point it undergoes a perfectly elastic collision with the angle of incidence equal to the angle of reflection and continues in a straight line until it next hits the boundary. It is usual to normalize and consider unit speed, as we do in this discussion for convenience. We take coordinates (x, v) given by the Euclidean coordinates in $x \in D$ together with a direction vector $v \in S^{d-1}$. A flow ϕ_t is defined with respect to Lebesgue almost every (x, v) by translating x a distance t defined by the direction vector v , taking account of reflections at boundaries. ϕ_t preserves a measure absolutely continuous with respect to Riemannian volume on (x, v) coordinates. The flow we have described is called a *billiard flow*. The corresponding *billiard map* is formed by taking the Poincaré map corresponding to the cross-section given by the boundary ∂D . We will describe the planar billiard map; the higher dimensional generalization is clear. The billiard map is a map $T: \partial D \rightarrow \partial D$, where ∂D is coordinatized by (s, θ) , $s \in [0, L]$, where L is the length of ∂D and $\theta \in (0, \pi)$ measures the angle that inward pointing vec-

tors make with the tangent line to ∂D at s . Given a point (s, θ) , the angle θ defines an oriented line $l(s, \theta)$ which intersects ∂D in two points s and s' . Reflecting l in the tangent line to ∂D at the point s' gives another oriented line passing through s' with angle θ' (measured with respect to the angular coordinate system based at s'). The billiard map is the map $T(s, \theta) = (s', \theta')$. T preserves a measure $\mu = \sin \theta ds \times d\theta$. The billiard flow may be modeled as a suspension flow over the billiard map (see Sect. “[Suspension Flows](#)”).

If the region D is a polygon in the plane (or polyhedron in \mathbb{R}^d), then ∂D consists of the faces of the polyhedron. The dynamical behavior of the billiard map or flow in regions with only flat (non-curved) boundaries is quite different to that of billiard flows or maps in regions D with strictly convex or strictly concave boundaries. The topological entropy of a flat polygonal billiard is zero. Research interest focuses on the existence and density of periodic or transitive orbits. It is known that if all the angles between sides are rational multiples of π then there are periodic orbits [17,74,112] and they are dense in the phase space [16]. It is also known that a residual set of polygonal billiards are topologically transitive and ergodic [58,117].

On the other hand, billiard maps in which ∂D has strictly convex components are physical examples of non-uniformly hyperbolic systems (with singularities). The meaning of concave or convex varies in the literature. We will consider a billiard flow inside a circle to be a system with a strictly concave boundary, while a billiard flow on the torus from which a circle has been excised to be a billiard flow with strictly convex boundary.

The class of billiards with some strictly convex boundary components, sometimes called *dispersing billiards* or *Sinai billiards*, was introduced by Sinai [106] who proved many of their fundamental properties. Lazutkin [63] proved that planar billiards with generic strictly concave boundary are not ergodic. Nevertheless Bunimovich [22,23] produced a large of billiard systems, *Bunimovich billiards*, with strictly concave boundary segments (perhaps with some flat boundaries as well) which were ergodic and non-uniformly hyperbolic. For more details see [26,54,66,109]. We will discuss possibly the simplest example of a dispersing billiard, namely a toral billiard with a single convex obstacle. Take the torus T^2 and consider a single strictly convex subdomain S with C^∞ boundary. The domain of the billiard map is $[0, L) \times (0, \pi)$, where L is the length of ∂S . The measure $\sin(\theta)ds \times d\theta$ is preserved. If the curvature of ∂S is everywhere non-zero, then the billiard map T has positive topological entropy, periodic points are dense, and in fact the system is isomorphic to a Bernoulli shift [41].

KAM-Systems and Stably Non-Ergodic Behavior

A celebrated theorem of Kolmogorov, Arnold and Moser (the KAM theorem) implies that the set of ergodic area-preserving diffeomorphisms of a compact surface without boundary is not dense in the C^r topology for $r \geq 4$. This has important implications, in that there are natural systems in which ergodicity is not generic. The constraint of perturbing in the class of area-preserving diffeomorphisms is an appropriate imposition in many physical models. We will take the version of the KAM theorem as given in Theorem 5.1 in [69] (original references include [3,59,79]). An elliptic fixed point for an area-preserving diffeomorphism T of a surface M is called a *non-degenerate elliptic fixed point* if there is a local C^r , $r \geq 4$, change of coordinates h so that in polar coordinates

$$hTh^{-1}(r, \theta) = (r, \theta + \alpha_0 + \alpha_1 r) + F(r, \theta),$$

where all derivatives of F up to order 3 vanish, $\alpha_1 \neq 0$ and $\alpha_0 \neq 0, \frac{\pm\pi}{2}, \pi, \frac{\pm 2\pi}{3}$. A map of the form

$$\tau(r, \theta) = (r, \theta + \alpha_0 + \alpha_1 r),$$

where $\alpha_1 \neq 0$, is called a *twist map*. Note that a twist map leaves invariant the circle $r = k$, for any constant k , and rotates each invariant curve by a rigid rotation $\alpha_1 r$, the magnitude of the rotation depending upon r . With respect to two-dimensional Lebesgue measure a twist map is certainly not ergodic.

Theorem 4 Suppose T is a volume-preserving diffeomorphism of class C^r , $r \geq 4$, of a surface M . If x is a non-degenerate elliptic fixed point, then for every $\epsilon > 0$ there exists a neighborhood U_ϵ of x and a set $U_{0,\epsilon} \subset U_\epsilon$ with the properties:

- (a) $U_{0,\epsilon}$ is a union of T -invariant simple closed curves of class C^{r-1} containing x in their interior.
- (b) The restriction of T to each such invariant curve is topologically conjugate to an irrational rotation.
- (c) $m(U_\epsilon - U_{0,\epsilon}) \leq \epsilon m(U_\epsilon)$, where m is Lebesgue measure on M .

It is possible to prove the existence of a C^r volume preserving diffeomorphism ($r \geq 4$) with a non-degenerate elliptic fixed point and also show that if T possesses a non-degenerate elliptic fixed point then there is a neighborhood V of T in the C^r topology on volume-preserving diffeomorphisms such that each $T' \in V$ possesses a non-degenerate elliptic fixed point Chapter II, Sect. 6 in [69]. As a corollary we have

Corollary 1 *Let M be a compact surface without boundary and $\text{Diff}^r(M)$ the space of C^r area-preserving diffeomorphisms with the C^r topology. Then the set of $T \in \text{Diff}^r(M)$ which are ergodic with respect to the probability measure determined by normalized area is not dense in $\text{Diff}^r(M)$ for $r \geq 4$.*

Smooth Uniformly Hyperbolic Diffeomorphisms and Flows

Time series of measurements on deterministic dynamical systems sometimes display limit laws exhibited by independent identically distributed random variables, such as the central limit theorem, and also various mixing properties. The models of hyperbolicity we discuss in this section have played a key role in showing how this phenomenon of ‘chaotic behavior’ arises in deterministic dynamical systems. Hyperbolic sets and their associated dynamics have also been pivotal in studies of structural stability. A smooth system is C^r *structurally stable* if a small perturbation in the C^r topology gives rise to a system which is topologically conjugate to the original. When modeling a physical system it is desirable that slight changes in the modeling parameters do not greatly affect the qualitative or quantitative behavior of the ensemble of orbits considered as a whole. The orbit of a point may change drastically under perturbation (especially if the system has sensitive dependence on initial conditions) but the collection of all orbits should ideally be ‘similar’ to the original unperturbed system. In the latter case one would hope that statistical properties also vary only slightly under perturbation. Structural stability is one, quite strong, notion of stability. The conclusion of a body of work on structural stability is that a system is C^1 structurally stable if and only if it is uniformly hyperbolic and satisfies a technical assumption called strong transversality (see below for details).

Suppose M is a C^1 compact Riemannian manifold equipped with metric d and tangent space TM with norm $\|\cdot\|$. Suppose also that $U \subset M$ is a non-empty open subset and $T: U \rightarrow T(U)$ is a C^1 diffeomorphism. A compact T invariant set $\Lambda \subset U$ is called a *hyperbolic set* if there is a splitting of the tangent space $T_p M$ at each point $p \in \Lambda$ into two invariant subspaces, $T_p M = E^u(p) \oplus E^s(p)$, and a number $0 < \lambda < 1$ such that for $n \geq 0$

$$\|D_p T^n v\| \leq C \lambda^n \|v\| \text{ for } v \in E^s(p),$$

$$\|D_p T^{-n} v\| < C \lambda^n \|v\| \text{ for } v \in E^u(p).$$

The subspace E^u is called the *unstable* or *expanding subspace* and the subspace E^s the *stable* or *contracting subspace*. The stable and unstable subspaces may be integrated

to produce *stable* and *unstable manifolds*

$$W^s(p) = \{y: d(T^n p, T^n y) \rightarrow 0\} \text{ as } n \rightarrow \infty,$$

$$W^u(p) = \{y: d(T^{-n} p, T^{-n} y) \rightarrow 0\} \text{ as } n \rightarrow \infty.$$

The stable and unstable manifolds are immersions of Euclidean spaces of the same dimension as $E^s(p)$ and $E^u(p)$, respectively, and are of the same differentiability as T . Moreover, $T_p(W^s(p)) = E^s(p)$ and $T_p(W^u(p)) = E^u(p)$. It is also useful to define *local stable manifolds* and *local unstable manifolds* by

$$W_\epsilon^s(p) = \{y \in W^s(p): d(T^n p, T^n y) < \epsilon\} \text{ for all } n \geq 0,$$

$$W_\epsilon^u(p) = \{y \in W^u(p): d(T^{-n} p, T^{-n} y) < \epsilon\} \\ \text{for all } n \geq 0.$$

Finally we discuss the notion of strong transversality. We say a point x is *non-wandering* if for each open neighborhood U of x there exists an $n > 0$ such that $T^n(U) \cap U \neq \emptyset$. The NW set of non-wandering points is called the *non-wandering set*. We say a dynamical system has the *strong transversal property* if $W^s(x)$ intersects $W^u(y)$ transversely for each pair of points $x, y \in NW$. In the C^r , $r \geq 1$ topology Robbin [94], de Melo [78] and Robinson [95,96] proved that dynamical systems with the strong transversal property are structurally stable, and Robinson [97] in addition showed that strong transversality was also necessary. Mañé [70] showed that a C^1 structurally stable diffeomorphism must be uniformly hyperbolic and Hayashi [47] extended this to flows. Thus a C^1 diffeomorphism or flow on a compact manifold is structurally stable if and only if it is uniformly hyperbolic and satisfies the strong transversality condition.

Geodesic Flow on Manifold of Negative Curvature The study of the geodesic flow on manifolds of negative sectional curvature by Hedlund and Hopf was pivotal to the development of the ergodic theory of hyperbolic systems. Suppose that M is a geodesically complete Riemannian manifold. Let $\gamma_{p,v}(t)$ be the geodesic with $\gamma_{p,v}(0) = p$ and $\dot{\gamma}_{p,v}(0) = v$, where $\dot{\gamma}_{p,v}$ denotes the derivative with respect to time t . The geodesic flow is a flow ϕ_t on the tangent bundle TM of M , $\phi_t: \mathbb{R} \times TM \rightarrow TM$, defined by

$$\phi_t(p, v) = (\gamma_{p,v}(t), \dot{\gamma}_{p,v}(t)).$$

where $(p, v) \in TM$. Since geodesics have constant speed, if $\|v\| = 1$ then $\|\dot{\gamma}_{p,v}(t)\| = 1$ for all t , and thus the unit tangent bundle $T^1 M = \{(p, v) \in TM: \|v\| = 1\}$ is preserved

under the geodesic flow. The geodesic flow and its restriction to the unit tangent bundle both preserve a volume form, Liouville measure. In 1934 Hedlund [48] proved that the geodesic flow on the unit tangent bundle of a surface of strictly negative constant sectional curvature is ergodic, and in 1939 Hopf [49] extended this result to manifolds of arbitrary dimension and strictly negative (not necessarily constant) curvature. Hopf's technique of proof of ergodicity (Hopf argument) was extremely influential and used the foliation of the tangent space into stable and unstable manifolds. For a clear exposition of this technique, and the property of absolute continuity of the foliations into stable and unstable manifolds, see [66]. The geodesic flow on manifolds of constant negative sectional curvature is an Anosov flow (see Sect. "Anosov Systems"). We remark that for surfaces sectional curvature is the same as Gaussian curvature. Recently the time-one map of the geodesic flow on the unit tangent bundle of a surface with constant negative curvature, which is a partially hyperbolic system (see Sect. "Partially Hyperbolic Dynamical Systems"), was shown to be stably ergodic [44], so the geodesic flow is still playing a major role in the development of ergodic theory.

Horocycle Flow All surfaces endowed with a Riemannian metric of constant negative curvature are quotients of the upper half-plane $\mathcal{H}^+ := \{x + iy \in \mathbb{C} : y > 0\}$ with the metric $ds^2 = \frac{dx^2 + dy^2}{y^2}$, whose sectional curvature is -1 . The orientation-preserving isometries of this metric are exactly the linear fractional (also known as Möbius) transformations:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \in SL(2, \mathbb{R}), \quad [z \in \mathcal{H}^+ \mapsto \frac{az + b}{cz + d} \in \mathcal{H}^+].$$

Since each matrix $\pm I$ corresponds to the identity transformation, we consider matrices in $PSL(2, \mathbb{R}) := SL(2, \mathbb{R})/\{\pm I\}$.

The unit tangent bundle, $S\mathcal{H}^+$, of the upper half-plane can be identified with $PSL(2, \mathbb{R})$. Then the geodesic flow corresponds to the transformations

$$t \in \mathbb{R} \mapsto \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix}$$

seen as acting on $PSL(2, \mathbb{R})$. The unstable foliation of an element $A \in PSL(2, \mathbb{R}) \cong S\mathcal{H}^+$ is given by

$$\begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} A, \quad t \in \mathbb{R},$$

and the flow along this foliation, given by

$$t \in \mathbb{R} \mapsto \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix},$$

is called the *horocycle flow*. Similarly for the flow induced on the unit tangent bundle of each quotient of the upper half-plane by a discrete group of linear fractional transformations.

The geodesic and horocycle flows acting on a (finite-volume) surface of constant negative curvature form the fundamental example of a transverse pair of actions. The geodesic flow often has many periodic orbits and many invariant measures, has positive entropy, and is in fact Bernoulli with respect to the natural measure [82], while the horocycle flow is often uniquely ergodic [39,72] and of entropy zero, although mixing of all orders [73]. See [46] for more details.

Markov Partitions and Coding If (X, T, \mathcal{B}, μ) is a dynamical system then a finite partition of X induces a coding of the orbits and a semi-conjugacy with a subshift on a symbol space (it may not of course be a full conjugacy). For hyperbolic systems a special class of partitions, *Markov partitions*, induce a conjugacy for the invariant dynamics to a subshift of finite type. A Markov partition \mathcal{P} for an invariant subset Λ of a diffeomorphism T of a compact manifold M is a finite collection of sets R_i , $1 \leq i \leq n$ called *rectangles*. The rectangles have the property, for some $\epsilon > 0$, if $x, y \in R_i$ then $W_\epsilon^s(x) \cap W_\epsilon^u(y) \in R_i$. This is sometimes described as being closed under *local product structure*. We let $W^u(x, R_i)$ denote $W_\epsilon^u(x) \cap R_i$ and $W^s(x, R_i)$ denote $W_\epsilon^s(x) \cap R_i$. Furthermore we require for all i, j :

- (1) Each R_i is the closure of its interior.
- (2) $\Lambda \subset \cup_i R_i$
- (3) $R_i \cap R_j = \partial R_i \cap \partial R_j$ if $i \neq j$
- (4) if $x \in R_i^o$ and $T(x) \in R_j^o$ then $W^u(T(x), R_j) \subset T(W^u(x, R_i))$ and $W^s(x, R_i) \subset T^{-1}(W^s(T(x), R_j))$

Anosov Systems An *Anosov diffeomorphism* [2] is a uniformly hyperbolic system in which the entire manifold is a hyperbolic set. Thus an Anosov diffeomorphism is a C^1 diffeomorphism T of M with a DT -invariant splitting (which is a continuous splitting) of the tangent space $TM(x)$ at each point p into a disjoint sum

$$T_p M = E^u(p) \oplus E^s(p)$$

and there exist constants $0 < \lambda < 1$, constant $C > 0$ such that $\|DT^n v\| < C\lambda^n \|v\|$ for all $v \in E^s(p)$ and $\|DT^{-n} w\| \leq C\lambda^n \|w\|$ for all $w \in E^u(p)$.

A similar definition holds for Anosov flows $\phi: \mathbb{R} \times M \rightarrow M$. A flow is *Anosov* if there is a splitting of the tangent bundle into flow-invariant subspaces E^u, E^s, E^c so

$D_p \phi_t E_p^s = E_{\phi_t(p)}^s$, $D_p \phi_t E_p^u = E_{\phi_t(p)}^u$ and $D_p \phi_t E_p^c = E_{\phi_t(p)}^c$, and at each point $p \in M$

$$T_p M = E_p^s \oplus E_p^u \oplus E_p^c$$

$$\|(D_p \phi_t)v\| < C\lambda^t \|v\| \text{ for } v \in E^s(p)$$

$$\|(D_p \phi_{-t})v\| < C\lambda^t \|v\| \text{ for } v \in E^u(p)$$

for some $0 < \lambda < 1$. The tangent to the flow direction $E^c(p)$ is a neutral direction:

$$\|(D_p \phi_t)v\| = \|v\| \text{ for } v \in E^c(p).$$

Anosov proved that Anosov flows and diffeomorphisms which preserve a volume form are ergodic [2] and are also structurally stable. Sinai [105] constructed Markov partitions for Anosov diffeomorphisms and hence coded trajectories via a subshift of finite type. Using ideas from statistical physics in [107] Sinai constructed Gibbs measures for Anosov systems. An SRB measure (see Sect. “Physically Relevant Measures and Strange Attractors”) is a type of Gibbs measure corresponding to the potential $-\log |\det(DT|_{E^u})|$ and is characterized by the property of absolutely continuous conditional measures on unstable manifolds.

The simplest examples of Anosov diffeomorphisms are perhaps the two-dimensional hyperbolic toral automorphisms (the $n > 2$ generalization is clear). Suppose A is a 2×2 matrix with integer entries

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

such that $\det(A) = 1$ and A has no eigenvalues of modulus 1. Then A defines a transformation of the two-dimensional torus $T^2 = S^1 \times S^1$ such that if $v \in T^2$,

$$v = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix},$$

then

$$Av = \begin{pmatrix} av_1 + bv_2 \\ cv_1 + dv_2 \end{pmatrix}.$$

A preserves Lebesgue (or Haar) measure and is ergodic. A prominent example of such a matrix is

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix},$$

which is sometimes called the Arnold Cat Map. Each point with rational coordinates $(p_1/q_1, p_2/q_2)$ is periodic. There are two eigenvalues $1/\lambda < 1 < \lambda = (3 + \sqrt{5})/2$ with orthogonal eigenvectors, and the projections of the eigenspaces from \mathbb{R}^2 to T^2 are the stable and unstable subspaces.

Axiom A Systems In the case of Anosov diffeomorphisms the splitting into contracting and expanding bundles holds on the entire phase space M . A system $T: M \rightarrow M$ is an *Axiom A system* if the non-wandering set NW is a hyperbolic set and periodic points are dense in the non-wandering set. $NW \subset M$ may have Lebesgue measure zero. A set $\Lambda \subset M$ is *locally maximal* if there exists an open set U such that $\Lambda = \bigcap_{n \in \mathbb{Z}} T^n(U)$. The solenoid and horseshoe discussed below are examples of locally maximal sets. Bowen [18] constructed Markov partitions for Axiom A diffeomorphisms. Ruelle imported ideas from statistical physics, in particular the idea of an equilibrium state and the variational principle, to the study of Axiom A systems (see [101,102]). This work extended the notion of Gibbs measure and other ideas from statistical mechanics, introduced by Sinai for Anosov systems [107], into Axiom A systems.

One achievement of the Axiom A program was the Smale Decomposition Theorem, which breaks the dynamics of Axiom A systems into locally maximal sets and describes the dynamics on each [18,19,108].

Theorem 5 (Spectral Decomposition Theorem) *If T is Axiom A then there is a unique decomposition of the non-wandering set NW of T*

$$NW = \Lambda_1 \cup \dots \cup \Lambda_k$$

as a disjoint union of closed, invariant, locally maximal hyperbolic sets Λ_i such that T is transitive on each Λ_i . Furthermore each Λ_i may be further decomposed into a disjoint union of closed sets Λ_i^j , $j = 1, \dots, n_i$ such that T^{n_i} is topologically mixing on each Λ_i^j and T cyclically permutes the Λ_i^j .

Horseshoe Maps This type of map was introduced by Steven Smale in the 1960's and has played a pivotal role in the development of dynamical systems theory. It is perhaps the canonical example of an Axiom A system [108] and is conjugate to a full shift on 2 symbols. Let S be a unit square in \mathbb{R}^2 and let T be a diffeomorphism of S onto its image such that $S \cap T(S)$ consists of two disjoint horizontal strips S_0 and S_1 . Think of stretching S uniformly in the horizontal direction and contracting uniformly in the vertical direction to form a long thin rectangle, and then bending the rectangle into the shape of a horseshoe and laying the straight legs of the horseshoe back on the unit square S . This transformation may be realized by a diffeomorphism and we may also require that T restricted to $T^{-1}S_i$, $i = 0, 1$, acts as a linear map. The restriction of T to the maximal invariant set $H = \bigcap_{i=-\infty}^{\infty} T^i S$ is a Smale horseshoe map. H is a Cantor set, the product of a Cantor

set in the horizontal direction and a Cantor set in the vertical direction. The conjugacy with the shift on two symbols is realized by mapping $x \in H$ to its itinerary with respect to the sets S_0 and S_1 under powers of T (positive and negative powers).

Solenoids The solenoid is defined on the solid torus X in \mathbb{R}^3 which we coordinatize as a circle of two-dimensional solid disks, so that

$$X = \{(\theta, z): \theta \in [0, 1) \text{ and } |z| \leq 1, z \in \mathbb{C}\}$$

The transformation $T: X \rightarrow X$ is given by

$$T(\theta, z) = \left(2\theta \pmod{1}, \frac{1}{4}z + \frac{1}{2}e^{2\pi i\theta}\right)$$

Geometrically the transformation stretches the torus to twice its length, shrinks its diameter by a factor of 4, then twists it and doubles it over, placing the resultant object without self-intersection back inside the original solid torus. $T(X)$ intersects each disk $D_c = \{(\theta, z): \theta = c\}$ in two smaller disks of $\frac{1}{4}$ the diameter. The transformation T contracts volume by a factor of 2 upon each application, yet there is expansion in the θ direction ($\theta \rightarrow 2\theta$). The solenoid $A = \bigcap_{n \geq 0} T^n(X)$ has zero Lebesgue measure, is T -invariant and is (locally) topologically a line segment cross a two-dimensional Cantor set (A intersects each disk D_c in a Cantor set). The set A is an *attractor*, in that all points inside X limit under iteration by T upon A . $T: A \rightarrow A$ is an Axiom A system.

Partially Hyperbolic Dynamical Systems

Partially hyperbolic dynamical systems are a generalization of uniformly hyperbolic systems in that an invariant central direction is allowed but the contraction in the central direction is strictly weaker than the contraction in the contracting direction and the expansion in the central direction is weaker than the expansion in the expanding direction. More precisely, suppose M is a C^1 compact (adapted) Riemannian manifold equipped with metric d and tangent space TM with norm $\|\cdot\|$. A C^1 diffeomorphism T of M is a partially hyperbolic diffeomorphism if there is a nontrivial continuous DT invariant splitting of the tangent space $T_p M$ at each point p into a disjoint sum

$$T_p M = E^u(p) \oplus E^c(p) \oplus E^s(p)$$

and continuous positive functions $m, M, \tilde{\gamma}, \gamma$ such that

- E^s is contracted:
if $v^s \in E^s(x) \setminus \{0\}$ then $\frac{\|D_p T^n v^s\|}{\|v^s\|} \leq m(p) < 1$;
- E^u is expanded:
if $v^u \in E^u(x) \setminus \{0\}$ then $\frac{\|D_p T^n v^u\|}{\|v^u\|} \geq M(p) > 1$;

- E^c is uniformly dominated by E^u and E^s :
if $v^c \in E^c(x) \setminus \{0\}$ then there are numbers $\tilde{\gamma}(p), \gamma(p)$ such that $m(p) < \tilde{\gamma}(p) \leq \frac{\|D_p T v^c\|}{\|v^c\|} \leq \gamma(p) < M(p)$.

The notion of partial hyperbolicity was introduced by Brin and Pesin [21] who proved existence and properties, including absolute continuity, of invariant foliations in this setting. There has been intense recent interest in partially hyperbolic systems primarily because significant progress has been made in establishing that certain volume-preserving partially hyperbolic systems are ‘stably ergodic’—that is, they are ergodic and under small (C^r topology) volume-preserving perturbations remain ergodic. This phenomenon had hitherto been restricted to uniformly hyperbolic systems. For recent developments, and precise statements, on stable ergodicity of partially hyperbolic systems see [24,92].

Compact Group Extensions of Uniformly Hyperbolic Systems A natural example of a partially hyperbolic system is given by a compact group extension of an Anosov diffeomorphism. If the following terms are not familiar see Sect. “Constructions” on standard constructions. Suppose that (T, M, μ) is an Anosov diffeomorphism, G is a compact connected Lie group and $h: M \rightarrow G$ is a differentiable map. The skew product $F: M \times G \rightarrow M \times G$ given by

$$F(x, g) = (Tx, h(x)g)$$

has a central direction in its tangent space corresponding to the Lie algebra LG of G (as a group element h acts isometrically on G so there is no expansion or contraction) and uniformly expanding and contracting bundles corresponding to those of the tangent space of $T: M \rightarrow M$. Thus $T(M \times G) = E^u \oplus LG \oplus E^s$.

Time-One Maps of Anosov Flows Another natural context in which partial hyperbolicity arises is in time-one maps of uniformly hyperbolic flows. Suppose $\phi_t: \mathbb{R} \times M \rightarrow M$ is an Anosov flow. The diffeomorphism $\phi_1: M \rightarrow M$ is a partially hyperbolic diffeomorphism with central direction given by the flow direction. There is no expansion or contraction in the central direction.

Non-Uniformly Hyperbolic Systems

The assumption of uniform hyperbolicity is quite restrictive and few ‘chaotic systems’ found in applications are likely to exhibit uniform hyperbolicity. A natural weakening of this assumption, and one that is non-trivial and

greatly extends the applicability of the theory, is to require the hyperbolic splitting (no longer uniform) to hold only at almost every point of phase space. A systematic theory was built by Pesin [89,90] on the assumption that the system has non-zero Lyapunov exponents μ almost everywhere where μ is a Lebesgue equivalent invariant probability measure. Recall that a number λ is a *Lyapunov exponent* for $p \in M$ if $\|D_p T^n v\| \sim e^{\lambda n}$ for some unit vector $v \in T_p M$. Oseledec's theorem [83] (see also p. 232 in [113]), which is also called the Multiplicative Ergodic Theorem, implies that if T is a C^1 diffeomorphism of M then for any T -invariant ergodic measure μ almost every point has well-defined Lyapunov exponents. One of the highlights of Pesin theory is the following structure theorem: If $T: M \rightarrow M$ is a $C^{1+\epsilon}$ diffeomorphism with a T -invariant Lebesgue equivalent Borel measure μ such that T has non-zero Lyapunov exponents with respect to μ then T has at most a countable number of ergodic components $\{C_i\}$ on each of which the restriction of T is either Bernoulli or Bernoulli times a rotation (by which we mean the support of $\mu_i = \mu|_{C_i}$ consists of a finite number n_i of sets $\{S_1^i, \dots, S_{n_i}^i\}$ cyclically permuted and T^{n_i} is Bernoulli when restricted to each S_j^i) [90,114]. This structure theorem has been generalized to SRB measures with non-zero Lyapunov exponents [64,90].

Physically Relevant Measures and Strange Attractors

(This paragraph is from the article on [► Measure Preserving Systems](#).) For Hamiltonian systems and other volume-preserving systems it is natural to consider ergodicity (and other statistical properties) of the system with respect to Lebesgue measure. In dissipative systems a measure equivalent to Lebesgue may not be invariant (for example the solenoid). Nevertheless Lebesgue measure has a distinguished role since sampling by experimenters is done with respect to Lebesgue measure. The idea of a physically relevant measure μ is that it determines the statistical behavior of a positive Lebesgue measure set of orbits, even though the support of μ may have zero Lebesgue measure. An example of such a situation in the uniformly hyperbolic setting is the solenoid Λ , where the attracting set Λ has Lebesgue measure zero and is (locally) topologically the product of a two-dimensional Cantor set and a line segment. Nevertheless Λ determines the behavior of all points in a solid torus in \mathbb{R}^3 . More generally, suppose that $T: M \rightarrow M$ is a diffeomorphism on a compact Riemannian manifold and that m is a version of Lebesgue measure on M , given by a smooth volume form. Although Lebesgue measure m is a distinguished physically relevant measure, m may not be invariant under T , and the system may even

be volume contracting in the sense that $m(T^n A) \rightarrow 0$ for all measurable sets A . Nevertheless an experimenter might observe long-term “chaotic” behavior whenever the state of the system gets close to some compact invariant set X which attracts a positive m -measure of orbits in the sense that these orbits limit on X . Possibly $m(X) = 0$, so that X is effectively invisible to the observer except through its effects on orbits not contained in X . The dynamics of T restricted to X can in fact be quite complicated—maybe a full shift, or a shift of finite type, or some other complicated topological dynamical system. Suppose there is a T -invariant measure μ supported on X such that for all continuous functions $\phi: M \rightarrow \mathbb{R}$

$$\frac{1}{n} \sum_{k=0}^{n-1} \phi \circ T^k(x) \rightarrow \int_X \phi \, d\mu, \quad (6)$$

for a positive m -measure of points $x \in M$. Then the long-term equilibrium dynamics of an observable set of points $x \in M$ (i.e. a set of points of positive m measure) is described by (X, T, μ) . In this situation μ is described as a *physical measure*. There has been a great deal of research on the properties of systems with attractors supporting physical measures.

In the dissipative non-uniformly hyperbolic setting the theory of ‘physically relevant’ measures is best developed in the theory of SRB (for Sinai, Ruelle and Bowen) measures. These dynamically invariant measures may be supported on a set of Lebesgue measure zero yet determine the asymptotic behavior of points in a set of positive Lebesgue measure.

If T is a diffeomorphism of M and μ is a T -invariant Borel probability measure with positive Lyapunov exponents which may be integrated to unstable manifolds, then we call μ an *SRB measure* if the conditional measure μ induces on the unstable manifolds is absolutely continuous with respect to the Riemannian volume element on these manifolds. The reason for this definition is technical but is motivated by the following observations. Suppose that the diffeomorphism has no zero Lyapunov exponents with respect to μ . Since T is a diffeomorphism, this implies T has negative Lyapunov exponents as well as positive Lyapunov exponents and corresponding local stable manifolds as well as local unstable manifolds. Suppose that a T -invariant set A consists of a union of unstable manifolds and is the support of an ergodic SRB measure μ and that $\phi: M \rightarrow \mathbb{R}$ is a continuous function. Since μ has absolutely continuous conditional measures on unstable manifolds with respect to conditional Lebesgue measure on the unstable manifolds, almost every point x in the

union of unstable manifolds U satisfies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=0}^{n-1} \phi \circ T^j(x) = \int \phi \, d\mu. \quad (7)$$

If $y \in W_\epsilon^s(x)$ for such an $x \in U$ then $d(T^n x, T^n y) \rightarrow 0$ and hence (7) implies

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{j=0}^{n-1} \phi \circ T^j(y) = \int \phi \, d\mu.$$

Furthermore, if the holonomy between unstable manifolds defined by sliding along stable manifolds is absolutely continuous (takes sets of zero Lebesgue measure on W^u to sets of zero Lebesgue measure on W^u), there is a positive Lebesgue measure of points (namely an unstable manifold and the union of stable manifolds through it) satisfying (7). Thus an SRB measure with absolutely continuous holonomy maps along stable manifolds is a physically relevant measure. If the stable foliation possesses this property it is called *absolutely continuous*. An Axiom A attractor for a C^2 diffeomorphism is an example of an SRB attractor [19,101,102,107]. The examples we have given of SRB measures and attractors and measures have been uniformly hyperbolic.

Recently much progress has been made in understanding the statistical properties of non-uniformly hyperbolic systems by using a tower (see Sect. “[Induced Transformations](#)”) to construct SRB measures. We refer to Young’s original papers [115,116], the book by Baladi [5] and to [114] for a recent survey on SRB measures in the non-uniformly hyperbolic setting.

Unimodal Maps Unimodal Maps of an interval are simple examples of non-uniformly hyperbolic dynamical systems that have played an important role in the development of dynamical systems theory. Suppose $I \subset \mathbb{R}$ is an interval; for simplicity we take $I = [0, 1]$. A *unimodal map* is a map $T: [0, 1] \rightarrow [0, 1]$ such that there exists a point $0 < c < 1$ and

- T is C^2 ;
- $T'(x) > 0$ for $x < c$, $T'(x) < 0$ for $x > c$;
- $T'(c) = 0$.

Such a map is clearly not uniformly expanding, as $|T'(x)| < 1$ for points in a neighborhood of c . The family of maps $T_\mu(x) = \mu x(1-x)$, $0 < \mu \leq 4$, is a family of unimodal maps with $c = 1/2$ and $T_2(1/2) = 1/2$, $T_4(1/2) = 1$.

We could have taken the interval I to be $[-1, 1]$ or indeed any interval with an obvious modification of the definition above. A well-studied family of unimodal maps

in this setting is the *logistic family* $f_a: [-1, 1] \rightarrow [-1, 1]$, $f_a(x) = 1 - ax^2$, $a \in (0, 2]$. The families f_a and T_μ are equivalent under a smooth coordinate change, so statements about one family may be translated into statements about the other.

Unimodal maps are studied because of the insights they offer into transitions from regular or periodic to chaotic behavior as a parameter (e.g. μ or a) is varied, the existence of absolutely continuous measures, and rates of decay of correlations of regular observations for non-uniformly hyperbolic systems.

Results of Jakobson [52] and Benedicks and Carleson [6] implies that in the case of the logistic family there is a positive Lebesgue measure set of a such that f_a has an absolutely continuous ergodic invariant measure μ_a . It has been shown by Keller and Nowicki [57] (see also Young [116]) that if f_a is mixing with respect to μ_a then the decay of correlations for Lipschitz observations on I is exponential. It is also known that the set of a such that f_a is mixing with respect to μ_a has positive Lebesgue measure. There is a well-developed theory concerning bifurcations the maps T_μ undergo as μ varies [27]. We briefly describe the period-doubling route to chaos in the family $T_\lambda(x) = \lambda x(1-x)$. For a nice account see [46]. We let c_λ denote the fixed point $\frac{\lambda-1}{\lambda}$. For $3 < \lambda \leq 1 + \sqrt{6}$, all points in $[0, 1]$ except for $0, c_\lambda$ and their preimages are attracted to a unique periodic orbit $O(p_\lambda)$ of period 2. There is a monotone sequence of parameter values λ_n ($\lambda_1 = 3$) such that for $\lambda_n < \lambda \leq \lambda_{n+1}$, T_λ has a unique attracting periodic orbit $O(\lambda_n)$ of period 2^n and for each $k = 1, 2, \dots, n-1$ a unique repelling orbit of period 2^k . All points in the interval $[0, 1]$ except for the repelling periodic orbits and their preimages are attracted to the attracting periodic orbit of period 2^n . At $\lambda = \lambda_n$ the periodic orbit $O(\lambda_n)$ undergoes a period-doubling bifurcation. Feigenbaum [36] found that the limit $\delta = \frac{\lambda_n - \lambda_{n-1}}{\lambda_{n+1} - \lambda_n} \sim 4.699 \dots$ exists and that in a wide class of unimodal maps this period-doubling cascade occurs and the differences between successive bifurcation parameters give the same limiting ratio, an example of *universality*. At the end of the period-doubling cascade at a parameter $\lambda_\infty \sim 3.569 \dots$, T_{λ_∞} has an invariant Cantor set C (the Feigenbaum attractor) which is topologically conjugate to the dyadic adding machine coexisting with isolated repelling orbits of period 2^n , $n = 0, 1, 2, \dots$. There is a unique repelling orbit of period 2^n for $n \geq 1$ along with two fixed points. The Cantor set C is the ω -limit set for all points that are not periodic or preimages of periodic orbits. C is the set of accumulation points of periodic orbits. Despite this picture of incredible complexity the topological entropy is zero for $\lambda \leq \lambda_\infty$. For $\lambda > \lambda_\infty$ the map T_λ

has positive topological entropy and infinitely many periodic orbits whose periods are not powers of 2. For each $\lambda \geq \lambda_\infty$, T_λ possesses an invariant Cantor set which is repelling for $\lambda > \lambda_\infty$. We say that T_λ is *hyperbolic* if there is only one attracting periodic orbit and the only recurrent sets are the attracting periodic orbit, repelling periodic orbits and possibly a repelling invariant Cantor set. It is known that the set of $\lambda \in [0, 4]$ for which T_λ is hyperbolic is open and dense [43]. Remarkably, by Jakobson's result [52] there is also a positive Lebesgue measure set of parameters λ for which T_λ has an absolutely continuous invariant measure μ_λ with a positive Lyapunov exponent.

Intermittent Maps Maps of the unit interval $T: [0, 1] \rightarrow [0, 1]$ which are expanding except at the point $x = 0$, where they are locally of form $T(x) \sim x + x^{1+\alpha}$, $\alpha > 0$, have been extensively studied both for the insights they give into rates of decay of correlations for non-uniformly hyperbolic systems (hyperbolicity is lost at the point $x = 0$, where the derivative is 1) and for their use as models of intermittent behavior in turbulence [71]. A fixed point where the derivative is 1 is sometimes called an *indifferent fixed point*. It is a model of *intermittency* in the sense that orbits close to 1 will stay close for many iterates (since the expansion is very weak there) and hence a time series of observations will be quite uniform for long periods of time before displaying chaotic type behavior after moving away from the indifferent fixed into that part of the domain where the map is uniformly expanding.

A particularly simple model [67] is provided by

$$T(x) = \begin{cases} x(1 + 2^\alpha x^\alpha) & \text{if } x \in [0, 1/2); \\ 2x - 1 & \text{if } x \in [1/2, 1]. \end{cases}$$

For $\alpha = 0$ the map is uniformly expanding and Lebesgue measure is invariant. In this case the rate of decay of correlations for Hölder observations is exponential. For $0 < \alpha < 1$ the map has an SRB measure μ_α with support the unit interval. For $\alpha \geq 1$ there are no absolutely continuous invariant probability measures though there are σ -finite absolutely continuous measures. Upper and lower polynomial bounds on the rate of decay of observations on such maps have been given as a function of $0 < \alpha < 1$ and the regularity of the observable. For details see [51, 67, 103].

Hénon Diffeomorphisms The Hénon family of diffeomorphisms was introduced and studied as Poincaré maps for the Lorenz system of equations. It is a two-parameter two-dimensional family which shares many characteristics with the logistic family and for small $b > 0$ may be considered a two-dimensional ‘perturbation’ of the logistic fam-

ily. The parametrized mapping is defined as

$$T_{a,b}(x, y) = (1 - ax^2 + y, bx),$$

so $T_{a,b}: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with $0 < a < 2$ and $b > 0$. Benedicks and Carleson [7] showed that for a positive-measure set of parameters (a, b) , $T_{a,b}$ has a topologically transitive attractor $\Lambda_{a,b}$. Benedicks and Young [8] later proved that for a positive-measure set of parameters (a, b) , $T_{a,b}$ has a topologically transitive SRB attractor $\Lambda_{a,b}$ with SRB measure $\mu_{a,b}$ and that $(T_{a,b}, \Lambda_{a,b}, \mu_{a,b})$ is isomorphic to a Bernoulli shift.

Complex Dynamics

Complex dynamics is concerned with the behavior of rational maps

$$\frac{\alpha_1 z^d + \alpha_2 z^{d-1} + \cdots + \alpha_{d+1}}{\beta_1 z^d + \beta_2 z^{d-1} + \cdots + \beta_{d+1}}$$

of the extended complex plane $\bar{\mathbb{C}}$ to itself, in which the domain is \mathbb{C} completed with the point at infinity (called the Riemann sphere). Recall that a family F of meromorphic functions is called *normal* on a domain D if every sequence possesses a subsequence that converges uniformly (in the spherical metric $\bar{\mathbb{C}} \sim S^2$) on compact subsets of D . A family is *normal at a point* $z \in \bar{\mathbb{C}}$ if it is normal on a neighborhood of z . The *Fatou set* $F(R) \subset \bar{\mathbb{C}}$ of a rational map $R: \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$ is the set of points $z \in \bar{\mathbb{C}}$ such that the family of forward iterates $\{R^n\}_{n \geq 0}$ is normal at z . The *Julia set* $J(R)$ is the complement of the Fatou set $F(R)$. The Fatou set is open and hence the Julia set is a closed set. Another characterization in the case $d > 1$ is that $J(R)$ is the closure of the set of all repelling periodic orbits of $R: \bar{\mathbb{C}} \rightarrow \bar{\mathbb{C}}$. Both $F(R)$ and $J(R)$ are invariant under R . The dynamics of greatest interest is the restriction $R: J(R) \rightarrow J(R)$. The Julia set often has a complicated fractal structure. In the case that $R_a(z) = z^2 - a$, $a \in \mathbb{C}$, the *Mandelbrot set* is defined as the set of a for which the orbit of the origin 0 is bounded. The topology of the Mandelbrot set has been the subject of intense research. The study of complex dynamics is important because of the fascinating and complicated dynamics displayed and also because techniques and results in complex dynamics have direct implications for the behavior of one-dimensional maps. For more details see [25].

Infinite Ergodic Theory

We may also consider a measure-preserving transformation (T, X, μ) of a measure space such that $\mu(X) = \infty$. For example X could be the real line equipped with Lebesgue measure. This setting also arises with compact X

in applications. For example, suppose $T: [0, 1] \rightarrow [0, 1]$ is the simple model of intermittency given in Sect. “Intermittent Maps” and $\gamma \in (1, 2)$. Then T possesses an absolutely continuous invariant measure μ with support $[0, 1]$, but $\mu([0, 1]) = \infty$. The Radon–Nikodym derivative of μ with respect to Lebesgue measure m exists but is not in $L^1(m)$.

In this setting we say a measurable set A is a *wandering set* for T if $\{T^{-n}A\}_{n=0}^{\infty}$ are disjoint. Let $D(T)$ be the measurable union of the collection of wandering sets for T . The transformation T is *conservative* with respect to μ if $(X \setminus D(T)) = X \pmod{\mu}$ (see the article on [Measure Preserving Systems](#)). It is usually necessary to assume T conservative with respect to μ to say anything interesting about its behavior. For example if $T(x) = x + \alpha$, $\alpha > 0$, is a translation of the real line then $D(T) = X$. The definition of ergodicity in this setting remains the same: T is *ergodic* if $A \in \mathcal{B}$ and $T^{-1}A = A \pmod{\mu}$ implies that $\mu(A) = 0$ or $\mu(A^c) = 0$. However the equivalence of ergodicity of T with respect to μ and the equality of time and space averages for $L^1(\mu)$ functions no longer holds. Thus in general μ ergodic does not imply that

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=0}^{n-1} \phi \circ T^i(x) = \int_X \phi \, d\mu \quad \mu \text{ a.e. } x \in X$$

for all $\phi \in L^1(\mu)$. In the example of the intermittent map with $\gamma \in (1, 2)$ the orbit of Lebesgue almost every $x \in X$ is dense in X , yet the fraction of time spent near the indifferent fixed point $x = 0$ tends to one for Lebesgue almost every $x \in X$. In fact it may be shown Sect. 2.4 in [1] that when $\mu(x) = \infty$ there are no constants $a_n > 0$ such that

$$\lim_{n \rightarrow \infty} \frac{1}{a_n} \sum_{i=0}^{n-1} \phi \circ T^i(x) = \int_X \phi \, d\mu \quad \mu \text{ a.e. } x \in X$$

Nevertheless it is sometimes possible to obtain distributional limits, rather than almost sure limits, of Birkhoff sums under suitable normalization. We refer the reader to Aaronson’s book [1] for more details.

Constructions

We give examples of some of the standard constructions in dynamical systems. Often these constructions appear in modeling situations (for example skew products are often used to model systems which react to inputs from other systems, continuous time systems are often modeled as suspension flows over discrete-time dynamics) or to reduce systems to simpler components (often a factor system or induced system is simpler to study). Unless stated otherwise, in the sequel we will be discussing measure-preserving

serving transformations on Lebesgue spaces (see the article on [Measure Preserving Systems](#)).

Products

Given measure-preserving systems (X, \mathcal{B}, μ, T) and (Y, \mathcal{C}, ν, S) , their *product* consists of their completed product measure space with the transformation $T \times S: X \times Y \rightarrow X \times Y$ defined by $(T \times S)(x, y) = (Tx, Sy)$ for all $(x, y) \in X \times Y$. Neither ergodicity nor transitivity is in general preserved by taking products; for example the product of an irrational rotation on the unit circle with itself is not ergodic. For a list of which mixing properties are preserved in various settings by forming a product see [113]. Given any countable family of measure-preserving transformations on probability spaces, their direct product is defined similarly.

Factors

We say that a measure-preserving system (Y, \mathcal{C}, ν, S) is a *factor* of a measure-preserving system (X, \mathcal{B}, μ, T) if (possibly after deleting a set of measure 0 from X) there is a measurable onto map $\phi: X \rightarrow Y$ such that

$$\begin{aligned} \phi^{-1}C &\subset \mathcal{B}, \\ \phi T &= S\phi, \quad \text{and} \\ \mu T^{-1} &= \nu. \end{aligned} \tag{8}$$

For Lebesgue spaces, there is a correspondence of factors of (X, \mathcal{B}, μ, T) and T -invariant complete sub- σ -algebras of \mathcal{B} . According to Rokhlin’s theory of Lebesgue spaces [98] factors also correspond to certain partitions of X (see the article on [Measure Preserving Systems](#)). A factor map $\phi: X \rightarrow Y$ between Lebesgue spaces is an *isomorphism* if and only if it has a measurable inverse, or equivalently $\phi^{-1}C = \mathcal{B}$ up to sets of measure 0.

Skew Products

If (X, \mathcal{B}, μ, T) is a measure-preserving system, (Y, \mathcal{C}, ν) is a measure-space, and $\{S_x: x \in X\}$ is a family of measure-preserving maps $Y \rightarrow Y$ such that the map that takes (x, y) to $S_x y$ is jointly measurable in the two variables x and y , then we may define a *skew product system* consisting of the product measure space of X and Y equipped with product measure $\mu \times \nu$ together with the measure-preserving map $T \ltimes S: X \times Y \rightarrow X \times Y$ defined by

$$(T \ltimes S)(x, y) = (Tx, S_x y). \tag{9}$$

The space Y is called the *fiber* of the skew product and the space X the *base*. Sometimes in the literature the word

skew product has a more general meaning and refers to the structure $(T \ltimes S)(x, y) = (Tx, S_x y)$ (without any assumption of measure-preservation), where the action of the map on the fiber Y is determined or ‘driven’ by the map $T: X \rightarrow X$.

Some common examples of skew products include:

Random Dynamical Systems Suppose that X indexes a collection of mappings $S_x: Y \rightarrow Y$. We may have a transformation $T: X \rightarrow X$ which is a full shift. Then the sequence of mappings $\{S_{T^n x}\}$ may be considered a (random) choice of a mapping $Y \rightarrow Y$ from the set $\{S_x: x \in X\}$. The projection onto Y of the orbits of $(Tx, S_x y)$ give the orbits of a point $y \in Y$ under a random composition of maps $S_{T^n x} \circ \cdots \circ S_{Tx} \circ S_x$. More generally we could consider the choice of maps S_x that are composed to come from any ergodic dynamical system, (T, X, μ) to model the effect of perturbations by a stationary ergodic ‘noise’ process.

Group Extensions of Dynamical Systems Suppose Y is a group, ν is a measure on Y invariant under a left group action, and $S_x y := g(x)y$ is given by a group-valued function $g: X \rightarrow Y$. In this setting g is often called a *cocycle*, since upon defining $g^{(n)}(x)$ by $(T \ltimes S)^{(n)}(x, y) = (T^n x, g^{(n)}(x)y)$ we have a cocycle relation, namely $g^{(m+n)}(x) = g^{(m)}(T^n x)g^{(n)}(x)$. Group extensions arise often in modeling systems with symmetry [37]. Common examples are provided by a random composition of matrices from a group of matrices (or more generally from a set of matrices which may form a group or not).

Induced Transformations

Since by the Poincaré Recurrence Theorem (see [113]) a measure-preserving transformation (T, X, μ, \mathcal{B}) on a probability space is recurrent, given any set B of positive measure, the return-time function

$$n_B(x) = \inf\{n \geq 1: T^n x \in B\} \quad (10)$$

is finite μ a.e. We may define the *first-return map* by

$$T_B x = T^{n_B(x)} x. \quad (11)$$

Then (after perhaps discarding as usual a set of measure 0) $T_B: B \rightarrow B$ is a measurable transformation which preserves the probability measure $\mu_B = \mu/\mu(B)$. The system $(B, \mathcal{B} \cap B, \mu_B, T_B)$ is called an *induced, first-return or derived transformation*. If (T, X, μ, \mathcal{B}) is ergodic then $(B, \mathcal{B} \cap B, \mu_B, T_B)$ is ergodic, but the converse is not in general true.

The construction of the transformation T_B allows us to represent the forward orbit of points in B via a *tower* or *skyscraper* over B . For each $n = 1, 2, \dots$, let

$$B_n = \{x \in B: n_B(x) = n\}. \quad (12)$$

Then $\{B_1, B_2, \dots\}$ form a partition of B , which we think of as the bottom floor or base of the tower. The next floor is made up of TB_2, TB_3, \dots , which form a partition of $TB \setminus B$, and so on. All these sets are disjoint. A *column* is a part of the tower of the form $B_n \cup TB_n \cup \cdots \cup T^{n-1}B_n$ for some $n = 1, 2, \dots$. The action of T on the entire tower is pictured as mapping each x not at the top of its column straight up to the point Tx above it on the next level, and mapping each point on the top level to $T^{n_B} x \in B$. An equivalent way to describe the transformation on the tower is to write for each n and $j < n$, $T^j B_n$ as $\{(x, j): x \in B_n\}$, and then the transformation F on the tower is

$$F(x, l) = \begin{cases} (x, l+1) & \text{if } l < n_B(x) - 1; \\ (T^{n_B(x)} x, 0) & \text{if } l = n_B(x) - 1. \end{cases}$$

If T preserves a measure μ , then F preserves $\mu \times dl$, where l is counting measure.

Sometimes the process of inducing yields an induced map which is easier to analyze (perhaps it has stronger hyperbolicity properties) than the original system. Sometimes it is possible to ‘lift’ ergodic or statistical properties from an induced system to the original system, so the process of inducing plays an important role in the study of statistical properties of dynamical systems [77].

It is possible to generalize the tower construction and relax the condition that $n_B(x)$ is the first-return time function. We may take a measurable set $B \subset X$ of positive μ measure and define for almost every point $x \in B$ a *height* or *ceiling* function $R: B \rightarrow \mathbb{N}$ and take a countable partition $\{X_n\}$ of B into the sets on which R is constant. We define the *tower* as the set $\Delta := \{(x, l): x \in B, 0 \leq l < R(x)\}$ and the *tower map* $F: \Delta \rightarrow \Delta$ by

$$F(x, l) = \begin{cases} (x, l+1) & \text{if } l < R(x) - 1; \\ (T^{R(x)} x, 0) & \text{if } l = R(x) - 1. \end{cases}$$

In this setting, if $\int_B R(x) d\mu < \infty$, we may define an F -invariant probability measure on Δ as $\frac{\mu}{C(R, B)} \times dl$, where dl is counting measure and $C(R, B)$ is the normalizing constant $C(R, B) = \mu(B) \int_B R(x) d\mu$. This viewpoint is connected with the construction of systems by cutting and stacking—see Sect. “Cutting and Stacking”.

Suspension Flows

The tower construction has an analogue in which the height function R takes values in \mathbb{R} rather than \mathbb{N} . Such towers are commonly used to model dynamical systems with continuous time parameter. Let (T, X, μ) be a measure-preserving system and $R: X \rightarrow (0, \infty)$ a measurable “ceiling” function on X . The set

$$X^R = \{(x, t): 0 \leq R(x) < t\}, \quad (13)$$

with measure ν given locally by the product of μ on X with Lebesgue measure m on \mathbb{R} , is a measure space in a natural way. If μ is a finite measure and R is integrable with respect to μ then ν is a finite measure. We define an action of \mathbb{R} on X^R by letting each point x flow at unit speed up the vertical lines $\{(x, t): 0 \leq t < R(x)\}$ under the graph of R until it hits the ceiling, then jump to Tx , and so on. More precisely, defining $R_n(x) = R(x) + \cdots + R(T^n x)$,

$$T_s(x, t) = \begin{cases} (x, s + t) & \text{if } 0 \leq s + t < R(x), \\ (Tx, s + t - R(x)) & \text{if } R(x) \leq s + t < R(x) + R(Tx) \\ \dots & \dots \\ (T^n x, s + t - [R(x) + \cdots + R(T^{n-1} x)]) & \text{if } R_{n-1}(x) \leq s + t < R_n(x). \end{cases} \quad (14)$$

Ergodicity of (T, X, μ) implies the ergodicity of (T_s, X^R, ν) .

Cutting and Stacking

Several of the most interesting examples in ergodic theory have been constructed by this method; in fact, because of Rokhlin’s Lemma (see Sect. “[Rokhlin’s Lemma](#)”) every ergodic measure-preserving transformation on a Lebesgue space is isomorphic to one constructed by cutting and stacking. For example, the von Neumann-Kakutani adding machine (or 2-odometer) (Sect. “[Adding Machines](#)”), the Chacon weakly mixing but not strongly mixing system (Sect. “Chacon System”), Ornstein’s mixing rank one examples (see e.g. p. 160 ff. in [80]), and many more.

We construct a Lebesgue measure-preserving transformation T on an interval X (bounded or maybe unbounded) by defining it as a translation on each of a pairwise disjoint countable collection of subintervals. The construction proceeds by stages, at each stage defining T on an additional part of X , until eventually T is defined a.e.

At each stage X is represented as a *tower*, which is defined to be a disjoint union of *columns*. A *column* is defined

to be a finite disjoint union of intervals of equal length, which are numbered from 0, for the “floor”, to the last one, for the “roof”, and which we picture as lying each above the preceding-numbered interval. T is defined on each *level* of a column (i. e. each interval in the column) except the roof by mapping it by translation to the next higher interval in the column.

At stage 0, we have just one column, consisting of all of X as the floor, and T is not defined anywhere. To pass from one stage to the next, the columns are *cut* and *stacked*. This means that each column is divided, by vertical cuts, into a disjoint union of subcolumns of equal height (but maybe not equal width), and then some of these subcolumns are stacked above others (of the same width) so as to form a new tower. This allows the definition of T to be extended to some parts of X that were previously tops of towers, since they now may have levels above them. (Sometimes columns of height 1 are thought of as forming a reservoir for “spacers” to be inserted between subcolumns that are being stacked.) If the measure of the union of the tops of the columns tends to 0, eventually T becomes defined a.e. This description in words can be made precise with cumbersome notation, but the process can also be given a neater graphical description, which we sketch in the next section.

Adic Transformations

A.M. Vershik has introduced a family of models, called *adic* or *Bratteli–Vershik* transformations, into ergodic theory and dynamical systems. One begins with a graph which is arranged in levels, finitely many vertices on each level, with connections only from each level to the adjacent ones. The space X consists of the set of all infinite paths in this graph; it is a compact metric space in a natural way. We are given an order on the set of edges into each vertex, and then X is partially ordered as follows: x and y are comparable if they agree from some point on, in which case we say that $x < y$ if at the last level n where they traverse different edges, the edge x_n of x is smaller than the edge y_n of y . A map T is defined by letting Tx be the smallest y that is larger than x , if there is one. In nice situations, T is a homeomorphism after defining it and its inverse on perhaps countably many maximal and minimal elements. Invariant measures can sometimes be defined by assigning weights to edges, which are then multiplied to define the measure of each cylinder set. This is a nice combinatorial way to present the cutting and stacking method of constructing m.p.t.’s, allows for more convenient analysis of questions such as orbit equivalence, and leads to the construction of many interesting examples, such as those

based on the Pascal or Euler graphs [4,38,76]. Odometers and generalizations are natural examples of adic systems. Vershik showed that in fact every ergodic measure-preserving transformation on a Lebesgue space is isomorphic to a uniquely ergodic adic transformation. See [111].

Rokhlin's Lemma

The following result is the fundamental starting point for many constructions in ergodic theory, from representing arbitrary systems in terms of cutting and stacking or adic systems, to constructing useful partitions and symbolic codings of abstract systems, to connecting convergence theorems in abstract ergodic theory with those in harmonic analysis. It allows us to picture arbitrarily long stretches of the action of a measure-preserving transformation as a translation within the set of integers. In the ergodic nonatomic case the statement follows readily from the construction of induced transformations.

Lemma 1 (Rokhlin's Lemma) *Let $T: X \rightarrow X$ be a measure-preserving transformation on a probability space (X, \mathcal{B}, μ) . Suppose that (X, \mathcal{B}, μ) is nonatomic and $T: X \rightarrow X$ is ergodic, or, more generally, (T, X, \mathcal{B}, μ) is aperiodic: that is to say, the set $\{x \in X: \text{there is } n \in \mathbb{N} \text{ such that } T^n x = x\}$ of periodic points has measure 0. Then given $n \in \mathbb{N}$ and $\epsilon > 0$, there is a measurable set $B \subset X$ such that the sets $B, TB, \dots, T^{n-1}B$ are pairwise disjoint and $\mu(\cup_{k=0}^{n-1} T^k B) > 1 - \epsilon$.*

Inverse Limits

Suppose that for each $i = 1, 2, \dots$ we have a Lebesgue probability space $(X_i, \mathcal{B}_i, \mu_i)$ and a measure-preserving transformation $T_i: X_i \rightarrow X_i$. Suppose also that for each $i \leq j$ there is a factor map $\phi_{ji}: (T_j, X_j, \mathcal{B}_j, \mu_j) \rightarrow (T_i, X_i, \mathcal{B}_i, \mu_i)$, such that each ϕ_{jj} is the identity on X_j and $\phi_{ji}\phi_{kj} = \phi_{ki}$ whenever $k \geq j \geq i$. Let

$$X = \{x \in \prod_{i=1}^{\infty} X_i: \phi_{ji}x_j = x_i \text{ for all } j \geq i\}. \quad (15)$$

For each j , let $\pi_j: X \rightarrow X_j$ be the projection defined by $\pi_j x = x_j$.

Let \mathcal{B} be the smallest σ -algebra of subsets of X which contains all the $\pi_j^{-1}\mathcal{B}_j$. Define μ on each $\pi_j^{-1}\mathcal{B}_j$ by

$$\mu(\pi_j^{-1}B) = \mu_j(B) \quad \text{for all } B \in \mathcal{B}_j. \quad (16)$$

Because $\phi_{ji}\pi_j = \pi_i$ for all $j \geq i$, the $\pi_j^{-1}\mathcal{B}_j$ are increasing, and so their union is an algebra. The set function μ can, with some difficulty, be shown to be countably additive on this algebra: since we are dealing with Lebesgue spaces, by means of measure-theoretic isomorphisms it is possible to replace the entire situation by compact metric spaces and

continuous maps, then use regularity of the measures involved—see p. 137 ff. in [88]. Thus by Carathéodory's Theorem (see the article on [Measure Preserving Systems](#)) μ extends to all of \mathcal{B} .

Define $T: X \rightarrow X$ by $T(x_j) = (T_j x_j)$. Then (T, X, \mathcal{B}, μ) is a measure-preserving system such that any system which has all the $(T_j, X_j, \mathcal{B}_j, \mu_j)$ as factors, also has (T, X, \mathcal{B}, μ) a factor.

Natural Extension

The natural extension is a way to produce an invertible system from a non-invertible system. The original system is a factor of its natural extension and its orbit structure and ergodic properties are captured by the natural extension, as will be seen from its construction. Let (T, X, \mathcal{B}, μ) be a measure-preserving transformation of a Lebesgue probability space. Define

$$\begin{aligned} \Omega &:= \{(x_0, x_1, x_2, \dots): x_n \\ &= T(x_{n+1}), x_n \in X, n = 0, 1, 2, \dots\} \end{aligned}$$

with $\sigma: \Omega \rightarrow \Omega$ defined by $\sigma((x_0, x_1, x_2, \dots)) = (T(x_0), x_0, x_1, \dots)$. The map σ is invertible on Ω . Given the invariant measure μ we define the invariant measure for the natural extension $\tilde{\mu}$ on Ω by defining it first on cylinder sets $C(A_0, A_1, \dots, A_k)$ by

$$\begin{aligned} \tilde{\mu}(C(A_0, A_1, \dots, A_k)) \\ = \mu(T^{-k}(A_0) \cap T^{-k+1}(A_1) \cap \dots \cap T^{-k+i}(A_i) \cap \dots \cap A_k) \end{aligned}$$

and then extending it to Ω using Kolmogorov's extension theorem. We think of (x_0, x_1, x_2, \dots) as being an inverse branch of $x_0 \in X$ under the mapping $T: X \rightarrow X$. The maps $\sigma, \sigma^{-1}: \Omega \rightarrow \Omega$ are ergodic with respect to $\tilde{\mu}$ if (T, X, \mathcal{B}, μ) is ergodic [113]. If $\pi: \Omega \rightarrow X$ is projection onto the first component i.e. $\pi(x_0, \dots, x_n, \dots) = x_0$ then $\pi \circ \sigma^n(x_0, \dots, x_n, \dots) = T^n(x_0)$ for all x_0 and thus the natural extension yields information about the orbits of X under T .

The natural extension is an inverse limit. Let (X, \mathcal{B}, μ) be a Lebesgue probability space and $T: X \rightarrow X$ a map such that $T^{-1}\mathcal{B} \subset \mathcal{B}$ and $\mu T^{-1} = \mu$. For each $i = 1, 2, \dots$ let $(T_i, X_i, \mathcal{B}_i, \mu_i) = (T, X, \mathcal{B}, \mu)$, and $\phi_{ji} = T^{j-i}$ for each $j > i$. Then the inverse limit $(\hat{T}, \hat{X}, \hat{\mathcal{B}}, \hat{\mu})$ of this system is an invertible measure-preserving system which is the natural extension of (T, X, \mathcal{B}, μ) . We have

$$\hat{T}^{-1}(x_1, x_2, \dots) = (x_2, x_3, \dots). \quad (17)$$

The original system (T, X, \mathcal{B}, μ) is a factor of $(\hat{T}, \hat{X}, \hat{\mathcal{B}}, \hat{\mu})$ (using any π_i as the factor map), and any factor mapping from an invertible system onto (T, X, \mathcal{B}, μ) consists

of a factor mapping onto $(\hat{T}, \hat{X}, \hat{B}, \hat{\mu})$ followed by projection onto the first coordinate.

Joinings

Given measure-preserving systems (T, X, \mathcal{B}, μ) and (S, Y, \mathcal{C}, ν) , a *joining* of the two systems is a $T \times S$ -invariant measure P on their product measurable space that projects to μ and ν , respectively, under the projections of $X \times Y$ to X and Y , respectively. That is, if $\pi_1: X \times Y \rightarrow X$ is the projection onto the first component i.e. $\pi_1(x, y) = x$ then $P(\pi_1^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$ and similarly for $\pi_2: X \times Y \rightarrow Y$.

This concept is the ergodic-theoretic version of the notion in probability theory of a *coupling*. The product measure $\mu \times \nu$ is always a joining of the two systems. If product measure is the only joining of the two systems, then we say that they are disjoint and write $X \perp Y$ [40]. If \mathcal{D} is any family of systems, we write \mathcal{D}^\perp for the family of all measure-preserving systems which are disjoint from every system in \mathcal{D} . Extensive recent accounts of the use of joinings in ergodic theory are in [42, 99, 110].

Future Directions

The basic examples and constructions presented here are idealized, and many of the underlying assumptions (such as uniform hyperbolicity) are seldom satisfied in applications, yet they have given important insights into the behavior of real-world physical systems. Recent developments have improved our understanding of the ergodic properties of non-uniformly and partially hyperbolic systems. The ergodic properties of deterministic systems will continue to be an active research area for the foreseeable future. The directions will include, among others: establishing statistical and ergodic properties under weakened dependence assumptions; the study of systems which display ‘anomalous statistics’; the study of the stability and typicality of ergodic behavior and mixing in dynamical systems; the ergodic theory of infinite-dimensional systems; advances in number theory (see the sections on Szemerédi and Ramsey theory); research into models with non-singular rather than invariant measures; and infinite-measure systems. Other chapters in this Encyclopedia discuss in more detail these and other topics.

Bibliography

Primary Literature

- Aaronson J (1997) An introduction to infinite ergodic theory, Mathematical Surveys and Monographs, vol 50. American Mathematical Society, Providence. MR 1450400 (99d:28025)
- Anosov DV (1967) Geodesic flows on closed riemannian manifolds of negative curvature. Trudy Mat Inst Steklov 90:209. MR 0224110 (36 #7157)
- Arnol'd VI (1963) Small denominators and problems of stability of motion in classical and celestial mechanics. Uspehi Mat Nauk 18(6 (114)):91–192. MR 0170705 (30 #943)
- Bailey S, Keane M, Petersen K, Salama IA (2006) Ergodicity of the adic transformation on the euler graph. Math Proc Cambridge Philos Soc 141(2):231–238. MR 2265871 (2007m:37010)
- Baladi V (2000) Positive transfer operators and decay of correlations. Advanced Series in Nonlinear Dynamics, vol 16. World Scientific Publishing Co Inc, River Edge. MR 1793194 (2001k:37035)
- Benedicks M, Carleson L (1985) On iterations of $1 - ax(X, \mathcal{B}, \mu)^2$ on $(-1, 1)$. Ann Math (2) 122(1):1–25. MR 799250 (87c:58058)
- Benedicks M, Carleson L (1991) The dynamics of the hénon map. Ann Math (2) 133(1):73–169. MR 1087346 (92d:58116)
- Benedicks M, Young LS (1993) Sinai-bowen-ruelle measures for certain hénon maps. Invent Math 112(3):541–576. MR 1218323 (94e:58074)
- Bertrand-Mathis A (1986) Développement en base θ ; répartition modulo un de la suite $(x\theta(X, \mathcal{B}, \mu)_n)_{n \geq 0}$; langages codés et θ -shift. Bull Soc Math France 114(3):271–323. MR 878240 (88e:11067)
- Billingsley P (1978) Ergodic Theory and Information. Robert E. Krieger Publishing Co, Huntington, N.Y., reprint of the 1965 original. MR 524567 (80b:28017)
- Bissinger BH (1944) A generalization of continued fractions. Bull Amer Math Soc 50:868–876. MR 0011338 (6,150h)
- Blanchard F (1989) β -expansions and symbolic dynamics. Theoret Comput Sci 65(2):131–141. MR 1020481 (90j:54039)
- Blanchard F, Hansel G (1986) Systèmes codés. Theoret Comput Sci 44(1):17–49. MR 858689 (88m:68029)
- Blanchard F, Hansel G (1986) Systèmes codés et limites de systèmes sofiques. C R Acad Sci Paris Sér I Math 303(10):475–477. MR 865864 (87m:94009)
- Blanchard F, Hansel G (1991) Sofic constant-to-one extensions of subshifts of finite type. Proc Amer Math Soc 112(1):259–265. MR 1050016 (91m:54050)
- Boshernitzan M, Galperin G, Krüger T, Troubetzkoy S (1998) Periodic billiard orbits are dense in rational polygons. Trans Amer Math Soc 350(9):3523–3535. MR 1458298 (98k:58179)
- Boshernitzan MD (1992) Billiards and rational periodic directions in polygons. Amer Math Monthly 99(6):522–529. MR 1166001 (93d:51043)
- Bowen R (1970) Markov partitions for axiom A diffeomorphisms. Amer J Math 92:725–747. MR 0277003 (43 #2740)
- Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. In: Lecture notes in mathematics, vol 470. Springer, Berlin. MR 0442989 (56 #1364)
- Boyarsky A, Góra P (1997) Laws of chaos: Invariant measures and dynamical systems in one dimension. Probability and its Applications. Birkhäuser, Boston. MR 1461536 (99a:58102)
- Brin MJ, Pesin JB (1974) Partially hyperbolic dynamical systems. Izv Akad Nauk SSSR Ser Mat 38:170–212. MR 0343316 (49 #8058)
- Bunimovich LA (1974) The ergodic properties of certain billiards. Funkcional Anal i Priložen 8(3):73–74. MR 0357736 (50 #10204)

23. Bunimovich LA (1979) On the ergodic properties of nowhere dispersing billiards. *Comm Math Phys* 65(3):295–312. MR 530154 (80h:58037)
24. Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. In: *Smooth ergodic theory and its applications*, Seattle, WA, 1999. *Proc Sympos Pure Math*, vol 69. Amer Math Soc, Providence, RI, pp 327–366. MR 1858538 (2002m:37042)
25. Carleson L, Gamelin TW (1993) *Complex dynamics*. Universitext: Tracts in Mathematics. Springer, New York. MR 1230383 (94h:30033)
26. Chernov N, Markarian R (2006) *Chaotic billiards*, Mathematical Surveys and Monographs, vol 127. American Mathematical Society, Providence, RI. MR 2229799 (2007f:37050)
27. Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical systems, *Progress in Physics*, vol 1. Birkhäuser, Boston. MR 613981 (82j:58078)
28. Cornfeld IP, Fomin SV, Sinai YG (1982) *Ergodic Theory, Grundlehren der mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences)*, vol 245. Springer, New York, translated from the Russian by A. B. Sossinskiĭ. MR 832433 (87f:28019)
29. Coven EM, Hedlund GA (1973) Sequences with minimal block growth. *Math Systems Theory* 7:138–153. MR 0322838 (48 #1199)
30. del Junco A (1978) A simple measure-preserving transformation with trivial centralizer. *Pacific J Math* 79(2):357–362. MR 531323 (80i:28034)
31. del Junco A, Rahe M, Swanson L (1980) Chacon's automorphism has minimal self-joinings. *J Analyse Math* 37:276–284. MR 583640 (81j:28027)
32. de Melo W, van Strien S (1993) One-dimensional dynamics. *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3))*, vol 25. Springer, Berlin. MR 1239171 (95a:58035)
33. de la Rue T (1993) Entropie d'un système dynamique gaussien: cas d'une action de $\mathbb{Z}(X, \mathcal{B}, \mu)$. *C R Acad Sci Paris Sér I Math* 317(2):191–194. MR 1231420 (94c:28022)
34. Downarowicz T (2005) Survey of odometers and Toeplitz flows. In: *Algebraic and topological dynamics*. *Contemp Math*, vol 385. Amer Math Soc, Providence, RI, pp 7–37. MR 2180227 (2006f:37009)
35. Everett CJ (1946) Representations for real numbers. *Bull Amer Math Soc* 52:861–869. MR 0018221 (8,259c)
36. Feigenbaum MJ (1978) Quantitative universality for a class of nonlinear transformations. *J Statist Phys* 19(1):25–52. MR 0501179 (58 #18601)
37. Field M, Nicol M (2004) Ergodic theory of equivariant diffeomorphisms: Markov partitions and stable ergodicity. *Mem Amer Math Soc* 169(803):viii+100. MR 2045641 (2005g:37041)
38. Frick SB, Petersen K () Random permutations and unique fully supported ergodicity for the Euler adic transformation. *Ann Inst H Poincaré Prob Stat*. To appear
39. Furstenberg H (1973) The unique ergodicity of the horocycle flow. In: *Recent advances in topological dynamics (Proc Conf, Yale Univ, New Haven, Conn, 1972; in honor of Gustav Arnold Hedlund)*. *Lecture Notes in Math*, vol 318. Springer, Berlin, pp 95–115. MR 0393339 (52 #14149)
40. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in diophantine approximation. *Math Systems Theory* 1:1–49. MR 0213508 (35 #4369)
41. Gallavotti G, Ornstein DS (1974) Billiards and bernoulli schemes. *Comm Math Phys* 38:83–101. MR 0355003 (50 #7480)
42. Glasner E (2003) *Ergodic Theory via Joinings*. Mathematical Surveys and Monographs, vol 101. American Mathematical Society, Providence, RI. MR 1958753 (2004c:37011)
43. Graczyk J, Świątek G (1997) Generic hyperbolicity in the logistic family. *Ann Math (2)* 146(1):1–52. MR 1469316 (99b:58079)
44. Grayson M, Pugh C, Shub M (1994) Stably ergodic diffeomorphisms. *Ann Math (2)* 140(2):295–329. MR 1298715 (95g:58128)
45. Guckenheimer J, Holmes P (1990) *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. *Applied Mathematical Sciences*, vol 42. Springer, New York, revised and corrected reprint of the 1983 original. MR 1139515 (93e:58046)
46. Hasselblatt B, Katok A (2003) *A First Course in Dynamics*. Cambridge University Press, with a panorama of recent developments. MR 1995704 (2004f:37001)
47. Hayashi S (1997) Connecting invariant manifolds and the solution of the $c(X, \mathcal{B}, \mu)$ stability and ω -stability conjectures for flows. *Ann Math (2)* 145(1):81–137. MR 1432037 (98b:58096)
48. Hedlund GA (1934) On the metrical transitivity of the geodesics on closed surfaces of constant negative curvature. *Ann Math (2)* 35(4):787–808. MR 1503197
49. Hopf E (1939) Statistik der geodätischen Linien in Mannigfaltigkeiten negativer Krümmung. *Ber Verh Sächs Akad Wiss Leipzig* 91:261–304. MR 0001464 (1,243a)
50. Host B (1995) Nombres normaux, entropie, translations. *Israel J Math* 91(1-3):419–428. MR 1348326 (96g:11092)
51. Hu H (2004) Decay of correlations for piecewise smooth maps with indifferent fixed points. *Ergodic Theory Dynam Systems* 24(2):495–524. MR 2054191 (2005a:37064)
52. Jakobson MV (1981) Absolutely continuous invariant measures for one-parameter families of one-dimensional maps. *Comm Math Phys* 81(1):39–88. MR 630331 (83j:58070)
53. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst Hautes Études Sci Publ Math* 51:137–173. MR 573822 (81i:28022)
54. Katok A, Strelcyn JM, Ledrappier F, Przytycki F (1986) Invariant manifolds, entropy and billiards; smooth maps with singularities. *Lecture Notes in Mathematics*, vol 1222. Springer, Berlin. MR 872698 (88k:58075)
55. Keane M (1968) Generalized morse sequences. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 10:335–353. MR 0239047 (39 #406)
56. Keane M (1977) Non-ergodic interval exchange transformations. *Israel J Math* 26(2):188–196. MR 0435353 (55 #8313)
57. Keller G, Nowicki T (1992) Spectral theory, zeta functions and the distribution of periodic points for Collet-Eckmann maps. *Comm Math Phys* 149(1):31–69. MR 1182410 (93i:58123)
58. Kerckhoff S, Masur H, Smillie J (1986) Ergodicity of billiard flows and quadratic differentials. *Ann Math (2)* 124(2):293–311. MR 855297 (88f:58122)
59. Kolmogorov AN (1954) On conservation of conditionally periodic motions for a small change in Hamilton's function. *Dokl Akad Nauk SSSR (NS)* 98:527–530. MR 0068687 (16,924c)
60. Krieger W (2000) On subshifts and topological Markov chains. In: *Numbers, information and complexity (Bielefeld, 1998)*. Kluwer, Boston, pp 453–472. MR 1755380 (2001g:37010)

61. Lagarias JC (1991) The Farey shift. Manuscript
62. Lagarias JC (1992) Number theory and dynamical systems. In: The unreasonable effectiveness of number theory (Orono, ME, 1991). Proc Sympos Appl Math, vol 46. Amer Math Soc, Providence, RI, pp 35–72. MR 1195841 (93m:11143)
63. Lazutkin VF (1973) Existence of caustics for the billiard problem in a convex domain. Izv Akad Nauk SSSR Ser Mat 37:186–216. MR 0328219 (48 #6561)
64. Ledrappier F (1984) Propriétés ergodiques des mesures de sinaï. Inst Hautes Études Sci Publ Math 59:163–188. MR 743818 (86f:58092)
65. Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge. MR 1369092 (97a:58050)
66. Liverani C, Wojtkowski MP (1995) Ergodicity in hamiltonian systems. In: Dynamics reported. Dynam Report Expositions Dynam Systems (N.S.), vol 4. Springer, Berlin, pp 130–202. MR 1346498 (96g:58144)
67. Liverani C, Saussol B, Vaienti S (1999) A probabilistic approach to intermittency. Ergodic Theory Dynam Systems 19(3):671–685. MR 1695915 (2000d:37029)
68. Lyons R (1988) On measures simultaneously 2- and 3-invariant. Israel J Math 61(2):219–224. MR 941238 (89e:28031)
69. Mañé R (1987) Ergodic theory and differentiable dynamics, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3)), vol 8. Springer, Berlin, translated from the Portuguese by Silvio Levy. MR 889254 (88c:58040)
70. Mañé R (1988) A proof of the $c(X, \mathcal{B}, \mu)_1$ stability conjecture. Inst Hautes Études Sci Publ Math 66:161–210. MR 932138 (89e:58090)
71. Manneville P, Pomeau Y (1980) Different ways to turbulence in dissipative dynamical systems. Phys D 1(2):219–226. MR 581352 (81h:58041)
72. Marcus B (1975) Unique ergodicity of the horocycle flow: variable negative curvature case. Israel J Math 21(2-3):133–144, Conference on Ergodic Theory and Topological Dynamics (Kibbutz Lavi, 1974). MR 0407902 (53 #11672)
73. Marcus B (1978) The horocycle flow is mixing of all degrees. Invent Math 46(3):201–209. MR 0488168 (58 #7731)
74. Masur H (1986) Closed trajectories for quadratic differentials with an application to billiards. Duke Math J 53(2):307–314. MR 850537 (87j:30107)
75. Mayer DH (1991) Continued fractions and related transformations. In: Ergodic theory, symbolic dynamics, and hyperbolic spaces (Trieste, 1989). Oxford Sci Publ. Oxford Univ Press, Oxford, pp 175–222. MR 1130177
76. Méla X, Petersen K (2005) Dynamical properties of the pascal adic transformation. Ergodic Theory Dynam Systems 25(1):227–256. MR 2122921 (2005k:37012)
77. Melbourne I, Török A (2004) Statistical limit theorems for suspension flows. Israel J Math 144:191–209. MR 2121540 (2006c:37005)
78. de Melo W (1973) Structural stability of diffeomorphisms on two-manifolds. Invent Math 21:233–246. MR 0339277 (49 #4037)
79. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. Nachr Akad Wiss Göttingen Math-Phys Kl II 1962:1–20. MR 0147741 (26 #5255)
80. Nadkarni MG (1998) Spectral Theory of Dynamical Systems. Birkhäuser Advanced Texts: Basler Lehrbücher. (Birkhäuser Advanced Texts: Basel Textbooks), Birkhäuser, Basel. MR 1719722 (2001d:37001)
81. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. Advances in Math 4:337–352. MR 0257322 (41 #1973)
82. Ornstein DS, Weiss B (1973) Geodesic flows are bernoullian. Israel J Math 14:184–198. MR 0325926 (48 #4272)
83. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Ljapunov exponents of dynamical systems. Trudy Moskov Mat Obšč 19:179–210. MR 0240280 (39 #1629)
84. Parry W (1960) On the β -expansions of real numbers. Acta Math Acad Sci Hungar 11:401–416. MR 0142719 (26 #288)
85. Parry W (1964) Representations for real numbers. Acta Math Acad Sci Hungar 15:95–105. MR 0166332 (29 #3609)
86. Parry W (1966) Symbolic dynamics and transformations of the unit interval. Trans Amer Math Soc 122:368–378. MR 0197683 (33 #5846)
87. Parry W (1996) Squaring and cubing the circle—rudolph's theorem. In: Ergodic theory of $\mathbf{Z}(X, \mathcal{B}, \mu)$ -actions (Warwick, 1993–1994). London Math Soc Lecture Note Ser, vol 228. Cambridge Univ Press, Cambridge, pp 177–183. MR 1411219 (97h:28009)
88. Parthasarathy KR (2005) Probability Measures on Metric Spaces. AMS Chelsea Publishing, Providence, RI, reprint of the 1967 original. MR 2169627 (2006d:60004)
89. Pesin JB (1976) Families of invariant manifolds that correspond to nonzero characteristic exponents. Izv Akad Nauk SSSR Ser Mat 40(6):1332–1379, 1440. MR 0458490 (56 #16690)
90. Pesin JB (1977) Characteristic Ljapunov exponents, and smooth ergodic theory. Uspehi Mat Nauk 32(4 (196)):55–112, 287. MR 0466791 (57 #6667)
91. Phillips E, Varadhan S (eds) (1975) Ergodic Theory. Courant Institute of Mathematical Sciences New York University, a seminar held at the Courant Institute of Mathematical Sciences, New York University, New York, 1973–1974; With contributions by S. Varadhan, E. Phillips, S. Alpern, N. Bitzenhofer and R. Adler. MR 0486431 (58 #6177)
92. Pugh C, Shub M (2004) Stable ergodicity. Bull Amer Math Soc (NS) 41(1):1–41 (electronic), with an appendix by Alexander Starkov. MR 2015448 (2005f:37011)
93. Rényi A (1957) Representations for real numbers and their ergodic properties. Acta Math Acad Sci Hungar 8:477–493. MR 0097374 (20 #3843)
94. Robbin JW (1971) A structural stability theorem. Ann Math (2) 94:447–493. MR 0287580 (44 #4783)
95. Robinson C (1975) Errata to: “structural stability of vector fields” (ann. of math. (2) 99:154–175 (1974)). Ann Math (2) 101:368. MR 0365630 (51 #1882)
96. Robinson C (1976) Structural stability of $c(X, \mathcal{B}, \mu)_1$ diffeomorphisms. J Differential Equations 22(1):28–73. MR 0474411 (57 #14051)
97. Robinson RC (1973) $c(X, \mathcal{B}, \mu)_r$ structural stability implies Kupka-Smale. In: Dynamical systems, (Proc Sympos, Univ Bahia, Salvador, 1971). Academic Press, New York, pp 443–449. MR 0334282 (48 #12601)
98. Rohlin VA (1952) On the fundamental ideas of measure theory. Amer Math Soc Translation 1952(71):55. MR 0047744 (13,924e)
99. Rudolph DJ (1990) Fundamentals of Measurable Dynamics: Ergodic theory on Lebesgue spaces. Oxford Science Publica-

- tions. The Clarendon Press Oxford University Press, New York. MR 1086631 (92e:28006)
100. Rudolph DJ (1990) $\times 2$ and $\times 3$ invariant measures and entropy. *Ergodic Theory Dynam Systems* 10(2):395–406. MR 1062766 (91g:28026)
 101. Ruelle D (1976) A measure associated with axiom-A attractors. *Amer J Math* 98(3):619–654. MR 0415683 (54 #3763)
 102. Ruelle D (1978) Thermodynamic formalism: the mathematical structures of classical equilibrium statistical mechanics. *Encyclopedia of Mathematics and its Applications*, vol 5. Addison-Wesley Publishing Co, Reading, MA, with a foreword by Giovanni Gallavotti and Gian-Carlo Rota. MR 511655 (80g:82017)
 103. Sarig O (2002) Subexponential decay of correlations. *Invent Math* 150(3):629–653. MR 1946554 (2004e:37010)
 104. Schweiger F (1995) *Ergodic Theory of Fibred Systems and Metric Number Theory*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1419320 (97h:11083)
 105. Sinai JG (1968) Markov partitions and u-diffeomorphisms. *Funkcional Anal i Prilozhen* 2(1):64–89. MR 0233038 (38 #1361)
 106. Sinai JG (1970) Dynamical systems with elastic reflections. Ergodic properties of dispersing billiards. *Uspehi Mat Nauk* 25(2 (152)):141–192. MR 0274721 (43 #481)
 107. Sinai JG (1972) Gibbs measures in ergodic theory. *Uspehi Mat Nauk* 27(4(166)):21–64. MR 0399421 (53 #3265)
 108. Smale S (1980) *The mathematics of time. Essays on dynamical systems, economic processes, and related topics*. Springer, New York. MR 607330 (83a:01068)
 109. Tabachnikov S (2005) *Geometry and billiards*, Student Mathematical Library, vol 30. American Mathematical Society, Providence, RI. MR 2168892 (2006h:51001)
 110. Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: *Ergodic theory and its connections with harmonic analysis* (Alexandria, 1993). *London Math Soc Lecture Note Ser*, vol 205. Cambridge Univ Press, Cambridge, pp 207–235. MR 1325699 (96d:28017)
 111. Vershik AM, Livshits AN (1992) Adic models of ergodic transformations, spectral theory, substitutions, and related topics. In: *Representation theory and dynamical systems*. *Adv Soviet Math*, vol 9. Amer Math Soc, Providence, RI, pp 185–204. MR 1166202 (93i:46131)
 112. Vorobets YB, Gal'perin GA, Stëpin AM (1992) Periodic billiard trajectories in polygons: generation mechanisms. *Uspekhi Mat Nauk* 47(3(285)):9–74, 207, (Russian with Russian summary), English translation: (1992) *Russian Math Surveys* 47(3):5–80. MR 1185299 (93h:58088)
 113. Walters P (1982) *An Introduction to Ergodic Theory*. *Graduate Texts in Mathematics*, vol 79. Springer, New York. MR 648108 (84e:28017)
 114. Young LS (1993) Ergodic theory of chaotic dynamical systems. In: *From Topology to Computation: Proceedings of the Smalefest* (Berkeley, CA, 1990). Springer, New York, pp 201–226. MR 1246120 (94i:58112)
 115. Young LS (1998) Statistical properties of dynamical systems with some hyperbolicity. *Ann Math* (2) 147(3):585–650. MR 1637655 (99h:58140)
 116. Young LS (1999) Recurrence times and rates of mixing. *Israel J Math* 110:153–188. MR 1750438 (2001j:37062)
 117. Zemljakov AN, Katok AB (1975) Topological transitivity of billiards in polygons. *Mat Zametki* 18(2):291–300. MR 0399423 (53 #3267)

Books and Reviews

- Baladi V (2000) *Positive transfer operators and decay of correlations*. *Advanced Series in Nonlinear Dynamics*, vol 16. World Scientific Publishing Co Inc, River Edge. MR 1793194 (2001k:37035)
- Billingsley P (1978) *Ergodic Theory and Information*. Robert E. Krieger Publishing Co, Huntington, N.Y., pp xiii+194, reprint of the 1965 original. MR 524567 (80b:28017)
- Billingsley P (1995) *Probability and Measure*. *Wiley Series in Probability and Mathematical Statistics*, 3rd edn. Wiley, New York, pp xiv+593, A Wiley-Interscience Publication. MR 1324786 (95k:60001)
- Bonatti C, Díaz LJ, Viana M (2005) Dynamics beyond uniform hyperbolicity: A global geometric and probabilistic perspective; *Mathematical Physics, III*. In: *Encyclopaedia of Mathematical Sciences*, vol. 102. Springer, Berlin, pp xviii+384. MR 2105774 (2005g:37001)
- Boyarsky A, Góra P (1997) *Laws of chaos: Invariant measures and dynamical systems in one dimension*. *Probability and its Applications*. Birkhäuser, Boston. MR 1461536 (99a:58102)
- Brin M, Stuck G (2002) *Introduction to dynamical systems*. Cambridge University Press, Cambridge. MR 1963683 (2003m:37001)
- Carleson L, Gamelin TW (1993) *Complex dynamics*. *Universitext: Tracts in Mathematics*. Springer, New York. MR 1230383 (94h:30033)
- Chernov N, Markarian R (2006) *Chaotic billiards*, *Mathematical Surveys and Monographs*, vol 127. American Mathematical Society, Providence, RI. MR 2229799 (2007f:37050)
- Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical systems, *Progress in Physics*, vol 1. Birkhäuser, Boston. MR 613981 (82j:58078)
- Cornfeld IP, Fomin SV, Sinai YG (1982) *Ergodic Theory*, *Grundlehren der mathematischen Wissenschaften (Fundamental Principles of Mathematical Sciences)*, vol 245. Springer, New York, translated from the Russian by A. B. Sosinskii. MR 832433 (87f:28019)
- Denker M, Grillenberger C, Sigmund K (1976) *Ergodic Theory on Compact Spaces*. *Lecture Notes in Mathematics*, vol 527. Springer, Berlin. MR 0457675 (56 #15879)
- Friedman NA (1970) *Introduction to Ergodic Theory*. Van Nostrand Reinhold Mathematical Studies, No 29. Van Nostrand Reinhold Co, New York. MR 0435350 (55 #8310)
- Glasner E (2003) *Ergodic Theory via Joinings*, *Mathematical Surveys and Monographs*, vol 101. American Mathematical Society, Providence, RI. MR 1958753 (2004c:37011)
- Halmos PR (1960) *Lectures on Ergodic Theory*. Chelsea Publishing Co, New York. MR 0111817 (22 #2677)
- Hasselblatt B, Katok A (2003) *A First Course in Dynamics: With a panorama of recent developments*. Cambridge University Press, Cambridge. MR 1995704 (2004f:37001)
- Hopf E (1937) *Ergodentheorie*, 1st edn. *Ergebnisse der Mathematik und ihrer Grenzgebiete*; 5. Bd, 2, J. Springer, Berlin
- Jacobs K (1965) Einige neuere Ergebnisse der Ergodentheorie. *Jber Deutsch Math-Verein* 67(Abt 1):143–182. MR 0186789 (32 #4244)
- Keller G (1998) *Equilibrium States in Ergodic Theory*. In: *London Mathematical Society Student Texts*, vol 42. Cambridge University Press, Cambridge, pp x+178. MR 1618769 (99e:28022)

- Mañé R (1987) Ergodic theory and differentiable dynamics. *Ergebnisse der Mathematik und ihrer Grenzgebiete (3) (Results in Mathematics and Related Areas (3))*, vol 8. Springer, Berlin, translated from the Portuguese by Silvio Levy. MR 889254 (88c:58040)
- Nadkarni MG (1998) *Spectral Theory of Dynamical Systems*. Birkhäuser Advanced Texts: Basler Lehrbücher. (Birkhäuser Advanced Texts: Basel Textbooks), Birkhäuser, Basel. MR 1719722 (2001d:37001)
- Katok A, Hasselblatt B (1995) *Introduction to the Modern Theory of Dynamical Systems*, *Encyclopedia of Mathematics and its Applications*, vol 54. Cambridge University Press, Cambridge, With a supplementary chapter by Katok and Leonardo Mendoza. MR 1326374 (96c:58055)
- Parry W, Pollicott M (1990) Zeta functions and the periodic orbit structure of hyperbolic dynamics. *Astérisque* (187-188):268. MR 1085356 (92f:58141)
- Ornstein DS, Rudolph DJ, Weiss B (1982) Equivalence of measure preserving transformations. *Mem Amer Math Soc* 37(262):xii+116. MR 653094 (85e:28026)
- Petersen K (1989) *Ergodic Theory*, *Cambridge Studies in Advanced Mathematics*, vol 2. Cambridge University Press, Cambridge. corrected reprint of the 1983 original. MR 1073173 (92c:28010)
- Royden HL (1988) *Real Analysis*, 3rd edn. Macmillan Publishing Company, New York. MR 1013117 (90g:00004)
- Rudolph DJ (1990) *Fundamentals of Measurable Dynamics: Ergodic theory on Lebesgue spaces*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1086631 (92e:28006)
- Schweiger F (1995) *Ergodic Theory of Fibred Systems and Metric Number Theory*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York. MR 1419320 (97h:11083)
- Thouvenot JP (1995) Some properties and applications of joinings in ergodic theory. In: *Ergodic theory and its connections with harmonic analysis* (Alexandria, 1993). *London Math Soc Lecture Note Ser*, vol 205. Cambridge Univ Press, Cambridge, pp 207–235. MR 1325699 (96d:28017)
- Walters P (1982) *An Introduction to Ergodic Theory*, *Graduate Texts in Mathematics*, vol 79. Springer, New York. MR 648108 (84e:28017)

Ergodic Theory of Cellular Automata

MARCUS PIVATO

Department of Mathematics, Trent University,
Peterborough, Canada

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Invariant Measures for CA](#)

[Limit Measures and Other Asymptotics](#)

[Measurable Dynamics](#)

[Entropy](#)

[Future Directions and Open Problems](#)

[Acknowledgments](#)

[Bibliography](#)

Glossary

Configuration space and the shift Let \mathbb{M} be a finitely generated group or monoid (usually abelian). Typically, $\mathbb{M} = \mathbb{N} := \{0, 1, 2, \dots\}$ or $\mathbb{M} = \mathbb{Z} := \{\dots, -1, 0, 1, 2, \dots\}$, or $\mathbb{M} = \mathbb{N}^E$, \mathbb{Z}^D , or $\mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \in \mathbb{N}$. In some applications, \mathbb{M} could be nonabelian (although usually amenable), but to avoid notational complexity we will generally assume \mathbb{M} is abelian and additive, with operation ‘+’.

Let \mathcal{A} be a finite set of symbols (called an *alphabet*). Let $\mathcal{A}^{\mathbb{M}}$ denote the set of all functions $\mathbf{a}: \mathbb{M} \rightarrow \mathcal{A}$, which we regard as \mathbb{M} -indexed *configurations* of elements in \mathcal{A} . We write such a configuration as $\mathbf{a} = [a_m]_{m \in \mathbb{M}}$, where $a_m \in \mathcal{A}$ for all $m \in \mathbb{M}$, and refer to $\mathcal{A}^{\mathbb{M}}$ as *configuration space*.

Treat \mathcal{A} as a discrete topological space; then \mathcal{A} is compact (because it is finite), so $\mathcal{A}^{\mathbb{M}}$ is compact in the Tychonoff product topology. In fact, $\mathcal{A}^{\mathbb{M}}$ is a *Cantor space*: it is compact, perfect, totally disconnected, and metrizable. For example, if $\mathbb{M} = \mathbb{Z}^D$, then the standard metric on $\mathcal{A}^{\mathbb{Z}^D}$ is defined $d(\mathbf{a}, \mathbf{b}) = 2^{-\Delta(\mathbf{a}, \mathbf{b})}$, where $\Delta(\mathbf{a}, \mathbf{b}) := \min \{|z|; a_z \neq b_z\}$.

Any $\mathbf{v} \in \mathbb{M}$, determines a continuous *shift map* $\sigma^{\mathbf{v}}: \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{A}^{\mathbb{M}}$ defined by $\sigma^{\mathbf{v}}(\mathbf{a})_m = a_{m+\mathbf{v}}$ for all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $m \in \mathbb{M}$. The set $\{\sigma^{\mathbf{v}}\}_{\mathbf{v} \in \mathbb{M}}$ is then a continuous \mathbb{M} -action on $\mathcal{A}^{\mathbb{M}}$, which we denote simply by “ σ ”.

If $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $\mathbb{U} \subset \mathbb{M}$, then we define $\mathbf{a}_{\mathbb{U}} \in \mathcal{A}^{\mathbb{U}}$ by $\mathbf{a}_{\mathbb{U}} := [a_u]_{u \in \mathbb{U}}$. If $m \in \mathbb{M}$, then strictly speaking, $\mathbf{a}_{m+\mathbb{U}} \in \mathcal{A}^{m+\mathbb{U}}$; however, it will often be convenient to ‘abuse notation’ and treat $\mathbf{a}_{m+\mathbb{U}}$ as an element of $\mathcal{A}^{\mathbb{U}}$ in the obvious way.

Cellular automata Let $\mathbb{H} \subset \mathbb{M}$ be some finite subset, and let $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ be a function (called a *local rule*). The *cellular automaton* (CA) determined by ϕ is the function $\Phi: \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{A}^{\mathbb{M}}$ defined by $\Phi(\mathbf{a})_m = \phi(\mathbf{a}_{m+\mathbb{H}})$ for all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ and $m \in \mathbb{M}$. Curtis, Hedlund and Lyndon showed that cellular automata are exactly the continuous transformations of $\mathcal{A}^{\mathbb{M}}$ which commute with all shifts (see Theorem 3.4 in [58]). We refer to \mathbb{H} as the *neighborhood* of Φ . For example, if $\mathbb{M} = \mathbb{Z}$, then typically $\mathbb{H} := [-\ell \dots r] := \{-\ell, 1-\ell, \dots, r-1, r\}$ for some *left radius* $\ell \geq 0$ and *right radius* $r \geq 0$. If $-\ell \geq 0$, then ϕ can either define CA on $\mathcal{A}^{\mathbb{N}}$ or define a *one-sided* CA on $\mathcal{A}^{\mathbb{Z}}$. If $\mathbb{M} = \mathbb{Z}^D$, then typically

$\mathbb{H} \subseteq [-R \dots R]^D$, for some *radius* $R \geq 0$. Normally we assume that ℓ , r , and R are chosen to be minimal. Several specific classes of CA will be important to us:

Linear CA Let $(\mathcal{A}, +)$ be a finite abelian group (e. g. $\mathcal{A} = \mathbb{Z}/p$, where $p \in \mathbb{N}$; usually p is prime). Then Φ is a *linear CA* (LCA) if the local rule ϕ has the form

$$\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{h \in \mathbb{H}} \varphi_h(a_h), \quad \forall \mathbf{a}_{\mathbb{H}} \in \mathcal{A}^{\mathbb{H}}, \quad (1)$$

where $\varphi_h: \mathcal{A} \rightarrow \mathcal{A}$ is an endomorphism of $(\mathcal{A}, +)$, for each $h \in \mathbb{H}$. We say that Φ has *scalar coefficients* if, for each $h \in \mathbb{H}$, there is some scalar $c_h \in \mathbb{Z}$, so that $\varphi_h(a_h) := c_h \cdot a_h$; then $\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{h \in \mathbb{H}} c_h a_h$. For example, if $\mathcal{A} = (\mathbb{Z}/p, +)$, then *all* endomorphisms are scalar multiplications, so *all* LCA have scalar coefficients.

If $c_h = 1$ for all $h \in \mathbb{H}$, then Φ has local rule $\phi(\mathbf{a}_{\mathbb{H}}) := \sum_{h \in \mathbb{H}} a_h$; in this case, Φ is called an *additive cellular automaton*; see ► [Additive Cellular Automata](#).

Affine CA If $(\mathcal{A}, +)$ is a finite abelian group, then an *affine CA* is one with a local rule $\phi(\mathbf{a}_{\mathbb{H}}) := c + \sum_{h \in \mathbb{H}} \varphi_h(a_h)$, where c is some constant and where $\varphi_h: \mathcal{A} \rightarrow \mathcal{A}$ are endomorphisms of $(\mathcal{A}, +)$. Thus, Φ is an LCA if $c = 0$.

Permutative CA Suppose $\Phi: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$ has local rule $\phi: \mathcal{A}^{[-\ell \dots r]} \rightarrow \mathcal{A}$. Fix $\mathbf{b} = [b_{1-\ell}, \dots, b_{r-1}, b_r] \in \mathcal{A}^{(-\ell \dots r]}$. For any $a \in \mathcal{A}$, define $[a \ \mathbf{b}] := [a, b_{1-\ell}, \dots, b_{r-1}, b_r] \in \mathcal{A}^{[-\ell \dots r]}$. We then define the function $\phi_{\mathbf{b}}: \mathcal{A} \rightarrow \mathcal{A}$ by $\phi_{\mathbf{b}}(a) := \phi([a \ \mathbf{b}])$. We say that Φ is *left-permutative* if $\phi_{\mathbf{b}}: \mathcal{A} \rightarrow \mathcal{A}$ is a permutation (i. e. a bijection) for all $\mathbf{b} \in \mathcal{A}^{(-\ell \dots r]}$. Likewise, given $\mathbf{b} = [b_{-\ell}, \dots, b_{r-1}] \in \mathcal{A}^{[-\ell \dots r)}$ and $c \in \mathcal{A}$, define $[\mathbf{b} \ c] := [b_{-\ell}, b_{1-\ell}, \dots, b_{r-1}, c] \in \mathcal{A}^{[-\ell \dots r]}$, and define ${}_{\mathbf{b}}\phi: \mathcal{A} \rightarrow \mathcal{A}$ by ${}_{\mathbf{b}}\phi(c) := \phi([\mathbf{b} \ c])$; then Φ is *right-permutative* if ${}_{\mathbf{b}}\phi: \mathcal{A} \rightarrow \mathcal{A}$ is a permutation for all $\mathbf{b} \in \mathcal{A}^{[-\ell \dots r)}$. We say Φ is *bipermutative* if it is both left- and right-permutative. More generally, if \mathbb{M} is any monoid, $\mathbb{H} \subset \mathbb{M}$ is any neighborhood, and $h \in \mathbb{H}$ is any fixed coordinate, then we define h -permutativity for a CA on $\mathcal{A}^{\mathbb{M}}$ in the obvious fashion.

For example, suppose $(\mathcal{A}, +)$ is an abelian group and Φ is an affine CA on $\mathcal{A}^{\mathbb{Z}}$ with local rule $\phi(\mathbf{a}_{\mathbb{H}}) = c + \sum_{h=-\ell}^r \varphi_h(a_h)$. Then Φ is left-permutative iff $\varphi_{-\ell}$ is an automorphism, and right-permutative iff φ_r is an automorphism. If $\mathcal{A} = \mathbb{Z}/p$, and p is prime, then *every* nontrivial endomorphism is an automorphism (because it is

multiplication by a nonzero element of \mathbb{Z}/p , which is a field), so in this case, *every* affine CA is permutative in every coordinate of its neighborhood (and in particular, bipermutative). If $\mathcal{A} \neq \mathbb{Z}/p$, however, then not all affine CA are permutative.

Permutative CA were introduced by Hedlund [58], §6, and are sometimes called *permutive CA*. Right permutative CA on $\mathcal{A}^{\mathbb{N}}$ are also called *toggle automata*. For more information, see Sect. 7 of

► [Topological Dynamics of Cellular Automata](#).

Subshifts A *subshift* is a closed, σ -invariant subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$. For any $\mathbb{U} \subset \mathbb{M}$, let $\mathbf{X}_{\mathbb{U}} := \{\mathbf{x}_{\mathbb{U}}; \mathbf{x} \in \mathbf{X}\} \subset \mathcal{A}^{\mathbb{U}}$. We say \mathbf{X} is a *subshift of finite type* (SFT) if there is some finite $\mathbb{U} \subset \mathbb{M}$ such that \mathbf{X} is entirely described by $\mathbf{X}_{\mathbb{U}}$, in the sense that $\mathbf{X} = \{\mathbf{x} \in \mathcal{A}^{\mathbb{M}}; \mathbf{x}_{\mathbb{U}+m} \in \mathbf{X}_{\mathbb{U}}, \forall m \in \mathbb{M}\}$.

In particular, if $\mathbb{M} = \mathbb{Z}$, then a (two-sided) *Markov subshift* is an SFT $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ determined by a set $\mathbf{X}_{\{0,1\}} \subset \mathcal{A}^{\{0,1\}}$ of *admissible transitions*; equivalently, \mathbf{X} is the set of all bi-infinite directed paths in a digraph whose vertices are the elements of \mathcal{A} , with an edge $a \leadsto b$ iff $(a, b) \in \mathbf{X}_{\{0,1\}}$. If $\mathbb{M} = \mathbb{N}$, then a *one-sided Markov subshift* is a subshift of $\mathcal{A}^{\mathbb{N}}$ defined in the same way.

If $D \geq 2$, then an SFT in $\mathcal{A}^{\mathbb{Z}^D}$ can be thought of as the set of admissible ‘tilings’ of \mathbb{R}^D by Wang tiles corresponding to the elements of $\mathbf{X}_{\mathbb{U}}$. (Wang tiles are unit squares (or (hyper)cubes) with various ‘notches’ cut into their edges (or (hyper)faces) so that they can only be juxtaposed in certain ways.)

A subshift $\mathbf{X} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ is *strongly irreducible* (or *topologically mixing*) if there is some $R \in \mathbb{N}$ such that, for any disjoint finite subsets $\mathbb{V}, \mathbb{U} \subset \mathbb{Z}^D$ separated by a distance of at least R , and for any $\mathbf{u} \in \mathbf{X}_{\mathbb{U}}$ and $\mathbf{v} \in \mathbf{X}_{\mathbb{V}}$, there is some $\mathbf{x} \in \mathbf{X}$ such that $\mathbf{x}_{\mathbb{U}} = \mathbf{u}$ and $\mathbf{x}_{\mathbb{V}} = \mathbf{v}$. Please see ► [Symbolic Dynamics](#) for more about subshifts.

Measures For any finite subset $\mathbb{U} \subset \mathbb{M}$, and any $\mathbf{b} \in \mathcal{A}^{\mathbb{U}}$, let $\langle \mathbf{b} \rangle := \{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}; \mathbf{a}_{\mathbb{U}} := \mathbf{b}\}$ be the *cylinder set* determined by \mathbf{b} . Let \mathfrak{B} be the sigma-algebra on $\mathcal{A}^{\mathbb{M}}$ generated by all cylinder sets. A (probability) measure μ on $\mathcal{A}^{\mathbb{M}}$ is a countably additive function $\mu: \mathfrak{B} \rightarrow [0, 1]$ such that $\mu[\mathcal{A}^{\mathbb{M}}] = 1$. A measure on $\mathcal{A}^{\mathbb{M}}$ is entirely determined by its values on cylinder sets. We will be mainly concerned with the following classes of measures:

Bernoulli measure Let β_0 be a probability measure on \mathcal{A} . The *Bernoulli measure* induced by β_0 is the measure β on $\mathcal{A}^{\mathbb{M}}$ such that, for any finite subset $\mathbb{U} \subset \mathbb{M}$, and any $\mathbf{a} \in \mathcal{A}^{\mathbb{U}}$, if $U := |\mathbb{U}|$, then $\beta[\langle \mathbf{a} \rangle] = \prod_{h \in \mathbb{H}} \beta_0(a_h)$.

Invariant measure Let μ be a measure on $\mathcal{A}^{\mathbb{M}}$, and let $\Phi: \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{A}^{\mathbb{M}}$ be a cellular automaton. The measure $\Phi\mu$ is defined by $\Phi\mu(\mathbf{B}) = \mu(\Phi^{-1}(\mathbf{B}))$, for any $\mathbf{B} \in \mathfrak{B}$. We say that μ is Φ -invariant (or that Φ is μ -preserving) if $\Phi\mu = \mu$. For more information, see ► [Ergodic Theory: Basic Examples and Constructions](#).

Uniform measure Let $A := |\mathcal{A}|$. The uniform measure η on $\mathcal{A}^{\mathbb{M}}$ is the Bernoulli measure such that, for any finite subset $U \subset \mathbb{M}$, and any $\mathbf{b} \in \mathcal{A}^U$, if $U := |U|$, then $\mu[\langle \mathbf{b} \rangle] = 1/A^U$.

The *support* of a measure μ is the smallest closed subset $X \subset \mathcal{A}^{\mathbb{M}}$ such that $\mu[X] = 1$; we denote this by $\text{supp}(\mu)$. We say μ has *full support* if $\text{supp}(\mu) = \mathcal{A}^{\mathbb{M}}$ – equivalently, $\mu[C] > 0$ for every cylinder subset $C \subset \mathcal{A}^{\mathbb{M}}$.

Notation Let $\mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ denote the set of all cellular automata on $\mathcal{A}^{\mathbb{M}}$. If $X \subset \mathcal{A}^{\mathbb{M}}$, then let $\mathcal{CA}(X)$ be the subset of all $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ such that $\Phi(X) \subseteq X$. Let $\mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}})$ be the set of all probability measures on $\mathcal{A}^{\mathbb{M}}$, and let $\mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}; \Phi)$ be the subset of Φ -invariant measures. If $X \subset \mathcal{A}^{\mathbb{M}}$, then let $\mathfrak{M}_{\text{ens}}(X)$ be the set of probability measures μ with $\text{supp}(\mu) \subseteq X$, and define $\mathfrak{M}_{\text{ens}}(X; \Phi)$ in the obvious way.

Font conventions Upper case calligraphic letters ($\mathcal{A}, \mathcal{B}, \mathcal{C}, \dots$) denote finite alphabets or groups. Upper-case bold letters ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \dots$) denote subsets of $\mathcal{A}^{\mathbb{M}}$ (e.g. subshifts), lowercase bold-faced letters ($\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$) denote elements of $\mathcal{A}^{\mathbb{M}}$, and Roman letters (a, b, c, \dots) are elements of \mathcal{A} or ordinary numbers. Lower-case sans-serif (\dots, m, n, p) are elements of \mathbb{M} , upper-case hollow font ($\mathbb{U}, \mathbb{V}, \mathbb{W}, \dots$) are subsets of \mathbb{M} . Upper-case Greek letters (Φ, Ψ, \dots) are functions on $\mathcal{A}^{\mathbb{M}}$ (e.g. CA, block maps), and lower-case Greek letters (ϕ, ψ, \dots) are other functions (e.g. local rules, measures.)

Acronyms in square brackets (e.g. ► [Topological Dynamics of Cellular Automata](#)) indicate cross-references to related entries in the Encyclopedia; these are listed at the end of this article.

Definition of the Subject

Loosely speaking, a *cellular automaton* (CA) is the ‘discrete’ analogue of a partial differential evolution equation: it is a spatially distributed, discrete-time, symbolic dynamical system governed by a local interaction rule which is invariant in space and time. In a CA, ‘space’ is discrete (usually the D -dimensional lattice, \mathbb{Z}^D) and the local statespace at each point in space is also discrete (a finite ‘alphabet’, usually denoted by \mathcal{A}).

A *measure-preserving dynamical system* (MPDS) is a dynamical system equipped with an invariant probability measure. Any MPDS can be represented as a stationary stochastic process (SSP) and vice versa; ‘chaos’ in the MPDS can be quantified via the information-theoretic ‘entropy’ of the corresponding SSP. An MPDS Φ on a statespace X also defines a unitary linear operator Φ_* on the Hilbert space $L^2(X)$; the spectral properties of Φ_* encode information about the global periodic structure and long-term informational asymptotics of Φ . *Ergodic theory* is the study of MPDSs and SSPs, and lies at the interface between dynamics, probability theory, information theory, and unitary operator theory.

Please refer to the Glossary for precise definitions of ‘CA’, ‘MPDS’, etc. Also, see ► [Ergodic Theory: Basic Examples and Constructions](#) for an introduction to ergodic theory.

Introduction

The study of CA as symbolic dynamical systems began with Hedlund [58], and the study of CA as MPDSs began with Coven and Paul [24] and Willson [144]. (Further historical details will unfold below, where appropriate.) The ergodic theory of CA is important for several reasons:

- CA are topological dynamical systems (► [Topological Dynamics of Cellular Automata](#), ► [Chaotic Behavior of Cellular Automata](#)). We can gain insight into the topological dynamics of a CA by identifying its invariant measures, and then studying the corresponding measurable dynamics.
- CA are often proposed as stylized models of spatially distributed systems in statistical physics – for example, as microscale models of hydrodynamics, or of atomic lattices (► [Cellular Automata Modeling of Physical Systems](#)). In this context, the distinct invariant measures of a CA correspond to distinct ‘phases’ of the physical system (► [Phase Transitions in Cellular Automata](#)).
- CA can also act as information-processing systems (► [Cellular Automata, Universality of](#), ► [Cellular Automata as Models of Parallel Computation](#)). Ergodic theory studies the ‘informational’ aspect of dynamical systems, so it is particularly suited to explicitly ‘informational’ dynamical systems like CA.

Article Roadmap

In Sect. “[Invariant Measures for CA](#)”, we characterize the invariant measures for various classes of CA. Then, in

Sect. “[Limit Measures and Other Asymptotics](#)”, we investigate which measures are ‘generic’ in the sense that they arise as the attractors for some large class of initial conditions. In Sect. “[Measurable Dynamics](#)” we study the mixing and spectral properties of CA as measure-preserving dynamical systems. Finally, in Sect. “[Entropy](#)”, we look at entropy. These sections are logically independent, and can be read in any order.

Invariant Measures for CA

The Uniform Measure vs. Surjective Cellular Automata

The uniform measure η plays a central role in the ergodic theory of cellular automata, because of the following result.

Theorem 1 *Let $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$, let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$ and let η be the uniform measure on $\mathcal{A}^{\mathbb{M}}$. Then $(\Phi \text{ preserves } \eta) \iff (\Phi \text{ is surjective})$.*

Proof sketch “ \implies ” If Φ preserves η , then Φ must map $\text{supp}(\eta)$ onto itself. But $\text{supp}(\eta) = \mathcal{A}^{\mathbb{M}}$; hence Φ is surjective.

“ \impliedby ” The case $D = 1$ follows from a result of W.A. Blankenship and Oscar S. Rothaus, which first appeared in Theorem 5.4 in [58]. The Blankenship–Rothaus Theorem states that, if $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ is surjective and has neighborhood $[-\ell \dots r]$, then for any $k \in \mathbb{N}$ and any $\mathbf{a} \in \mathcal{A}^k$, the Φ -preimage of the cylinder set $\langle \mathbf{a} \rangle$ is a disjoint union of exactly $A^{r+\ell}$ cylinder sets of length $k + r + \ell$; it follows that $\mu[\Phi^{-1}(\langle \mathbf{a} \rangle)] = A^{r+\ell}/A^{k+r+\ell} = A^{-k} = \mu\langle \mathbf{a} \rangle$. This result was later reproved by Kleaveland (see Theorem 5.1 in [74]). The special case $\mathcal{A} = \{0, 1\}$ also appeared in Theorem 2.4 in [131].

The case $D \geq 2$ follows from the multidimensional version of the Blankenship–Rothaus Theorem, which was proved by Maruoka and Kimura (see Theorem 2 in [93]) (their proof assumes that $D = 2$ and that Φ has a ‘quiescent’ state, but neither hypothesis is essential). Alternatively, “ \impliedby ” follows from recent, more general results of Meester, Burton, and Steif; see Example 9 below. \square

Example 2 Let $\mathbb{M} = \mathbb{Z}$ or \mathbb{N} and consider CA on $\mathcal{A}^{\mathbb{M}}$.

- (a) Say that Φ is *bounded-to-one* if there is some $B \in \mathbb{N}$ such that every $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ has at most B preimages. Then $(\Phi \text{ is bounded-to-one}) \iff (\Phi \text{ is surjective})$.
- (b) Any posexpansive CA on $\mathcal{A}^{\mathbb{M}}$ is surjective (see Subsect. “[Posexpansive and Permutative CA](#)” below).
- (c) Any left- or right-permutative CA on $\mathcal{A}^{\mathbb{Z}}$ (or right-permutative CA on $\mathcal{A}^{\mathbb{N}}$) is surjective. This includes, for example, most linear CA.

Hence, in any of these cases, Φ preserves the uniform measure.

Proof For (a), see Theorem 5.9 in [58], or Corollary 8.1.20, p. 271 in [81]. For (b), see Proposition 2.2 in [9], in the case $\mathcal{A}^{\mathbb{N}}$; their argument also works for $\mathcal{A}^{\mathbb{Z}}$.

Part (c) follows from (b) because any permutative CA is posexpansive (Proposition 11 below). There is also a simple direct proof for a right-permutative CA on $\mathcal{A}^{\mathbb{N}}$: using right-permutativity, you can systematically construct a preimage of any desired image sequence, one entry at a time. See Theorem 6.6 in [58] for the proof in $\mathcal{A}^{\mathbb{Z}}$. \square

The surjectivity of a one-dimensional CA can be determined in finite time using certain combinatorial tests ([► Topological Dynamics of Cellular Automata](#)). However, for $D \geq 2$, it is formally undecidable whether an arbitrary CA on $\mathcal{A}^{\mathbb{Z}^D}$ is surjective ([► Tiling Problem and Undecidability in Cellular Automata](#)). This problem is sometimes referred to as the *Garden of Eden problem*, because an element of $\mathcal{A}^{\mathbb{Z}^D}$ with no Φ -preimage is called a *Garden of Eden* (GOE) configuration for Φ (because it could only ever occur at the ‘beginning of time’). However, it is known that a CA is surjective if it is ‘almost injective’ in a certain sense, which we now specify.

Let $(\mathbb{M}, +)$ be any monoid, and let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$ have neighborhood $\mathbb{H} \subset \mathbb{M}$. If $\mathbb{B} \subset \mathbb{M}$ is any subset, then we define

$$\begin{aligned} \overline{\mathbb{B}} &:= \mathbb{B} + \mathbb{H} = \{\mathbf{b} + \mathbf{h}; \mathbf{b} \in \mathbb{B}, \mathbf{h} \in \mathbb{H}\}; \\ \text{and } \partial\mathbb{B} &:= \overline{\mathbb{B}} \cap \overline{\mathbb{B}^c}. \end{aligned}$$

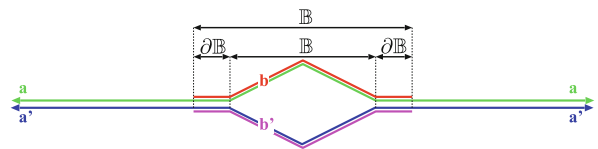
If \mathbb{B} is finite, then so is $\overline{\mathbb{B}}$ (because \mathbb{H} is finite). If Φ has local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$, then ϕ induces a function $\Phi_{\mathbb{B}}: \mathcal{A}^{\overline{\mathbb{B}}} \rightarrow \mathcal{A}^{\mathbb{B}}$ in the obvious fashion. A \mathbb{B} -*bubble* (or \mathbb{B} -*diamond*) is a pair $\mathbf{b}, \mathbf{b}' \in \mathcal{A}^{\overline{\mathbb{B}}}$ such that:

$$\mathbf{b} \neq \mathbf{b}'; \quad \mathbf{b}|_{\partial\mathbb{B}} = \mathbf{b}'|_{\partial\mathbb{B}}; \quad \text{and} \quad \Phi_{\mathbb{B}}(\mathbf{b}) = \Phi_{\mathbb{B}}(\mathbf{b}').$$

Suppose $\mathbf{a}, \mathbf{a}' \in \mathcal{A}^{\mathbb{M}}$ are two configurations such that

$$\mathbf{a}|_{\overline{\mathbb{B}}} = \mathbf{b}, \quad \mathbf{a}'|_{\overline{\mathbb{B}}} = \mathbf{b}', \quad \text{and} \quad \mathbf{a}|_{\mathbb{B}^c} = \mathbf{a}'|_{\mathbb{B}^c}.$$

Then it is easy to verify that $\Phi(\mathbf{a}) = \Phi(\mathbf{a}')$. We say that \mathbf{a} and \mathbf{a}' form a *mutually erasable pair* (because Φ ‘erases’ the difference between \mathbf{a} and \mathbf{a}'). Figure 1 is a schematic representation of this structure in the case $D = 1$ (hence



Ergodic Theory of Cellular Automata, Figure 1
A ‘diamond’ in $\mathcal{A}^{\mathbb{Z}}$

the term ‘diamond’). If $D = 2$, then \mathbf{a} and \mathbf{a}' are like two membranes which are glued together everywhere except for a \mathbb{B} -shaped ‘bubble’. We say that Φ is *pre-injective* if any (and thus, all) of the following three conditions hold:

- Φ admits no bubbles.
- Φ admits no mutually erasable pairs.
- For any $\mathbf{c} \in \mathcal{A}^{\mathbb{M}}$, if $\mathbf{a}, \mathbf{a}' \in \Phi^{-1}\{\mathbf{c}\}$ are distinct, then \mathbf{a} and \mathbf{a}' must differ in infinitely many locations.

For example, any injective CA is preinjective (because a mutually erasable pair for Φ gives two distinct Φ -preimages for some point). More to the point, however, if \mathbb{B} is finite, and Φ admits a \mathbb{B} -bubble $(\mathbf{b}, \mathbf{b}')$, then we can embed N disjoint copies of \mathbb{B} into \mathbb{M} , and thus, by making various choices between \mathbf{b} and \mathbf{b}' on different translates, we obtain a configuration with 2^N distinct Φ -preimages (where N is arbitrarily large). But if some configurations in $\mathcal{A}^{\mathbb{M}}$ have such a large number of preimages, then other configurations in $\mathcal{A}^{\mathbb{M}}$ must have very few preimages, or even none. This leads to the following result:

Theorem 3 (Garden of Eden) *Let \mathbb{M} be a finitely generated amenable group (e.g. $\mathbb{M} = \mathbb{Z}^D$). Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$.*

- (a) *Φ is surjective if and only if Φ is pre-injective.*
 (b) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be a strongly irreducible SFT such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. Then $\Phi(\mathbf{X}) = \mathbf{X}$ if and only if $\Phi|_{\mathbf{X}}$ is pre-injective.*

Proof (a) The case $\mathbb{M} = \mathbb{Z}^2$ was originally proved by Moore [103] and Myhill [104]; see ► [Cellular Automata and Groups](#). The case $\mathbb{M} = \mathbb{Z}$ was implicit Hedlund (Lemma 5.11, and Theorems 5.9 and 5.12 in [58]). The case when \mathbb{M} is a finite-dimensional group was proved by Machi and Mignosi [92]. Finally, the general case was proved by Ceccherini-Silberstein, Machi, and Scarabotti (see Theorem 3 in [20]), see ► [Cellular Automata and Groups](#).

(b) The case $\mathbb{M} = \mathbb{Z}$ is Corollary 8.1.20 in [81] (actually this holds for any sofic subshift); see also Fiorenzi [38]. The general case is Corollary 4.8 in [39]. \square

Corollary 4 (Incompressibility) *Suppose \mathbb{M} is a finitely generated amenable group and $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$. If Φ is injective, then Φ is surjective.*

Remark 5 (a) A *cellular network* is a CA-like system defined on an infinite, locally finite digraph, with different local rules at different nodes. By assuming a kind of ‘amenability’ for this digraph, and then imposing some weak global statistical symmetry conditions on the local rules, Gromov (see Theorem 8.F’ in [51]) has generalized the GOE Theorem 3 to a large class of such cellular net-

works (which he calls ‘endomorphisms of symbolic algebraic varieties’). See also [19].

(b) In the terminology suggested by Gottschalk [46], Incompressibility Corollary 4 says that the group \mathbb{M} is *surjunctive*; Gottschalk claims that ‘surjunctivity’ was first proved for all residually finite groups by Lawton (unpublished); see ► [Cellular Automata and Groups](#). For a recent direct proof (not using the GOE theorem), see Weiss (see Theorem 1.6 in [143]). Weiss also defines *sofic* groups (a class containing both residually finite groups and amenable groups) and shows that Corollary 4 holds whenever \mathbb{M} is a sofic group (see Theorem 3.2 in [143]); see also ► [Cellular Automata and Groups](#).

(c) If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is an SFT such that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$, then Corollary 4 holds as long as \mathbf{X} is ‘semi-strongly irreducible’; see Fiorenzi (see Corollary 4.10 in [40]).

Invariance of Maxentropy Measures

If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is any subshift with topological entropy $h_{\text{top}}(\mathbf{X}, \sigma)$, and $\mu \in \mathfrak{M}_{\text{inv}}(\mathbf{X}, \sigma)$ has measurable entropy $h(\mu, \sigma)$, then in general, $h(\mu, \sigma) \leq h_{\text{top}}(\mathbf{X}, \sigma)$; we say μ is a *measure of maximal entropy* (or *maxentropy measure*) if $h(\mu, \sigma) = h_{\text{top}}(\mathbf{X}, \sigma)$. (See Example 75(a) for definitions.)

Every subshift admits one or more maxentropy measures. If $D = 1$ and $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible subshift of finite type (SFT), then Parry (see Theorem 10 in [110]) showed that \mathbf{X} admits a *unique* maxentropy measure $\eta_{\mathbf{X}}$ (now called the *Parry measure*); see Theorem 8.10, p. 194 in [142] or Sect. 13.3, pp. 443–444 in [81]. Theorem 1 is then a special case of the following result:

Theorem 6 (Coven, Paul, Meester and Steif) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be an SFT having a unique maxentropy measure $\eta_{\mathbf{X}}$, and let $\Phi \in \mathcal{CA}(\mathbf{X})$. Then Φ preserves $\eta_{\mathbf{X}}$ if and only if $\Phi(\mathbf{X}) = \mathbf{X}$.*

Proof The case $D = 1$ is Corollary 2.3 in [24]. The case $D \geq 2$ follows from Theorem 2.5(iii) in [95], which states: if \mathbf{X} and \mathbf{Y} are SFTs, and $\Phi: \mathbf{X} \rightarrow \mathbf{Y}$ is a factor mapping, and μ is a maxentropy measure on \mathbf{X} , then $\Phi(\mu)$ is a maxentropy measure on \mathbf{Y} . \square

For example, if $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible SFT and $\eta_{\mathbf{X}}$ is its Parry measure, and $\Phi(\mathbf{X}) = \mathbf{X}$, then Theorem 6 says $\Phi(\eta_{\mathbf{X}}) = \eta_{\mathbf{X}}$, as observed by Coven and Paul (see Theorem 5.1 in [24]). Unfortunately, higher-dimensional SFTs do *not*, in general, have unique maxentropy measures. Burton and Steif [14] provided a plethora of examples of such nonuniqueness, but they also gave a sufficient condition for uniqueness of the maxentropy measure, which we now explain.

Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be an SFT and let $\mathbb{U} \subset \mathbb{Z}^D$. For any $\mathbf{x} \in \mathbf{X}$, let $\mathbf{x}_{\mathbb{U}} := [x_u]_{u \in \mathbb{U}}$ be its ‘projection’ to $\mathcal{A}^{\mathbb{U}}$, and let $\mathbf{X}_{\mathbb{U}} := \{\mathbf{x}_{\mathbb{U}}; \mathbf{x} \in \mathbf{X}\} \subseteq \mathcal{A}^{\mathbb{U}}$. Let $\mathbb{V} := \mathbb{U}^c \subset \mathbb{Z}^D$. For any $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$ and $\mathbf{v} \in \mathcal{A}^{\mathbb{V}}$, let $[\mathbf{uv}]$ denote the element of $\mathcal{A}^{\mathbb{Z}^D}$ such that $[\mathbf{uv}]_{\mathbb{U}} = \mathbf{u}$ and $[\mathbf{uv}]_{\mathbb{V}} = \mathbf{v}$. Let

$$\mathbf{X}^{(\mathbf{u})} := \{\mathbf{v} \in \mathcal{A}^{\mathbb{V}}; [\mathbf{uv}] \in \mathbf{X}\}$$

be the set of all “ \mathbf{X} -admissible completions” of \mathbf{u} (thus, $\mathbf{X}^{(\mathbf{u})} \neq \emptyset \Leftrightarrow \mathbf{u} \in \mathbf{X}_{\mathbb{U}}$). If $\mu \in \mathcal{M}_{\text{inv}}(\mathcal{A}^{\mathbb{Z}^D})$, and $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$, then let $\mu^{(\mathbf{u})}$ denote the *conditional measure* on $\mathcal{A}^{\mathbb{V}}$ induced by \mathbf{u} . If \mathbb{U} is finite, then $\mu^{(\mathbf{u})}$ is just the restriction of μ to the cylinder set $\langle \mathbf{u} \rangle$. If \mathbb{U} is infinite, then the precise definition of $\mu^{(\mathbf{u})}$ involves a ‘disintegration’ of μ into ‘fibre measures’ (we will suppress the details).

Let $\mu_{\mathbb{U}}$ be the projection of μ onto $\mathcal{A}^{\mathbb{U}}$. If $\text{supp}(\mu) \subseteq \mathbf{X}$, then $\text{supp}(\mu_{\mathbb{U}}) \subseteq \mathbf{X}_{\mathbb{U}}$, and for any $\mathbf{u} \in \mathcal{A}^{\mathbb{U}}$, $\text{supp}(\mu^{(\mathbf{u})}) \subseteq \mathbf{X}^{(\mathbf{u})}$. We say that μ is a *Burton–Steif measure* on \mathbf{X} if:

- (1) $\text{supp}(\mu) = \mathbf{X}$; and
- (2) For any $\mathbb{U} \subset \mathbb{Z}^D$ whose complement \mathbb{U}^c is finite, and for $\mu_{\mathbb{U}}$ -almost any $\mathbf{u} \in \mathbf{X}_{\mathbb{U}}$, the measure $\mu^{(\mathbf{u})}$ is uniformly distributed on the (finite) set $\mathbf{X}^{(\mathbf{u})}$.

For example, if $\mathbf{X} = \mathcal{A}^{\mathbb{Z}^D}$, then the only Burton–Steif measure is the uniform Bernoulli measure. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}$ is an irreducible SFT, then the only Burton–Steif measure is the Parry measure. If $r > 0$ and $\mathbb{B} := [-r \dots r]^D \subset \mathbb{Z}^D$, and \mathbf{X} is an SFT determined by a set of admissible words $\mathbf{X}_{\mathbb{B}} \subset \mathcal{A}^{\mathbb{B}}$, then it is easy to check that any Burton–Steif measure μ on \mathbf{X} must be a Markov random field with interaction range r .

Theorem 7 (Burton and Steif) *Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ be a subshift of finite type.*

- (a) *Any maxentropy measure on \mathbf{X} is a Burton–Steif measure.*
- (b) *If \mathbf{X} is strongly irreducible, then any Burton–Steif measure on \mathbf{X} is a maxentropy measure for \mathbf{X} .*

Proof (a) and (b) are Propositions 1.20 and 1.21 of [15], respectively. For a proof in the case when \mathbf{X} is a symmetric nearest-neighbor subshift of finite type, see Propositions 1.19 and 4.1 of [14], respectively. \square

Any subshift admits at least one maxentropy measure, so any SFT admits at least one Burton–Steif measure. Theorems 6 and 7 together imply:

Corollary 8 *If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is an SFT which admits a unique Burton–Steif measure $\eta_{\mathbf{X}}$, then $\eta_{\mathbf{X}}$ is the unique maxentropy measure for \mathbf{X} . Thus, if $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ and $\Phi(\mathbf{X}) = \mathbf{X}$, then $\Phi(\eta_{\mathbf{X}}) = \eta_{\mathbf{X}}$.*

Example 9 If $\mathbf{X} = \mathcal{A}^{\mathbb{Z}^D}$, then we get Theorem 1, because the unique Burton–Steif measure on $\mathcal{A}^{\mathbb{Z}^D}$ is the uniform Bernoulli measure.

Remark If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift admitting a unique maxentropy measure μ , and $\text{supp}(\mu) = \mathbf{X}$, then Wiess (see Theorem 4.2 in [143]) has observed that \mathbf{X} automatically satisfies Incompressibility Corollary 4. In particular, this applies to any SFT having a unique Burton–Steif measure.

Periodic Invariant Measures

If $P \in \mathbb{N}$, then a sequence $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ is *P-periodic* if $\sigma^P(\mathbf{a}) = \mathbf{a}$. If $A := |\mathcal{A}|$, then there are exactly A^P such sequences, and a measure μ on $\mathcal{A}^{\mathbb{Z}}$ is called *P-periodic* if μ is supported entirely on these *P-periodic* sequences. More generally, if \mathbb{M} is any monoid and $\mathbb{P} \subset \mathbb{M}$ is any submonoid, then a configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ is *P-periodic* if $\sigma^{\mathbf{p}}(\mathbf{a}) = \mathbf{a}$ for all $\mathbf{p} \in \mathbb{P}$. (For example, if $\mathbb{M} = \mathbb{Z}$ and $\mathbb{P} := P\mathbb{Z}$, then the \mathbb{P} -periodic configurations are the *P-periodic* sequences). Let $\mathcal{A}^{\mathbb{M}/\mathbb{P}}$ denote the set of \mathbb{P} -periodic configurations. If $P := |\mathbb{M}/\mathbb{P}|$, then $|\mathcal{A}^{\mathbb{M}/\mathbb{P}}| = A^P$. A measure μ is called *P-periodic* if $\text{supp}(\mu) \subseteq \mathcal{A}^{\mathbb{M}/\mathbb{P}}$.

Proposition 10 *Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$. If $\mathbb{P} \subset \mathbb{M}$ is any submonoid and $|\mathbb{M}/\mathbb{P}|$ is finite, then there exists a \mathbb{P} -periodic, Φ -invariant measure.*

Proof sketch If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$, then $\Phi(\mathcal{A}^{\mathbb{M}/\mathbb{P}}) \subseteq \mathcal{A}^{\mathbb{M}/\mathbb{P}}$. Thus, if μ is \mathbb{P} -periodic, then $\Phi^t(\mu)$ is \mathbb{P} -periodic for all $t \in \mathbb{N}$. Thus, the Cesàro limit of the sequence $\{\Phi^t(\mu)\}_{t=1}^{\infty}$ is \mathbb{P} -periodic and Φ -invariant. This Cesàro limit exists because $\mathcal{A}^{\mathbb{M}/\mathbb{P}}$ is finite. \square

These periodic measures have finite (hence discrete) support, but by convex-combining them, it is easy to obtain (nonergodic) Φ -invariant measures with countable, dense support. When studying the invariant measures of CA, we usually regard these periodic measures (and their convex combinations) as somewhat trivial, and concentrate instead on invariant measures supported on aperiodic configurations.

Posexpansive and Permutative CA

Let $\mathbb{B} \subset \mathbb{M}$ be a finite subset, and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$, then we define a continuous function $\Phi_{\mathbb{B}}^{\mathbb{N}}: \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{B}^{\mathbb{N}}$ by

$$\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathbf{a}) := [\mathbf{a}_{\mathbb{B}}; \Phi(\mathbf{a})_{\mathbb{B}}; \Phi^2(\mathbf{a})_{\mathbb{B}}; \Phi^3(\mathbf{a})_{\mathbb{B}}; \dots] \in \mathcal{B}^{\mathbb{N}}. \quad (2)$$

Clearly, $\Phi_{\mathbb{B}}^{\mathbb{N}} \circ \Phi = \sigma \circ \Phi_{\mathbb{B}}^{\mathbb{N}}$. We say that Φ is *B-posexpansive* if $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is injective. Equivalently, for any $\mathbf{a}, \mathbf{a}' \in$

$\mathcal{A}^{\mathbb{M}}$, if $\mathbf{a} \neq \mathbf{a}'$, then there is some $t \in \mathbb{N}$ such that $\Phi^t(\mathbf{a})_{\mathbb{B}} \neq \Phi^t(\mathbf{a}')_{\mathbb{B}}$. We say Φ is *positively expansive* (or *posexpansive*) if Φ is \mathbb{B} -posexpansive for some finite \mathbb{B} (it is easy to see that this is equivalent to the usual definition of positive expansiveness a topological dynamical system). For more information see Sect. 8 ► [Topological Dynamics of Cellular Automata](#).

Thus, if $\mathbf{X} := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{M}}) \subset \mathcal{B}^{\mathbb{N}}$, then \mathbf{X} is a compact, shift-invariant subset of $\mathcal{B}^{\mathbb{N}}$, and $\Phi_{\mathbb{B}}^{\mathbb{N}}: \mathcal{A}^{\mathbb{M}} \rightarrow \mathbf{X}$ is an isomorphism from the system $(\mathcal{A}^{\mathbb{M}}, \Phi)$ to the one-sided subshift (\mathbf{X}, σ) , which is sometimes called the *canonical factor* or *column shift* of Φ . The easiest examples of posexpansive CA are one-dimensional, permutative automata.

Proposition 11 (a) Suppose $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{N}})$ has neighborhood $[r \dots R]$, where $0 \leq r < R$. Let $\mathbb{B} := [0 \dots R]$ and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. Then

$$(\Phi \text{ is right permutative}) \iff (\Phi \text{ is } \mathbb{B}\text{-posexpansive, and } \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{N}}) = \mathcal{B}^{\mathbb{N}}).$$

(b) Suppose $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ has neighborhood $[-L \dots R]$, where $-L < 0 < R$. Let $\mathbb{B} := [-L \dots R]$, and let $\mathcal{B} := \mathcal{A}^{\mathbb{B}}$. Then

$$(\Phi \text{ is bipermutative}) \iff (\Phi \text{ is } \mathbb{B}\text{-posexpansive, and } \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}}) = \mathcal{B}^{\mathbb{N}}).$$

Thus, one-sided, right-permutative CA and two-sided, bipermutative CA are both topologically conjugate to the one-sided full shift $(\mathcal{B}^{\mathbb{N}}, \sigma)$, where \mathcal{B} is an alphabet with $|\mathcal{A}|^{R+L}$ symbols (setting $L = 0$ in the one-sided case).

Proof Suppose $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ (where $\mathbb{M} = \mathbb{N}$ or \mathbb{Z}). Draw a picture of the spacetime diagram for Φ . For any $t \in \mathbb{N}$, and any $\mathbf{b} \in \mathcal{B}^{[0 \dots t]}$, observe how (bi)permutativity allows you to reconstruct a unique $\mathbf{a}_{[-tL \dots tR]} \in \mathcal{A}^{[-tL \dots tR]}$ such that $\mathbf{b} = (\mathbf{a}_{\mathbb{B}}, \Phi(\mathbf{a})_{\mathbb{B}}, \Phi^2(\mathbf{a})_{\mathbb{B}}, \dots, \Phi^{t-1}(\mathbf{a})_{\mathbb{B}})$. By letting $t \rightarrow \infty$, we see that the function $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is a bijection between $\mathcal{A}^{\mathbb{M}}$ and $\mathcal{B}^{\mathbb{N}}$. \square

Remark 12 (a) The idea of Proposition 11 is implicit in Theorem 6.7 in [58], but it was apparently first stated explicitly by Shereshevsky and Afraïmovich (see Theorem 1 in [130]). It was later rediscovered by Kleveland (see Corollary 7.3 in [74]) and Fagnani and Margara (see Theorem 3.2 in [35]).

(b) Proposition 11(b) has been generalized to higher dimensions by Allouche and Skordev (see Proposition 1 in [3]), who showed that, if $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ is permutative in the ‘corner’ entries of its neighborhood, then Φ is conjugate to a full shift $(\mathcal{K}^{\mathbb{N}}, \sigma)$, where \mathcal{K} is an uncountable, compact space.

Proposition 11 is quite indicative of the general case. Posexpansiveness occurs only in one-dimensional CA, in which it takes a very specific form. To explain this, suppose (\mathbb{M}, \cdot) is a group with finite generating set $\mathbb{G} \subset \mathbb{M}$. For any $r > 0$, let $\mathbb{B}(r) := \{g_1 \cdot g_2 \cdots g_r; g_1, \dots, g_r \in \mathbb{G}\}$. The *dimension* (or *growth degree*) of (\mathbb{M}, \cdot) is defined $\dim(\mathbb{M}, \cdot) := \limsup_{r \rightarrow \infty} \log |\mathbb{B}(r)| / \log(r)$; see [50] or [49]. It can be shown that this number is independent of the choice of generating set \mathbb{G} , and is always an integer. For example, $\dim(\mathbb{Z}^D, +) = D$. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then we define its topological entropy $h_{\text{top}}(\mathbf{X})$ with respect to $\dim(\mathbb{M})$ in the obvious fashion (see Example 75(a)).

Theorem 13 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$.

- (a) If $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$ with $D + E \geq 2$, then Φ cannot be posexpansive.
- (b) If \mathbb{M} is any group with $\dim(\mathbb{M}) \geq 2$, and $\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}$ is any subshift with $h_{\text{top}}(\mathbf{X}) > 0$, and $\Phi(\mathbf{X}) \subseteq \mathbf{X}$, then the system (\mathbf{X}, Φ) cannot be posexpansive.
- (c) Suppose $\mathbb{M} = \mathbb{Z}$ or \mathbb{N} , and Φ has neighborhood $[-L \dots R] \subset \mathbb{M}$. Let $\bar{L} := \max\{0, L\}$, $\bar{R} := \max\{0, R\}$ and $\mathbb{B} := [-\bar{L} \dots \bar{R}]$. If Φ is posexpansive, then Φ is \mathbb{B} -posexpansive.

Proof (a) is Corollary 2 in [127]; see also Theorem 4.4 in [37]. Part (b) follows by applying Theorem 1.1 in [128] to the natural extension of (\mathbf{X}, Φ) .

(c) The case $\mathbb{M} = \mathbb{Z}$ is Proposition 7 in [75]. The case $\mathbb{M} = \mathbb{N}$ is Proposition 2.3 in [9]. \square

Proposition 11 says bipermutative CA on $\mathcal{A}^{\mathbb{Z}}$ are conjugate to full shifts. Using his formidable theory of *textile systems*, Nasu extended this to *all* posexpansive CA on $\mathcal{A}^{\mathbb{Z}}$.

Theorem 14 (Nasu’s) Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mathbb{B} \subset \mathbb{Z}$. If Φ is \mathbb{B} -posexpansive, then $\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}}) \subseteq \mathcal{B}^{\mathbb{N}}$ is a one-sided SFT which is conjugate to a one-sided full shift $\mathcal{C}^{\mathbb{N}}$ for some alphabet \mathcal{C} with $|\mathcal{C}| \geq 3$.

Proof sketch The fact that $\mathbf{X} := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}})$ is an SFT follows from Theorem 10 in [75] or Theorem 10.1 in [76]. Next, Theorem 3.12(1) on p. 49 of [105] asserts that, if Φ is any surjective endomorphism of an irreducible, aperiodic, SFT $\mathbf{Y} \subseteq \mathcal{A}^{\mathbb{Z}}$, and (\mathbf{Y}, Φ) is itself conjugate to an SFT, then (\mathbf{Y}, Φ) is actually conjugate to a full shift $(\mathcal{C}^{\mathbb{N}}, \sigma)$ for some alphabet \mathcal{C} with $|\mathcal{C}| \geq 3$. Let $\mathbf{Y} := \mathcal{A}^{\mathbb{Z}}$ and invoke K urka’s result.

For a direct proof not involving textile systems, see Theorem 4.9 in [86]. \square

Remark 15 (a) See Theorem 60(d) for an ‘ergodic’ version of Theorem 14.

(b) In contrast to Proposition 11, Nasu's Theorem 14 does *not* say that $\Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{Z}})$ itself is a full shift – only that it is conjugate to one.

If $(X, \mu; \Psi)$ is a measure-preserving dynamical system (MPDS) with sigma-algebra \mathfrak{B} , then a *one-sided generator* is a finite partition $\mathfrak{P} \subset \mathfrak{B}$ such that $\bigvee_{i=0}^{\infty} \Psi^{-i} \mathfrak{P} \stackrel{\mu}{=} \mathfrak{B}$. If \mathfrak{P} has C elements, and C is a finite set with $|C| = C$, then \mathfrak{P} induces an essentially injective function $p: X \rightarrow C^{\mathbb{N}}$ such that $p \circ \Psi = \sigma \circ p$. Thus, if $\lambda := p(\mu)$, then $(X, \mu; \Psi)$ is measurably isomorphic to the (one-sided) stationary stochastic process $(C^{\mathbb{N}}, \lambda; \sigma)$. If Ψ is invertible, then a (two-sided) *generator* is a finite partition $\mathfrak{P} \subset \mathfrak{B}$ such that $\bigvee_{i=-\infty}^{\infty} \Psi^i \mathfrak{P} \stackrel{\mu}{=} \mathfrak{B}$. The Krieger Generator Theorem says every finite-entropy, invertible MPDS has a generator; indeed, if $h(\Psi, \mu) \leq \log_2(C)$, then $(X, \mu; \Psi)$ has a generator with C or less elements (► [Ergodic Theory: Basic Examples and Constructions](#)). If $|C| = C$, then once again, \mathfrak{P} induces a measurable isomorphism from $(X, \mu; \Psi)$ to a two-sided stationary stochastic process $(C^{\mathbb{Z}}, \lambda; \sigma)$, for some stationary measure λ on $\mathcal{A}^{\mathbb{Z}}$.

Corollary 16 (Universal Representation) *Let $\mathbb{M} = \mathbb{N}$ or \mathbb{Z} , and let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$ have neighborhood $\mathbb{H} \subset \mathbb{M}$. Suppose that*

- either $\mathbb{M} = \mathbb{N}$, Φ is right-permutative, and $\mathbb{H} = [r \dots R]$ for some $0 \leq r < R$, and then let $C := R \log_2 |\mathcal{A}|$;*
- or $\mathbb{M} = \mathbb{Z}$, Φ is bipermutative, and $\mathbb{H} = [-L \dots R]$, and then let $C := (\bar{L} + \bar{R}) \log_2 |\mathcal{A}|$ where $\bar{L} := \max\{0, L\}$ and $\bar{R} := \max\{0, R\}$;*
- or $\mathbb{M} = \mathbb{Z}$ and Φ is positively expansive, and $h_{\text{top}}(\mathcal{A}^{\mathbb{M}}, \Phi) = \log_2(C)$ for some $C \in \mathbb{N}$.*

- (a) *Let $(X, \mu; \Psi)$ be any MPDS with a one-sided generator having at most C elements. Then there exists $\nu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$ such that the system $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is measurably isomorphic to $(X, \mu; \Psi)$.*
- (b) *Let $(X, \mu; \Psi)$ be an invertible MPDS, with measurable entropy $h(\mu, \phi) \leq \log_2(C)$. Then there exists $\nu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$ such that the natural extension of the system $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is measurably isomorphic to (X, μ, Ψ) .*

Proof Under each of the three hypotheses, Proposition 11 or Theorem 14 yields a topological conjugacy $\Gamma: (C^{\mathbb{N}}, \sigma) \rightarrow (\mathcal{A}^{\mathbb{M}}, \Phi)$, where C is a set of cardinality C .

(a) As discussed above, there is a measure λ on $C^{\mathbb{N}}$ such that $(C^{\mathbb{N}}, \lambda; \sigma)$ is measurably isomorphic to (X, μ, Ψ) . Thus, $\nu := \Gamma[\lambda]$ is a Φ -invariant measure on $\mathcal{A}^{\mathbb{M}}$, and $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to $(C^{\mathbb{N}}, \mu, \Psi)$ via Γ .

(b) As discussed above, there is a measure λ on $C^{\mathbb{Z}}$ such that $(C^{\mathbb{Z}}, \lambda; \sigma)$ is measurably isomorphic to (X, μ, Ψ) .

Let $\lambda_{\mathbb{N}}$ be the projection of λ to $C^{\mathbb{N}}$; then $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$ is a one-sided stationary process. Thus, $\nu := \Gamma[\lambda_{\mathbb{N}}]$ is a Φ -invariant measure on $\mathcal{A}^{\mathbb{M}}$, and $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$ via Γ . Thus, the natural extension of $(\mathcal{A}^{\mathbb{M}}, \nu; \Phi)$ is isomorphic to the natural extension of $(C^{\mathbb{N}}, \lambda_{\mathbb{N}}; \sigma)$, which is $(C^{\mathbb{Z}}, \lambda; \sigma)$, which is in turn isomorphic to $(X, \mu; \Psi)$. \square

Remark 17 The Universal Representation Corollary implies that studying the measurable dynamics of the CA Φ with respect to some arbitrary Φ -invariant measure ν will generally tell us nothing whatsoever about Φ . For these measurable dynamics to be meaningful, we must pick a measure on $\mathcal{A}^{\mathbb{M}}$ which is somehow ‘natural’ for Φ . First, this measure should be shift-invariant (because one of the defining properties of CA is that they commute with the shift). Second, we should seek a measure which has maximal Φ -entropy or is distinguished in some other way. (In general, the measures ν given by the Universal Representation Corollary will neither be σ -invariant, nor have maximal entropy for Φ .)

If $\Phi_{\mathbb{N}} \in \text{CA}(\mathcal{A}^{\mathbb{N}})$, and $\Phi_{\mathbb{Z}} \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ is the CA obtained by applying the same local rule to all coordinates in \mathbb{Z} , then $\Phi_{\mathbb{Z}}$ can *never* be posexpansive: if $\mathbb{B} = [-B \dots B]$, and $\mathbf{a}, \mathbf{a}' \in \mathcal{A}^{\mathbb{Z}}$ are any two sequences such that $\mathbf{a}_{(-\infty \dots -B)} \neq \mathbf{a}'_{(-\infty \dots -B)}$, then $\Phi^t(\mathbf{a})_{\mathbb{B}} = \Phi^t(\mathbf{a}')_{\mathbb{B}}$ for all $t \in \mathbb{N}$, because the local rule of Φ only propagates information to the left. Thus, in particular, the posexpansive CA on $\mathcal{A}^{\mathbb{Z}}$ are completely unrelated to the posexpansive CA on $\mathcal{A}^{\mathbb{N}}$. Nevertheless, posexpansive CA on $\mathcal{A}^{\mathbb{N}}$ behave quite similarly to those on $\mathcal{A}^{\mathbb{Z}}$.

Theorem 18 *Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{N}})$ have neighborhood $[r \dots R]$, where $0 \leq r < R$, and let $\mathbb{B} := [0 \dots R]$. Suppose Φ is posexpansive. Then:*

- (a) $\mathbf{X} := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mathcal{A}^{\mathbb{N}}) \subseteq \mathcal{B}^{\mathbb{N}}$ is a topologically mixing SFT.
- (b) The topological entropy of Φ is $\log_2(k)$ for some $k \in \mathbb{N}$.
- (c) If η is the uniform measure on $\mathcal{A}^{\mathbb{N}}$, then $\Phi_{\mathbb{B}}^{\mathbb{N}}(\eta)$ is the Parry measure on \mathbf{X} . Thus, η is the maxentropy measure for Φ .

Proof See Corollary 3.7 and Theorems 3.8 and 3.9 in [9] or Theorem 4.8(1,2,4) in [86]. \square

Remark (a) See Theorem 58 for an ‘ergodic’ version of Theorem 18.

(b) The analog of Nasu's Theorem 14 (i. e. conjugacy to a full shift) is *not* true for posexpansive CA on $\mathcal{A}^{\mathbb{N}}$. See [13] for a counterexample.

(c) If $\Phi: \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{A}^{\mathbb{N}}$ is invertible, then we define the function $\Phi_{\mathbb{B}}^{\mathbb{Z}}: \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{B}^{\mathbb{Z}}$ by extending the defini-

tion of $\Phi_{\mathbb{B}}^{\mathbb{N}}$ to negative times. We say that Φ is *expansive* if $\Phi_{\mathbb{B}}^{\mathbb{Z}}$ is bijective for some finite $\mathbb{B} \subset \mathbb{N}$. Expansiveness is a much weaker condition than *positive* expansiveness. Nevertheless, the analog of Theorem 18(a) is true: if $\Phi: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$ is invertible and expansive, then $\mathcal{B}^{\mathbb{Z}}$ is conjugate to a (two-sided) subshift of finite type; see Theorem 1.3 in [106].

Measure Rigidity in Algebraic CA

Theorem 1 makes the uniform measure η a ‘natural’ invariant measure for a surjective CA Φ . However, Proposition 10 and Corollary 16 indicate that there are many other (unnatural) Φ -invariant measures as well. Thus, it is natural to seek conditions under which the uniform measure η is the unique (or almost unique) measure which is Φ -invariant, shift-invariant, and perhaps ‘nondegenerate’ in some other sense – a phenomenon which is sometimes called *measure rigidity*. Measure rigidity has been best understood when Φ is compatible with an underlying algebraic structure on $\mathcal{A}^{\mathbb{M}}$.

Let $\star: \mathcal{A}^{\mathbb{M}} \times \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{A}^{\mathbb{M}}$ be a binary operation (‘multiplication’) and let $\bullet^{-1}: \mathcal{A}^{\mathbb{M}} \rightarrow \mathcal{A}^{\mathbb{M}}$ be an unary operation (‘inversion’) such that $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group, and suppose both operations are continuous and commute with all \mathbb{M} -shifts; then $(\mathcal{A}^{\mathbb{M}}, \star)$ is called a *group shift*. For example, if (\mathcal{A}, \cdot) is itself a finite group, and $\mathcal{A}^{\mathbb{M}}$ is treated as a Cartesian product and endowed with componentwise multiplication, then $(\mathcal{A}^{\mathbb{M}}, \cdot)$ is a group shift. However, not all group shifts arise in this manner; see [70,71,72,73,124]. If $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group shift, then a *subgroup shift* is a closed, shift-invariant subgroup $\mathbf{G} \subset \mathcal{A}^{\mathbb{M}}$ (i. e. \mathbf{G} is both a subshift and a subgroup).

If (\mathbf{G}, \star) is a subgroup shift, then the *Haar measure* on \mathbf{G} is the unique probability measure $\eta_{\mathbf{G}}$ on \mathbf{G} which is invariant under translation by all elements of \mathbf{G} . That is, if $\mathbf{g} \in \mathbf{G}$, and $\mathbf{U} \subset \mathbf{G}$ is any measurable subset, and $\mathbf{U} \star \mathbf{g} := \{\mathbf{u} \star \mathbf{g}; \mathbf{u} \in \mathbf{U}\}$, then $\eta_{\mathbf{G}}[\mathbf{U} \star \mathbf{g}] = \eta_{\mathbf{G}}[\mathbf{U}]$. In particular, if $\mathbf{G} = \mathcal{A}^{\mathbb{M}}$, then $\eta_{\mathbf{G}}$ is just the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{M}}$. The Haar measure is a maxentropy measure on \mathbf{G} (see Subsect. “[Invariance of Maxentropy Measures](#)”).

If $(\mathcal{A}^{\mathbb{M}}, \star)$ is a group shift, and $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{M}}$ is a subgroup shift, and $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$, then Φ is called an *endomorph*ic (or *algebraic*) CA on \mathbf{G} if $\Phi(\mathbf{G}) \subseteq \mathbf{G}$ and $\Phi: \mathbf{G} \rightarrow \mathbf{G}$ is an endomorphism of (\mathbf{G}, \star) as a topological group. Let $\mathcal{ECA}(\mathbf{G}, \star)$ denote the set of endomorphic CA on \mathbf{G} . For example, suppose $(\mathcal{A}, +)$ is abelian, and let $(\mathbf{G}, \star) := (\mathcal{A}^{\mathbb{M}}, +)$ with the product group structure; then the endomorphic CA on $\mathcal{A}^{\mathbb{M}}$ are exactly the linear CA. However, if (\mathcal{A}, \cdot) is a *nonabelian* group, then

endomorphic CA on $(\mathcal{A}^{\mathbb{M}}, \cdot)$ are *not* the same as multiplicative CA.

Even in this context, CA admit many nontrivial invariant measures. For example, it is easy to check the following:

Proposition 19 *Let $\mathcal{A}^{\mathbb{M}}$ be a group shift and let $\Phi \in \mathcal{ECA}(\mathcal{A}^{\mathbb{M}}, \star)$. Let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{M}}$ be any Φ -invariant subgroup shift; then the Haar measure on \mathbf{G} is Φ -invariant.*

For example, if $(\mathcal{A}, +)$ is any nonsimple abelian group, and $(\mathcal{A}^{\mathbb{M}}, +)$ has the product group structure, then $\mathcal{A}^{\mathbb{M}}$ admits many nontrivial subgroup shifts; see [73]. If Φ is any linear CA on $\mathcal{A}^{\mathbb{M}}$ with scalar coefficients, then *every* subgroup shift of $\mathcal{A}^{\mathbb{M}}$ is Φ -invariant, so Proposition 19 yields many nontrivial Φ -invariant measures. To isolate η as a unique measure, we must impose further restrictions. The first nontrivial results in this direction were by Host, Maass, and Martínez [61]. Let $h(\Phi, \mu)$ be the entropy of Φ relative to the measure μ (see Sect. “[Entropy](#)” for definition).

Proposition 20 *Let $\mathcal{A} := \mathbb{Z}/p$, where p is prime. Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be a linear CA with neighborhood $\{0, 1\}$, and let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. If μ is σ -ergodic, and $h(\Phi, \mu) > 0$, then μ is the Haar measure η on $\mathcal{A}^{\mathbb{Z}}$.*

Proof See Theorem 12 in [61]. □

A similar idea is behind the next result, only with the roles of Φ and σ reversed. If μ is a measure on $\mathcal{A}^{\mathbb{N}}$, and $\mathbf{b} \in \mathcal{A}^{[1 \dots \infty]}$, then we define the conditional measure $\mu^{(\mathbf{b})}$ on \mathcal{A} by $\mu^{(\mathbf{b})}(a) := \mu[x_0 = a | \mathbf{x}_{[1 \dots \infty]} = \mathbf{b}]$, where \mathbf{x} is a μ -random sequence. For example, if μ is a Bernoulli measure, then $\mu^{(\mathbf{b})}(a) = \mu[x_0 = a]$, independent of \mathbf{b} ; if μ is a Markov measure, then $\mu^{(\mathbf{b})}(a) = \mu[x_0 = a | x_1 = b_1]$.

Proposition 21 *Let (\mathcal{A}, \cdot) be any finite (possibly nonabelian) group, and let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{N}})$ have multiplicative local rule $\phi: \mathcal{A}^{\{0,1\}} \rightarrow \mathcal{A}$ defined by $\phi(a_0, a_1) := a_0 \cdot a_1$. Let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. If μ is Φ -ergodic, then there is some subgroup $C \subset \mathcal{A}$ such that, for every $\mathbf{b} \in \mathcal{A}^{[1 \dots \infty]}$, $\text{supp}(\mu^{(\mathbf{b})})$ is a right coset of C , and $\mu^{(\mathbf{b})}$ is uniformly distributed on this coset.*

Proof See Theorem 3.1 in [113]. □

Example 22 Let Φ and μ be as in Proposition 21. Let η be the Haar measure on $\mathcal{A}^{\mathbb{N}}$.

(a) μ has *complete connections* if $\text{supp}(\mu^{(\mathbf{b})}) = \mathcal{A}$ for μ -almost all $\mathbf{b} \in \mathcal{A}^{[1 \dots \infty]}$. Thus, if μ has complete connections in Proposition 21, then $\mu = \eta$.

(b1) Suppose $h(\mu, \sigma) > h_0 := \max\{\log_2 |C|; C \text{ a proper subgroup of } \mathcal{A}\}$. Then $\mu = \eta$.

(b2) In particular, suppose $\mathcal{A} = (\mathbb{Z}/p, +)$, where p is prime; then $h_0 = 0$. Thus, if Φ has local rule $\phi(a_0, a_1) := a_0 + a_1$, and μ is any σ -invariant, Φ -ergodic measure with $h(\mu, \sigma) > 0$, then $\mu = \eta$. This is closely analogous to Proposition 20, but ‘dual’ to it, because the roles of Φ and σ are reversed in the ergodicity and entropy hypotheses.

(c) If $C \subset \mathcal{A}$ is a subgroup, and μ is the Haar measure on the subgroup shift $C^{\mathbb{N}} \subset \mathcal{A}^{\mathbb{N}}$, then μ satisfies the conditions of Proposition 21. Other, less trivial possibilities also exist (see Examples 3.2(b,c) in [113]).

If μ is a measure on $\mathcal{A}^{\mathbb{Z}}$, and $\mathbf{X}, \mathbf{Y} \subset \mathcal{A}^{\mathbb{Z}}$, then we say \mathbf{X} *essentially equals* \mathbf{Y} and write $\mathbf{X} \overline{\mu} \mathbf{Y}$ if $\mu[\mathbf{X} \Delta \mathbf{Y}] = 0$. If $n \in \mathbb{N}$, then let

$$\mathfrak{I}_n(\mu) := \{\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}}; \sigma^n(\mathbf{X}) \overline{\mu} \mathbf{X}\}$$

be the sigma-algebra of subsets of $\mathcal{A}^{\mathbb{Z}}$ which are ‘essentially’ σ^n -invariant. Thus, μ is σ -ergodic if and only if $\mathfrak{I}_1(\mu)$ is trivial (i. e. contains only sets of measure zero or one). We say μ is *totally σ -ergodic* if $\mathfrak{I}_n(\mu)$ is trivial for all $n \in \mathbb{N}$ (► [Ergodicity and Mixing Properties](#)).

Let $(\mathcal{A}^{\mathbb{Z}}, *)$ be any group shift. The identity element \mathbf{e} of $(\mathcal{A}^{\mathbb{Z}}, *)$ is a constant sequence. Thus, if $\Phi \in \text{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$ is surjective, then $\ker(\Phi) := \{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}; \Phi(\mathbf{a}) = \mathbf{e}\}$ is a finite, shift-invariant subgroup of $\mathcal{A}^{\mathbb{Z}}$ (i. e. a finite collection of σ -periodic sequences).

Proposition 23 *Let $(\mathcal{A}^{\mathbb{Z}}, *)$ be a (possibly nonabelian) group shift, and let $\Phi \in \text{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$ be bipermutative, with neighborhood $\{0, 1\}$. Let $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$. Suppose that:*

(IE) μ is totally ergodic for σ ; (H) $h(\Phi, \mu) > 0$; and
(K) $\ker(\Phi)$ contains no nontrivial σ -invariant subgroups.
Then μ is the Haar measure on $\mathcal{A}^{\mathbb{Z}}$.

Proof See Theorem 5.2 in [113]. \square

Example 24 If $\mathcal{A} = \mathbb{Z}/p$ and $(\mathcal{A}^{\mathbb{Z}}, +)$ is the product group, then Φ is a linear CA and condition (c) is automatically satisfied, so Proposition 23 becomes a special case of Proposition 20.

If $\Phi \in \text{ECA}(\mathcal{A}^{\mathbb{Z}}, *)$, then we have an increasing sequence of finite, shift-invariant subgroups $\ker(\Phi) \subseteq \ker(\Phi^2) \subseteq \ker(\Phi^3) \subseteq \dots$. If $\mathbf{K}(\Phi) := \bigcup_{n=1}^{\infty} \ker(\Phi^n)$, then $\mathbf{K}(\Phi)$ is a countable, shift-invariant subgroup of $(\mathcal{A}^{\mathbb{Z}}, *)$.

Theorem 25 *Let $(\mathcal{A}^{\mathbb{Z}}, +)$ be an abelian group shift, and let $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}}$ be a subgroup shift. Let $\Phi \in \text{ECA}(\mathbf{G}, +)$ be bipermutative, and let $\mu \in \mathfrak{M}_{\text{ens}}(\mathbf{G}; \Phi, \sigma)$. Suppose:*

(I) $\mathfrak{I}_{kP}(\mu) = \mathfrak{I}_1(\mu)$, where P is the lowest common multiple of the σ -periods of all elements in $\ker(\Phi)$, and

$k \in \mathbb{N}$ is any common multiple of all prime factors of $|\mathcal{A}|$.

(H) $h(\Phi, \mu) > 0$.

Furthermore, suppose that either:

(E1) μ is ergodic for the $\mathbb{N} \times \mathbb{Z}$ action (Φ, σ) ;

(K1) Every infinite, σ -invariant subgroup of $\mathbf{K}(\Phi) \cap \mathbf{G}$ is dense in \mathbf{G} ;

or:

(E2) μ is σ -ergodic;

(K2) Every infinite, (Φ, σ) -invariant subgroup of $\mathbf{K}(\Phi) \cap \mathbf{G}$ is dense in \mathbf{G} .

Then μ is the Haar measure on \mathbf{G} .

Proof See Theorems 3.3 and 3.4 of [122], or Théorèmes V.4 and V.5 on p. 115 of [120]. In the special case when \mathbf{G} is irreducible and has topological entropy $\log_2(p)$ (where p is prime), Sobottka has given a different and simpler proof, by using his theory of ‘quasigroup shifts’ to establish an isomorphism between Φ and a linear CA on \mathbb{Z}/p , and then invoking Theorem 7. See Theorems 7.1 and 7.2 of [136], or Teoremas IV.3.1 and IV.3.2 on pp. 100–101 of [134]. \square

Example 26 (a) Let $\mathcal{A} := \mathbb{Z}/p$, where p is prime. Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be linear, and suppose that $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$ is (Φ, σ) -ergodic, $h(\Phi, \mu) > 0$, and $\mathfrak{I}_{p(p-1)}(\mu) = \mathfrak{I}_1(\mu)$. Setting $k = p$ and $P = p - 1$ in Theorem 25, we conclude that μ is the Haar measure on $\mathcal{A}^{\mathbb{Z}}$. For Φ with neighborhood $\{0, 1\}$, this result first appeared as Theorem 13 in [61].

(b) If $(\mathcal{A}^{\mathbb{Z}}, *)$ is abelian, then Proposition 23 is a special case of Theorem 25 [hypothesis (IE) of the former implies hypotheses (I) and (E2) of the latter, while (K) implies (K2)]. Note, however, that Proposition 23 also applies to nonabelian groups.

An *algebraic \mathbb{Z}^D -action* is an action of \mathbb{Z}^D by automorphisms on a compact abelian group \mathbf{G} . For example, if $\mathbf{G} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ is an abelian subgroup shift, then σ is an algebraic \mathbb{Z}^D -action. The invariant measures of algebraic \mathbb{Z}^D -actions have been studied in Schmidt (see Sect. 29 in [124]), Silberger (see Sect. 7 in [132]), and Einsiedler [28, 29].

If $\Phi \in \text{CA}(\mathbf{G})$, then a *complete history* for Φ is a sequence $(\mathbf{g}_t)_{t \in \mathbb{Z}} \in \mathbf{G}^{\mathbb{Z}}$ such that $\Phi(\mathbf{g}_t) = \mathbf{g}_{t+1}$ for all $t \in \mathbb{Z}$. Let $\Phi^{\mathbb{Z}}(\mathbf{G}) \subset \mathbf{G}^{\mathbb{Z}} \subseteq (\mathcal{A}^{\mathbb{Z}^D})^{\mathbb{Z}} \cong \mathcal{A}^{\mathbb{Z}^{D+1}}$ be the set of all complete histories for Φ ; then $\Phi^{\mathbb{Z}}(\mathbf{G})$ is a subshift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$. If $\Phi \in \text{ECA}[\mathbf{G}]$, then $\Phi^{\mathbb{Z}}(\mathbf{G})$ is itself an abelian subgroup shift, and the shift action of \mathbb{Z}^{D+1} on $\Phi^{\mathbb{Z}}(\mathbf{G})$ is thus an algebraic \mathbb{Z}^{D+1} -action. Any (Φ, σ) -invariant measure on \mathbf{G} extends in the obvious way to a σ -invariant measure on $\Phi^{\mathbb{Z}}(\mathbf{G})$. Thus, any result about the invariant measures (or rigidity) of algebraic \mathbb{Z}^{D+1} -actions can

be translated immediately into a result about the invariant measures (or rigidity) of endomorphic cellular automata.

Proposition 27 *Let $G \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift and let $\Phi \in \text{ECA}(G)$. Suppose $\mu \in \mathcal{M}_{\text{ens}}(G; \Phi, \sigma)$ is (Φ, σ) -totally ergodic, and has entropy dimension $d \in [1 \dots D]$ (see Subsect. “Entropy Geometry and Expansive Subdynamics”). If the system $(G, \mu; \Phi, \sigma)$ admits no factors whose d -dimensional measurable entropy is zero, then there is a Φ -invariant subgroup shift $G' \subseteq G$ and some element $\mathbf{x} \in G$ such that μ is the translated Haar measure on the ‘affine’ subset $G' + \mathbf{x}$.*

Proof This follows from Corollary 2.3 in [29]. \square

If we remove the requirement of ‘no zero-entropy factors’, and instead require G and Φ to satisfy certain technical algebraic conditions, then μ must be the Haar measure on G (see Theorem 1.2 in [29]). These strong hypotheses are probably necessary, because in general, the system (G, σ, Φ) admits uncountably many distinct nontrivial invariant measures, even if (G, σ, Φ) is *irreducible*, meaning that G contains no proper, infinite, Φ -invariant subgroup shifts:

Proposition 28 *Let $G \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift, let $\Phi \in \text{ECA}(G)$, and suppose (G, σ, Φ) is irreducible. For any $s \in [0, 1)$, there exists a (Φ, σ) -ergodic measure $\mu \in \mathcal{M}_{\text{ens}}(G; \Phi, \sigma)$ such that $h(\mu, \Phi^n \circ \sigma^z) = s \cdot h_{\text{top}}(G, \Phi^n \circ \sigma^z)$ for every $n \in \mathbb{N}$ and $z \in \mathbb{Z}^D$.*

Proof This follows from Corollary 1.4 in [28]. \square

Let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}; \sigma)$ and let $\mathbb{H} \subset \mathbb{M}$ be a finite subset. We say that μ is \mathbb{H} -mixing if, for any \mathbb{H} -indexed collection $\{\mathbf{U}_h\}_{h \in \mathbb{H}}$ of measurable subsets of $\mathcal{A}^{\mathbb{M}}$,

$$\lim_{n \rightarrow \infty} \mu \left[\bigcap_{h \in \mathbb{H}} \sigma^{nh}(\mathbf{U}_h) \right] = \prod_{h \in \mathbb{H}} \mu[\mathbf{U}_h] .$$

For example, if $|\mathbb{H}| = H$, then any H -multiply σ -mixing measure (see Subsect. “Mixing and Ergodicity”) is \mathbb{H} -mixing.

Proposition 29 *Let $G \subseteq \mathcal{A}^{\mathbb{Z}^D}$ be an abelian subgroup shift and let $\Phi \in \text{ECA}(G)$ have neighborhood \mathbb{H} (with $|\mathbb{H}| \geq 2$). Suppose (G, σ, Φ) is irreducible, and let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$. Then μ is \mathbb{H} -mixing if and only if μ is the Haar measure of G .*

Proof This follows from [124], Corollary 29.5, p. 289 (note that Schmidt uses ‘almost minimal’ to mean ‘irreducible’). A significant generalization of Proposition 29 appears in [115] \square

The Furstenberg Conjecture

Let $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ be the circle group, which we identify with the interval $[0, 1)$. Define the functions $\times_2, \times_3: \mathbb{T}^1 \rightarrow \mathbb{T}^1$ by $\times_2(t) = 2t \pmod{1}$ and $\times_3(t) = 3t \pmod{1}$. Clearly, these maps commute, and preserve the Lebesgue measure on \mathbb{T}^1 . Furstenberg [44] speculated that the *only* nonatomic \times_2 - and \times_3 -invariant measure on \mathbb{T}^1 was the Lebesgue measure. Rudolph [119] showed that, if ρ is (\times_2, \times_3) -invariant measure and *not* Lebesgue, then the systems $(\mathbb{T}^1, \rho, \times_2)$ and $(\mathbb{T}^1, \rho, \times_3)$ have zero entropy; this was later generalized in [60, 69]. It is not known whether any nonatomic measures exist on \mathbb{T}^1 which satisfy Rudolph’s conditions; this is considered an outstanding problem in abstract ergodic theory.

To see the connection between Furstenberg’s Conjecture and cellular automata, let $\mathcal{A} = \{0, 1, 2, 3, 4, 5\}$, and define the surjection $\Psi: \mathcal{A}^{\mathbb{N}} \rightarrow \mathbb{T}^1$ by mapping each $\mathbf{a} \in \mathcal{A}^{\mathbb{N}}$ to the element of $[0, 1)$ having \mathbf{a} as its base-6 expansion. That is:

$$\Psi(a_0, a_1, a_2, \dots) := \sum_{n=0}^{\infty} \frac{a_n}{6^{n+1}} .$$

The map Ψ is injective everywhere except on the countable set of sequences ending in $[000 \dots]$ or $[555 \dots]$ (on this set, Ψ is 2-to-1). Furthermore, Ψ defines a semiconjugacy from \times_2 and \times_3 into two CA on $\mathcal{A}^{\mathbb{N}}$. Let $\mathbb{H} := \{0, 1\}$, and define local maps $\xi_2, \xi_3: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ as follows:

$$\begin{aligned} \xi_2(a_0, a_1) &= \left[2a_0 \right]_6 + \left[\frac{a_1}{3} \right] \text{ and} \\ \xi_3(a_0, a_1) &= \left[3a_0 \right]_6 + \left[\frac{a_1}{2} \right] , \end{aligned}$$

where, $[a]_6$ is the least residue of a , mod 6. If $\mathcal{E}_p \in \text{CA}(\mathcal{A}^{\mathbb{N}})$ has local map ξ_p (for $p = 2, 3$), then it is easy to check that \mathcal{E}_p corresponds to multiplication by p in base-6 notation. In other words, $\Psi \circ \times_p = \mathcal{E}_p \circ \Psi$ for $p = 2, 3$.

If λ is the Lebesgue measure on \mathbb{T}^1 , then $\Psi(\lambda) = \eta$, where η is the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{N}}$. Thus, η is \mathcal{E}_2 - and \mathcal{E}_3 -invariant, and Furstenberg’s Conjecture asserts that η is the *only* nonatomic measure on $\mathcal{A}^{\mathbb{N}}$ which is both \mathcal{E}_2 - and \mathcal{E}_3 -invariant. The shift map $\sigma: \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{A}^{\mathbb{N}}$ corresponds to multiplication by 6 in base-6 notation. Hence, $\mathcal{E}_2 \circ \mathcal{E}_3 = \sigma$. From this it follows that a measure μ is $(\mathcal{E}_2, \mathcal{E}_3)$ -invariant if and only if μ is (\mathcal{E}_2, σ) -invariant if and only if μ is (σ, \mathcal{E}_3) -invariant. Thus, Furstenberg’s Conjecture equivalently asserts that η is the only stationary, \mathcal{E}_3 -invariant nonatomic measure on $\mathcal{A}^{\mathbb{N}}$, and Rudolph’s result asserts that η is the only such nonatomic measure with nonzero entropy; this is analogous to the

‘measure rigidity’ results of Subsect. “[Measure Rigidity in Algebraic CA](#)”. The existence of zero-entropy, (σ, \mathcal{E}_3) -invariant, nonatomic measures remains an open question.

Remark 30 (a) There is nothing special about 2 and 3; the same results hold for any pair of prime numbers.

(b) Lyons [85] and Rudolph and Johnson [68] have also established that a wide variety of \times_2 -invariant probability measures on \mathbb{T}^1 will weak* converge, under the iteration of \times_3 , to the Lebesgue measure (and vice versa). In the terminology of Subsect. “[Asymptotic Randomization by Linear Cellular Automata](#)”, these results immediately translate into equivalent statements about the ‘asymptotic randomization’ of initial probability measures on $\mathcal{A}^{\mathbb{N}}$ under the iteration of \mathcal{E}_2 or \mathcal{E}_3 .

Domains, Defects, and Particles

Suppose $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$, and there is a collection of Φ -invariant subshifts $\mathbf{P}_1, \mathbf{P}_2, \dots, \mathbf{P}_N \subset \mathcal{A}^{\mathbb{Z}}$ (called *phases*). Any sequence \mathbf{a} can be expressed a finite or infinite concatenation

$$\mathbf{a} = [\dots \mathbf{a}_{-2} \mathbf{d}_{-2} \mathbf{a}_{-1} \mathbf{d}_{-1} \mathbf{a}_0 \mathbf{d}_0 \mathbf{a}_1 \mathbf{d}_1 \mathbf{a}_2 \dots],$$

where each *domain* \mathbf{a}_k is a finite word (or half-infinite sequence) which is admissible to phase \mathbf{P}_n for some $n \in [1 \dots N]$, and where each *defect* \mathbf{d}_k is a (possibly empty) finite word (note that this decomposition may not be unique). Thus, $\Phi(\mathbf{a}) = \mathbf{a}'$, where

$$\mathbf{a}' = [\dots \mathbf{a}'_{-2} \mathbf{d}'_{-2} \mathbf{a}'_{-1} \mathbf{d}'_{-1} \mathbf{a}'_0 \mathbf{d}'_0 \mathbf{a}'_1 \mathbf{d}'_1 \mathbf{a}'_2 \dots],$$

and, for every $k \in \mathbb{Z}$, \mathbf{a}'_k belongs to the same phase as \mathbf{a}_k . We say that Φ has *stable phases* if, for any such \mathbf{a} and \mathbf{a}' in $\mathcal{A}^{\mathbb{Z}}$, it is the case that, for all $k \in \mathbb{Z}$, $|\mathbf{d}'_k| \leq |\mathbf{d}_k|$. In other words, the defects do not grow over time. However, they may propagate sideways; for example, \mathbf{d}'_k may be slightly to the right of \mathbf{d}_k , if the domain \mathbf{a}'_k is larger than \mathbf{a}_k , while the domain \mathbf{a}'_{k+1} is slightly smaller than \mathbf{a}_{k+1} . If \mathbf{a}_k and \mathbf{a}_{k+1} belong to different phases, then the defect \mathbf{d}_k is sometimes called a *domain boundary* (or ‘wall’, or ‘edge particle’). If \mathbf{a}_k and \mathbf{a}_{k+1} belong to the same phase, then the defect \mathbf{d}_k is sometimes called a *dislocation* (or ‘kink’).

Often $\mathbf{P}_n = \{\mathbf{p}\}$ where $\mathbf{p} = [\dots ppp \dots]$ is a constant sequence, or each \mathbf{P}_n consists of the σ -orbit of a single periodic sequence. More generally, the phases $\mathbf{P}_1, \dots, \mathbf{P}_N$ may be subshifts of finite type. In this case, most sequences in $\mathcal{A}^{\mathbb{Z}}$ can be fairly easily and unambiguously decomposed into domains separated by defects. However, if the phases are more complex (e.g. sofic shifts), then the exact definition of a ‘defect’ is actually fairly complicated – see [114] for a rigorous discussion.

Example 31 Let $\mathcal{A} = \{0, 1\}$ and let $\mathbb{H} = \{-1, 0, 1\}$. *Elementary cellular automaton* (ECA) #184 is the CA $\Phi: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$ with local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ given as follows: $\phi(a_{-1}, a_0, a_1) = 1$ if $a_0 = a_1 = 1$, or if $a_{-1} = 1$ and $a_0 = 0$. On the other hand, $\phi(a_{-1}, a_0, a_1) = 0$ if $a_{-1} = a_0 = 0$, or if $a_1 = 0$ and $a_0 = 1$. Heuristically, each ‘1’ represents a ‘car’ moving cautiously to the right on a single-lane road. During each iteration, each car will advance to the site in front of it, unless that site is already occupied, in which case the car will remain stationary. ECA#184 exhibits one stable phase \mathbf{P} , given by the 2-periodic sequence $[\dots 0101.0101 \dots]$ and its translate $[\dots 1010.1010 \dots]$ (here the decimal point indicates the zeroth coordinate), and Φ acts on \mathbf{P} like the shift. The phase \mathbf{P} admits two dislocations of width 2. The dislocation $\mathbf{d}_0 = [00]$ moves uniformly to the right, while the dislocation $\mathbf{d}_1 = [11]$ moves uniformly to the left. In the traffic interpretation, \mathbf{P} represents freely flowing traffic, \mathbf{d}_0 represents a stretch of empty road, and \mathbf{d}_1 represents a traffic jam.

Example 32 Let $\mathcal{A} := \mathbb{Z}/N$, and let $\mathbb{H} := [-1 \dots 1]$. The one-dimensional, N -color *cyclic cellular automaton* (CCA_N) $\Phi: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$ has local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ defined:

$$\phi(\mathbf{a}) := \begin{cases} a_0 + 1 & \text{if there is some } h \in \mathbb{H} \\ & \text{with } a_h = a_0 + 1; \\ a_0 & \text{otherwise.} \end{cases}$$

(here, addition is mod N). The CCA has phases $\mathbf{P}_0, \mathbf{P}_1, \dots, \mathbf{P}_{N-1}$, where $\mathbf{P}_a = \{[\dots aaa \dots]\}$ for each $a \in \mathcal{A}$. A domain boundary between \mathbf{P}_a and \mathbf{P}_{a-1} moves with constant velocity towards the \mathbf{P}_{a-1} side. All other domain boundaries are stationary.

In a *particle cellular automaton* (PCA), $\mathcal{A} = \{\emptyset\} \sqcup \mathcal{P}$, where \mathcal{P} is a set of ‘particle types’ and \emptyset represents a vacant site. Each particle $p \in \mathcal{P}$ is assigned some (constant) velocity vector $\mathbf{v}(p) \in (-\mathbb{H})$ (where \mathbb{H} is the neighborhood of the automaton). Particles propagate with constant velocity through \mathbb{M} until two particles try to simultaneously enter the same site in the lattice, at which point the outcome is determined by a *collision rule*: a stylized ‘chemical reaction equation’. For example, an equation “ $p_1 + p_2 \rightsquigarrow p_3$ ” means that, if particle types p_1 and p_2 collide, they coalesce to produce a particle of type p_3 . On the other hand, “ $p_1 + p_2 \rightsquigarrow \emptyset$ ” means that the two particles annihilate on contact. Formally, given a set of velocities and collision

rules, the local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ is defined

$$\phi(\mathbf{a}) := \begin{cases} p & \text{if there is a unique } \mathbf{h} \in \mathbb{H} \text{ and } p \in \mathcal{P} \text{ with} \\ & a_{\mathbf{h}} = p \text{ and } v(p) = -\mathbf{h} : \\ q & \text{if } \{p \in \mathcal{P}; a_{-\mathbf{v}(p)} = p\} = \{p_1, p_2, \dots, p_n\}, \\ & \text{and } p_1 + \dots + p_n \leadsto q. \end{cases}$$

Example 33 The one-dimensional *ballistic annihilation model* (BAM) contains two particle types: $\mathcal{P} = \{\pm 1\}$, with the following rules:

$$v(1) = 1, \quad v(-1) = -1, \quad \text{and} \quad -1 + 1 \leadsto \emptyset.$$

(This CA is sometimes also called *Just Gliders*.) Thus, $a_z = 1$ if the cell z contains a particle moving to the right with velocity 1, whereas $a_z = -1$ if the cell z contains a particle moving left with velocity -1, and $a_z = \emptyset$ if cell z is vacant. Particles move with constant velocity until they collide with oncoming particles, at which point both particles are annihilated. If $\mathcal{B} := \{\pm 1, \emptyset\}$ and $\mathbb{H} = [-1 \dots 1] \subset \mathbb{Z}$, then we can represent the BAM using $\psi \in \mathcal{CA}(\mathcal{B}^{\mathbb{Z}})$ with local rule $\psi: \mathcal{B}^{\mathbb{H}} \rightarrow \mathcal{B}$ defined:

$$\psi(b_{-1}, b_0, b_1) := \begin{cases} -1 & \text{if } b_1 = -1 \text{ and} \\ & b_{-1}, b_0 \in \{-1, \emptyset\}; \\ 1 & \text{if } b_{-1} = 1 \text{ and } b_0, b_1 \in \{1, \emptyset\}; \\ \emptyset & \text{otherwise.} \end{cases}$$

Particle CA can be seen as ‘toy models’ of particle physics or microscale chemistry. More interestingly, however, one-dimensional PCA often arise as factors of coalescent-domain CA, with the ‘particles’ tracking the motion of the defects.

Example 34 (a) Let $\mathcal{A} := \{0, 1\}$ and let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be ECA#184. Let $\mathcal{B} := \{\pm 1, 0\}$, and let $\Psi \in \mathcal{CA}(\mathcal{B}^{\mathbb{Z}})$ be the BAM. Let $\mathbb{G} := \{0, 1\}$, and let $\Gamma: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{B}^{\mathbb{Z}}$ be the block map with local rule $\gamma: \mathcal{A}^{\mathbb{G}} \rightarrow \mathcal{B}$ defined

$$\gamma(a_0, a_1) := 1 - a_0 - a_1 = \begin{cases} 1 & \text{if } [a_0, a_1] = [0, 0] = \mathbf{d}_0; \\ -1 & \text{if } [a_0, a_1] = [1, 1] = \mathbf{d}_1; \\ 0 & \text{otherwise.} \end{cases}$$

Then $\Gamma \circ \Phi = \Psi \circ \Gamma$; in other words, the BAM is a factor of ECA#184, and tracks the motion of the dislocations.

(b) Again, let $\Psi \in \mathcal{CA}(\mathcal{B}^{\mathbb{Z}})$ be the BAM. Let $\mathcal{A} = \mathbb{Z}/3$, and let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be the 3-color CCA. Let $\mathbb{G} := \{0, 1\}$, and let $\Gamma: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{B}^{\mathbb{Z}}$ be the block map with local rule $\gamma: \mathcal{A}^{\mathbb{G}} \rightarrow \mathcal{B}$ defined

$$\gamma(a_0, a_1) := (a_0 - a_1) \bmod 3.$$

Then $\Gamma \circ \Phi = \Psi \circ \Gamma$; in other words, the BAM is a factor of CCA₃, and tracks the motion of the domain boundaries.

Thus, it is often possible to translate questions about coalescent domain CA into questions about particle CA, which are generally easier to study. For example, the invariant measures of the BAM have been completely characterized.

Proposition 35 Let $\mathcal{B} = \{\pm 1, 0\}$, and let $\Psi: \mathcal{B}^{\mathbb{Z}} \rightarrow \mathcal{B}^{\mathbb{Z}}$ be the BAM.

- (a) The sets $\mathbf{R} := \{0, 1\}^{\mathbb{Z}}$ and $\mathbf{L} := \{0, -1\}^{\mathbb{Z}}$ are Ψ -invariant, and Ψ acts as a right-shift on \mathbf{R} and as a left-shift on \mathbf{L} .
 (b) Let $\mathbf{L}^+ := \{0, -1\}^{\mathbb{N}}$ and $\mathbf{R}^- := \{0, 1\}^{-\mathbb{N}}$, and let

$$\mathbf{X} := \{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}; \exists z \in \mathbb{Z} \text{ such that } \mathbf{a}_{(-\infty \dots z]} \in \mathbf{R}^- \text{ and } \mathbf{a}_{[z \dots \infty)} \in \mathbf{L}^+\}.$$

Then \mathbf{X} is Ψ -invariant. For any $\mathbf{x} \in \mathbf{X}$, Ψ acts as a right shift on $\mathbf{a}_{(-\infty \dots z]}$, and as a left-shift on $\mathbf{x}_{(z \dots \infty)}$. (The boundary point z executes some kind of random walk.)

- (c) Any Ψ -invariant measure on $\mathcal{A}^{\mathbb{Z}}$ can be written in a unique way as a convex combination of four measures δ_0, ρ, λ , and μ , where: δ_0 is the point mass on the ‘vacuum’ configuration $[\dots 0 0 0 \dots]$, ρ is any shift-invariant measure on \mathbf{R} , λ is any shift-invariant measure on \mathbf{L} , and μ is a measure on \mathbf{X} .

Furthermore, there exist shift-invariant measures μ_- and μ_+ on \mathbf{R}^- and \mathbf{L}^+ , respectively, such that, for μ -almost all $\mathbf{x} \in \mathbf{X}$, $\mathbf{x}_{(-\infty \dots z]}$ is μ_- -distributed and $\mathbf{x}_{[z \dots \infty)}$ is μ_+ -distributed.

Proof (a) and (b) are obvious; (c) is Theorem 1 in [8]. \square

Remark 36 (a) Proposition 35(c) can be immediately translated into a complete characterization of the invariant measures of ECA#184, via the factor map Γ in Example 34(a); see [8], Theorem 3. Likewise, using the factor map in Example 34(b) we get a complete characterization of the invariant measures for CCA₃.

(b) Proposition 48 and Corollaries 49 and 50 describe the limit measures of the BAM, CCA₃, and ECA#184. Also, Blank [10] has characterized invariant measures for a broad class of multilane, multi-speed traffic models (including ECA#184); see Remark 51(b).

(c) Kůrka [78] has defined, for any $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$, a construction similar to the set \mathbf{X} in Proposition 35(b). For any $n \in \mathbb{N}$ and $z \in \mathbb{Z}$, let $\mathbf{S}_{z,n}$ be the set of fixed points of $\Phi^n \circ \sigma^z$; then $\mathbf{S}_{z,n}$ is a subshift of finite type, which Kůrka calls a *signal subshift* with *velocity* $v = z/n$. (For example, if Φ is the BAM, then $\mathbf{R} = \mathbf{S}_{1,1}$ and $\mathbf{L} = \mathbf{S}_{-1,1}$.)

Now, suppose that $z_1/n_1 > z_2/n_2 > \dots > z_J/n_J$. The *join* of the signal subshifts $\mathbf{S}_{z_1, n_1}, \mathbf{S}_{z_2, n_2}, \dots, \mathbf{S}_{z_J, n_J}$ is the set \mathbf{S} of all infinite sequences $[\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_J]$, where for all $j \in [1..J]$, \mathbf{a}_j is a (possibly empty) finite word or (half-)infinite sequence admissible to the subshift \mathbf{S}_{z_j, n_j} . (For example, if \mathbf{S} is the join of $\mathbf{S}_{1,1} = \mathbf{R}$ and $\mathbf{S}_{-1,1} = \mathbf{L}$ from Proposition 35(a), then $\mathbf{S} = \mathbf{L} \cup \mathbf{X} \cup \mathbf{R}$.) It follows that $\mathbf{S} \subseteq \Phi(\mathbf{S}) \subseteq \Phi^2(\mathbf{S}) \subseteq \dots$. If we define $\Phi^\infty(\mathbf{S}) := \bigcup_{t=0}^\infty \Phi^t(\mathbf{S})$, then $\Phi^\infty(\mathbf{S}) \subseteq \Phi^\infty(\mathcal{A}^\mathbb{Z})$, where $\Phi^\infty(\mathcal{A}^\mathbb{Z}) := \bigcap_{t=0}^\infty \Phi^t(\mathcal{A}^\mathbb{Z})$ is the omega limit set of Φ (see Proposition 5 in [78]). The support of any Φ -invariant measure must be contained in $\Phi^\infty(\mathcal{A}^\mathbb{Z})$, so invariant measures may be closely related to the joins of signal subshifts. See [► Topological Dynamics of Cellular Automata](#) for more information.

In the case of the BAM, it is not hard to check that $\Phi^\infty(\mathbf{S}) = \mathbf{S} = \Phi^\infty(\mathcal{A}^\mathbb{Z})$; this suggests an alternate proof of Proposition 35(c). It would be interesting to know whether a conclusion analogous to Proposition 35(c) holds for other $\Phi \in \mathbf{CA}(\mathcal{A}^\mathbb{Z})$ such that $\Phi^\infty(\mathcal{A}^\mathbb{Z})$ is a join of signal subshifts.

Limit Measures and Other Asymptotics

Asymptotic Randomization by Linear Cellular Automata

The results of Subsect. “[Measure Rigidity in Algebraic CA](#)” suggest that the uniform Bernoulli measure η is the ‘natural’ measure for algebraic CA, because η is the unique invariant measure satisfying any one of several collections of reasonable criteria. In this section, we will see that η is ‘natural’ in quite another way: it is the unique limit measure for linear CA from a large set of initial conditions.

If $\{\mu_n\}_{n=1}^\infty$ is a sequence of measures on $\mathcal{A}^\mathbb{M}$, then this sequence *weak* converges* to the measure μ_∞ (“ $wk * \lim \mu_n = \mu_\infty$ ”) if, for all cylinder sets $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$, $\lim_{n \rightarrow \infty} \mu_n[\mathbf{B}] = \mu_\infty[\mathbf{B}]$. Equivalently, for all continuous functions $f: \mathcal{A}^\mathbb{M} \rightarrow \mathbb{C}$, we have

$$\lim_{n \rightarrow \infty} \int_{\mathcal{A}^\mathbb{M}} f d\mu_n = \int_{\mathcal{A}^\mathbb{M}} f d\mu_\infty.$$

The *Cesàro average* (or *Cesàro limit*) of $\{\mu_n\}_{n=1}^\infty$ is $wk * \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mu_n$, if this limit exists.

Let $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^\mathbb{M})$ and let $\Phi \in \mathbf{CA}(\mathcal{A}^\mathbb{M})$. For any $t \in \mathbb{N}$, the measure $\Phi^t \mu$ is defined by $\Phi^t \mu(\mathbf{B}) = \mu(\Phi^{-t}(\mathbf{B}))$, for any measurable subset $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$. We say that Φ *asymptotically randomizes* μ if the Cesàro average of the sequence $\{\Phi^n \mu\}_{n=1}^\infty$ is η . Equivalently, there is a sub-

set $\mathbb{J} \subset \mathbb{N}$ of density 1, such that

$$wk * \lim_{\substack{j \rightarrow \infty \\ j \in \mathbb{J}}} \Phi^j \mu = \eta.$$

The uniform measure η is the measure of maximal entropy on $\mathcal{A}^\mathbb{M}$. Thus, asymptotic randomization is kind of ‘Second Law of Thermodynamics’ for CA.

Let $(\mathcal{A}, +)$ be a finite abelian group, and let Φ be a linear cellular automaton (LCA) on $\mathcal{A}^\mathbb{M}$. Recall that Φ has *scalar coefficients* if there is some finite $\mathbb{H} \subset \mathbb{M}$, and integer coefficients $\{c_h\}_{h \in \mathbb{H}}$ so that Φ has a local rule of the form

$$\Phi(\mathbf{a}_\mathbb{H}) := \sum_{h \in \mathbb{H}} c_h a_h, \quad (3)$$

An LCA Φ is *proper* if Φ has scalar coefficients as in Eq. (3), and if, furthermore, for any prime divisor p of $|\mathcal{A}|$, there are at least two $h, h' \in \mathbb{H}$ such that $c_h \not\equiv 0 \not\equiv c_{h'} \pmod p$. For example, if $\mathcal{A} = \mathbb{Z}/n$ for some $n \in \mathbb{N}$, then every LCA on $\mathcal{A}^\mathbb{M}$ has scalar coefficients; in this case, Φ is proper if, for every prime p dividing n , at least two of these coefficients are coprime to p . In particular, if $\mathcal{A} = \mathbb{Z}/p$ for some prime p , then Φ is proper as long as $|\mathbb{H}| \geq 2$.

Let $\text{PLCA}(\mathcal{A}^\mathbb{M})$ be the set of proper linear CA on $\mathcal{A}^\mathbb{M}$. If $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^\mathbb{M})$, recall that μ has *full support* if $\mu[\mathbf{B}] > 0$ for every cylinder set $\mathbf{B} \subset \mathcal{A}^\mathbb{M}$.

Theorem 37 *Let $(\mathcal{A}, +)$ be a finite abelian group, let $\mathbb{M} := \mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \geq 0$, and let $\Phi \in \text{PLCA}(\mathcal{A}^\mathbb{M})$. Let μ be any Bernoulli measure or Markov random field on $\mathcal{A}^\mathbb{M}$ having full support. Then Φ asymptotically randomizes μ .*

History Theorem 37 was first proved for simple one-dimensional LCA randomizing Bernoulli measures on $\mathcal{A}^\mathbb{Z}$, where \mathcal{A} was a cyclic group. In the case $\mathcal{A} = \mathbb{Z}/2$, Theorem 37 was independently proved for the *nearest-neighbor XOR* CA (having local rule $\phi(a_{-1}, a_0, a_1) = a_{-1} + a_1 \pmod 2$) by Miyamoto [100] and Lind [82]. This result was then generalized to $\mathcal{A} = \mathbb{Z}/p$ for any prime p by Cai and Luo [16]. Next, Maass and Martínez [87] extended the Miyamoto/Lind result to the *binary Ledrappier* CA (local rule $\phi(a_0, a_1) = a_0 + a_1 \pmod 2$). Soon after, Ferrari et al. [36] considered the case when \mathcal{A} was an abelian group of order p^k (p prime), and proved Theorem 37 for any *Ledrappier* CA (local rule $\phi(a_0, a_1) = c_0 a_0 + c_1 a_1$, where $c_0, c_1 \not\equiv 0 \pmod p$) acting on any measure on $\mathcal{A}^\mathbb{Z}$ having full support and ‘rapidly decaying correlations’ (see Part II(a) below). For example, this includes any Markov measure on $\mathcal{A}^\mathbb{Z}$ with full support. Next, Pivato and Yasawii [116] generalized Theorem 37 to any PLCA acting on any fully supported N -step Markov chain on $\mathcal{A}^\mathbb{Z}$.

or any nontrivial Bernoulli measure on $\mathcal{A}^{\mathbb{Z}^D \times \mathbb{N}^E}$, where $\mathcal{A} = \mathbb{Z}_{/p^k}$ (p prime). Finally, Pivato and Yassawi [117] proved Theorem 37 in full generality, as stated above.

The proofs of Theorem 37 and its variations all involve two parts:

Part I. A careful analysis of the local rule of Φ^t (for all $t \in \mathbb{N}$), showing that the neighborhood of Φ^t grows large as $t \rightarrow \infty$ (and in some cases, contains large ‘gaps’).

Part II. A demonstration that the measure μ exhibits ‘rapidly decaying correlations’ between widely separated elements of \mathbb{M} ; hence, when these elements are combined using Φ^t , it is as if we are summing independent random variables.

Part I Any linear CA with scalar coefficients can be written as a ‘Laurent polynomial of shifts’. That is, if Φ has local rule (3), then for any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$,

$$\Phi(\mathbf{a}) := \sum_{\mathbf{h} \in \mathbb{H}} c_{\mathbf{h}} \sigma^{\mathbf{h}}(\mathbf{a}) \quad (\text{where we add configurations componentwise}).$$

We indicate this by writing “ $\Phi = F(\sigma)$ ”, where $F \in \mathbb{Z}[x_1^{\pm 1}, x_2^{\pm 1}, \dots, x_D^{\pm 1}]$ is the D -variable Laurent polynomial defined:

$$F(x_1, \dots, x_D) := \sum_{(h_1, \dots, h_D) \in \mathbb{H}} c_{\mathbf{h}} x_1^{h_1} x_2^{h_2} \dots x_D^{h_D}.$$

For example, if Φ is the *nearest-neighbor XOR* CA, then $\Phi = \sigma^{-1} + \sigma^1 = F(\sigma)$, where $F(x) = x^{-1} + x$. If Φ is a *Ledrappier* CA, then $\Phi = c_0 \text{Id} + c_1 \sigma^1 = F(\sigma)$, where $F(x) = c_0 + c_1 x$.

It is easy to verify that, if F and G are two such polynomials, and $\Phi = F(\sigma)$ while $\Gamma = G(\sigma)$, then $\Phi \circ \Gamma = (F \cdot G)(\sigma)$, where $F \cdot G$ is the product of F and G in the polynomial ring $\mathbb{Z}[x_1^{\pm 1}, x_2^{\pm 1}, \dots, x_D^{\pm 1}]$. In particular, this means that $\Phi^t = F^t(\sigma)$ for all $t \in \mathbb{N}$. Thus, iterating an LCA is equivalent to computing the powers of a polynomial.

If $\mathcal{A} = \mathbb{Z}_{/p}$, then we can compute the coefficients of F^t modulo p . If p is prime, then this can be done using a result of Lucas [84], which provides a formula for the binomial coefficient $\binom{a}{b}$ in terms of the base- p expansions of a and b . For example, if $p = 2$, then Lucas’ theorem says that Pascal’s triangle, modulo 2, looks like a ‘discrete Sierpinski triangle’, made out of 0’s and 1’s. (This is why fragments of the Sierpinski triangle appear frequently in the spacetime diagrams of linear CA on $\mathcal{A} = \mathbb{Z}_{/2}$, a phenomenon which has inspired much literature on ‘fractals and automatic sequences in cellular

automata’; see [2,4,5,6,7,52,53,54,55,56,57,94,138,139,140,145,146,147,148,149].) Thus, Lucas’ Theorem, along with some combinatorial lemmas about the structure of base- p expansions, provides the machinery for Part I.

Part II There are two approaches to analyzing probability measures on $\mathcal{A}^{\mathbb{M}}$; one using renewal theory, and the other using harmonic analysis.

II(a) Renewal theory This approach was developed by Maass, Martínez and their collaborators. Loosely speaking, if $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$ has sufficiently large support and sufficiently rapid decay of correlations (e. g. a Markov chain), and $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ is a μ -random sequence, then we can treat \mathbf{a} as if there is a sparse, randomly distributed set of ‘renewal times’ when the normal stochastic evolution of \mathbf{a} is interrupted by independent, random ‘errors’. By judicious use of Part I described above, one can use this ‘renewal process’ to make it seem as though Φ^t is summing independent random variables.

For example, if $(\mathcal{A}, +)$ be an abelian group of order p^k where p is prime, and $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$ has *complete connections* (see Example 22(a)) and *summable decay* (which means that a certain sequence of coefficients (measuring long-range correlation) decays fast enough that its sum is finite), and $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ is a Ledrappier CA, then Ferrari et al. (see Theorem 1.3 in [36]) showed that Φ asymptotically randomizes μ . (For example, this applies to any N -step Markov chain with full support on $\mathcal{A}^{\mathbb{Z}}$.) Furthermore, if $\mathcal{A} = \mathbb{Z}_{/p} \times \mathbb{Z}_{/p}$, and $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ has linear local rule $\phi\left(\begin{bmatrix} x_0 \\ y_0 \end{bmatrix}, \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}\right) = (y_0, x_0 + y_1)$, then Maass and Martínez [88] showed that Φ randomizes any Markov measure with full support on $\mathcal{A}^{\mathbb{Z}}$. Maass and Martínez again handled Part II using renewal theory. However, in this case, Part I involves some delicate analysis of the (noncommutative) algebra of the matrix-valued coefficients; unfortunately, their argument does not generalize to other LCA with noncommuting, matrix-valued coefficients. (However, Proposition 8 of [117] suggests a general strategy for dealing with such LCA).

II(b) Harmonic analysis This approach to Part II was implicit in the early work of Lind [82] and Cai and Luo [16], but was developed in full generality by Pivato and Yassawi [116,117,118]. We regard $\mathcal{A}^{\mathbb{M}}$ as a direct product of copies of the group $(\mathcal{A}, +)$, and endow it with the product group structure; then $(\mathcal{A}^{\mathbb{M}}, +)$ is a compact abelian topological group. A *character* on $(\mathcal{A}^{\mathbb{M}}, +)$ is a continuous group homomorphism $\chi: \mathcal{A}^{\mathbb{M}} \rightarrow \mathbb{T}$, where $\mathbb{T} := \{c \in \mathbb{C}; |c| = 1\}$ is the unit circle group. If μ is

a measure on $\mathcal{A}^{\mathbb{M}}$, then the *Fourier coefficients* of μ are defined: $\hat{\mu}[\chi] = \int_{\mathcal{A}^{\mathbb{M}}} \chi d\mu$, for every character χ .

If $\chi: \mathcal{A}^{\mathbb{M}} \rightarrow \mathbb{T}$ is any character, then there is a unique finite subset $\mathbb{K} \subset \mathbb{M}$ (called the *support* of χ) and a unique collection of nontrivial characters $\chi_k: \mathcal{A} \rightarrow \mathbb{T}$ for all $k \in \mathbb{K}$, such that,

$$\chi(\mathbf{a}) = \prod_{k \in \mathbb{K}} \chi_k(a_k), \quad \forall \mathbf{a} \in \mathcal{A}^{\mathbb{M}}. \quad (4)$$

We define $\text{rank}[\chi] := |\mathbb{K}|$. The measure μ is called *harmonically mixing* if, for all $\epsilon > 0$, there is some R such that for all characters χ , $(\text{rank}[\chi] \geq R) \implies (|\hat{\mu}[\chi]| < \epsilon)$.

The set $\mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ of harmonically mixing measures on $\mathcal{A}^{\mathbb{M}}$ is quite inclusive. For example, if μ is any (N -step) Markov chain with full support on $\mathcal{A}^{\mathbb{Z}}$, then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ (see Propositions 8 and 10 in [116]), and if $\nu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}})$ is absolutely continuous with respect to this μ , then $\nu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ also (see Corollary 9 in [116]). If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime) then any nontrivial Bernoulli measure on $\mathcal{A}^{\mathbb{M}}$ is harmonically mixing (see Proposition 6 in [116]). Furthermore, if $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$ has complete connections and summable decay, then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{Z}})$ (see Theorem 23 in [61]). If $\mathfrak{M} := \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}; \mathbb{C})$ is the set of all complex-valued measures on $\mathcal{A}^{\mathbb{M}}$, then \mathfrak{M} is *Banach algebra* (i. e. it is a vector space under the obvious definition of addition and scalar multiplication for measures, and a Banach space under the total variation norm, and finally, since $\mathcal{A}^{\mathbb{M}}$ is a topological group, \mathfrak{M} is a ring under convolution). Then $\mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ is an ideal in \mathfrak{M} , is closed under the total variation norm, and is dense in the weak* topology on \mathfrak{M} (see Propositions 4 and 7 in [116]).

Finally, if μ is any Markov random field on $\mathcal{A}^{\mathbb{M}}$ which is *locally free* (which roughly means that the boundary of any finite region does not totally determine the interior of that region), then $\mu \in \mathfrak{Sm}(\mathcal{A}^{\mathbb{M}})$ (see Theorem 1.3 in [118]). In particular, this implies:

Proposition 38 *If $(\mathcal{A}, +)$ is any finite group, and $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}})$ is any Markov random field with full support, then μ is harmonically mixing.*

Proof This follows from Theorem 1.3 in [118]. It is also a special case of Theorem 15 in [117]. \square

If χ is a character, and Φ is a LCA, then $\chi \circ \Phi^t$ is also a character, for any $t \in \mathbb{N}$ (because it is a composition of two continuous group homomorphisms). We say Φ is *diffusive* if there is a subset $\mathbb{J} \subset \mathbb{N}$ of density 1, such that, for every character χ of $\mathcal{A}^{\mathbb{M}}$,

$$\lim_{\mathbb{J} \ni j \rightarrow \infty} \text{rank}[\chi \circ \Phi^j] = \infty.$$

Proposition 39 *Let $(\mathcal{A}, +)$ be any finite abelian group and let \mathbb{M} be any monoid. If μ is harmonically mixing and Φ is diffusive, then Φ asymptotically randomizes μ .*

Proof See Theorem 12 in [117]. \square

Proposition 40 *Let $(\mathcal{A}, +)$ be any abelian group and let $\mathbb{M} := \mathbb{Z}^D \times \mathbb{N}^E$ for some $D, E \geq 0$. If $\Phi \in \text{PLCA}(\mathcal{A}^{\mathbb{M}})$, then Φ is diffusive.*

Proof The proof uses Lucas' theorem, as described in Part I above. See Theorem 15 in [116] for the case $\mathcal{A} = \mathbb{Z}_{/p}$ when p prime. See Theorem 6 in [117] for the case when \mathcal{A} is any cyclic group. That proof easily extends to any finite abelian group \mathcal{A} : write \mathcal{A} as a product of cyclic groups and decompose Φ into separate automata over these cyclic factors. \square

Proof of Theorem 37 Combine Propositions 38, 39, and 40. \square

Remark (a) Proposition 40 can be generalized: we do not need the coefficients of Φ to be integers, but merely to be a collection of automorphisms of \mathcal{A} which commute with one another (so that Lucas' theorem from Part I is still applicable). See Theorem 9 in [117].

(b) For simplicity, we stated Theorem 37 for measures with full support; however, Proposition 39 actually applies to many Markov random fields *without* full support, because harmonic mixing only requires 'local freedom' (see Theorem 1.3 in [118]). For example, the support of a Markov chain on $\mathcal{A}^{\mathbb{Z}}$ is Markov subshift. If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime), then Proposition 39 yields asymptotic randomization of the Markov chain as long as the transition digraph of the underlying Markov subshift admits at least *two* distinct paths of length 2 between any pair of vertices in \mathcal{A} . More generally, if $\mathbb{M} = \mathbb{Z}^D$, then the support of any Markov random field on $\mathcal{A}^{\mathbb{Z}^D}$ is an SFT, which we can regard as the set of all tilings of \mathbb{R}^D by a certain collection of Wang tiles. If $\mathcal{A} = \mathbb{Z}_{/p}$ (p prime), then Proposition 39 yields asymptotic randomization of the Markov random field as long as the underlying Wang tiling is flexible enough that any hole can always be filled in at least two ways; see Sect. 1 in [118].

Remark 41 (Generalizations and Extensions) (a) Pivato and Yassawi (see Thm 3.1 in [118]) proved a variation of Theorem 60 where diffusion (of Φ) is replaced with a slightly stronger condition called *dispersion*, so that harmonic mixing (of μ) can be replaced with a slightly weaker condition called *dispersion mixing* (DM). It is unknown whether all proper linear CA are dispersive, but a very large class are (including, for example, $\Phi = \text{Id} + \sigma$).

Any uniformly mixing measure with positive entropy is DM (see Theorem 5.2 in [118]); this includes, for example, any mixing *quasimarkov measure* (i.e. the image of a Markov measure under a block map; these are the natural measures supported on sofic shifts). Quasimarkov measures are not, in general, harmonically mixing (see Sect. 2 in [118]), but this result shows they are still asymptotically randomized by most linear CA.

(b) Suppose $\mathbf{G} \subset \mathcal{A}^{\mathbb{Z}^D}$ is a σ -transitive *subgroup shift* (see Subsect. “[Measure Rigidity in Algebraic CA](#)” for definition), and let $\Phi \in \text{PLCA}(\mathbf{G})$. If \mathbf{G} satisfies an algebraic condition called the *follower lifting property* (FLP) and μ is any Markov random field with $\text{supp}(\mu) = \mathbf{G}$, then Maass, Martínez, Pivato, and Yassawi [89] have shown that Φ asymptotically randomizes μ to a maxentropy measure on \mathbf{G} . Furthermore, if $D = 1$, then this maxentropy measure is the Haar measure on \mathbf{G} . In particular, if \mathcal{A} is an abelian group of prime-power order, then *any* transitive Markov subgroup $\mathbf{G} \subset \mathcal{A}^{\mathbb{Z}}$ satisfies the FLP, so this result holds for any multistep Markov measure on \mathbf{G} . See also [90] for the special case when $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ has local rule $\phi(x_0, x_1) = x_0 + x_1$. In the special case when Φ has local rule $\phi(x_0, x_1) = c_0 x_0 + c_1 x_1 + a$, the result has been extended to measures with complete connections and summable decay; see Teorema III.2.1, p. 71 in [134] or see Theorem 1 in [91].

(c) All the aforementioned results concern asymptotic randomization of initial measures with nonzero entropy. Is nonzero entropy either necessary or sufficient for asymptotic randomization? First let $\mathbf{X}_N \subset \mathcal{A}^{\mathbb{Z}}$ be the set of N -periodic points (see Subsect. “[Periodic Invariant Measures](#)”) and suppose $\text{supp}(\mu) \subseteq \mathbf{X}_N$. Then the Cesàro limit of $\{\Phi^t(\mu)\}_{t \in \mathbb{N}}$ will also be a measure supported on \mathbf{X}_N , so μ_∞ cannot be the uniform measure on $\mathcal{A}^{\mathbb{Z}}$. Nor, in general, will μ_∞ be the uniform measure on \mathbf{X}_N ; this follows from Jen’s (1988) exact characterization of the limit cycles of linear CA acting on \mathbf{X}_N .

What if μ is a *quasiperiodic* measure, such as the unique σ -invariant measure on a Sturmian shift? There exist quasiperiodic measures on $(\mathbb{Z}/2)^{\mathbb{Z}}$ which are *not* asymptotically randomized by the Ledrappier CA (see Sect. 15 in [112]). But it is unknown whether this extends to all quasiperiodic measures or all linear CA.

There is also a measure μ on $\mathcal{A}^{\mathbb{Z}}$ which has zero σ -entropy, yet is still asymptotically randomized by Φ (see Sect. 8 in [118]). Loosely speaking, μ is a Toeplitz measure with a very low density of ‘bit errors’. Thus, μ is ‘almost’ deterministic (so it has zero entropy), but by sufficiently increasing the density of ‘bit errors’, we can introduce just enough randomness to allow asymptotic randomization to occur.

(d) Suppose (G, \cdot) is a *nonabelian* group and $\Phi: G^{\mathbb{Z}} \rightarrow G^{\mathbb{Z}}$ has *multiplicative* local rule $\phi(\mathbf{g}) := g_{h_1}^{n_1} g_{h_2}^{n_2} \cdots g_{h_J}^{n_J}$, for some $\{h_1, \dots, h_J\} \subset \mathbb{Z}$ (possibly not distinct) and $n_1, \dots, n_J \in \mathbb{N}$. If G is nilpotent, then G can be decomposed into a tower of abelian group extensions; this induces a structural decomposition of Φ into a tower of skew products of ‘relative’ linear CA. This strategy was first suggested by Moore [102], and was developed by Pivato (see Theorem 21 in [111]), who proved a version of Theorem 37 in this setting.

(e) Suppose (Q, \star) is a *quasigroup* – that is, \star is a binary operation such that for any $q, r, s \in Q$, $(q \star r = q \star s) \iff (r = s) \iff (r \star q = s \star q)$. Any finite *associative* quasigroup has an identity, and any associative quasigroup with an identity is a group. However there are also many nonassociative finite quasigroups. If we define a ‘multiplicative’ CA $\Phi: Q^{\mathbb{Z}} \rightarrow Q^{\mathbb{Z}}$ with local rule $\phi: Q^{\{0,1\}} \rightarrow Q$ given by $\phi(q_0, q_1) = q_0 \star q_1$, then it is easy to see that Φ is bipermutative if and only if (Q, \star) is a quasigroup. Thus, quasigroups seem to provide the natural algebraic framework for studying bipermutative CA; this was first proposed by Moore [101], and later explored by Host, Maass, and Martínez (see Sect. 3 in [61]), Pivato (see Sect. 2 in [113]), and Sobottka [134,135,136].

Note that $Q^{\mathbb{Z}}$ is a quasigroup under component-wise \star -multiplication. A *quasigroup shift* is a subshift $\mathbf{X} \subset Q^{\mathbb{Z}}$ which is also a subquasigroup; it follows that $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. If \mathbf{X} and Φ satisfy certain strong algebraic conditions, and $\mu \in \mathfrak{M}_{\text{ces}}(\mathbf{X}; \sigma)$ has complete connections and summable decay, then the sequence $\{\Phi^t \mu\}_{t=1}^\infty$ Cesàro-converges to a maxentropy measure μ_∞ on \mathbf{X} (thus, if \mathbf{X} is irreducible, then μ_∞ is the Parry measure; see Subsect. “[Invariance of Maxentropy Measures](#)”). See Theorem 6.3(i) in [136], or Teorema IV.5.3, p. 107 in [134].

Hybrid Modes of Self-Organization

Most cellular automata do not asymptotically randomize; instead they seem to weak* converge to limit measures concentrated on small (i.e. low-entropy) subsets of the statespace $\mathcal{A}^{\mathbb{M}}$ – a phenomenon which can be interpreted as a form of ‘self-organization’. Exact limit measures have been computed for a few CA. For example, let $\mathcal{A} = \{0, 1, 2\}$ and let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}^D})$ be the *Greenberg–Hastings* model (a simple model of an excitable medium). Durrett and Steif [26] showed that, if $D \geq 2$ and μ is any Bernoulli measure on $\mathcal{A}^{\mathbb{Z}^D}$, then $\mu_\infty := w k * \lim_{t \rightarrow \infty} \Phi^t \mu$ exists; μ_∞ -almost all points are 3-periodic for Φ , and although μ_∞ is not a Bernoulli measure, the system $(\mathcal{A}^{\mathbb{Z}^D}, \mu_\infty, \sigma)$ is measurably isomorphic to a Bernoulli system.

In other cases, the limit measure cannot be exactly computed, but can still be estimated. For example, let $\mathcal{A} = \{\pm 1\}$, $\theta \in (0, 1)$, and $R > 0$, and let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be the (R, θ) -threshold voter CA (where each cell computes the fraction of its radius- R neighbors which disagree with its current sign, and negates its sign if this fraction is at least θ). Durret and Steif [27] and Fisch and Gravner [43] have described the long-term behavior of Φ in the limit as $R \rightarrow \infty$. If $\theta < 1/2$, then every initial condition falls into a two-periodic orbit (and if $\theta < 1/4$, then every cell simply alternates its sign). Let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$; if $1/2 < \theta$, then for any finite subset $\mathbb{B} \subset \mathbb{Z}$, if R is large enough, then ‘most’ initial conditions (relative to η) converge to orbits that are fixed inside \mathbb{B} . Indeed, there is a critical value $\theta_c \approx 0.6469076$ such that, if $\theta_c < \theta$, and R is large enough, then ‘most’ initial conditions (for η) are already fixed inside \mathbb{B} ; see also [137] for an analysis of behavior at the critical value.

In still other cases, the limit measure is known to exist, but is still mysterious; this is true for the Cesàro limit measures of Coven CA, for example see [87], Theorem 1. However, for most CA, it is difficult to even show that limit measures exist. Except for the linear CA of Subsect. “Asymptotic Randomization by Linear Cellular Automata”, there is no large class of CA whose limit measures have been exactly characterized. Often, it is much easier to study the dynamical asymptotics of CA at a purely topological level.

If $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$, then $\mathcal{A}^{\mathbb{M}} \supseteq \Phi(\mathcal{A}^{\mathbb{M}}) \supseteq \Phi^2(\mathcal{A}^{\mathbb{M}}) \supseteq \dots$. The *limit set* of Φ is the nonempty subshift $\Phi^\infty(\mathcal{A}^{\mathbb{M}}) := \bigcap_{t=1}^{\infty} \Phi^t(\mathcal{A}^{\mathbb{M}})$. For any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, the *omega-limit set* of \mathbf{a} is the set $\omega(\mathbf{a}, \Phi)$ of all cluster points of the Φ -orbit $\{\Phi^t(\mathbf{a})\}_{t=1}^{\infty}$. A closed subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a (Conley) *attractor* if there exists a clopen subset $\mathbf{U} \supseteq \mathbf{X}$ such that $\Phi(\mathbf{U}) \subseteq \mathbf{U}$ and $\mathbf{X} = \bigcap_{t=1}^{\infty} \Phi^t(\mathbf{U})$. It follows that $\omega(\Phi, \mathbf{u}) \subseteq \mathbf{X}$ for all $\mathbf{u} \in \mathbf{U}$. For example, $\Phi^\infty(\mathcal{A}^{\mathbb{M}})$ is an attractor (let $\mathbf{U} := \mathcal{A}^{\mathbb{M}}$). The topological attractors of CA were analyzed by Hurley [63,65,66], who discovered severe constraints on the possible attractor structures a CA could exhibit; see Sect. 9 of ▶ [Topological Dynamics of Cellular Automata](#) and ▶ [Cellular Automata, Classification of](#).

Within pure topological dynamics, attractors and (omega) limit sets are the natural formalizations of the heuristic notion of ‘self-organization’. The corresponding formalization in pure ergodic theory is the weak* limit measure. However, both weak* limit measures and topological attractors fail to adequately describe the sort of self-organization exhibited by many CA. Thus, several ‘hybrid’ notions self-organization have been developed, which combine topological and measurable criteria. These hybrid notions are more flexible and inclusive than purely

topological notions. However, they do not require the explicit computation (or even the existence) of weak* limit measures, so in practice they are much easier to verify than purely ergodic notions.

Milnor–Hurley μ -attractors If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a closed subset, then for any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, we define $d(\mathbf{a}, \mathbf{X}) := \inf_{\mathbf{x} \in \mathbf{X}} d(\mathbf{a}, \mathbf{x})$. If $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$, then the *basin* (or *realm*) of \mathbf{X} is the set

$$\begin{aligned} \text{Basin}(\mathbf{X}) &:= \left\{ \mathbf{a} \in \mathcal{A}^{\mathbb{M}} ; \lim_{t \rightarrow \infty} d(\Phi^t(\mathbf{a}), \mathbf{X}) = 0 \right\} \\ &= \left\{ \mathbf{a} \in \mathcal{A}^{\mathbb{M}} ; \omega(\mathbf{a}, \Phi) \subseteq \mathbf{X} \right\}. \end{aligned}$$

Suppose $\Phi(\mathbf{X}) \subseteq \mathbf{X}$. If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}})$, then \mathbf{X} is a μ -*attractor* if $\mu[\text{Basin}(\mathbf{X})] > 0$; we call \mathbf{X} a *lean μ -attractor* if in addition, $\mu[\text{Basin}(\mathbf{X})] > \mu[\text{Basin}(\mathbf{Y})]$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. Finally, a μ -attractor \mathbf{X} is *minimal* if $\mu[\text{Basin}(\mathbf{Y})] = 0$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. For example, if \mathbf{X} is a μ -attractor, and (\mathbf{X}, Φ) is minimal as a dynamical system, then \mathbf{X} is a minimal μ -attractor. This concept was introduced by Milnor [96,97] in the context of smooth dynamical systems; its ramifications for CA were first explored by Hurley [64,65].

If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}, \sigma)$, then μ is *weakly σ -mixing* if, for any measurable sets $\mathbf{U}, \mathbf{V} \subset \mathcal{A}^{\mathbb{Z}^D}$, there is a subset $\mathbb{J} \subset \mathbb{Z}^D$ of density 1 such that $\lim_{\mathbb{J} \ni j \rightarrow \infty} \mu[\sigma^j(\mathbf{U}) \cap \mathbf{V}] = \mu[\mathbf{U}] \cdot \mu[\mathbf{V}]$ (see Subsect. “Mixing and Ergodicity”). For example, any Bernoulli measure is weakly mixing. A subshift $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is σ -*minimal* if \mathbf{X} contains no proper non-empty subshifts. For example, if \mathbf{X} is just the σ -orbit of some σ -periodic point, then \mathbf{X} is σ -minimal.

Proposition 42 Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, and let \mathbf{X} be a μ -attractor.

- (a) If μ is σ -ergodic, and $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then $\mu[\text{Basin}(\mathbf{X})] = 1$.
- (b) If \mathbb{M} is countable, and \mathbf{X} is σ -minimal subshift with $\mu[\text{Basin}(\mathbf{X})] = 1$, then \mathbf{X} is lean.
- (c) Suppose $\mathbb{M} = \mathbb{Z}^D$ and μ is weakly σ -mixing.
 - (i) If \mathbf{X} is a minimal μ -attractor, then \mathbf{X} is a subshift, so $\mu[\text{Basin}(\mathbf{X})] = 1$, and thus \mathbf{X} is the only lean μ -attractor of Φ .
 - (ii) If \mathbf{X} is a Φ -periodic orbit which is also a lean μ -attractor, then \mathbf{X} is minimal, $\mu[\text{Basin}(\mathbf{X})] = 1$, and \mathbf{X} contains only constant configurations.

Proof (a) If \mathbf{X} is σ -invariant, then $\text{Basin}(\mathbf{X})$ is also σ -invariant; hence $\mu[\text{Basin}(\mathbf{X})] = 1$ because μ is σ -ergodic.

(b) Suppose $\mathbf{Y} \subsetneq \mathbf{X}$ was a proper closed subset with $\mu[\text{Basin}(\mathbf{Y})] = 1$. For any $\mathbf{m} \in \mathbb{M}$, it is easy

to check that $\text{Basin}(\sigma^m[\mathbf{Y}]) = \sigma^m[\text{Basin}(\mathbf{Y})]$. Thus, if $\tilde{\mathbf{Y}} := \bigcap_{m \in \mathbb{M}} \sigma^m(\mathbf{Y})$, then $\text{Basin}(\tilde{\mathbf{Y}}) = \bigcap_{m \in \mathbb{M}} \sigma^m[\text{Basin}(\mathbf{Y})]$, so $\mu[\text{Basin}(\tilde{\mathbf{Y}})] = 1$ (because \mathbb{M} is countable). Thus, $\tilde{\mathbf{Y}}$ is nonempty, and is a subshift of \mathbf{X} . But \mathbf{X} is σ -minimal, so $\tilde{\mathbf{Y}} = \mathbf{X}$, which means $\mathbf{Y} = \mathbf{X}$. Thus, \mathbf{X} is a lean μ -attractor.

(c) In the case when μ is a Bernoulli measure, (c)[i] is Theorem B in [64] or Proposition 2.7 in [65], while (c)[ii] is Theorem A in [65]. Hurley's proofs easily extend to the case when μ is weakly σ -mixing. The only property we require of μ is this: for any nontrivial measurable sets $\mathbf{U}, \mathbf{V} \subset \mathcal{A}^{\mathbb{Z}^D}$, and any $\mathbf{z} \in \mathbb{Z}^D$, there is some $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^D$ with $\mathbf{z} = \mathbf{x} - \mathbf{y}$, such that $\mu[\sigma^{\mathbf{y}}(\mathbf{U}) \cap \mathbf{V}] > 0$ and $\mu[\sigma^{\mathbf{x}}(\mathbf{U}) \cap \mathbf{V}] > 0$. This is clearly true if μ is weakly mixing (because if $\mathbb{J} \subset \mathbb{Z}^D$ has density 1, then $\mathbb{J} \cap (\mathbf{z} + \mathbb{J}) \neq \emptyset$ for any $\mathbf{z} \in \mathbb{Z}^D$).

Proof sketch for (c)[i] If \mathbf{X} is a (minimal) μ -attractor, then so is $\sigma^{\mathbf{y}}(\mathbf{X})$, and $\text{Basin}[\sigma^{\mathbf{y}}(\mathbf{X})] = \sigma^{\mathbf{y}}[\text{Basin}(\mathbf{X})]$. Thus, weak mixing yields $\mathbf{x}, \mathbf{y} \in \mathbb{Z}^D$ such that $\text{Basin}[\sigma^{\mathbf{x}}(\mathbf{X})] \cap \text{Basin}[\mathbf{X}]$ and $\text{Basin}[\sigma^{\mathbf{y}}(\mathbf{X})] \cap \text{Basin}[\mathbf{X}]$ are both nontrivial. But the basins of distinct minimal μ -attractors must be disjoint; thus $\sigma^{\mathbf{x}}(\mathbf{X}) = \mathbf{X} = \sigma^{\mathbf{y}}(\mathbf{X})$. But $\mathbf{x} - \mathbf{y} = \mathbf{z}$, so this means $\sigma^{\mathbf{z}}(\mathbf{X}) = \mathbf{X}$. This holds for all $\mathbf{z} \in \mathbb{Z}^D$, so \mathbf{X} is a subshift, so (a) implies $\mu[\text{Basin}(\mathbf{X})] = 1$. \square

Section 4 of [64] contains several examples showing that the minimal topological attractor of Φ can be different from its minimal μ -attractor. For example, a CA can have different minimal μ -attractors for different choices of μ . On the other hand, there is a CA possessing a minimal topological attractor but with no minimal μ -attractors for any Bernoulli measure μ .

Hilmy–Hurley Centers

Let $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$. For any closed subset $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$, we define

$$\mu_{\mathbf{a}}[\mathbf{X}] := \liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mathbf{1}_{\mathbf{X}}(\Phi^n(\mathbf{a})).$$

(Thus, if μ is a Φ -ergodic measure on $\mathcal{A}^{\mathbb{M}}$, then Birkhoff's Ergodic Theorem asserts that $\mu_{\mathbf{a}}[\mathbf{X}] = \mu[\mathbf{X}]$ for μ -almost all $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$). The *center* of \mathbf{a} is the set:

$$\text{Cent} < (\mathbf{a}, \Phi) := \bigcap \left\{ \text{closed subsets } \mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}; \mu_{\mathbf{a}}[\mathbf{X}] = 1 \right\}.$$

Thus, $\text{Cent}(\mathbf{a}, \Phi)$ is the smallest closed subset such that $\mu_{\mathbf{a}}[\text{Cent}(\mathbf{a}, \Phi)] = 1$. If $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is closed, then the *well* of \mathbf{X} is the set

$$\text{Well}(\mathbf{X}) := \left\{ \mathbf{a} \in \mathcal{A}^{\mathbb{M}}; \text{Cent}(\mathbf{a}, \Phi) \subseteq \mathbf{X} \right\}.$$

If $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}})$, then \mathbf{X} is a μ -center if $\mu[\text{Well}(\mathbf{X})] > 0$; we call \mathbf{X} a *lean μ -center* if in addition, $\mu[\text{Well}(\mathbf{X})] > \mu[\text{Well}(\mathbf{Y})]$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. Finally, a μ -center \mathbf{X} is *minimal* if $\mu[\text{Well}(\mathbf{Y})] = 0$ for any proper closed subset $\mathbf{Y} \subsetneq \mathbf{X}$. This concept was introduced by Hilmy [59] in the context of smooth dynamical systems; its ramifications for CA were first explored by Hurley [65].

Proposition 43 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, and let \mathbf{X} be a μ -center.

- (a) If μ is σ -ergodic, and $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ is a subshift, then $\mu[\text{Well}(\mathbf{X})] = 1$.
- (b) If \mathbb{M} is countable, and \mathbf{X} is σ -minimal subshift with $\mu[\text{Well}(\mathbf{X})] = 1$, then \mathbf{X} is lean.
- (c) Suppose $\mathbb{M} = \mathbb{Z}^D$ and μ is weakly σ -mixing. If \mathbf{X} is a minimal μ -center, then \mathbf{X} is a subshift, \mathbf{X} is the only lean μ -center, and $\mu[\text{Well}(\mathbf{X})] = 1$.

Proof (a) and (b) are very similar to the proofs of Proposition 42(a,b).

(c) is proved for Bernoulli measures as Theorem B in [65]. The proof is quite similar to Proposition 42(c)[i], and again, we only need μ to be weakly mixing. \square

Section 4 of [65] contains several examples of minimal μ -centers which are not μ -attractors. In particular, the analogue of Proposition 42(c)[ii] is false for μ -centers.

Kürka–Maass μ -limit Sets If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \sigma)$, then Kürka and Maass define the μ -limit set of Φ :

$$\Lambda(\Phi, \mu) := \bigcap \left\{ \text{closed subsets } \mathbf{X} \subset \mathcal{A}^{\mathbb{M}}; \lim_{t \rightarrow \infty} \Phi^t \mu(\mathbf{X}) = 1 \right\}.$$

It suffices to take this intersection only over all *cylinder* sets \mathbf{X} . By doing this, we see that $\Lambda(\Phi, \mu)$ is a subshift of $\mathcal{A}^{\mathbb{M}}$, and is defined by the following property: for any finite $\mathbb{B} \subset \mathbb{M}$ and any word $\mathbf{b} \in \mathcal{A}^{\mathbb{B}}$, \mathbf{b} is admissible to $\Lambda(\Phi, \mu)$ if and only if $\liminf_{t \rightarrow \infty} \Phi^t \mu[\mathbf{b}] > 0$.

Proposition 44 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \sigma)$.

- (a) If $\text{wk} * \lim_{t \rightarrow \infty} \Phi^t \mu = \nu$, then $\Lambda(\Phi, \mu) = \text{supp}(\nu)$. Suppose $\mathbb{M} = \mathbb{Z}$.
- (b) If Φ is surjective and has an equicontinuous point, and μ has full support on $\mathcal{A}^{\mathbb{Z}}$, then $\Lambda(\Phi, \mu) = \mathcal{A}^{\mathbb{Z}}$.
- (c) If Φ is left- or right-permutative and μ is connected (see below), then $\Lambda(\Phi, \mu) = \mathcal{A}^{\mathbb{Z}}$.

Proof For (a), see Proposition 2 in [79]. For (b,c), see Theorems 2 and 3 in [77]; for earlier special cases of these results, see also Propositions 4 and 5 in [79]. \square

Remark 45 (a) In Proposition 44(c), the measure μ is *connected* if there is some constant $C > 0$ such that, for any finite word $\mathbf{b} \in \mathcal{A}^*$, and any $a \in \mathcal{A}$, we have $\mu[\mathbf{b} a] \geq C \cdot \mu[\mathbf{b}]$ and $\mu[a \mathbf{b}] \geq C \cdot \mu[\mathbf{b}]$.

For example, any Bernoulli, Markov, or N -step Markov measure with full support is connected. Also, any measure with ‘complete connections’ (see Example 22(a)) is connected.

(b) Proposition 44(a) shows that μ -limit sets are closely related to the weak* limits of measures. Recall from Subsect. “Asymptotic Randomization by Linear Cellular Automata” that the uniform Bernoulli measure η is the weak* limit of a large class of initial measures under the action of linear CA. Presumably the same result should hold for a much larger class of *permutative* CA, but so far this is unproven, except in some special cases [see Remarks 41(d,e)]. Proposition 44(a,c) implies that the limit measure of a permutative CA (if it exists) must have full support – hence it can’t be ‘too far’ from η .

Kůrka’s Measure Attractors Let $\mathcal{M}_{\text{lim}}^{\sigma} := \mathcal{M}_{\text{lim}}^{\sigma}(\mathcal{A}^{\mathbb{M}}, \sigma)$ have the weak* topology, and define $\Phi_*: \mathcal{M}_{\text{lim}}^{\sigma} \rightarrow \mathcal{M}_{\text{lim}}^{\sigma}$ by $\Phi_*(\mu) = \mu \circ \Phi^{-1}$. Then Φ_* is continuous, so we can treat $(\mathcal{M}_{\text{lim}}^{\sigma}, \Phi_*)$ itself as a compact topological dynamical system. The “weak* limit measures” of Φ are simply the attracting fixed points of $(\mathcal{M}_{\text{lim}}^{\sigma}, \Phi_*)$. However, even if the Φ_* -orbit of a measure μ does not weak* converge to a fixed point, we can still consider the omega-limit set of μ . In particular, the limit set $\Phi_*^{\infty}(\mathcal{M}_{\text{lim}}^{\sigma})$ is the union of the omega-limit sets of all σ -invariant initial measures under Φ_* . Kůrka defines the *measure attractor* of Φ :

$$\text{MeasAttr}(\Phi) := \overline{\bigcup \{\text{supp}(\mu); \mu \in \Phi_*^{\infty}(\mathcal{M}_{\text{lim}}^{\sigma})\}} \subseteq \mathcal{A}^{\mathbb{M}}.$$

(The bar denotes topological closure.) A configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}^D}$ is *densely recurrent* if any word which occurs in \mathbf{a} does so with nonzero frequency. Formally, for any finite $\mathbb{B} \subset \mathbb{Z}^D$

$$\limsup_{N \rightarrow \infty} \frac{\#\{z \in [-N \dots N]^D; \mathbf{a}_{\mathbb{B}+z} = \mathbf{a}_{\mathbb{B}}\}}{(2N+1)^D} > 0.$$

If $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^D}$ is a subshift, then the *densely recurrent subshift* of \mathbf{X} is the closure \mathbf{D} of the set of all densely recurrent points in \mathbf{X} . If $\mu \in \mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{X})$, then the Birkhoff Ergodic Theorem implies that $\text{supp}(\mu) \subseteq \mathbf{D}$; see Proposition 8.8 in [1], p. 164. From this it follows that $\mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{X}) = \mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{D})$. On the other hand, $\mathbf{D} = \bigcup \{\text{supp}(\mu); \mu \in \mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{D})\}$. In other words, densely recurrent subshifts are the only subshifts which are ‘covered’ by their own set of shift-invariant measures.

Proposition 46 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})[\mathcal{A}^{\mathbb{Z}^D}]$. Let \mathbf{D} be the densely recurrent subshift of $\Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$. Then $\mathbf{D} = \text{MeasAttr}(\Phi)$, and $\Phi^{\infty}(\mathcal{M}_{\text{lim}}^{\sigma}) = \mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{D}, \sigma)$.

Proof Case $D = 1$ is Proposition 13 in [78]. The same proof works for $D \geq 2$. \square

Synthesis

The various hybrid modes of self-organization are related as follows:

Proposition 47 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$.

(a) Let $\mu \in \mathcal{M}_{\text{lim}}^{\sigma}(\mathcal{A}^{\mathbb{M}}, \sigma)$ and let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be any closed set.

- [i] If \mathbf{X} is a topological attractor and μ has full support, then \mathbf{X} is a μ -attractor.
- [ii] If \mathbf{X} is a μ -attractor, then \mathbf{X} is a μ -center.
- [iii] Suppose $\mathbb{M} = \mathbb{Z}^D$, and that μ is weakly σ -mixing. Let \mathbf{Y} be the intersection of all topological attractors of Φ . If Φ has a minimal μ -attractor \mathbf{X} , then $\mathbf{X} \subseteq \mathbf{Y}$.
- [iv] If μ is σ -ergodic, then $\Lambda(\Phi, \mu) \subseteq \bigcap \{\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}; \mathbf{X} \text{ a subshift and } \mu\text{-attractor}\} \subseteq \Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$.
- [v] Thus, if μ is σ -ergodic and has full support, then

$$\Lambda(\Phi, \mu) \subseteq \bigcap \{\mathbf{X} \subseteq \mathcal{A}^{\mathbb{M}}; \mathbf{X} \text{ a subshift and a topological attractor}\}.$$

- [vi] If \mathbf{X} is a subshift, then $(\Lambda(\Phi, \mu) \subseteq \mathbf{X}) \iff (\omega(\Phi_*, \mu) \subseteq \mathcal{M}_{\text{lim}}^{\sigma}(\mathbf{X}))$.

(b) Let $\mathbb{M} = \mathbb{Z}^D$. Let \mathcal{B} be the set of all Bernoulli measures on $\mathcal{A}^{\mathbb{Z}^D}$, and for any $\beta \in \mathcal{B}$, let \mathbf{X}_{β} be the minimal β -attractor for Φ (if it exists).

There is a comeager subset $\mathbf{A} \subset \mathcal{A}^{\mathbb{Z}^D}$ such that $\bigcup_{\beta \in \mathcal{B}} \mathbf{X}_{\beta} \subseteq \bigcap_{\mathbf{a} \in \mathbf{A}} \omega(\mathbf{a}, \Phi)$.

(c) $\text{MeasAttr}(\Phi) = \bigcup \{\Lambda(\Phi, \mu); \mu \in \mathcal{M}_{\text{lim}}^{\sigma}(\mathcal{A}^{\mathbb{M}})\}$.

(d) If $\mathbb{M} = \mathbb{Z}^D$, then $\text{MeasAttr}(\Phi) \subseteq \Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$.

Proof (a)[i]: If \mathbf{U} is a clopen subset and $\Phi^{\infty}(\mathbf{U}) = \mathbf{X}$, then $\mathbf{U} \subseteq \text{Basin}(\mathbf{X})$; thus, $0 < \mu[\mathbf{U}] \leq \mu[\text{Basin}(\mathbf{X})]$, where the “ $<$ ” is because μ has full support.

(a)[ii]: For any $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$, it is easy to see that $\text{Cent}(\mathbf{a}, \Phi) \subseteq \omega(\mathbf{a}, \Phi)$. Thus, $\text{Well}(\mathbf{X}) \supseteq \text{Basin}(\mathbf{X})$. Thus, $\mu[\text{Well}(\mathbf{X})] \geq \mu[\text{Basin}(\mathbf{X})] > 0$.

(a)[iii] is Proposition 3.3 in [64]. (Again, Hurley states and proves this in the case when μ is a Bernoulli measure, but his proof only requires weak mixing.)

(a)[iv]: Let \mathbf{X} be a subshift and a μ -attractor; we claim that $\Lambda(\Phi, \mu) \subseteq \mathbf{X}$. Proposition 42(a) says $\mu[\text{Basin}(\mathbf{X})] = 1$.

Let $\mathbb{B} \subset \mathbb{M}$ be any finite set. If $\mathbf{b} \in \mathcal{A}^{\mathbb{B}} \setminus \mathbf{X}_{\mathbb{B}}$, then

$$\{\mathbf{a} \in \mathcal{A}^{\mathbb{M}}; \exists T \in \mathbb{N} \text{ such that } \forall t \geq T, \Phi^t(\mathbf{a})_{\mathbb{B}} \neq \mathbf{b}\} \supseteq \text{Basin}(\mathbf{X}).$$

It follows that the left-hand set has μ -measure 1, which implies that $\lim_{t \rightarrow \infty} \Phi^t \mu(\mathbf{b}) = 0$ – hence \mathbf{b} is a forbidden word in $\Lambda(\Phi, \mu)$.

Thus, all the words forbidden in \mathbf{X} are also forbidden in $\Lambda(\Phi, \mu)$. Thus $\Lambda(\Phi, \mu) \subseteq \mathbf{X}$. (The case $\mathbb{M} = \mathbb{Z}$ of (a)[iv] appears as Prop. 1 in [79] and as Prop. II.27, p. 67 in [120]; see also Cor. II.30 in [120] for a slightly stronger result.)

(a)[v] follows from (a)[iv] and (a)[i].

(a)[vi] is Proposition 1 in [77] or Proposition 10 in [78]; the argument is fairly similar to (a)[iv]. (Kůrka assumes $\mathbb{M} = \mathbb{Z}$, but this is not necessary.)

(b) is Proposition 5.2 in [64].

(c) Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{M}}$ be a subshift and let $\mathfrak{M}_{\text{attr}}^{\sigma} = \mathfrak{M}_{\text{attr}}^{\sigma}(\mathcal{A}^{\mathbb{M}})$. Then

$$\begin{aligned} (\text{Meas} \mathbf{A}_{\text{tr}}(\Phi) \subseteq \mathbf{X}) \\ \iff (\text{supp}(\nu) \subseteq \mathbf{X}, \forall \nu \in \Phi_*^{\infty}(\mathfrak{M}_{\text{attr}}^{\sigma})) \\ \iff (\nu \in \mathfrak{M}_{\text{attr}}^{\sigma}(\mathbf{X}), \forall \nu \in \Phi_*^{\infty}(\mathfrak{M}_{\text{attr}}^{\sigma})) \\ \iff (\Phi_*^{\infty}(\mathfrak{M}_{\text{attr}}^{\sigma}) \subseteq \mathfrak{M}_{\text{attr}}^{\sigma}(\mathbf{X})) \\ \iff (\omega(\Phi_*, \mu) \subseteq \mathfrak{M}_{\text{attr}}^{\sigma}(\mathbf{X}), \forall \mu \in \mathfrak{M}_{\text{attr}}^{\sigma}) \\ \iff (\Lambda(\Phi, \mu) \subseteq \mathbf{X}, \forall \mu \in \mathfrak{M}_{\text{attr}}^{\sigma}) \\ (*) \\ \iff (\bigcup \{\Lambda(\Phi, \mu) \mid \mu \in \mathfrak{M}_{\text{attr}}^{\sigma}\} \subseteq \mathbf{X}). \end{aligned}$$

where (*) is by (a)[vi]. It follows that $\text{Meas} \mathbf{A}_{\text{tr}}(\Phi) = \bigcup \{\Lambda(\Phi, \mu); \mu \in \mathfrak{M}_{\text{attr}}^{\sigma}(\mathcal{A}^{\mathbb{M}})\}$.

(d) follows immediately from Proposition 46. \square

Examples and Applications The most natural examples of these hybrid modes of self-organization arise in the *particle cellular automata* (PCA) introduced in Subsect. “**Domains, Defects, and Particles**”. The long-term dynamics of a PCA involves a steady reduction in particle density, as particles coalesce or annihilate one another in collisions. Thus, presumably, for almost any initial configuration $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, the sequence $\{\Phi^t(\mathbf{a})\}_{t=1}^{\infty}$ should converge to the subshift \mathbf{Z} of configurations containing no particles (or at least, no particles of certain types), as $t \rightarrow \infty$. Unfortunately, this presumption is generally *false* if we interpret ‘convergence’ in the strict topological dynamical sense: the occasional particles will continue to wander near the origin at arbitrarily large times in the future orbit of \mathbf{a} (albeit with diminishing frequency), so $\omega(\mathbf{a}, \Phi)$ will *not* be contained in \mathbf{Z} . However, the presumption becomes true if we

instead employ one of the more flexible hybrid notions introduced above. For example, most initial probability measures μ should converge, under iteration of Φ to a measure concentrated on configurations with few or no particles; hence we expect that $\Lambda(\Phi, \mu) \subseteq \mathbf{Z}$. As discussed in Subsect. “**Domains, Defects, and Particles**”, a result about self-organization in a PCA can sometimes be translated into an analogous result about self-organization in associated coalescent-domain CA.

Proposition 48 *Let $\mathcal{A} = \{0, \pm 1\}$ and let $\Psi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be the Ballistic Annihilation Model (BAM) from Example 33. Let $\mathbf{R} := \{0, 1\}^{\mathbb{Z}}$ and $\mathbf{L} := \{0, -1\}^{\mathbb{Z}}$.*

- (a) *If $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$, then $\nu = \text{wk} * \lim_{t \rightarrow \infty} \Psi^t(\mu)$ exists, and has one of three forms: either $\nu \in \mathfrak{M}_{\text{ens}}(\mathbf{R}, \sigma)$, or $\nu \in \mathfrak{M}_{\text{ens}}(\mathbf{L}, \sigma)$, or $\nu = \delta_0$, the point mass on the sequence $\mathbf{0} = [\dots 000 \dots]$.*
- (b) *Thus, the measure attractor of Φ is $\mathbf{R} \cup \mathbf{L}$ (note that $\mathbf{R} \cap \mathbf{L} = \{\mathbf{0}\}$).*
- (c) *In particular, if μ is a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$ with $\mu[+1] = \mu[-1]$, then $\nu = \delta_0$.*
- (d) *Let μ be a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$.*

- [i] *If $\mu[+1] > \mu[-1]$, then \mathbf{R} is a μ -attractor – i. e. $\mu[\text{Basin}(\mathbf{R})] > 0$.*
- [ii] *If $\mu[+1] < \mu[-1]$, then \mathbf{L} is a μ -attractor.*
- [iii] *If $\mu[+1] = \mu[-1]$, then $\{\mathbf{0}\}$ is not a μ -attractor, because $\mu[\text{Basin}\{\mathbf{0}\}] = 0$. However, $\Lambda(\Phi, \mu) = \{\mathbf{0}\}$.*

Proof (a) is Theorem 6 of [8], and (b) follows from (a). (c) follows from Theorem 2 of [42]. (d)[i,ii] were first observed by Gilman (see Sect. 3, pp. 111–112 in [45], and later by Kůrka and Maass (see Example 4 in [80]). (d)[iii] follows immediately from (c): the statement $\Lambda(\Phi, \mu) = \{\mathbf{0}\}$ is equivalent to asserting that $\lim_{t \rightarrow \infty} \Phi^t \mu[\pm 1] = 0$, which a consequence of (c). Another proof of (d)[iii] is Proposition 11 in [80]; see also Example 3 in [79] or Prop. II.32, p. 70 in [120]. \square

Corollary 49 *Let $\mathcal{A} = \mathbb{Z}/3$, let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CCA_3 (see Example 32), and let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$. Then $\text{wk} * \lim_{t \rightarrow \infty} \Phi^t(\eta) = \frac{1}{3}(\delta_0 + \delta_1 + \delta_2)$, where δ_a is the point mass on the sequence $[\dots aaa \dots]$ for each $a \in \mathcal{A}$.*

Proof Combine Proposition 48(c) with the factor map Γ in Example 34(b). See Theorem 1 of [42] for details. \square

Corollary 50 *Let $\mathcal{A} = \{0, 1\}$, let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ be ECA#184 (see Example 31).*

- (a) $\text{MeasAttr}(\Phi) = \mathbf{R} \cup \mathbf{L}$, where $\mathbf{R} \subset \mathcal{A}^{\mathbb{Z}}$ is the set of sequences not containing [11], and $\mathbf{L} \subset \mathcal{A}^{\mathbb{Z}}$ is the set of sequences not containing [00].
- (b) If η is the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$, then $wk * \lim_{t \rightarrow \infty} \Phi^t(\eta) = \frac{1}{2}(\delta_0 + \delta_1)$, where δ_0 and δ_1 are the point masses on $[\dots 010.101 \dots]$ and $[\dots 101.010 \dots]$.
- (c) Let μ be a Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$.
- [i] If $\mu[0] > \mu[1]$, then \mathbf{R} is a μ -attractor – i. e. $\mu[\text{Basin}(\mathbf{R})] > 0$.
- [ii] If $\mu[0] < \mu[1]$, then \mathbf{L} is a μ -attractor.

Proof sketch Let Γ be the factor map from Example 34(a). To prove (a), apply Γ to Proposition 48(b); see Example 26, Sect. 9 in [78] for details. To prove (b), apply Γ to Proposition 48(c); see Proposition 12 in [80] for details. To prove (c), apply Γ to Proposition 48(d)[i,ii]; \square

Remark 51 (a) The other parts of Proposition 48 can likewise be translated into equivalent statements about the measure attractors and μ -attractors of CCA_3 and $\text{ECA}\#184$.

(b) Recall that $\text{ECA}\#184$ is a model of single-lane, traffic, where each car is either stopped or moving rightwards at unit speed. Blank [10] has extended Corollary 50(c) to a much broader class of CA models of multi-lane, multi-speed traffic. For any such model, let $\mathbf{R} \subset \mathcal{A}^{\mathbb{Z}}$ be the set of ‘free flowing’ configurations where each car has enough space to move rightwards at its maximum possible speed. Let $\mathbf{L} \subset \mathcal{A}^{\mathbb{Z}}$ be the set of ‘jammed’ configurations where the cars are so tightly packed that the jammed clusters can propagate (leftwards) through the cars at maximum speed. If μ is any Bernoulli measure, then $\mu[\text{Basin}(\mathbf{R})] = 1$ if the μ -average density of cars is greater than $1/2$, whereas $\mu[\text{Basin}(\mathbf{L})] = 1$ if the density is less than $1/2$ Theorems 1.2 and 1.3 in [10]. Thus, $\mathbf{L} \sqcup \mathbf{R}$ is a (non-lean) μ -attractor, although not a topological attractor Lemma 2.13 in [10].

Example 52 A cyclic addition and ballistic annihilation model (CABAM) contains the same ‘moving’ particles ± 1 as the BAM (Example 33), but also has one or more ‘stationary’ particle types. Let $3 \leq N \in \mathbb{N}$, and let $\mathcal{P} = \{1, 2, \dots, N-1\} \subset \mathbb{Z}_N$, where we identify $N-1$ with -1 , modulo N . It will be convenient to represent the ‘vacant’ state \emptyset as 0; thus, $\mathcal{A} = \mathbb{Z}_N$. The particles 1 and -1 have velocities and collisions as in the BAM, namely:

$$v(1) = 1, \quad v(-1) = -1, \quad \text{and} \quad -1 + 1 \rightsquigarrow \emptyset.$$

We set $v(p) = 0$, for all $p \in [2 \dots N-2]$, and employ the following collision rule:

If $p_{-1} + p_0 + p_1 \equiv q, \pmod{N}$, then $p_{-1} + p_0 + p_1 \rightsquigarrow q$. (5)

(here, any one of p_{-1}, p_0, p_1 , or q could be 0, signifying vacancy). For example, if $N = 5$ and a (rightward moving) type $+1$ particle strikes a (stationary) type 3 particle, then the $+1$ particle is annihilated and the 3 particle turns into a (stationary) 4 particle. If another $+1$ particle hits the 4 particle, then both are annihilated, leaving a vacancy (0).

Let $\mathcal{B} = \mathbb{Z}_N$, and let $\Psi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CABAM. Then the set of fixed points of Ψ is $\mathbf{F} = \{\mathbf{f} \in \mathcal{B}^{\mathbb{Z}}; f_z \neq \pm 1, \forall z \in \mathbb{Z}\}$. Note that, if $\mathbf{b} \in \text{Basin}[\mathbf{F}]$ – that is, if $\omega(\mathbf{b}, \Psi) \subseteq \mathbf{F}$ – then in fact $\lim_{t \rightarrow \infty} \Psi^t(\mathbf{b})$ exists and is a Ψ -fixed point.

Proposition 53 Let $\mathcal{B} = \mathbb{Z}_N$, let $\Psi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be the CABAM, and let η be the uniform Bernoulli measure on $\mathcal{B}^{\mathbb{Z}}$. If $N \geq 5$, then \mathbf{F} is a ‘global’ η -attractor – that is, $\eta[\text{Basin}(\mathbf{F})] = 1$. However, if $N \leq 4$, then $\eta[\text{Basin}(\mathbf{F})] = 0$.

Proof See Theorem 1 of [41]. \square

Let $\mathcal{A} = \mathbb{Z}_N$ and let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be the N -color CCA from Example 32. Then the set of fixed points of Φ is $\mathbf{F} = \{\mathbf{f} \in \mathcal{A}^{\mathbb{Z}}; f_z - f_{z+1} \neq \pm 1, \forall z \in \mathbb{Z}\}$. Note that, if $\mathbf{a} \in \text{Basin}[\mathbf{F}]$, then in fact $\lim_{t \rightarrow \infty} \Phi^t(\mathbf{a})$ exists and is a Φ -fixed point.

Corollary 54 Let $\mathcal{A} = \mathbb{Z}_N$, let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ be the N -color CCA, and let η be the uniform Bernoulli measure on $\mathcal{A}^{\mathbb{Z}}$. If $N \geq 5$, then \mathbf{F} is a ‘global’ η -attractor – that is, $\eta[\text{Basin}(\mathbf{F})] = 1$. However, if $N \leq 4$, then $\eta[\text{Basin}(\mathbf{F})] = 0$.

Proof sketch Let $\mathcal{B} = \mathbb{Z}_N$ and let $\Psi \in \text{CA}(\mathcal{B}^{\mathbb{Z}})$ be the N -particle CABAM. Construct a factor map $\Gamma: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{B}^{\mathbb{Z}}$ with local rule $\gamma(a_0, a_1) := (a_0 - a_1) \bmod N$, similar to Example 34(b). Then $\Gamma \circ \Phi = \Psi \circ \Gamma$, and the Ψ -particles track the Φ -domain boundaries. Now apply Γ to Proposition 53. \square

Example 55 Let $\mathcal{A} = \{0, 1\}$ and let $\mathbb{H} = \{-1, 0, 1\}$. Elementary Cellular Automaton #18 is the one-dimensional CA with local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ given: $\phi[100] = 1 = \phi[001]$, and $\phi(\mathbf{a}) = 0$ for all other $\mathbf{a} \in \mathcal{A}^{\mathbb{H}}$.

Empirically, $\text{ECA}\#18$ has one stable phase: the *odd sofic shift S*, defined by the \mathcal{A} -labeled digraph $\textcircled{1} \rightleftharpoons \textcircled{0} \rightleftharpoons \textcircled{0}$. In other words, a sequence is admissible to \mathbf{S} as long as an pair of consecutive ones are separated by an *odd* number of zeroes. Thus, a *defect* is any word of the form $10^{2m}1$ (where 0^{2m} represents $2m$ zeroes) for any $m \in \mathbb{N}$. Thus, defects can be arbitrarily large, they can grow and move arbitrarily quickly, and they can coalesce across arbitrarily large distances. Thus, it is impossible to construct a particle CA which tracks the motion of these defects. Nevertheless, in computer simulations, one can visually follow the moving defects through time, and they appear to per-

form random walks. Over time, the density of defects decreases as they randomly collide and annihilate. This was empirically observed by Grassberger [47,48] and Boccara et al. [11]. Lind (see Sect. 5 in [82]) conjectured that this gradual elimination of defects caused almost all initial conditions to converge, in some sense, to \mathbf{S} under application of Φ .

Eloranta and Numelin [34] proved that the defects of Φ individually perform random walks. However, the motions of neighboring defects are highly correlated. They are not *independent* random walks, so one cannot use standard results about stochastic interacting particle systems to conclude that the defect density converges to zero. To solve problems like this, Kůrka [77] developed a theory of ‘particle weight functions’ for CA.

Let \mathcal{A}^* be the set of all finite words in the alphabet \mathcal{A} . A *particle weight function* is a bounded function $p: \mathcal{A}^* \rightarrow \mathbb{N}$, so that, for any $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, we interpret

$$\#_p(\mathbf{a}) := \sum_{r=0}^{\infty} \sum_{z \in \mathbb{Z}} p(\mathbf{a}_{[z \dots z+r]}) \quad \text{and} \\ \delta_p(\mathbf{a}) := \sum_{r=0}^{\infty} \lim_{N \rightarrow \infty} \frac{1}{2N} \sum_{z=-N}^N p(\mathbf{a}_{[z \dots z+r]})$$

to be, respectively the ‘number of particles’ and ‘density of particles’ in configuration \mathbf{a} (clearly $\#_p(\mathbf{a})$ is finite if and only if $\delta_p(\mathbf{a}) = 0$). The function p can count the single-letter ‘particles’ of a PCA, or the short-length ‘domain boundaries’ found in ECA#184 and the CCA of Examples 31 and 32. However, p can also track the arbitrarily large defects of ECA#18. For example, define $p_{18}(10^{2m}1) = 1$ (for any $m \in \mathbb{N}$), and define $p_{18}(\mathbf{a}) = 0$ for all other $\mathbf{a} \in \mathcal{A}^*$.

Let $Z_p := \{\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}; \#_p(\mathbf{a}) = 0\}$ be the set of *vacuum configurations*. (For example, if $p = p_{18}$ as above, then Z_p is just the odd sofic shift \mathbf{S} .) If the iteration of a CA Φ decreases the number (or density) of particles, then one expects Z_p to be a limit set for Φ in some sense. Indeed, if $\mu \in \mathcal{M}_{\text{ltv}}^{\sigma} := \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$, then we define $\Delta_p(\mu) := \int_{\mathcal{A}^{\mathbb{Z}}} \delta_p d\mu$. If Φ is ‘ p -decreasing’ in a certain sense, then Δ_p acts as a Lyapunov function for the dynamical system $(\mathcal{M}_{\text{ltv}}^{\sigma}, \Phi_*)$. Thus, with certain technical assumptions, we can show that, if $\mu \in \mathcal{M}_{\text{ltv}}^{\sigma}$ is connected, then $\Lambda(\Phi, \mu) \subseteq Z_p$ (see Theorem 8 in [77]). Furthermore, under certain conditions, $\text{MeasAtt}(\Phi) \subseteq Z_p$ (see Theorem 7 in [77]). Using this machinery, Kůrka proved:

Proposition 56 *Let $\Phi: \mathcal{A}^{\mathbb{Z}} \rightarrow \mathcal{A}^{\mathbb{Z}}$ be ECA#18, and let $\mathbf{S} \subset \mathcal{A}^{\mathbb{Z}}$ be the odd sofic shift. If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$ is connected, then $\Lambda(\Phi, \mu) \subseteq \mathbf{S}$.*

Proof See Example 6.3 of [77]. \square

Measurable Dynamics

If $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then the triple $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *measure-preserving dynamical system* (MPDS), and thus, amenable to the methods of classical ergodic theory.

Mixing and Ergodicity

If $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$, then the topological dynamical system $(\mathcal{A}^{\mathbb{M}}, \Phi)$ is *topologically transitive* (or *topologically ergodic*) if, if, for any open subsets $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$, there exists $t \in \mathbb{N}$ such that $\mathbf{U} \cap \Phi^{-t}(\mathbf{V}) \neq \emptyset$. Equivalently, there exists some $\mathbf{a} \in \mathcal{A}^{\mathbb{M}}$ whose orbit $\mathcal{O}(\mathbf{a}) := \{\Phi^t(\mathbf{a})\}_{t=0}^{\infty}$ is dense in $\mathcal{A}^{\mathbb{M}}$. If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then the system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *ergodic* if, for any nontrivial measurable $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$, there exists some $t \in \mathbb{N}$ such that $\mu[\mathbf{U} \cap \Phi^{-t}(\mathbf{V})] > 0$. The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *totally ergodic* if $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi^n)$ is ergodic for every $n \in \mathbb{N}$. The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *(strongly) mixing* if, for any nontrivial measurable $\mathbf{U}, \mathbf{V} \subseteq \mathcal{A}^{\mathbb{M}}$,

$$\lim_{t \rightarrow \infty} \mu[\mathbf{U} \cap \Phi^{-t}(\mathbf{V})] = \mu[\mathbf{U}] \cdot \mu[\mathbf{V}]. \quad (6)$$

The system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *weakly mixing* if the limit (6) holds as $n \rightarrow \infty$ along an increasing subsequence $\{t_n\}_{n=1}^{\infty}$ of *density one* – i.e. such that $\lim_{n \rightarrow \infty} t_n/n = 1$. For any $M \in \mathbb{N}$, we say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *M-mixing* if, for any measurable $\mathbf{U}_0, \mathbf{U}_1, \dots, \mathbf{U}_M \subseteq \mathcal{A}^{\mathbb{M}}$,

$$\lim_{\substack{|t_n - t_m| \rightarrow \infty \\ \forall n \neq m \in \{0 \dots M\}}} \mu \left[\bigcap_{m=0}^M \Phi^{-t_m}(\mathbf{U}_m) \right] = \prod_{m=0}^M \mu[\mathbf{U}_m] \quad (7)$$

(thus, ‘strong’ mixing is 1-mixing). We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *multimixing* (or *mixing of all orders*) if $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is *M-mixing* for all $M \in \mathbb{N}$.

We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *Bernoulli endomorphism* if its natural extension (► [Ergodic Theory: Basic Examples and Constructions](#)) is measurably isomorphic to a system $(\mathcal{B}^{\mathbb{Z}}, \beta; \sigma)$, where $\beta \in \mathcal{M}_{\text{ens}}(\mathcal{B}^{\mathbb{Z}}; \sigma)$ is a Bernoulli measure. We say $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$ is a *Kolmogorov endomorphism* if its natural extension is a Kolmogorov (or ‘K’) automorphism; see ► [Ergodicity and Mixing Properties](#). The following chain of implications is well-known; see ► [Ergodicity and Mixing Properties](#).

Theorem 57 *Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{M}})$, let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, and let $\mathbf{X} = \text{supp}(\mu)$. Then \mathbf{X} is a compact, Φ -invariant set. Furthermore:*

(μ, Φ) is Bernoulli $\implies (\mu, \Phi)$ is Kolmogorov $\implies (\mu, \Phi)$ is multimixing $\implies (\mu, \Phi)$ is mixing

$\implies (\mu, \Phi)$ is weakly mixing $\implies (\mu, \Phi)$ is totally ergodic $\implies (\mu, \Phi)$ is ergodic \implies The system (\mathbf{X}, Φ) is topologically transitive $\implies \Phi: \mathbf{X} \rightarrow \mathbf{X}$ is surjective.

Theorem 58 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{N}})$ be posexpansive (see Subsect. “Posexpansive and Permutative CA”). Then $(\mathcal{A}^{\mathbb{N}}, \Phi)$ has topological entropy $\log_2(k)$ for some $k \in \mathbb{N}$, Φ preserves the uniform measure η , and $(\mathcal{A}^{\mathbb{N}}, \eta; \Phi)$ is a uniformly distributed Bernoulli endomorphism on an alphabet of cardinality k .

Proof Extend the argument of Theorem 18. See Corollary 3.10 in [9] or Theorem 4.8(5) in [86]. \square

Example 59 Suppose $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{N}})$ is right-permutative, with neighborhood $[r \dots R]$, where $0 \leq r < R$. Then $h_{\text{top}}(\Phi) = \log_2(|\mathcal{A}|^R)$, so Theorem 58 says that $(\mathcal{A}^{\mathbb{N}}, \eta; \Phi)$ is a uniformly distributed Bernoulli endomorphism on the alphabet $\mathcal{B} := \mathcal{A}^R$.

In this case, it is easy to see this directly. If $\Phi_{\mathbb{B}}^{\mathbb{N}}: \mathcal{A}^{\mathbb{N}} \rightarrow \mathcal{B}^{\mathbb{N}}$ is as in Eq. (2), then $\beta := \Phi_{\mathbb{B}}^{\mathbb{N}}(\eta)$ is the uniform Bernoulli measure on $\mathcal{B}^{\mathbb{N}}$, and $\Phi_{\mathbb{B}}^{\mathbb{N}}$ is an isomorphism from $(\mathcal{A}^{\mathbb{N}}, \mu; \Phi)$ to $(\mathcal{B}^{\mathbb{N}}, \beta; \sigma)$.

Theorem 60 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ have neighborhood $[L \dots R]$. Suppose that

- either** (a) $0 \leq L < R$ and Φ is right-permutative;
- or** (b) $L < R \leq 0$ and Φ is left-permutative;
- or** (c) $L < R$ and Φ is bipermutative;
- or** (d) Φ is posexpansive.

Then Φ preserves the uniform measure η , and $(\mathcal{A}^{\mathbb{Z}}, \eta; \Phi)$ is a Bernoulli endomorphism.

Proof For cases (a) and (b), see Theorem 2.2 in [125]. For case (c), see Theorem 2.7 in [125] or Corollary 7.3 in [74]. For (d), extend the argument of Theorem 14; see Theorem 4.9 in [86]. \square

Remark Theorem 60(c) can be extended to some higher-dimensional permutative CA using Proposition 1 in [3]; see Remark 12(b).

Theorem 61 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ have neighborhood $[L \dots R]$. Suppose that

- either** (a) Φ is surjective and $0 < L \leq R$;
- or** (b) Φ is surjective and $L \leq R < 0$;
- or** (c) Φ is right-permutative and $R \neq 0$;
- or** (d) Φ is left-permutative and $L \neq 0$.

Then Φ preserves η , and $(\mathcal{A}^{\mathbb{Z}}, \eta; \Phi)$ is a Kolmogorov endomorphism.

Proof Cases (a) and (b) are Theorem 2.4 in [125]. Cases (c) and (d) are from [129]. \square

Corollary 62 Any CA satisfying the hypotheses of Theorem 61 is multimixing.

Proof This follows from Theorems 57 and 61. See also Theorem 3.2 in [131] for a direct proof that any CA satisfying hypotheses (a) or (b) is 1-mixing. See Theorem 6.6 in [74] for a proof that any CA satisfying hypotheses (c) or (d) is multimixing. \square

Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ have neighborhood \mathbb{H} . An element $\mathbf{x} \in \mathbb{H}$ is *extremal* if $\langle \mathbf{x}, \mathbf{x} \rangle > \langle \mathbf{x}, \mathbf{h} \rangle$ for all $\mathbf{h} \in \mathbb{H} \setminus \{\mathbf{x}\}$. We say Φ is *extremally permutative* if Φ is permutative in some extremal coordinate.

Theorem 63 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let η be the uniform measure. If Φ is extremally permutative, then $(\mathcal{A}^{\mathbb{Z}^D}, \eta; \Phi)$ is mixing.

Proof See Theorem A in [144] for the case $D = 2$ and $\mathcal{A} = \mathbb{Z}/2$. Willson described Φ as ‘linear’ in an extremal coordinate (which is equivalent to permutative when $\mathcal{A} = \mathbb{Z}/2$), and then concluded that Φ was ‘ergodic’ – however, he did this by explicitly showing that Φ was mixing. His proof technique easily generalizes to any extremally permutative CA on any alphabet, and any $D \geq 1$. \square

Theorem 64 Let $\mathcal{A} = \mathbb{Z}/m$. Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ have linear local rule $\phi: \mathcal{A}^{\mathbb{H}} \rightarrow \mathcal{A}$ given by $\phi(\mathbf{a}_{\mathbb{H}}) = \sum_{h \in \mathbb{H}} c_h \cdot a_h$, where $c_h \in \mathbb{Z}$ for all $h \in \mathbb{H}$. Let η be the uniform measure on $\mathcal{A}^{\mathbb{Z}^D}$. The following are equivalent:

- (a) Φ preserves η and $(\mathcal{A}^{\mathbb{Z}^D}, \eta, \Phi)$ is ergodic.
- (b) $(\mathcal{A}^{\mathbb{Z}^D}, \Phi)$ is topologically transitive.
- (c) $\gcd\{c_h\}_{0 \neq h \in \mathbb{H}}$ is coprime to m .
- (d) For all prime divisors p of m , there is some nonzero $h \in \mathbb{H}$ such that c_h is not divisible by p .

Proof Theorem 3.2 in [18]; see also [17]. For a different proof in the case $D = 2$, see Theorem 6 in [123]. \square

Spectral Properties

If $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}})$, then let $\mathbf{L}_{\mu}^2 = \mathbf{L}^2(\mathcal{A}^{\mathbb{M}}, \mu)$ be the set of measurable functions $f: \mathcal{A}^{\mathbb{M}} \rightarrow \mathbb{C}$ such that $\|f\|_2 := (\int_{\mathcal{A}^{\mathbb{M}}} |f|^2 d\mu)^{1/2}$ is finite. If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{M}})$ and $\mu \in \mathfrak{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, then Φ defines a unitary linear operator $\Phi_*: \mathbf{L}_{\mu}^2 \rightarrow \mathbf{L}_{\mu}^2$ by $\Phi_*(f) = f \circ \Phi$ for all $f \in \mathbf{L}_{\mu}^2$. If $f \in \mathbf{L}_{\mu}^2$, then f is an *eigenfunction* of Φ , with *eigenvalue* $c \in \mathbb{C}$, if $\Phi_*(f) = c \cdot f$. By definition of Φ_* , any eigenvalue must be an element of the unit circle $\mathbb{T} := \{c \in \mathbb{C} : |c| = 1\}$. Let $\mathbb{S}_{\Phi} \subset \mathbb{T}$ be the set of all eigenvalues of Φ , and for any $s \in \mathbb{S}_{\Phi}$, let $\mathbf{E}_s(\Phi) := \{f \in \mathbf{L}_{\mu}^2 : \Phi_* f = sf\}$ be the corresponding eigenspace. For example, if f is constant μ -almost everywhere, then $f \in \mathbf{E}_1(\Phi)$. Let $\mathbf{E}(\Phi) := \bigcup_{s \in \mathbb{S}_{\Phi}} \mathbf{E}_s(\Phi)$. Note that \mathbb{S}_{Φ} is a group. Indeed, if $s_1, s_2 \in \mathbb{S}_{\Phi}$, and $f_1 \in \mathbf{E}_{s_1}$ and $f_2 \in \mathbf{E}_{s_2}$,

then $(f_1 f_2) \in E_{s_1 s_2}$ and $(1/f_1) \in E_{1/s_1}$. Thus, S_Φ is called the *spectral group* of Φ .

If $s \in S_\Phi$, then heuristically, an s -eigenfunction is an ‘observable’ of the dynamical system $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ which exhibits quasiperiodically recurrent behavior. Thus, the spectral properties of Φ characterize the ‘recurrent aspect’ of its dynamics (or the lack thereof). For example:

- $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is ergodic $\iff E_1(\Phi)$ contains only constant functions $\iff \dim[E_s(\Phi)] = 1$ for all $s \in S_\Phi$.
- $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is weakly mixing (see Subsect. “Mixing and Ergodicity”) $\iff E(\Phi)$ contains only constant functions $\iff (\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is ergodic and $S_\Phi = \{1\}$.

We say $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ has *discrete spectrum* if L^2_μ is spanned by $E(\Phi)$. In this case, $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is measurably isomorphic to an MPDS defined by translation on a compact abelian group (e.g. an irrational rotation of a torus, an odometer, etc.).

If $\mu \in \mathcal{M}_{\text{inv}}(\mathcal{A}^\mathbb{M}, \sigma)$, then there is a natural unitary \mathbb{M} -action on L^2_μ , where $\sigma_*^m(f) = f \circ \sigma_*^m$. A *character* of \mathbb{M} is a monoid homomorphism $\chi: \mathbb{M} \rightarrow \mathbb{T}$. The set $\widehat{\mathbb{M}}$ of all characters is a group under pointwise multiplication, called the *dual group* of \mathbb{M} . If $f \in L^2_\mu$ and $\chi \in \widehat{\mathbb{M}}$, then f is a χ -eigenfunction of $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ if $\sigma_*^m(f) = \chi(m) \cdot f$ for all $m \in \mathbb{M}$; then χ is called a *eigencharacter*. The *spectral group* of $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is then the subgroup $S_\sigma \subset \widehat{\mathbb{M}}$ of all eigencharacters. For any $\chi \in S_\sigma$, let $E_\chi(\sigma)$ be the corresponding eigenspace, and let $E(\sigma) := \bigsqcup_{\chi \in S_\sigma} E_\chi(\sigma)$.

- $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is ergodic $\iff E_1(\sigma)$ contains only constant functions $\iff \dim[E_\chi(\sigma)] = 1$ for all $\chi \in S_\sigma$.
- $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is weakly mixing $\iff E(\sigma)$ contains only constant functions $\iff (\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is ergodic and $S_\sigma = \{1\}$.

$(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ has *discrete spectrum* if L^2_μ is spanned by $E(\sigma)$. In this case, the system $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is measurably isomorphic to an action of \mathbb{M} by translations on a compact abelian group.

Example 65 Let $\mathbb{M} = \mathbb{Z}$; then any character $\chi: \mathbb{Z} \rightarrow \mathbb{T}$ has the form $\chi(n) = c^n$ for some $c \in \mathbb{T}$, so a χ -eigenfunction is just a eigenfunction with eigenvalue c . In this case, the aforementioned spectral properties for the \mathbb{Z} -action by shifts are equivalent to the corresponding spectral properties of the CA $\Phi = \sigma^1$. Bernoulli measures and irreducible Markov chains are weakly mixing. On other hand, several important classes of symbolic dynamical systems have dis-

crete spectrum, including Sturmian shifts, constant-length substitution shifts, and regular Toeplitz shifts; see ► [Symbolic Dynamics](#) and ► [Dynamics of Cellular Automata in Non-compact Spaces](#).

Proposition 66 Let $\Phi \in \mathcal{CA}(\mathcal{A}^\mathbb{M})$, and let $\mu \in \mathcal{M}_{\text{inv}}(\mathcal{A}^\mathbb{M}; \sigma)$ be σ -ergodic.

- (a) $E(\sigma) \subseteq E(\Phi)$.
- (b) If $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ has discrete spectrum, then so does $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$.
- (c) Suppose μ is Φ -ergodic. If $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is weakly mixing, then so is $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$.

Proof (a) Suppose $\chi \in \widehat{\mathbb{M}}$ and $f \in E_\chi$. Then $f \circ \Phi \in E_\chi$ also, because for all $m \in \mathbb{M}$, $f \circ \Phi \circ \sigma_*^m = f \circ \sigma_*^m \circ \Phi = \chi(m) \cdot f \circ \Phi$. But if $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is ergodic, then $\dim[E_\chi(\sigma)] = 1$; hence $f \circ \Phi$ must be a scalar multiple of f . Thus, f is also an eigenfunction for Φ . (b) follows from (a).

(c) By reversing the roles of Φ and σ in (a), we see that $E(\Phi) \subseteq E(\sigma)$. But if $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is weakly mixing, then $E(\sigma) = \{\text{constant functions}\}$. Thus, $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is also weakly mixing. \square

Example 67 (a) Let μ be any Bernoulli measure on $\mathcal{A}^\mathbb{M}$. If μ is Φ -invariant and Φ -ergodic, then $(\mathcal{A}^\mathbb{M}, \mu; \Phi)$ is weakly mixing (because $(\mathcal{A}^\mathbb{M}, \mu; \sigma)$ is weakly mixing).

(b) Let $P \in \mathbb{N}$ and suppose μ is a Φ -invariant measure supported on the set X_P of P -periodic sequences (see Proposition 10). Then $(\mathcal{A}^\mathbb{Z}, \mu; \sigma)$ has discrete spectrum (with rational eigenvalues). But X_P is finite, so the system (X_P, Φ) is also periodic; hence $(\mathcal{A}^\mathbb{Z}, \mu; \Phi)$ also has discrete spectrum (with rational eigenvalues).

(c) Downarowicz [25] has constructed an example of a regular Toeplitz shift $X \subset \mathcal{A}^\mathbb{Z}$ and $\Phi \in \mathcal{CA}(\mathcal{A}^\mathbb{Z})$ (not the shift) such that $\Phi(X) \subseteq X$. Any regular Toeplitz shift is uniquely ergodic, and the unique shift-invariant measure μ has discrete spectrum; thus, $(\mathcal{A}^\mathbb{Z}, \mu; \Phi)$ also has discrete spectrum.

Aside from Examples 67(b,c), the literature contains no examples of discrete-spectrum, invariant measures for CA; this is an interesting area for future research.

Entropy

Let $\Phi \in \mathcal{CA}(\mathcal{A}^\mathbb{M})$. For any finite $\mathbb{B} \subset \mathbb{M}$, let $\mathcal{B} := \mathcal{A}^\mathbb{B}$, let $\Phi_\mathbb{B}^\mathbb{N}: \mathcal{A}^\mathbb{N} \rightarrow \mathcal{B}^\mathbb{N}$ be as in Eq. (2), and let $X := \Phi_\mathbb{B}^\mathbb{N}(\mathcal{A}^\mathbb{M}) \subseteq \mathcal{A}^\mathbb{N}$; then define

$$H_{\text{top}}(\mathbb{B}; \Phi) := h_{\text{top}}(X) = \lim_{T \rightarrow \infty} \frac{1}{T} \log_2(\#X_{[0..T)}).$$

If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)$, let $\nu := \Phi_{\mathbb{B}}^{\mathbb{N}}(\mu)$; then ν is a σ -invariant measure on $\mathcal{B}^{\mathbb{N}}$. Define

$$H_{\mu}(\mathbb{B}; \Phi) := h_{\nu}(\sigma) = - \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{\mathbf{b} \in \mathcal{B}^{[0 \dots T]}} \nu[\mathbf{b}] \log_2(\nu[\mathbf{b}]).$$

The *topological entropy* of $(\mathcal{A}^{\mathbb{M}}, \Phi)$ and the *measurable entropy* of $(\mathcal{A}^{\mathbb{M}}, \Phi, \mu)$ are then defined

$$h_{\text{top}}(\Phi) := \sup_{\substack{\mathbb{B} \subset \mathbb{M} \\ \text{finite}}} H_{\text{top}}(\mathbb{B}; \Phi) \quad \text{and}$$

$$h_{\mu}(\Phi) := \sup_{\substack{\mathbb{B} \subset \mathbb{M} \\ \text{finite}}} H_{\mu}(\mathbb{B}; \Phi).$$

The famous *Variational Principle* states that $h_{\text{top}}(\Phi) = \sup \{h_{\mu}(\Phi); \mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{M}}, \Phi)\}$; see ► [Entropy in Ergodic Theory](#) or Sect. 10 ► [Topological Dynamics of Cellular Automata](#).

If \mathbb{M} has more than one dimension (e.g. $\mathbb{M} = \mathbb{Z}^D$ or \mathbb{N}^D for $D \geq 2$) then most CA on $\mathcal{A}^{\mathbb{M}}$ have infinite entropy. Thus, entropy is mainly of interest in the case $\mathbb{M} = \mathbb{Z}$ or \mathbb{N} . Coven [23] was the first to compute the topological entropy of a CA; he showed that $h_{\text{top}}(\Phi) = 1$ for a large class of left-permutative, one-sided CA on $\{0, 1\}^{\mathbb{N}}$ (which have since been called *Coven CA*). Later, Lind [83] showed how to construct CA whose topological entropy was any element of a countable dense subset of \mathbb{R}_+ , consisting of logarithms of certain algebraic numbers. Theorems 14 and 18(b) above characterize the topological entropy of posexexpansive CA. However, Hurd et al. [62] showed that there is no algorithm which can compute the topological entropy of an arbitrary CA; see ► [Tiling Problem and Undecidability in Cellular Automata](#).

Measurable entropy has also been computed for a few special classes of CA. For example, if $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$ is bipermutative with neighborhood $\{0, 1\}$ and $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \Phi; \sigma)$ is σ -ergodic, then $h_{\mu}(\Phi) = \log_2(K)$ for some integer $K \leq |\mathcal{A}|$ (see Thm 4.1 in [113]). If η is the uniform measure, and Φ is posexexpansive, then Theorems 58 and 60 above characterize $h_{\eta}(\Phi)$. Also, if Φ satisfies the conditions of Theorem 61, then $h_{\eta}(\Phi) > 0$, and furthermore, all factors of the MPDS $(\mathcal{A}^{\mathbb{Z}}, \mu; \Phi)$ also have positive entropy.

However, unlike abstract dynamical systems, CA come with an explicit spatial ‘geometry’. The most fruitful investigations of CA entropy are those which have interpreted entropy in terms of how information propagates through this geometry.

Lyapunov Exponents

Wolfram [150] suggested that the propagation speed of ‘perturbations’ in a one-dimensional CA Φ could trans-

form ‘spatial’ entropy [i.e. $h(\sigma)$] into ‘temporal’ entropy [i.e. $h(\Phi)$]. He compared this propagation speed to the ‘Lyapunov exponent’ of a smooth dynamical system: it determines the exponential rate of divergence between two initially close Φ -orbits (see pp. 172, 261 and 514 in [151]). Shereshevsky [126] formalized Wolfram’s intuition and proved the conjectured entropy relationship; his results were later improved by Tisseur [141]. Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$, let $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, and let $z \in \mathbb{Z}$. Define

$$\mathbf{W}_z^+(\mathbf{a}) := \left\{ \mathbf{w} \in \mathcal{A}^{\mathbb{Z}}; \mathbf{w}_{[z \dots \infty)} = \mathbf{a}_{[z \dots \infty)} \right\}, \quad \text{and}$$

$$\mathbf{W}_z^-(\mathbf{a}) := \left\{ \mathbf{w} \in \mathcal{A}^{\mathbb{Z}}; \mathbf{w}_{(-\infty \dots z]} = \mathbf{a}_{(-\infty \dots z]} \right\}.$$

Thus, we obtain each $\mathbf{w} \in \mathbf{W}_z^+(\mathbf{a})$ (respectively $\mathbf{W}_z^-(\mathbf{a})$) by ‘perturbing’ \mathbf{a} somewhere to the left (resp. right) of coordinate z . Next, for any $t \in \mathbb{N}$, define

$$\widetilde{\Lambda}_t^+(\mathbf{a}) := \min \{z \in \mathbb{N}; \Phi^t[\mathbf{W}_0^+(\mathbf{a})] \subseteq \mathbf{W}_z^+(\Phi^t[\mathbf{a}])\},$$

and

$$\widetilde{\Lambda}_t^-(\mathbf{a}) := \min \{z \in \mathbb{N}; \Phi^t[\mathbf{W}_0^-(\mathbf{a})] \subseteq \mathbf{W}_{-z}^-(\Phi^t[\mathbf{a}])\}.$$

Thus, $\widetilde{\Lambda}_t^{\pm}$ measures the farthest distance which any perturbation of \mathbf{a} at coordinate 0 could have propagated by time t . Next, define $\Lambda_t^{\pm}(\mathbf{a}) := \max_{z \in \mathbb{Z}} \widetilde{\Lambda}_t^{\pm}(\sigma^z(\mathbf{a}))$. Then Shereshevsky [126] defined the (*maximum*) *Lyapunov exponents*

$$\lambda^+(\Phi, \mathbf{a}) := \lim_{t \rightarrow \infty} \frac{1}{t} \Lambda_t^+(\mathbf{a}), \quad \text{and}$$

$$\lambda^-(\Phi, \mathbf{a}) := \lim_{t \rightarrow \infty} \frac{1}{t} \Lambda_t^-(\mathbf{a}),$$

whenever these limits exist. Let $\mathbf{G}(\Phi) := \{\mathbf{g} \in \mathcal{A}^{\mathbb{Z}}; \lambda^{\pm}(\Phi, \mathbf{g}) \text{ both exist}\}$. The subset $\mathbf{G}(\Phi)$ is ‘generic’ within $\mathcal{A}^{\mathbb{Z}}$ in a very strong sense, and the Lyapunov exponents detect ‘chaotic’ topological dynamics.

Proposition 68 *Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}})$.*

- (a) *Let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}, \sigma)$. Suppose that either: [i] μ is also Φ -invariant; or: [ii] μ is σ -ergodic and $\text{supp}(\mu)$ is a Φ -invariant subset. Then $\mu(\mathbf{G}) = 1$.*
- (b) *The set \mathbf{G} and the functions $\lambda^{\pm}(\Phi, \bullet)$ are (Φ, σ) -invariant. Thus, if μ is either Φ -ergodic or σ -ergodic, then there exist constants $\lambda_{\mu}^{\pm}(\Phi) \geq 0$ such that $\lambda^{\pm}(\Phi, \mathbf{g}) = \lambda_{\mu}^{\pm}(\Phi)$ for μ -all $\mathbf{g} \in \mathbf{G}$.*
- (c) *If Φ is posexexpansive, then there is a constant $c > 0$ such that $\lambda^{\pm}(\Phi, \mathbf{g}) \geq c$ for all $\mathbf{g} \in \mathbf{G}$.*
- (d) *Let η be the uniform Bernoulli measure. If Φ is surjective, then $h_{\text{top}}(\Phi) \leq (\lambda_{\eta}^+(\Phi) + \lambda_{\eta}^-(\Phi)) \cdot \log |\mathcal{A}|$.*

Proof (a) follows from the fact that, for any $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$, the sequence $[\Lambda_t^{\pm}(\mathbf{a})]_{t \in \mathbb{N}}$ is subadditive in t . Condition [i]

is Theorem 1 in [126], and follows from Kingman's sub-additive ergodic theorem. Condition [ii] is Proposition 3.1 in [141].

(b) is clear by definition of λ^\pm . (c) is Theorem 5.2 in [37]. (d) is Proposition 5.3 in [141]. \square

For any Φ -ergodic $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$, Shereshevsky (see Theorem 2 in [126]) showed that $h_\mu(\Phi) \leq (\lambda_\mu^+(\Phi) + \lambda_\mu^-(\Phi)) \cdot h_\mu(\sigma)$. Tisseur later improved this estimate. For any $T \in \mathbb{N}$, let

$$\begin{aligned} \tilde{I}_T^+(\mathbf{a}) &:= \min \{z \in \mathbb{N} ; \forall t \in [1 \dots T], \\ &\quad \Phi^t[\mathbf{W}_z^+(\mathbf{a})] \subseteq \mathbf{W}_0^+(\Phi^t[\mathbf{a}])\} \quad \text{and} \\ \tilde{I}_T^-(\mathbf{a}) &:= \min \{z \in \mathbb{N} ; \forall t \in [1 \dots T], \\ &\quad \Phi^t[\mathbf{W}_z^-(\mathbf{a})] \subseteq \mathbf{W}_0^-(\Phi^t[\mathbf{a}])\}. \end{aligned}$$

Next, for any $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$, define $\hat{I}_T^\pm(\mu) := \int_{\mathcal{A}^{\mathbb{Z}}} \tilde{I}_T^\pm(\mathbf{a}) d\mu[\mathbf{a}]$.

Tisseur then defined the *average Lyapunov exponents*: $I_\mu^\pm(\Phi) := \liminf_{T \rightarrow \infty} \hat{I}_T^\pm(\mu)/T$.

Theorem 69 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \sigma)$.

- (a) If $\text{supp}(\mu)$ is Φ -invariant, then $I_\mu^+(\Phi) \leq \lambda_\mu^+(\Phi)$ and $I_\mu^-(\Phi) \leq \lambda_\mu^-(\Phi)$, and one or both inequalities are sometimes strict.
- (b) If μ is σ -ergodic and Φ -invariant, then $h_\mu(\Phi) \leq (I_\mu^+(\Phi) + I_\mu^-(\Phi)) \cdot h_\mu(\sigma)$, and this inequality is sometimes strict.
- (c) If $\text{supp}(\mu)$ contains Φ -equicontinuous points, then $I_\mu^+(\Phi) = I_\mu^-(\Phi) = h_\mu(\Phi) = 0$.

Proof See [141]: (a) is Proposition 3.2 and Example 6.1; (b) is Theorem 5.1 and Example 6.2; and (c) is Proposition 5.2. \square

Directional Entropy

Milnor [98,99] introduced directional entropy to capture the intuition that information in a CA propagates in particular directions with particular 'velocities', and that different CA 'mix' information in different ways. Classical entropy is unable to detect this informational anisotropy. For example, if $\mathcal{A} = \{0, 1\}$ and $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ has local rule $\phi(a_0, a_1) = a_0 + a_1 \pmod{2}$, then $h_{\text{top}}(\Phi) = 1 = h_{\text{top}}(\sigma)$, despite the fact that Φ vigorously 'mixes' information together and propagates any 'perturbation' outwards in an expanding cone, whereas σ merely shifts information to the left in a rigid and essentially trivial fashion.

If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$, then a *complete history* for Φ is a sequence $(\mathbf{a}_t)_{t \in \mathbb{Z}} \in (\mathcal{A}^{\mathbb{Z}^D})^{\mathbb{Z}} \cong \mathcal{A}^{\mathbb{Z}^{D+1}}$ such that $\Phi(\mathbf{a}_t) =$

\mathbf{a}_{t+1} for all $t \in \mathbb{Z}$. Let $\mathbf{X}^{\text{Hist}} := \mathbf{X}^{\text{Hist}}(\Phi) \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be the subshift of all complete histories for Φ , and let σ be the \mathbb{Z}^{D+1} shift action on \mathbf{X}^{Hist} , then $(\mathbf{X}^{\text{Hist}}; \sigma)$ is conjugate to the natural extension of the system $(\mathbf{Y}; \Phi, \sigma)$, where $\mathbf{Y} := \Phi^\infty(\mathcal{A}^{\mathbb{M}}) := \bigcap_{t=1}^\infty \Phi^t(\mathcal{A}^{\mathbb{Z}^D})$ is the omega-limit set of Φ . If $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$, then $\text{supp}(\mu) \subseteq \mathbf{Y}$, and μ extends to a σ -invariant measure $\tilde{\mu}$ on \mathbf{X}^{Hist} in the obvious way.

Let $\vec{v} = (v_0; v_1, \dots, v_D) \in \mathbb{R} \times \mathbb{R}^D \cong \mathbb{R}^{D+1}$. For any bounded open subset $B \subset \mathbb{R}^{D+1}$ and $T > 0$, let $B(T\vec{v}) := \{b + t\vec{v}; b \in B \text{ and } t \in [0, T]\}$ be the 'sheared cylinder' in \mathbb{R}^{D+1} with cross-section B and length $T|\vec{v}|$ in the direction \vec{v} , and let $\mathbb{B}(T\vec{v}) := B(T\vec{v}) \cap \mathbb{Z}^{D+1}$. Let $\mathbf{X}_{\mathbb{B}(T\vec{v})}^{\text{Hist}} := \{\mathbf{x}_{\mathbb{B}(T\vec{v})}; \mathbf{x} \in \mathbf{X}^{\text{Hist}}(\Phi)\}$. We define

$$H_{\text{top}}(\Phi; B, \vec{v}) := \limsup_{T \rightarrow \infty} \frac{1}{T} \log_2 [\# \mathbf{X}_{\mathbb{B}(T\vec{v})}^{\text{Hist}}]; \quad \text{and}$$

$$H_\mu(\Phi; B, \vec{v}) := - \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{\mathbf{x} \in \mathbf{X}_{\mathbb{B}(T\vec{v})}^{\text{Hist}}} \tilde{\mu}[\mathbf{x}] \log_2(\tilde{\mu}[\mathbf{x}]).$$

We then define the \vec{v} -directional topological entropy and \vec{v} -directional μ -entropy of Φ by

$$h_{\text{top}}(\Phi; \vec{v}) := \sup_{\substack{B \subset \mathbb{R}^{D+1} \\ \text{open \& bounded}}} h_{\text{top}}(\Phi; B, \vec{v}); \quad \text{and} \quad (8)$$

$$h_\mu(\Phi; \vec{v}) := \sup_{\substack{B \subset \mathbb{R}^{D+1} \\ \text{open \& bounded}}} h_\mu(\Phi; B, \vec{v}). \quad (9)$$

Proposition 70 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$.

- (a) Directional entropy is homogeneous. That is, for any $\vec{v} \in \mathbb{R}^{D+1}$ and $r > 0$, $h_{\text{top}}(\Phi, r\vec{v}) = r \cdot h_{\text{top}}(\Phi, \vec{v})$ and $h_\mu(\Phi, r\vec{v}) = r \cdot h_\mu(\Phi, \vec{v})$.
- (b) If $\vec{v} = (t; \mathbf{z}) \in \mathbb{Z} \times \mathbb{Z}^D$, then $h_{\text{top}}(\Phi, \vec{v}) = h_{\text{top}}(\Phi^t \circ \sigma^{\mathbf{z}})$ and $h_\mu(\Phi, \vec{v}) = h_\mu(\Phi^t \circ \sigma^{\mathbf{z}})$.
- (c) There is an extension of the $\mathbb{Z} \times \mathbb{Z}^D$ -system $(\mathbf{X}^{\text{Hist}}, \Phi; \sigma)$ to an $\mathbb{R} \times \mathbb{R}^D$ -system $(\tilde{\mathbf{X}}, \tilde{\Phi}, \tilde{\sigma})$ such that, for any $\vec{v} = (t; \vec{u}) \in \mathbb{R} \times \mathbb{R}^D$ we have $h_{\text{top}}(\Phi, \vec{v}) = h_{\text{top}}(\tilde{\Phi}^t \circ \tilde{\sigma}^{\vec{u}})$.

For any $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi, \sigma)$, there is an extension $\tilde{\mu} \in \mathcal{M}_{\text{ens}}(\tilde{\mathbf{X}}; \tilde{\Phi}, \tilde{\sigma})$ such that for any $\vec{v} = (t; \vec{u}) \in \mathbb{R} \times \mathbb{R}^D$ we have $h_\mu(\Phi, \vec{v}) = h_{\tilde{\mu}}(\tilde{\Phi}^t \circ \tilde{\sigma}^{\vec{u}})$.

Proof (a,b) follow from the definition. (c) is Proposition 2.1 in [109]. \square

Remark 71 Directional entropy can actually be defined for any continuous \mathbb{Z}^{D+1} -action on a compact metric space, and in particular, for any subshift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$. The directional entropy of a CA Φ is then just the directional

entropy of the subshift $\mathbf{X}^{\text{Hist}}(\Phi)$. Proposition 70 holds for any subshift.

Directional entropy is usually infinite for multidimensional CA (for the same reason that classical entropy is usually infinite). Thus, most of the analysis has been for one-dimensional CA. For example, Kitchens and Schmidt (see Sect. 1 in [72]) studied the directional topological entropy of one-dimensional linear CA, while Smilie (see Proposition 1.1 in [133]) computed the directional topological entropy for ECA#184. If Φ is linear, then the function $\vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v})$ is piecewise linear and convex, but if Φ is ECA#184, it is neither.

If \vec{v} has rational entries, then Proposition 70(a,b) shows that $h(\Phi, \vec{v})$ is a rational multiple of the classical entropy of some composite CA, which can be computed through classical methods. However, if \vec{v} is irrational, then $h(\Phi, \vec{v})$ is quite difficult to compute using the formulae (8) and (9), and Proposition 70(c), while theoretically interesting, is not very computationally useful. Can we compute $h(\Phi, \vec{v})$ as the limit of $h(\Phi, \vec{v}_k)$ where $\{\vec{v}_k\}_{k=1}^{\infty}$ is a sequence of rational vectors tending to \vec{v} ? In other words, is directional entropy *continuous* as a function of \vec{v} ? What other properties has $h(\Phi, \vec{v})$ as a function of \vec{v} ?

Theorem 72 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ and let $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}}; \Phi)$.

- (a) The function $\mathbb{R}^2 \ni \vec{v} \mapsto h_{\mu}(\Phi, \vec{v}) \in \mathbb{R}$ is continuous.
- (b) Suppose there is some $(t, z) \in \mathbb{N} \times \mathbb{Z}$ with $t \geq 1$, such that $\Phi^t \circ \sigma^z$ is *posexpansive*. Then the function $\mathbb{R}^2 \ni \vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v}) \in \mathbb{R}$ is convex, and thus, Lipschitz-continuous.
- (c) However, there exist other $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$ for which the function $\mathbb{R}^2 \ni \vec{v} \mapsto h_{\text{top}}(\Phi, \vec{v}) \in \mathbb{R}$ is not continuous.
- (d) Suppose Φ has neighborhood $[-\ell \dots r] \subset \mathbb{Z}$. If $\vec{v} = (t; x) \in \mathbb{R}^2$, then let $z_{\ell} := x - \ell t$ and $z_r := x + rt$. Let $L := \log |\mathcal{A}|$.

[i] Suppose $z_{\ell} \cdot z_r \geq 0$. Then $h_{\mu}(\Phi; \vec{v}) \leq \max\{|z_{\ell}|, |z_r|\} \cdot L$. Furthermore:

- If Φ is *right-permutative*, and $|z_{\ell}| \leq |z_r|$, then $h_{\mu}(\Phi; \vec{v}) = |z_r| \cdot L$.
- If Φ is *left-permutative*, and $|z_r| \leq |z_{\ell}|$, then $h_{\mu}(\Phi; \vec{v}) = |z_{\ell}| \cdot L$.

[ii] Suppose $z_{\ell} \cdot z_r \leq 0$. Then $h_{\mu}(\Phi; \vec{v}) \leq |z_r - z_{\ell}| \cdot L$. Furthermore, if Φ is *bipermutative* in this case, then $h_{\mu}(\Phi; \vec{v}) = |z_r - z_{\ell}| \cdot L$.

Proof (a) is Corollary 3.3 in [109], while (b) is Théorème III.11 and Corollaire III.12, pp. 79–80 in [120]. (c) is Proposition 1.2 in [133].

(d) summarizes the main results of [21]. See also Example 6.2 in [99] for an earlier analysis of permutative

CA in the case $r = \ell = 1$; see also Example 6.4 in [12] and Sect. 1 in Sect. 1 in [72] for the special case when Φ is linear. \square

Remark 73 (a) In fact, the conclusion of Theorem 72(b) holds as long as Φ has any *posexpansive directions* (even irrational ones). A posexpansive direction is analogous to an *expansive subspace* (see Subsect. “Entropy Geometry and Expansive Subdynamics”), and is part of Sablik’s theory of ‘directional dynamics’ for one-dimensional CA; see Remark 84(b) below. Using this theory, Sablik has also shown that $h_{\mu}(\Phi; \vec{v}) = 0 = h_{\text{top}}(\Phi, \vec{v})$ whenever \vec{v} is an *equicontinuous direction* for Φ , whereas $h_{\mu}(\Phi; \vec{v}) \neq 0 \neq h_{\text{top}}(\Phi, \vec{v})$ whenever \vec{v} is a *right- or left posexpansive direction* for Φ . See Sect. §III.4.5–Sect. §III.4.6, pp. 86–88 in [120].

(b) Courbage and Kamiński have defined a ‘directional’ version of the Lyapunov exponents introduced in Subsect. “Lyapunov Exponents”. If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$, $\mathbf{a} \in \mathcal{A}^{\mathbb{Z}}$ and $\vec{v} = (t; z) \in \mathbb{N} \times \mathbb{Z}$, then $\lambda_{\vec{v}}^{\pm}(\Phi, \mathbf{a}) := \lambda^{\pm}(\Phi^t \circ \sigma^z, \mathbf{a})$, where λ^{\pm} are defined as in Subsect. “Lyapunov Exponents”. If $\vec{v} \in \mathbb{R}^2$ is irrational, then the definition of $\lambda_{\vec{v}}^{\pm}(\Phi, \mathbf{a})$ is somewhat more subtle. For any Φ and \mathbf{a} , the function $\mathbb{R}^2 \ni \vec{v} \mapsto \lambda_{\vec{v}}^{\pm}(\Phi, \mathbf{a}) \in \mathbb{R}$ is homogeneous and continuous (see Lemma 2 and Proposition 3 in [22]). If $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}}; \Phi, \sigma)$ is σ -ergodic, then $\lambda_{\vec{v}}^{\pm}(\Phi, \bullet)$ is constant μ -almost everywhere, and is related to $h_{\mu}(\Phi; \vec{v})$ through an inequality exactly analogous to Theorem 69(b); see Theorem 1 in [22].

Cone Entropy For any $\vec{v} \in \mathbb{R}^{D+1}$, any angle $\theta > 0$, and any $N > 0$, we define

$$\mathbb{K}(N\vec{v}, \theta) := \left\{ \mathbf{z} \in \mathbb{Z}^{D+1}; |\mathbf{z}| \leq N|\vec{v}| \text{ and } \mathbf{z} \bullet \vec{v} / |\mathbf{z}||\vec{v}| \geq \cos(\theta) \right\}.$$

Geometrically, this is the set of all \mathbb{Z}^{D+1} -lattice points in a cone of length $N|\vec{v}|$ which subtends an angle of 2θ around an axis parallel to \vec{v} , and which has its apex at the origin. If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$, then let $\mathbf{X}^{\text{Hist}}(N\vec{v}, \theta) := \{\mathbf{x}_{\mathbb{K}(N\vec{v}, \theta)}; \mathbf{x} \in \mathbf{X}^{\text{Hist}}(\Phi)\}$. If $\mu \in \mathfrak{M}_{\text{cons}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi)$, and $\tilde{\mu}$ is the extension of μ to \mathbf{X}^{Hist} , then the *cone entropy* of (Φ, μ) in direction \vec{v} is defined

$$h_{\mu}^{\text{cone}}(\Phi, \vec{v}) := - \lim_{\theta \searrow 0} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{\mathbf{x} \in \mathbf{X}^{\text{Hist}}(N\vec{v}, \theta)} \tilde{\mu}[\mathbf{x}] \log_2(\tilde{\mu}[\mathbf{x}]).$$

Park [107,108] attributes this concept to Doug Lind. Like directional entropy, cone entropy can be defined for any

continuous \mathbb{Z}^{D+1} -action, and is generally infinite for multidimensional CA. However, for one-dimensional CA, Park has proved:

Theorem 74 If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$, $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}}; \Phi)$ and $\vec{v} \in \mathbb{R}^2$, then $h_{\mu}^{\text{cone}}(\Phi, \vec{v}) = h_{\mu}(\Phi, \vec{v})$.

Proof See Theorem 1 in [108]. \square

Entropy Geometry and Expansive Subdynamics

Directional entropy is the one-dimensional version of a multidimensional ‘entropy density’ function, which was introduced by Milnor [99] to address the fact that classical and directional entropy are generally infinite for multidimensional CA. Milnor’s ideas were then extended by Boyle and Lind [12], using their theory of *expansive subdynamics*.

Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, and let $\mu \in \mathcal{M}_{\text{ens}}(\mathbf{X}; \sigma)$. For any bounded $B \subset \mathbb{R}^{D+1}$, let $\mathbb{B} := B \cap \mathbb{Z}^{D+1}$, let $\mathbf{X}_{\mathbb{B}} := \mathbf{X}_{\mathbb{B}}$, and then define

$$H_{\mathbf{X}}(B) := \log_2 |\mathbf{X}_{\mathbb{B}}| \quad \text{and}$$

$$H_{\mu}(B) := - \sum_{\mathbf{x} \in \mathbf{X}_{\mathbb{B}}} \mu[\mathbf{x}] \log_2(\mu[\mathbf{x}]).$$

The *topological entropy dimension* $\dim(\mathbf{X})$ is the smallest $d \in [0 \dots D+1]$ having some constant $c > 0$ such that, for any finite $B \subset \mathbb{R}^{D+1}$, $H_{\mathbf{X}}(B) \leq c \cdot \text{diam}[B]^d$. The *measurable entropy dimension* $\dim(\mu)$ is defined similarly, only with H_{μ} in place of $H_{\mathbf{X}}$. Note that $\dim(\mu) \leq \dim(\mathbf{X})$, because $H_{\mu}(B) \leq H_{\mathbf{X}}(B)$ for all $B \subset \mathbb{R}^{D+1}$.

For any bounded $B \subset \mathbb{R}^{D+1}$ and ‘scale factor’ $s > 0$, let $sB := \{sb; b \in B\}$. For any radius $r > 0$, let $(sB)^r := \{\mathbf{x} \in \mathbb{R}^{D+1}; d(\mathbf{x}, sB) \leq r\}$. Define the *d-dimensional topological entropy density* of B by

$$h_{\mathbf{X}}^d(B) := \sup_{r>0} \limsup_{s \rightarrow \infty} H_{\mathbf{X}}[(sB)^r] / s^d. \quad (10)$$

Define *d-dimensional measurable entropy density* $h_{\mu}^d(B)$ similarly, only using H_{μ} instead of $H_{\mathbf{X}}$. Note that, for any $d < \dim(\mathbf{X})$ [respectively, $d < \dim(\mu)$], $h_{\mathbf{X}}^d(B)$ [resp. $h_{\mu}^d(B)$] will be infinite, whereas for any $d > \dim(\mathbf{X})$ [resp. $d > \dim(\mu)$], $h_{\mathbf{X}}^d(B)$ [resp. $h_{\mu}^d(B)$] will be zero; hence $\dim(\mathbf{X})$ [resp. $\dim(\mu)$] is the unique value of d for which the function $h_{\mathbf{X}}^d$ [resp. h_{μ}^d] defined in Eq. (10) could be nontrivial.

Example 75 (a) If $d = D + 1$, and B is a unit cube centered at the origin, then $h_{\mathbf{X}}^{D+1}(B)$ (resp. $h_{\mu}^{D+1}(B)$) is just the classical $(D+1)$ -dimensional topological (resp. measurable) entropy of \mathbf{X} (resp. μ) as a $(D+1)$ -dimensional

subshift (resp. random field); see [► Entropy in Ergodic Theory](#).

(b) However, the most important case for Milnor [99] (and us) is when $\mathbf{X} = \mathbf{X}^{\text{Hist}}(\Phi)$ for some $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$. In this case, $\dim(\mu) \leq \dim(\mathbf{X}) \leq D < D+1$. In particular, if $d = 1$, then for any $\vec{v} \in \mathbb{R}^{D+1}$, if $B := \{r\vec{v}; r \in [0, 1]\}$, then $h_{\mathbf{X}}^1(B) = h_{\text{top}}(\Phi; \vec{v})$ and $h_{\mu}^1(B) = h_{\mu}(\Phi; \vec{v})$ are *directional entropies* of Subsect. “[Directional Entropy](#)”.

For any $d \in [0 \dots D+1]$, let λ^d be the d -dimensional Hausdorff measure on \mathbb{R}^{D+1} such that, if $P \subset \mathbb{R}^{D+1}$ is any *d-plane* (i.e. a d -dimensional linear subspace of \mathbb{R}^{D+1}), then λ^d restricts to the d -dimensional Lebesgue measure on P .

Theorem 76 Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, and let $\mu \in \mathcal{M}_{\text{ens}}(\mathbf{X}; \sigma)$. Let $d = \dim(\mathbf{X})$ (or $\dim(\mu)$) and let h^d be $h_{\mathbf{X}}^d$ (or h_{μ}^d). Let $B, C \subset \mathbb{R}^{D+1}$ be compact sets. Then

- (a) $h^d(B)$ is well-defined and finite.
- (b) If $B \subseteq C$ then $h^d(B) \leq h^d(C)$.
- (c) $h^d(B \cup C) \leq h^d(B) + h^d(C)$.
- (d) $h^d(B + \vec{v}) = h^d(B)$ for any $\vec{v} \in \mathbb{R}^{D+1}$.
- (e) $h^d(sB) = s^d \cdot h^d(B)$ for any $s > 0$.
- (f) There is some constant c such that $h_d(B) \leq c\lambda^d(B)$ for all compact $B \subset \mathbb{R}^{D+1}$.
- (g) If $d \in \mathbb{N}$, then for any d -plane $P \subset \mathbb{R}^{D+1}$, there is some $\mathfrak{S}^d(P) \geq 0$ such that $h^d(B) = \mathfrak{S}^d(P) \cdot \lambda^d(B)$ for any compact subset $B \subset P$ with $\lambda^d(\partial B) = 0$.
- (h) There is a constant $\overline{H}_{\mathbf{X}}^d < \infty$ such that $\mathfrak{S}_{\mathbf{X}}^d(P) \leq \overline{H}_{\mathbf{X}}^d$ for all d -planes P .

Proof See Theorems 1 and 2 and Corollary 1 in [99], or see Theorems 6.2, 6.3, and 6.13 in [12]. \square

Example 77 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and let $\mathbf{X} := \mathbf{X}^{\text{Hist}}(\Phi)$. If $P := \{0\} \times \mathbb{R}^D$, then $\mathfrak{S}^D(P)$ is the classical D -dimensional entropy of the omega limit set $\mathbf{Y} := \Phi^{\infty}(\mathcal{A}^{\mathbb{Z}^D})$; heuristically, this measures the asymptotic level of ‘spatial disorder’ in \mathbf{Y} . If $P \subset \mathbb{R}^{D+1}$ is some other D -plane, then $\mathfrak{S}^D(P)$ measures some combination of the ‘spatial disorder’ of \mathbf{Y} with the dynamical entropy of Φ .

Let $d \in [1 \dots D+1]$, and let $P \subset \mathbb{R}^{D+1}$ be a d -plane. For any $r > 0$, let $P(r) := \{z \in \mathbb{Z}^{D+1}; d(z, P) < r\}$. We say P is *expansive* for \mathbf{X} if there is some $r > 0$ such that, for any $\mathbf{x}, \mathbf{y} \in \mathbf{X}$, $(\mathbf{x}_{P(r)} = \mathbf{y}_{P(r)}) \iff (\mathbf{x} = \mathbf{y})$. If P is spanned by d rational vectors, then $P \cap \mathbb{Z}^{D+1}$ is a rank- d sublattice $\mathbb{L} \subset \mathbb{Z}^{D+1}$, and P is expansive if and only if the induced \mathbb{L} -action on \mathbf{X} is expansive. However, if P is ‘irrational’, then expansiveness is a more subtle concept; see Sect. 2 in [12] for more information.

If $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ and $\mathbf{X} = \mathbf{X}^{\text{Hist}}(\Phi)$, then Φ is *quasi-invertible* if \mathbf{X} admits an expansive D -plane P (this is a natural extension of Milnor’s (1988; §7) definition in terms of

‘causal cones’). Heuristically, if we regard \mathbb{Z}^{D+1} as ‘space-time’ (in the spirit of special relativity), then \mathcal{P} can be seen as ‘space’, and any direction transversal to \mathcal{P} can be interpreted as the flow of ‘time’.

Example 78 (a) If Φ is invertible, then it is quasi-invertible, because $\{0\} \times \mathbb{R}^D$ is an expansive D -plane (recall that the zeroth coordinate is time).

(b) Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}})$, so that $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^2}$. Let Φ have neighborhood $[-\ell \dots r]$, with $-\ell \leq 0 \leq r$, and let $\mathcal{L} \subset \mathbb{R}^2$ be a line with slope S through the origin (Fig. 2).

- [i] If Φ is right-permutative, and $0 < S \leq \frac{1}{\ell+1}$, then \mathcal{L} is expansive for \mathbf{X} .
- [ii] If Φ is left-permutative, and $\frac{-1}{r+1} \leq S < 0$, then \mathcal{L} is expansive for \mathbf{X} .
- [iii] If Φ is bipermutative, and $\frac{-1}{r+1} \leq S < 0$ or $0 < S \leq \frac{1}{\ell+1}$, then \mathcal{L} is expansive for \mathbf{X} .
- [iv] If Φ is posexexpansive (see Subsect. “[Posexexpansive and Permutative CA](#)”) then the ‘time’ axis $\mathcal{L} = \mathbb{R} \times \{0\}$ is expansive for \mathbf{X} .

Hence, in any of these cases, Φ is quasi-invertible. (Presumably, something similar is true for multidimensional permutative CA.)

Proposition 79 Let $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$, let $\mathbf{X} = \mathbf{X}^{\text{Hist}}(\Phi)$, let $\mu \in \mathfrak{M}_{\text{ens}}(\mathbf{X}; \sigma)$, and let \mathfrak{S}^d and \overline{H}_X^d be as in Theorem 76(g,h).

(a) If $\mathfrak{S}_X^D(\{0\} \times \mathbb{R}^D) = 0$, then $\overline{H}_X^D = 0$.

(b) Let $d \in [1 \dots D]$, and suppose that \mathbf{X} admits an expansive d -plane. Then:

- [i] $\dim(\mathbf{X}) \leq d$;
- [ii] There is a constant $\overline{H}_\mu^d < \infty$ such that $\mathfrak{S}_\mu^d(P) \leq \overline{H}_\mu^d$ for all d -planes P ;
- [iii] If $\mathfrak{S}^d(P) = 0$ for some expansive d -plane P , then $\overline{H}^d = 0$.

Proof (a) is Corollary 3 in [99], (b)[i] is Corollary 1.4 in [128], and (b)[ii] is Theorem 6.19(2) in [12].

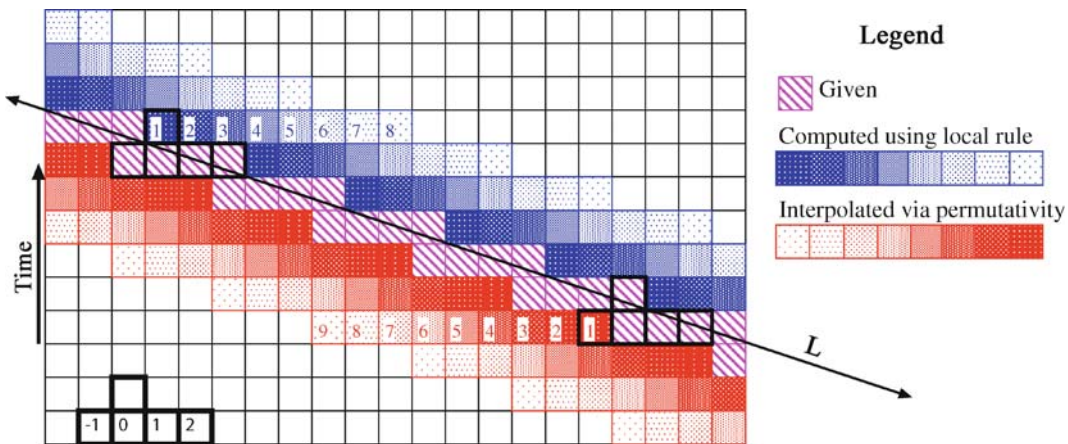
(b)[iii]: See Theorem 6.3(4) in [12] for “ $\overline{H}_X^d = 0$ ”. See Theorem 6.19(1) in [12] for “ $\overline{H}_\mu^d = 0$ ”. \square

If $d \in [1 \dots D+1]$, then a d -frame in \mathbb{R}^{D+1} is a d -tuple $\mathbf{F} := (\vec{v}_1, \dots, \vec{v}_d)$, where $\vec{v}_1, \dots, \vec{v}_d \in \mathbb{R}^{D+1}$ are linearly independent. Let $\mathfrak{F}_{\text{time}}(D+1, d)$ be the set of all d -frames in \mathbb{R}^{D+1} ; then $\mathfrak{F}_{\text{time}}(D+1, d)$ is an open subset of $\mathbb{R}^{D+1} \times \dots \times \mathbb{R}^{D+1} := \mathbb{R}^{(D+1) \times d}$. Let

$$\mathfrak{E}_{\text{expansive}}(\mathbf{X}, d) := \{\mathbf{F} \in \mathfrak{F}_{\text{time}}(D+1, d) ; \text{span}(\mathbf{F}) \text{ is expansive for } \mathbf{X}\}.$$

Then $\mathfrak{E}_{\text{expansive}}(\mathbf{X}, d)$ is an open subset of $\mathfrak{F}_{\text{time}}(D+1, d)$, by Lemma 3.4 in [12]. A connected component of $\mathfrak{E}_{\text{expansive}}(\mathbf{X}, d)$ is called an *expansive component* for \mathbf{X} . For any $\mathbf{F} \in \mathfrak{F}_{\text{time}}(D+1, d)$, let $[\mathbf{F}]$ be the d -dimensional parallelepiped spanned by \mathbf{F} , and let $\mathfrak{h}_X^d(\mathbf{F}) := h_X^d([\mathbf{F}]) = \mathfrak{S}_X^d(\text{span}(\mathbf{F})) \cdot \lambda^d([\mathbf{F}])$, where the last equality is by Theorem 76(g). The next result is a partial extension of Theorem 72(b).

Proposition 80 Let $\mathbf{X} \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift, suppose $d := \dim(\mathbf{X}) \in \mathbb{N}$, and let $\mathcal{C} \subset \mathfrak{E}_{\text{expansive}}(\mathbf{X}, d)$ be an expan-



Ergodic Theory of Cellular Automata, Figure 2

Example 78(b)[iii]: A left permutative CA Φ is quasi-invertible. In this picture, $[-\ell \dots r] = [-1 \dots 2]$, and \mathcal{L} is a line of slope $-1/3$. If $\mathbf{x} \in \mathbf{X}$ and we know the entries of \mathbf{x} in a neighborhood of \mathcal{L} , then we can reconstruct the rest of \mathbf{x} as shown. Entries above \mathcal{L} are directly computed using the local rule of Φ . Entries below \mathcal{L} are interpolated via left-permutativity. In both cases, the reconstruction occurs in consecutive diagonal lines, whose order is indicated by shading from darkest to lightest in the figure

sive component. Then the function $h_X^d: \mathbb{C} \rightarrow \mathbb{R}$ is convex in each of its d distinct \mathbb{R}^{D+1} -valued arguments. Thus, h_X^d is Lipschitz-continuous on \mathbb{C} .

Proof See Theorem 6.9(1,4) in [12]. \square

For measurable entropy, we can say much more. Recall that a d -linear form is a function $\omega: \mathbb{R}^{(D+1) \times d} \rightarrow \mathbb{R}$ which is linear in each of its d distinct \mathbb{R}^{D+1} -valued arguments and antisymmetric.

Theorem 81 Let $X \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift and let $\mu \in \mathcal{M}_{\text{ens}}(X; \sigma)$. Suppose $d := \dim(\mu) \in \mathbb{N}$, and let $\mathbb{C} \subset \mathbb{E}_{\text{expansive}}(X, d)$ be an expansive component for X . Then there is a d -linear form $\omega: \mathbb{R}^{(D+1) \times d} \rightarrow \mathbb{R}$ such that h_μ^d agrees with ω on \mathbb{C} .

Proof Theorem 6.16 in [12]. \square

If $\bar{H}_\mu^d \neq 0$, then Theorem 81 means that there is an orthogonal $(D+1-d)$ -frame $\bar{W} := (\bar{w}_{d+1}, \dots, \bar{w}_{D+1})$ (transversal to all frames in \mathbb{C}) such that, for any d -frame $\bar{V} := (\bar{v}_1, \dots, \bar{v}_d) \in \mathbb{C}$,

$$h_\mu^d(\bar{V}) = \det(\bar{v}_1, \dots, \bar{v}_d; \bar{w}_{d+1}, \dots, \bar{w}_{D+1}). \quad (11)$$

Thus, the d -plane orthogonal to $\{\bar{w}_{d+1}, \dots, \bar{w}_{D+1}\}$ is the d -plane which maximizes \mathcal{S}_μ^d – this is the d -plane manifesting the most rapid decay of correlation with distance. On the other hand, $\text{span}(\bar{W})$ is the $(D+1-d)$ -plane along which correlations decay the most slowly. Also, if $\bar{V} \in \mathbb{C}$, then Eq. (11) implies that \mathbb{C} cannot contain any frame spanning $\text{span}(\bar{V})$ with reversed orientation (e.g. an odd permutation of \bar{V}), because entropy is nonnegative.

Example 82 Let $\Phi \in \text{CA}(\mathcal{A}^{\mathbb{Z}^D})$ be quasi-invertible, and let P be an expansive D -plane for $X := X^{\text{Hist}}(\Phi)$ (see Example 78). The D -frames spanning P fall into two expansive components (related by orientation-reversal); let \mathbb{C} be union of these two components. Let $\mu \in \mathcal{M}_{\text{ens}}(\mathcal{A}^{\mathbb{Z}^D}; \Phi)$, and extend μ to a σ -invariant measure on X . In this case, Theorem 81 is equivalent to Theorem 4 in [99], which says there a vector $\bar{w} \in \mathbb{R}^{D+1}$ such that, for any D -frame $(\bar{v}_1, \dots, \bar{v}_D) \in \mathbb{C}$, $h_\mu^d(F) = |\det(\bar{v}_1, \dots, \bar{v}_D; \bar{w})|$. Thus, $\mathcal{S}_\mu^d(P)$ is maximized when P is the hyperplane orthogonal to \bar{w} . Heuristically, \bar{w} points in the direction of minimum correlation decay (or maximum ‘causality’) – the direction which could most properly be called ‘time’ for the MPDS (Φ, μ) .

Theorem 81 yields the following generalization the Variational Principle:

Theorem 83 Let $X \subset \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift and suppose $d := \dim(X) \in \mathbb{N}$.

- (a) If $F \in \mathbb{E}_{\text{expansive}}(X, d)$, then there exists $\mu \in \mathcal{M}_{\text{ens}}(X; \sigma)$ such that $h_X^d(F) = h_\mu^d(F)$.
- (b) Let $\mathbb{C} \subset \mathbb{E}_{\text{expansive}}(X, d)$ be an expansive component for X . There exists some $\mu \in \mathcal{M}_{\text{ens}}(X; \sigma)$ such that $h_X^d = h_\mu^d$ on \mathbb{C} if and only if h_X^d is a d -linear form on \mathbb{C} .

Proof Proposition 6.24 and Theorem 6.25 in [12]. \square

Remark 84 (a) If $G \subset \mathcal{A}^{\mathbb{Z}^D}$ is an abelian subgroup shift and $\Phi \in \text{ECA}(G)$, then $X^{\text{Hist}}(\Phi)$ is a subgroup shift of $\mathcal{A}^{\mathbb{Z}^{D+1}}$, which can be viewed as an algebraic \mathbb{Z}^{D+1} -action (see discussion prior to Proposition 27). In this context, the expansive subspaces of $X^{\text{Hist}}(\Phi)$ have been completely characterized by Einsiedler et al. (see Theorem 8.4 in [33]). Furthermore, certain dynamical properties (such as positive entropy, completely positive entropy, or Bernoullicity) are common amongst all elements of each expansive component of $X^{\text{Hist}}(\Phi)$ (see Theorem 9.8 in [33]) (this sort of ‘commonality’ within expansive components was earlier emphasized by Boyle and Lind (see [12])). If $X^{\text{Hist}}(\Phi)$ has entropy dimension 1 (e.g. Φ is a one-dimensional linear CA), the structure of $X^{\text{Hist}}(\Phi)$ has been thoroughly analyzed by Einsiedler and Lind [30]. Finally, if G_1 and G_2 are subgroup shifts, and $\Phi_k \in \text{ECA}(G_k)$ and $\mu_k \in \mathcal{M}_{\text{ens}}(G_k; \Phi, \sigma)$ for $k = 1, 2$, with $\dim(\mu_1) = \dim(\mu_2) = 1$, then Einsiedler and Ward [32] have given conditions for the measure-preserving systems $(G_1, \mu_1; \Phi_1, \sigma)$ and $(G_2, \mu_2; \Phi_2, \sigma)$ to be disjoint.

(b) Boyle and Lind’s ‘expansive subdynamics’ concerns expansiveness along certain directions in the space-time diagram of a CA. Recently, M. Sablik has developed a theory of *directional dynamics*, which explores other topological dynamical properties (such as equicontinuity and sensitivity to initial conditions) along spatiotemporal directions in a CA; see [120], Chapitre II or [121].

Future Directions and Open Problems

1. We now have a fairly good understanding of the ergodic theory of linear and/or ‘abelian’ CA. The next step is to extend these results to CA with nonlinear and/or nonabelian algebraic structures. In particular:
 - (a) Almost all the measure rigidity results of Subsect. “[Measure Rigidity in Algebraic CA](#)” are for endomorphic CA on abelian group shifts, except for Propositions 21 and 23. Can we extend these results to CA on nonabelian group shifts or other permutative CA?
 - (b) Likewise, the asymptotic randomization results of Subsect. “[Asymptotic Randomization by Linear Cellular Automata](#)” are almost exclusively

for linear CA with scalar coefficients, and for $\mathbb{M} = \mathbb{Z}^D \times \mathbb{N}^E$. Can we extend these results to LCA with noncommuting, matrix-valued coefficients? (The problem is: if the coefficients do not commute, then the ‘polynomial representation’ and Lucas’ theorem become inapplicable.) Also, can we obtain similar results for multiplicative CA on *nonabelian* groups? (See Remark 41(d).) What about other permutative CA? (See Remark 41(e).) Finally, what if \mathbb{M} is a nonabelian group? (For example, Lind and Schmidt (unpublished) [31] have recently investigated algebraic actions of the discrete Heisenberg group.)

2. Cellular automata are often seen as models of spatially distributed computation. Meaningful ‘computation’ could possibly occur when a CA interacts with a highly structured initial configuration (e.g. a substitution sequence), whereas such computation is probably impossible in the roiling cauldron of noise arising from a mixing, positive entropy measure (e.g. a Bernoulli measure or Markov random field). Yet almost all the results in this article concern the interaction of CA with such mixing, positive-entropy measures. We are starting to understand the topological dynamics of CA acting on non-mixing and/or zero-entropy symbolic dynamical systems, (e.g. substitution shifts, automatic shifts, regular Toeplitz shifts, and quasisturmian shifts); see ► [Dynamics of Cellular Automata in Non-compact Spaces](#). However, almost nothing is known about the interaction of CA with the natural invariant measures on these systems. In particular:
 - (a) The invariant measures discussed in Sect. “[Invariant Measures for CA](#)” all have nonzero entropy (see, however, Example 67(c)). Are there any nontrivial zero-entropy measures for interesting CA?
 - (b) The results of Subsect. “[Asymptotic Randomization by Linear Cellular Automata](#)” all concern the asymptotic randomization of initial measures with nonzero entropy, except for Remark 41(c). Are there similar results for zero-entropy measures?
 - (c) Zero-entropy systems often have an appealing combinatorial description via cutting-and-stacking constructions, Bratteli diagrams, or finite state machines. Likewise, CA admit a combinatorial description (via local rules). How do these combinatorial descriptions interact?
3. As we saw in Subsect. “[Domains, Defects, and Particles](#)”, and also in Propositions 48–56, emergent defect dynamics can be a powerful tool for analyzing the measurable dynamics of CA. Defects in one-dimensional CA generally act like ‘particles’, and their ‘kinematics’

is fairly well-understood. However, in higher dimensions, defects can be much more topologically complicated (e.g. they can look like curves or surfaces), and their evolution in time is totally mysterious. Can we develop a theory of multidimensional defect dynamics?

4. Almost all the results about mixing and ergodicity in Subsect. “[Mixing and Ergodicity](#)” are for one-dimensional (mostly permutative) CA and for the uniform measure on $\mathcal{A}^{\mathbb{Z}}$. Can similar results be obtained for other CA and/or measures on $\mathcal{A}^{\mathbb{Z}}$? What about CA in $\mathcal{A}^{\mathbb{Z}^D}$ for $D \geq 2$?
5. Let μ be a (Φ, σ) -invariant measure on $\mathcal{A}^{\mathbb{M}}$. Proposition 66 suggests an intriguing correspondence between certain spectral properties (namely, weak mixing and discrete spectrum) for the system $(\mathcal{A}^{\mathbb{M}}, \mu; \sigma)$ and those for the system $(\mathcal{A}^{\mathbb{M}}, \mu; \Phi)$. Does a similar correspondence hold for other spectral properties, such as continuous spectrum, Lebesgue spectral type, spectral multiplicity, rigidity, or mild mixing?
6. Let $\mathbf{X} \in \mathcal{A}^{\mathbb{Z}^{D+1}}$ be a subshift admitting an expansive D -plane $P \subset \mathbb{R}^{D+1}$. As discussed in Subsect. “[Entropy Geometry and Expansive Subdynamics](#)”, if we regard \mathbb{Z}^{D+1} as ‘spacetime’, then we can treat P as a ‘space’, and a transversal direction as ‘time’. Indeed, if P is spanned by rational vectors, then the Curtis–Hedlund–Lyndon theorem implies that \mathbf{X} is isomorphic to the history shift of some invertible $\Phi \in \mathcal{CA}(\mathcal{A}^{\mathbb{Z}^D})$ acting on some Φ -invariant subshift $\mathbf{Y} \subseteq \mathcal{A}^{\mathbb{Z}^D}$ (where we embed \mathbb{Z}^D in P). If P is irrational, then this is not the case; however, \mathbf{X} still seems very much like the history shift of a spatially distributed symbolic dynamical system, closely analogous to a CA, except with a continually fluctuating ‘spatial distribution’ of state information, and perhaps with occasional nonlocal interactions. For example, Proposition 79(b)[i] implies that $\dim(\mathbf{X}) \leq D$, just as for a CA. How much of the theory of invertible CA can be generalized to such systems?

I will finish with the hardest problem of all. Cellular automata are tractable mainly because of their *homogeneity*: CA are embedded in a highly regular spatial geometry (i.e. a lattice or other Cayley digraph) with the same local rule everywhere. However, many of the most interesting spatially distributed symbolic dynamical systems are not nearly this homogeneous. For example:

- CA are often proposed as models of spatially distributed physical systems. Yet in many such systems (e.g. living tissues, quantum ‘foams’), the underlying geometry is not a flat Euclidean space, but a curved manifold. A good discrete model of such a manifold

can be obtained through a Voronoi tessellation of sufficient density; a realistic symbolic dynamical model would be a CA-like system defined on the dual graph of this Voronoi tessellation.

- As mentioned in question #3, defects in multidimensional CA may have the geometry of curves, surfaces, or other embedded submanifolds (possibly with varying nonzero thickness). To model the evolution of such a defect, we could treat it as a CA-like object whose underlying geometry is an (evolving) manifold, and whose local rules (although partly determined by the local rule of the original CA) are spatially heterogeneous (because they are also influenced by incoming information from the ambient ‘nondefective’ space).
- The CA-like system arising in question #6 has a D -dimensional planar geometry, but the distribution of ‘cells’ within this plane (and, presumably, the local rules between them) are constantly fluctuating.

More generally, any topological dynamical system on a Cantor space can be represented as a *cellular network*: a CA-like system defined on an infinite digraph, with different local rules at different nodes. Gromov [51] has generalized the Garden of Eden Theorem 3 to this setting (see Remark 5(a)). However, other than Gromov’s work, basically nothing is known about such systems. Can we generalize any of the theory of cellular automata to cellular networks? Is it possible to develop a nontrivial ergodic theory for such systems?

Acknowledgments

I would like to thank François Blanchard, Mike Boyle, Maurice Courbage, Doug Lind, Petr Kůrka, Servet Martínez, Kyewon Koh Park, Mathieu Sablik, Jeffrey Steif, and Marcelo Sobottka, who read draft versions of this article and made many invaluable suggestions, corrections, and comments. (Any errors which remain are mine.) To Reem.

Bibliography

1. Akin E (1993) The general topology of dynamical systems, Graduate Studies in Mathematics, vol 1. American Mathematical Society, Providence
2. Allouche JP (1999) Cellular automata, finite automata, and number theory. In: Cellular automata (Saissac, 1996), Math. Appl., vol 460. Kluwer, Dordrecht, pp 321–330
3. Allouche JP, Skordev G (2003) Remarks on permutative cellular automata. J Comput Syst Sci 67(1):174–182
4. Allouche JP, von Haeseler F, Peitgen HO, Skordev G (1996) Linear cellular automata, finite automata and Pascal’s triangle. Discret Appl Math 66(1):1–22
5. Allouche JP, von Haeseler F, Peitgen HO, Petersen A, Skordev G (1997) Automaticity of double sequences generated by one-dimensional linear cellular automata. Theoret Comput Sci 188(1-2):195–209
6. Barbé A, von Haeseler F, Peitgen HO, Skordev G (1995) Coarse-graining invariant patterns of one-dimensional two-state linear cellular automata. Internat J Bifur Chaos Appl Sci Eng 5(6):1611–1631
7. Barbé A, von Haeseler F, Peitgen HO, Skordev G (2003) Rescaled evolution sets of linear cellular automata on a cylinder. Internat J Bifur Chaos Appl Sci Eng 13(4):815–842
8. Belitsky V, Ferrari PA (2005) Invariant measures and convergence properties for cellular automaton 184 and related processes. J Stat Phys 118(3-4):589–623
9. Blanchard F, Maass A (1997) Dynamical properties of expansive one-sided cellular automata. Israel J Math 99:149–174
10. Blank M (2003) Ergodic properties of a simple deterministic traffic flow model. J Stat Phys 111(3-4):903–930
11. Boccara N, Naser J, Roger M (1991) Particle-like structures and their interactions in spatiotemporal patterns generated by one-dimensional deterministic cellular automata. Phys Rev A 44(2):866–875
12. Boyle M, Lind D (1997) Expansive subdynamics. Trans Amer Math Soc 349(1):55–102
13. Boyle M, Fiebig D, Fiebig UR (1997) A dimension group for local homeomorphisms and endomorphisms of onesided shifts of finite type. J Reine Angew Math 487:27–59
14. Burton R, Steif JE (1994) Non-uniqueness of measures of maximal entropy for subshifts of finite type. Ergodic Theory Dynam Syst 14(2):213–235
15. Burton R, Steif JE (1995) New results on measures of maximal entropy. Israel J Math 89(1-3):275–300
16. Cai H, Luo X (1993) Laws of large numbers for a cellular automaton. Ann Probab 21(3):1413–1426
17. Cattaneo G, Formenti E, Manzini G, Margara L (1997) On ergodic linear cellular automata over Z_m . In: STACS 97 (Lübeck). Lecture Notes in Computer Science, vol 1200. Springer, Berlin, pp 427–438
18. Cattaneo G, Formenti E, Manzini G, Margara L (2000) Ergodicity, transitivity, and regularity for linear cellular automata over Z_m . Theoret Comput Sci 233(1-2):147–164
19. Ceccherini-Silberstein T, Fiorenzi F, Scarabotti F (2004) The Garden of Eden theorem for cellular automata and for symbolic dynamical systems. In: Random walks and geometry, de Gruyter, Berlin, pp 73–108
20. Ceccherini-Silberstein TG, Machi A, Scarabotti F (1999) Amenable groups and cellular automata. Ann Inst Fourier (Grenoble) 49(2):673–685
21. Courbage M, Kamiński B (2002) On the directional entropy of \mathbb{Z}^2 -actions generated by cellular automata. Studia Math 153(3):285–295
22. Courbage M, Kamiński B (2006) Space-time directional Lyapunov exponents for cellular automata. J Stat Phys 124(6):1499–1509
23. Coven EM (1980) Topological entropy of block maps. Proc Amer Math Soc 78(4):590–594
24. Coven EM, Paul ME (1974) Endomorphisms of irreducible subshifts of finite type. Math Syst Theory 8(2):167–175
25. Downarowicz T (1997) The royal couple conceals their mutual relationship: a noncoalescent Toeplitz flow. Israel J Math 97:239–251

26. Durrett R, Steif JE (1991) Some rigorous results for the Greenberg–Hastings model. *J Theoret Probab* 4(4):669–690
27. Durrett R, Steif JE (1993) Fixation results for threshold voter systems. *Ann Probab* 21(1):232–247
28. Einsiedler M (2004) Invariant subsets and invariant measures for irreducible actions on zero-dimensional groups. *Bull London Math Soc* 36(3):321–331
29. Einsiedler M (2005) Isomorphism and measure rigidity for algebraic actions on zero-dimensional groups. *Monatsh Math* 144(1):39–69
30. Einsiedler M, Lind D (2004) Algebraic \mathbb{Z}^d -actions on entropy rank one. *Trans Amer Math Soc* 356(5):1799–1831 (electronic)
31. Einsiedler M, Rindler H (2001) Algebraic actions of the discrete Heisenberg group and other non-abelian groups. *Aequationes Math* 62(1–2):117–135
32. Einsiedler M, Ward T (2005) Entropy geometry and disjointness for zero-dimensional algebraic actions. *J Reine Angew Math* 584:195–214
33. Einsiedler M, Lind D, Miles R, Ward T (2001) Expansive subdynamics for algebraic \mathbb{Z}^d -actions. *Ergodic Theory Dynam Systems* 21(6):1695–1729
34. Eloranta K, Nummelin E (1992) The kink of cellular automaton Rule 18 performs a random walk. *J Stat Phys* 69(5–6):1131–1136
35. Fagnani F, Margara L (1998) Expansivity, permutivity, and chaos for cellular automata. *Theory Comput Syst* 31(6):663–677
36. Ferrari PA, Maass A, Martínez S, Ney P (2000) Cesàro mean distribution of group automata starting from measures with summable decay. *Ergodic Theory Dynam Syst* 20(6):1657–1670
37. Finelli M, Manzini G, Margara L (1998) Lyapunov exponents versus expansivity and sensitivity in cellular automata. *J Complexity* 14(2):210–233
38. Fiorenzi F (2000) The Garden of Eden theorem for sofic shifts. *Pure Math Appl* 11(3):471–484
39. Fiorenzi F (2003) Cellular automata and strongly irreducible shifts of finite type. *Theoret Comput Sci* 299(1–3):477–493
40. Fiorenzi F (2004) Semi-strongly irreducible shifts. *Adv Appl Math* 32(3):421–438
41. Fisch R (1990) The one-dimensional cyclic cellular automaton: a system with deterministic dynamics that emulates an interacting particle system with stochastic dynamics. *J Theoret Probab* 3(2):311–338
42. Fisch R (1992) Clustering in the one-dimensional three-color cyclic cellular automaton. *Ann Probab* 20(3):1528–1548
43. Fisch R, Gravner J (1995) One-dimensional deterministic Greenberg–Hastings models. *Complex Systems* 9(5):329–348
44. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math Systems Theory* 1:1–49
45. Gilman RH (1987) Classes of linear automata. *Ergodic Theory Dynam Syst* 7(1):105–118
46. Gottschalk W (1973) Some general dynamical notions. In: *Recent advances in topological dynamics* (Proc Conf Topological Dynamics, Yale Univ, New Haven, 1972; in honor of Gustav Arnold Hedlund), *Lecture Notes in Math*, vol 318. Springer, Berlin, pp 120–125
47. Grassberger P (1984) Chaos and diffusion in deterministic cellular automata. *Phys D* 10(1–2):52–58, cellular automata (Los Alamos, 1983)
48. Grassberger P (1984) New mechanism for deterministic diffusion. *Phys Rev A* 28(6):3666–3667
49. Grigorchuk RI (1984) Degrees of growth of finitely generated groups and the theory of invariant means. *Izv Akad Nauk SSSR Ser Mat* 48(5):939–985
50. Gromov M (1981) Groups of polynomial growth and expanding maps. *Inst Hautes Études Sci Publ Math* 53:53–73
51. Gromov M (1999) Endomorphisms of symbolic algebraic varieties. *J Eur Math Soc (JEMS)* 1(2):109–197
52. von Haeseler F, Peitgen HO, Skordev G (1992) Pascal’s triangle, dynamical systems and attractors. *Ergodic Theory Dynam Systems* 12(3):479–486
53. von Haeseler F, Peitgen HO, Skordev G (1993) Cellular automata, matrix substitutions and fractals. *Ann Math Artificial Intelligence* 8(3–4):345–362, theorem proving and logic programming (1992)
54. von Haeseler F, Peitgen HO, Skordev G (1995) Global analysis of self-similarity features of cellular automata: selected examples. *Phys D* 86(1–2):64–80, chaos, order and patterns: aspects of nonlinearity—the “gran finale” (Como, 1993)
55. von Haeseler F, Peitgen HO, Skordev G (1995) Multifractal decompositions of rescaled evolution sets of equivariant cellular automata. *Random Comput Dynam* 3(1–2):93–119
56. von Haeseler F, Peitgen HO, Skordev G (2001) Self-similar structure of rescaled evolution sets of cellular automata. I. *Internat J Bifur Chaos Appl Sci Eng* 11(4):913–926
57. von Haeseler F, Peitgen HO, Skordev G (2001) Self-similar structure of rescaled evolution sets of cellular automata. II. *Internat J Bifur Chaos Appl Sci Eng* 11(4):927–941
58. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. *Math Syst Theory* 3:320–375
59. Hilmy H (1936) Sur les centres d’attraction minimaux des systèmes dynamiques. *Compositio Mathematica* 3:227–238
60. Host B (1995) Nombres normaux, entropie, translations. *Israel J Math* 91(1–3):419–428
61. Host B, Maass A, Martínez S (2003) Uniform Bernoulli measure in dynamics of permutative cellular automata with algebraic local rules. *Discret Contin Dyn Syst* 9(6):1423–1446
62. Hurd LP, Kari J, Culik K (1992) The topological entropy of cellular automata is uncomputable. *Ergodic Theory Dynam Syst* 12(2):255–265
63. Hurley M (1990) Attractors in cellular automata. *Ergodic Theory Dynam Syst* 10(1):131–140
64. Hurley M (1990) Ergodic aspects of cellular automata. *Ergodic Theory Dynam Syst* 10(4):671–685
65. Hurley M (1991) Varieties of periodic attractor in cellular automata. *Trans Amer Math Soc* 326(2):701–726
66. Hurley M (1992) Attractors in restricted cellular automata. *Proc Amer Math Soc* 115(2):563–571
67. Jen E (1988) Linear cellular automata and recurring sequences in finite fields. *Comm Math Phys* 119(1):13–28
68. Johnson A, Rudolph DJ (1995) Convergence under \times_q of \times_p invariant measures on the circle. *Adv Math* 115(1):117–140
69. Johnson ASA (1992) Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of the integers. *Israel J Math* 77(1–2):211–240
70. Kitchens B (2000) Dynamics of $\text{mathbf{Z}^d}$ actions on Markov subgroups. In: *Topics in symbolic dynamics and applications* (Temuco, 1997), *London Math Soc Lecture Note Ser*, vol 279. Cambridge Univ Press, Cambridge, pp 89–122

71. Kitchens B, Schmidt K (1989) Automorphisms of compact groups. *Ergodic Theory Dynam Syst* 9(4):691–735
72. Kitchens B, Schmidt K (1992) Markov subgroups of $(\mathbb{Z}/2\mathbb{Z})^{\mathbb{Z}^2}$. In: *Symbolic dynamics and its applications* (New Haven, 1991), *Contemp Math*, vol 135. Amer Math Soc, Providence, pp 265–283
73. Kitchens BP (1987) Expansive dynamics on zero-dimensional groups. *Ergodic Theory Dynam Syst* 7(2):249–261
74. Kleveand R (1997) Mixing properties of one-dimensional cellular automata. *Proc Amer Math Soc* 125(6):1755–1766
75. Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. *Ergodic Theory Dynam Systems* 17(2):417–433
76. Kůrka P (2001) Topological dynamics of cellular automata. In: *Codes, systems, and graphical models* (Minneapolis, 1999), *IMA vol Math Appl*, vol 123. Springer, New York, pp 447–485
77. Kůrka P (2003) Cellular automata with vanishing particles. *Fund Inform* 58(3-4):203–221
78. Kůrka P (2005) On the measure attractor of a cellular automaton. *Discret Contin Dyn Syst (suppl.)*:524–535
79. Kůrka P, Maass A (2000) Limit sets of cellular automata associated to probability measures. *J Stat Phys* 100(5-6):1031–1047
80. Kůrka P, Maass A (2002) Stability of subshifts in cellular automata. *Fund Inform* 52(1-3):143–155, special issue on cellular automata
81. Lind D, Marcus B (1995) *An introduction to symbolic dynamics and coding*. Cambridge University Press, Cambridge
82. Lind DA (1984) Applications of ergodic theory and sofic systems to cellular automata. *Phys D* 10(1-2):36–44, cellular automata (Los Alamos, 1983)
83. Lind DA (1987) Entropies of automorphisms of a topological Markov shift. *Proc Amer Math Soc* 99(3):589–595
84. Lucas E (1878) Sur les congruences des nombres eulériens et les coefficients différentiels des fonctions trigonométriques suivant un module premier. *Bull Soc Math France* 6:49–54
85. Lyons R (1988) On measures simultaneously 2- and 3-invariant. *Israel J Math* 61(2):219–224
86. Maass A (1996) Some dynamical properties of one-dimensional cellular automata. In: *Dynamics of complex interacting systems* (Santiago, 1994), *Nonlinear Phenom. Complex Systems*, vol 2. Kluwer, Dordrecht, pp 35–80
87. Maass A, Martínez S (1998) On Cesàro limit distribution of a class of permutative cellular automata. *J Stat Phys* 90(1-2):435–452
88. Maass A, Martínez S (1999) Time averages for some classes of expansive one-dimensional cellular automata. In: *Cellular automata and complex systems* (Santiago, 1996), *Nonlinear Phenom. Complex Systems*, vol 3. Kluwer, Dordrecht, pp 37–54
89. Maass A, Martínez S, Pivato M, Yassawi R (2006) Asymptotic randomization of subgroup shifts by linear cellular automata. *Ergodic Theory Dynam Syst* 26(4):1203–1224
90. Maass A, Martínez S, Pivato M, Yassawi R (2006) Attractiveness of the Haar measure for the action of linear cellular automata in abelian topological Markov chains. In: *Dynamics and Stochastics: Festschrift in honour of Michael Keane*, vol 48 of, *Lecture Notes Monograph Series of the IMS*, vol 48. Institute for Mathematical Statistics, Beachwood, pp 100–108
91. Maass A, Martínez S, Sobottka M (2006) Limit measures for affine cellular automata on topological Markov subgroups. *Nonlinearity* 19(9):2137–2147, <http://stacks.iop.org/0951-7715/19/2137>
92. Machi A, Mignosi F (1993) Garden of Eden configurations for cellular automata on Cayley graphs of groups. *SIAM J Discret Math* 6(1):44–56
93. Maruoka A, Kimura M (1976) Condition for injectivity of global maps for tessellation automata. *Inform Control* 32(2):158–162
94. Mauldin RD, Skordev G (2000) Random linear cellular automata: fractals associated with random multiplication of polynomials. *Japan J Math (NS)* 26(2):381–406
95. Meester R, Steif JE (2001) Higher-dimensional subshifts of finite type, factor maps and measures of maximal entropy. *Pacific J Math* 200(2):497–510
96. Milnor J (1985) Correction and remarks: “On the concept of attractor”. *Comm Math Phys* 102(3):517–519
97. Milnor J (1985) On the concept of attractor. *Comm Math Phys* 99(2):177–195
98. Milnor J (1986) Directional entropies of cellular automaton-maps. In: *Disordered systems and biological organization* (Les Houches, 1985), *NATO Adv Sci Inst Ser F Comput Syst Sci*, vol 20. Springer, Berlin, pp 113–115
99. Milnor J (1988) On the entropy geometry of cellular automata. *Complex Syst* 2(3):357–385
100. Miyamoto M (1979) An equilibrium state for a one-dimensional life game. *J Math Kyoto Univ* 19(3):525–540
101. Moore C (1997) Quasilinear cellular automata. *Phys D* 103(1-4):100–132, *lattice dynamics* (Paris, 1995)
102. Moore C (1998) Predicting nonlinear cellular automata quickly by decomposing them into linear ones. *Phys D* 111(1-4):27–41
103. Moore EF (1963) Machine models of self reproduction. *Proc Symp Appl Math* 14:17–34
104. Myhill J (1963) The converse of Moore’s Garden-of-Eden theorem. *Proc Amer Math Soc* 14:685–686
105. Nasu M (1995) Textile systems for endomorphisms and automorphisms of the shift. *Mem Amer Math Soc* 114(546):viii+215
106. Nasu M (2002) The dynamics of expansive invertible one-sided cellular automata. *Trans Amer Math Soc* 354(10):4067–4084 (electronic)
107. Park KK (1995) Continuity of directional entropy for a class of \mathbb{Z}^2 -actions. *J Korean Math Soc* 32(3):573–582
108. Park KK (1996) Entropy of a skew product with a \mathbb{Z}^2 -action. *Pacific J Math* 172(1):227–241
109. Park KK (1999) On directional entropy functions. *Israel J Math* 113:243–267
110. Parry W (1964) Intrinsic Markov chains. *Trans Amer Math Soc* 112:55–66
111. Pivato M (2003) Multiplicative cellular automata on nilpotent groups: structure, entropy, and asymptotics. *J Stat Phys* 110(1-2):247–267
112. Pivato M (2005) Cellular automata versus quasisturmian shifts. *Ergodic Theory Dynam Syst* 25(5):1583–1632
113. Pivato M (2005) Invariant measures for bipermutative cellular automata. *Discret Contin Dyn Syst* 12(4):723–736
114. Pivato M (2007) Spectral domain boundaries cellular automata. *Fundamenta Informaticae* 77(special issue), available at: <http://arxiv.org/abs/math.DS/0507091>
115. Pivato M (2008) Module shifts and measure rigidity in linear cellular automata. *Ergodic Theory Dynam Syst* (to appear)

116. Pivato M, Yassawi R (2002) Limit measures for affine cellular automata. *Ergodic Theory Dynam Syst* 22(4):1269–1287
117. Pivato M, Yassawi R (2004) Limit measures for affine cellular automata. II. *Ergodic Theory Dynam Syst* 24(6):1961–1980
118. Pivato M, Yassawi R (2006) Asymptotic randomization of sofic shifts by linear cellular automata. *Ergodic Theory Dynam Syst* 26(4):1177–1201
119. Rudolph DJ (1990) $\times 2$ and $\times 3$ invariant measures and entropy. *Ergodic Theory Dynam Syst* 10(2):395–406
120. Sablik M (2006) Étude de l'action conjointe d'un automate cellulaire et du décalage: Une approche topologique et ergodique. Ph D thesis, Université de la Méditerranée, Faculté des sciences de Luminy, Marseille
121. Sablik M (2008) Directional dynamics for cellular automata: A sensitivity to initial conditions approach. submitted to *Theor Comput Sci* 400(1–3):1–18
122. Sablik M (2008) Measure rigidity for algebraic bipermutative cellular automata. *Ergodic Theory Dynam Syst* 27(6):1965–1990
123. Sato T (1997) Ergodicity of linear cellular automata over \mathbf{Z}_m . *Inform Process Lett* 61(3):169–172
124. Schmidt K (1995) Dynamical systems of algebraic origin, *Progress in Mathematics*, vol 128. Birkhäuser, Basel
125. Shereshevsky MA (1992) Ergodic properties of certain surjective cellular automata. *Monatsh Math* 114(3–4):305–316
126. Shereshevsky MA (1992) Lyapunov exponents for one-dimensional cellular automata. *J Nonlinear Sci* 2(1):1–8
127. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. *Indag Math (NS)* 4(2):203–210
128. Shereshevsky MA (1996) On continuous actions commuting with actions of positive entropy. *Colloq Math* 70(2):265–269
129. Shereshevsky MA (1997) K-property of permutative cellular automata. *Indag Math (NS)* 8(3):411–416
130. Shereshevsky MA, Afraimovich VS (1992/93) Bipermutative cellular automata are topologically conjugate to the one-sided Bernoulli shift. *Random Comput Dynam* 1(1):91–98
131. Shirvani M, Rogers TD (1991) On ergodic one-dimensional cellular automata. *Comm Math Phys* 136(3):599–605
132. Silberger S (2005) Subshifts of the three dot system. *Ergodic Theory Dynam Syst* 25(5):1673–1687
133. Smillie J (1988) Properties of the directional entropy function for cellular automata. In: *Dynamical systems (College Park, 1986–87)*, *Lecture Notes in Math*, vol 1342. Springer, Berlin, pp 689–705
134. Sobottka M (2005) Representación y aleatorización en sistemas dinámicos de tipo algebraico. Ph D thesis, Universidad de Chile, Facultad de ciencias físicas y matemáticas, Santiago
135. Sobottka M (2007) Topological quasi-group shifts. *Discret Continuous Dyn Syst* 17(1):77–93
136. Sobottka M (to appear 2007) Right-permutative cellular automata on topological Markov chains. *Discret Continuous Dyn Syst*. Available at <http://arxiv.org/abs/math/0603326>
137. Steif JE (1994) The threshold voter automaton at a critical point. *Ann Probab* 22(3):1121–1139
138. Takahashi S (1990) Cellular automata and multifractals: dimension spectra of linear cellular automata. *Phys D* 45(1–3):36–48, cellular automata: theory and experiment (Los Alamos, NM, 1989)
139. Takahashi S (1992) Self-similarity of linear cellular automata. *J Comput Syst Sci* 44(1):114–140
140. Takahashi S (1993) Cellular automata, fractals and multifractals: space-time patterns and dimension spectra of linear cellular automata. In: *Chaos in Australia (Sydney, 1990)*, World Sci Publishing, River Edge, pp 173–195
141. Tisseur P (2000) Cellular automata and Lyapunov exponents. *Nonlinearity* 13(5):1547–1560
142. Walters P (1982) An introduction to ergodic theory, *Graduate Texts in Mathematics*, vol 79. Springer, New York
143. Weiss B (2000) Sofic groups and dynamical systems. *Sankhyā Ser A* 62(3):350–359
144. Willson SJ (1975) On the ergodic theory of cellular automata. *Math Syst Theory* 9(2):132–141
145. Willson SJ (1984) Cellular automata can generate fractals. *Discret Appl Math* 8(1):91–99
146. Willson SJ (1984) Growth rates and fractional dimensions in cellular automata. *Phys D* 10(1–2):69–74, cellular automata (Los Alamos, 1983)
147. Willson SJ (1986) A use of cellular automata to obtain families of fractals. In: *Chaotic dynamics and fractals (Atlanta, 1985)*, *Notes Rep Math Sci Eng*, vol 2. Academic Press, Orlando, pp 123–140
148. Willson SJ (1987) Computing fractal dimensions for additive cellular automata. *Phys D* 24(1–3):190–206
149. Willson SJ (1987) The equality of fractional dimensions for certain cellular automata. *Phys D* 24(1–3):179–189
150. Wolfram S (1985) Twenty problems in the theory of cellular automata. *Physica Scripta* 9:1–35
151. Wolfram S (1986) *Theory and Applications of Cellular Automata*. World Scientific, Singapore

Ergodic Theory: Fractal Geometry

JÖRG SCHMELING

Center for Mathematical Sciences, Lund University,
Lund, Sweden

Article Outline

Glossary

Definition of the Subject

Introduction

Preliminaries

Brief Tour Through Some Examples

Dimension Theory of Low-Dimensional Dynamical
Systems – Young's Dimension Formula

Dimension Theory
of Higher-Dimensional Dynamical Systems

Hyperbolic Measures

General Theory

Endomorphisms

Multifractal Analysis

Future Directions

Bibliography

Glossary

Dynamical system A (discrete time) dynamical system describes the time evolution of a point in phase space. More precisely a space X is given and the time evolution is given by a map $T: X \rightarrow X$. The main interest is to describe the asymptotic behavior of the trajectories (orbits) $T^n(x)$, i. e. the evolution of an initial point $x \in X$ under the iterates of the map T . More generally one is interested in obtaining information on the geometrically complicated invariant sets or measures which describe the asymptotic behavior.

Fractal geometry Many objects of interest (invariant sets, invariant measures etc.) exhibit a complicated structure that is far from being smooth or regular. The aim of fractal geometry is to study those objects. One of the main tools is the fractal dimension theory that helps to extract important properties of geometrically “irregular” sets.

Definition of the Subject

The connection between fractal geometry and dynamical system theory is very diverse. There is no unified approach and many of the ideas arose from significant examples. Also the dynamical system theory has been shown to have a strong impact on classical fractal geometry. In this article there are first presented some examples showing nontrivial results coming from the application of dimension theory. Some of these examples require a deeper knowledge of the theory of smooth dynamical systems then can be provided here. Nevertheless, the flavor of these examples can be understood. Then there is a brief overview of some of the most developed parts of the application of fractal geometry to dynamical system theory. Of course a rigorous and complete treatment of the theory cannot be given. The cautious reader may wish to check the original papers. Finally, there is an outlook over the most recent developments. This article is by no means meant to be complete. It is intended to give some of the ideas and results from this field.

Introduction

In this section some of the aspects of fractal geometry in dynamical systems are pointed out. Some notions that are used will be defined later on and I intend only to give a flavor of the applications. The nonfamiliar reader will find the definitions in the corresponding sections and can return to this section later. The *geometry* of many invariant sets or invariant measures of dynamical systems (including attractors, measures defining the statistical properties)

look very complicated at *all scales*, and their geometry is impossible to describe using standard geometric tools. For some important classes of dynamical systems, these complicated structures are intensively studied using notions of dimension. In many cases it becomes possible to relate these notions of dimension to other fundamental dynamical characteristics, such as Lyapunov exponents, entropies, pressure, etc.

On the other hand tools from dynamical systems, especially from ergodic theory and thermodynamic formalism, are extremely useful to explore the fractal properties of the objects in question. This includes dimensions of limit sets of geometric constructions (the standard Cantor set being the most famous example), which a priori, are not related to dynamical systems [46,100]. Many dimension formulas for asymptotic sets of dynamical systems are obtained by means of Bowen-type formulas, i. e. as roots of some functionals arising from thermodynamic formalism.

The dimension of a set is a subtle characteristic which measures the *geometric complexity* of the set at arbitrarily fine scales. There are many notions of dimension, and most definitions involve a measurement of geometric complexity at scale ε (which ignores the irregularities of the set at size less than ε) and then considers the limiting measurement as $\varepsilon \rightarrow 0$. A priori (and in general) these different notions can be different. An important result is the affirmative solution of the Eckmann–Ruelle conjecture by Barreira, Pesin and Schmeling [17], which says that for smooth nonuniformly hyperbolic systems, the pointwise dimension is almost everywhere constant with respect to a hyperbolic measure. This result implies that many dimension characteristics for the measure coincide.

The deep connection between dynamical systems and dimension theory seems to have been first discovered by Billingsley [21] through several problems in number theory.

Another link between dynamical systems and dimension theory is through Pesin’s theory of dimension-like characteristics. This general theory is a unification of many notions of dimension along with many fundamental quantities in dynamical system such as entropies, pressure, etc.

However, there are numerous examples of dynamical systems exhibiting pathological behavior with respect to fractal geometrical characteristics. In particular higher-dimensional systems seem to be as complicated as general objects considered in geometric measure theory. Therefore, a clean and unified theory is still not available.

The study of characteristic notions like entropy, exponents or dimensions is an essential issue in the theory of dynamical systems. In many cases it helps to classify or to understand the dynamics. Most of these charac-

teristics were introduced for different questions and concepts. For example, entropy was introduced to distinguish nonisomorphic systems and appeared to be a complete invariant for Bernoulli systems (Ornstein). Later, the thermodynamic formalism (see [107]) introduced new quantities like the pressure. Bowen [27] and also Ruelle discovered a remarkable connection between the thermodynamic formalism and the dimension theory for invariant sets. Since then many efforts were taken to find the relations between all these different quantities. It occurred that the dimension of invariant sets or measures carries lots of information about the system, combining its combinatorial complexity with its geometric complexity. Unfortunately it is extremely difficult to compute the dimension in general. The general flavor is that local divergence of orbits and global recurrence cause complicated global behavior (chaos). It is impossible to study the exact (infinite) trajectory of all orbits. One way out is to study the statistical properties of “typical” orbits by means of an invariant measure. Although the underlying system might be smooth the invariant measures may often be singular.

Preliminaries

Throughout the article the following situation is considered. Let M be a compact Riemannian manifold without boundary. On M is acting a dynamical system generated by a $C^{1+\alpha}$ diffeomorphism $T: M \rightarrow M$. The presence of a dynamical system provides several important additional tools and methods for the theory of fractal dimensions. Also the theory of fractal dimensions allows one to draw deep conclusions about the dynamical system. The importance and relevance of the study of fractal dimension will be explained in later sections.

In the next sections some of the most important tools in the fractal theory of dynamical systems are considered. The definitions given here are not necessarily the original definitions but rather the ones which are closer to contemporary use. More details can be found in [95].

Some Ergodic Theory

Ergodic theory is a powerful method to analyze statistical properties of dynamical systems. All the following facts can be found in standard books on ergodic theory like [103, 124].

The main idea in ergodic theory is to relate global quantities to observations along single orbits. Let us consider an **invariant measure**: $\mu(f^{-1}A) = \mu(A)$ for all measurable sets A . Such a measure “selects” typical trajectories. It is important to note that the properties vary with the in-

variant measures. Any such invariant measure can be decomposed into elementary parts (ergodic components).

An invariant measure is called **ergodic** if for any invariant set $A = T^{-1}A$ one has $\mu(A)\mu(M \setminus A) = 0$ (with the agreement $0 \cdot \infty = 0$), i. e. from the measure-theoretic point of view there are no nontrivial invariant subsets.

The importance of ergodic **probability** measures (i. e. $\mu(M) = 1$) lies in the following theorem of Birkhoff

Theorem 1 (Birkhoff) *Let μ be an ergodic probability measure and $\varphi \in L^1(\mu)$. Then*

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \varphi(T^k x) = \int_M \varphi d\mu \quad \mu - a.e.$$

Hausdorff Dimension

With $Z \subset \mathbb{R}^N$ and $s \geq 0$ one defines

$$m_H(s, Z) = \lim_{\delta \rightarrow 0} \inf_{\{B_i\}} \left\{ \sum_i \text{diam}(B_i)^s : \sup_i \text{diam}(B_i) < \varepsilon \text{ and } \bigcup_i B_i \supset Z \right\}.$$

Note that this limit exists. $m_H(s, Z)$ is called the **s-dimensional outer Hausdorff measure of Z**. It is immediate that there exists a unique value s^* , called the **Hausdorff dimension of Z**, at which $m_H(s, Z)$ jumps from ∞ to 0.

In general it is very hard to find optimal coverings and hence it is often impossible to compute the Hausdorff dimension of a set. Therefore a simpler notion – the lower and upper **Box dimension** – was introduced. The difference to the Hausdorff dimension is that the covering balls are assumed to have the same radius ε . Since then the limit as $\varepsilon \rightarrow 0$ does not have to exist one arrives at the notion of the upper and lower dimension.

Dimension of a Measure

Definition 1 Let $Z \subset \mathbb{R}^N$ and let μ be a probability measure supported on Z . Define the **Hausdorff dimension of the measure μ** by

$$\dim_H(\mu) \equiv \inf_{K \subset Z: \mu(K)=1} \dim_H(K).$$

Pointwise Dimension

Most invariant sets or measures are not strongly self-similar, i. e. the local geometry at arbitrarily fine scales might look different from point to point. Therefore, the notion

of pointwise dimension with respect to a Borel probability measure is defined.

Let μ be a Borel probability measure. By $B(x, \varepsilon)$ the ball with center x and radius ε is denoted. The **pointwise dimension** of the measure μ at the point x is defined as

$$d_\mu(x) := \lim_{\varepsilon \rightarrow 0} \frac{\log \mu(B(x, \varepsilon))}{\log \varepsilon}$$

if the above limit exists. If $d_\mu(x) = d$, then for small ε the measure of small balls scales as $\mu(B(x, \varepsilon)) \asymp \varepsilon^d$.

Proposition 1 Suppose μ is a probability measure supported on $Z \subset \mathbb{R}^N$. Then $d_\mu(x) = d$ for μ - almost all $x \in Z$ implies $\dim_H(\mu) = d$.

One should not take the existence of a local dimension (even for good measures) for granted. Later on it will be seen that the spectrum of the pointwise dimension (dimension spectrum) is a main object of study in classical multifractal analysis.

The dimension of a measure or a set of measures is its geometric complexity. However, under the presence of a dynamical system one also wants to measure the dynamical (combinatorial) complexity of the system. This leads to the notion of entropy.

Dimension-Like Characteristics and Topological Entropy

Pesin's theory of dimension-like characteristics provides a unified treatment of dimensions and important dynamical quantities like entropies and pressure. The topological entropy of a continuous map f with respect to a subset Z in a metric space (X, ρ) (in particular $X = M$ - a Riemannian manifold) can be defined as a dimension-like characteristic. For each $n \in \mathbb{N}$ and $\varepsilon > 0$, define the **Bowen ball** $B_n(x, \varepsilon) = \{y \in X: \rho(T^i(x), T^i(y)) \leq \varepsilon \text{ for } 0 \leq i \leq n\}$. Then let

$$m_h(Z, \alpha, \varepsilon, n) := \lim_{n \rightarrow \infty} \inf \left\{ \sum_i e^{\alpha n_i} : n_i > n, \bigcup_i B_{n_i}(x, \varepsilon) \supset Z \right\}.$$

This gives rise to an outer measure that jumps from ∞ to 0 at some value α^* . This threshold value α^* is called the **topological entropy of Z** (at scale ε). However, in many situations this value does not depend on ε . The topological entropy is denoted by $h_{\text{top}}(T|Z)$.

If Z is f - invariant and compact, this definition of topological entropy coincides with the usual definition of topological entropy [124].

The **entropy h_μ of a measure μ** is defined as $h_\mu = \inf \{h_{\text{top}}(T|Z) : \mu(Z) = 1\}$. For ergodic measures this definition coincides with the Kolmogorov-Sinai entropy (see [95]).

One has to note that in the definition of entropy metric ("round") balls are substituted by Bowen balls and the metric diameter by the "depth" of the Bowen ball. Therefore, the relation between entropy and dimension is determined by the relation of "round" balls to "oval" dynamical Bowen balls. If one understands how metric balls can be efficiently used to cover dynamical balls one can use the dynamical and relatively easy relation to compute notion of entropy to determine the dimension. However, in higher dimensions this relation is by far nontrivial. A heuristic argument for comparing "round" balls with dynamical balls is given in Subsect. "The Kaplan-Yorke Conjecture".

The Pressure Functional

A useful tool in the dimension analysis of dynamical systems is the pressure functional. It was originally defined by means of statistical physics (thermodynamic formalism) as the free energy (or pressure) of a potential φ (see for example [107]). However, in this article a dimension-like definition (see [95]) is more suitable. Again an outer measure using Bowen balls will be used. Let $\varphi: M \rightarrow \mathbb{R}$ be a continuous function and

$$m_P(Z, \alpha, \varepsilon, n, \varphi) = \lim_{n \rightarrow \infty} \inf \left\{ \sum_i \exp(-\alpha n_i) + \sup_{x \in B_{n_i}(x, \varepsilon)} \sum_{k=0}^n \varphi(T^k x) \right\}.$$

This defines an outer measure that jumps from ∞ to 0 as α increases. The threshold value α^* is called the **topological pressure** of the potential φ denoted by $P(\varphi)$. In many situations it does not depend on ε .

There is also a third way of defining the pressure in terms of a variational principle (see [124]):

$$P(\varphi) = \sup_{\mu \text{ - invariant}} \left(h_\mu + \int_M \varphi d\mu \right)$$

Brief Tour Through Some Examples

Before describing the fractal theory of dynamical systems in more detail some ideas about its role are presented. The application of dimension theory has many different aspects. At this point some (but by far not all) important examples are considered that should give the reader some

feeling about the importance and wide use of dimension theory in dynamical system theory.

Dimension of Conformal Repellers:

Ruelle's Pressure Formula

Computing or estimating a dimension via a *pressure formula* is a fundamental technique. Explicit properties of pressure help to analyze subtle characteristics. For example, the smooth dependence of the Hausdorff dimension of basic sets for Axiom- A surface diffeomorphisms on the derivative of the map follows from smoothness of pressure.

Ruelle proved the following pressure formula for the Hausdorff dimension of a conformal repeller. A conformal repeller J is an invariant set $T(J) = J = \{x \in M: f^n x \in V \forall n \in \mathbb{N} \text{ and some neighborhood } V \text{ of } J\}$ such that for any $x \in J$ the differential $D_x T = a(x) \text{Iso}_x$ where $a(x)$ is a scalar with $|a(x)| > 1$ and Iso_x an isometry of the tangent space $T_x M$.

Theorem 2 ([108]) *Let $T: J \rightarrow J$ be a conformal repeller, and consider the function $t \rightarrow P(-t \log |D_x T|)$, where P denotes pressure. Then*

$$P(-s \log |D_x T|) = 0 \iff s = \dim_H(J)$$

Iterated Function Systems

In fractal geometry one of the most-studied objects are iterated function systems, where there are given n contractions F_1, \dots, F_n of \mathbb{R}^d . The unique compact set J fulfilling $J = \bigcup_i F_i(J)$ is called the attractor of the iterated function system (see [42]). One question is to evaluate its Hausdorff dimension. Often (for example under the open set condition) the attractor J can be represented as the repeller of a piecewise expanding map $T: \mathbb{R}^d \rightarrow \mathbb{R}^d$ where the F_i are the inverse branches of the map T . In general it is by far not trivial to determine the dimension of J or even the dimension of a measure sitting on J . The following example explains some of those difficulties.

Let $1/2 < \lambda < 1$ and consider the maps $F_i: [0, 1] \rightarrow [0, 1]$ given by $F_1(x) = \lambda x$ and $F_2(x) = \lambda x + (1 - \lambda)$. Then the images of F_1, F_2 have an essential overlap and $J = [0, 1]$. If one randomizes this construction in the way that one applies both maps each with probability $1/2$ a probability measure is induced on J . This measure might be absolute continuous with respect to Lebesgue measure or not. Already Erdős realized that for some special values of λ (for example for the inverse of the golden mean) the induced measure is singular. In a breakthrough paper B. Solomyak ([119]) proved that for a.e. λ the in-

duced measure is absolutely continuous. A main ingredient in the proof is a transversality condition in the parameter space: the images of arbitrary two random samples of the (infinite) applications of the maps F_i have to cross with nonzero speed when the parameter λ changes. This is a general mechanism which allows one to handle more general situations.

Homoclinic Bifurcations

for Dissipative Surface Diffeomorphisms

Homoclinic tangencies and their bifurcations play a fundamental role in the theory of dynamical systems [87,88,89]. Systems with homoclinic tangencies have a complicated and subtle quasi-local behavior. Newhouse showed that homoclinic tangencies can persist under small perturbations, and that horseshoes may co-exist with infinitely many sinks in a neighborhood of the homoclinic orbit and hence the system is not hyperbolic (**Newhouse phenomenon**).

Let $T_\mu: M^2 \rightarrow M^2$ be a smooth parameter family of surface diffeomorphisms that exhibits for $\mu = 0$ an invariant hyperbolic set Λ_0 (horseshoe) and undergoes a homoclinic bifurcation. The Hausdorff dimension of the hyperbolic set Λ_0 for T_0 determines whether hyperbolicity is the *typical* dynamical phenomenon near T_0 or not. If $\dim_H \Lambda_0 < 1$, then hyperbolicity is the prevalent dynamical phenomenon near f_0 . This is not the case if $\dim_H \Lambda_0 > 1$.

More precisely, let \mathcal{NW}_μ denote the set of nonwandering points of T_μ in an open neighborhood of Λ_0 after the homoclinic bifurcation. Let ℓ denote Lebesgue measure.

Theorem 3 ([87,88]) *If $\dim_H \Lambda_0 < 1$, then*

$$\lim_{\mu_0 \rightarrow 0} \frac{\ell\{\mu \in [0, \mu_0]: \mathcal{NW}_\mu \text{ is hyperbolic}\}}{\mu_0} = 1.$$

The hyperbolicity co-exists with the Newhouse phenomena for a residual set of parameter values.

Theorem 4 (Palis–Yoccoz) *If $\dim_H \Lambda_0 > 1$, then*

$$\lim_{\mu_0 \rightarrow 0} \frac{\ell\{\mu \in [0, \mu_0]: \mathcal{NW}_\mu \text{ is hyperbolic}\}}{\mu_0} < 1.$$

Some Applications to Number Theory

Sometimes dimension problems in number theory can be transferred to dynamical systems and attacked using tools from dynamics.

Example Diadic expansion of numbers:

Consider a real number expanded in base 2, i.e. $x = \sum_{n=1}^{\infty} x_n/2^n$. Let

$$X_p = \left\{ x: \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} x_k = p \right\}.$$

Borel showed that $\ell(X_{1/2}) = 1$, where ℓ denotes Lebesgue measure. This result is an easy consequence of the Birkhoff ergodic theorem applied to the characteristic function of the digit 1 (which in this simple case is the Strong Law of Large Numbers for i.i.d. processes).

One can ask how large is the set X_p in general. Eggleston [39] discovered the following wonderful dimension formula, which Billingsley [21] interpreted in terms of dynamics and reproved using tools from ergodic theory.

Theorem 5 ([39]) *The Hausdorff dimension of X_p is given by*

$$\dim_H(X_p) = (1/\log 2)[-p \log p - (1-p) \log(1-p)].$$

The underlying dynamical system is $E_2(x) = 2x \bmod 1$ and the dimension is the dimension of the Bernoulli p measure. Other cases included that by Rényi who proposed generalization of the base d expansion from integer base d to noninteger base β . In this case the underlying dynamical system is the (in general non-Markovian) beta shift $\beta(x) = \beta x \bmod 1$ studied in [110].

There are many investigations concerning the approximation of real numbers by rationals by using dynamical methods. The underlying dynamical system in this case is the Gauss map $T(x) = (1/x) \bmod 1$. This map is uniformly expanding, but has infinitely many branches.

Example Continued fraction approximation of numbers

Consider the continued fraction expansion of $x \in [0, 1]$, i.e.,

$$x = \frac{1}{a_1 + \frac{1}{a_2 + \frac{1}{a_3 + \frac{1}{a_4 + \ddots}}}} = [a_1, a_2, a_3, a_4, \dots],$$

and the approximants $p_n(x)/q_n(x) = [a_1, a_2, \dots, a_n]$ (i.e. the approximation given by the finite continued fraction after step n).

The set of numbers which admit a *faster approximation* by rational numbers is defined as follows. For $\tau \geq 2$, let \mathcal{F}_τ

$$\mathcal{F}_\tau = \left\{ x \in [0, 1]: \left| x - \frac{p}{q} \right| \leq \frac{1}{q^\tau} \text{ for infinitely many } p/q \right\}.$$

It is well known that this set has zero measure for each $\tau \geq 2$. Jarník [58] computed the Hausdorff dimension of \mathcal{F}_τ and showed that $\dim_H(\mathcal{F}_\tau) = 2/\tau$. However, nowadays there are methods from dynamical systems which not only allow unified proofs but also can handle more subtle subsets of the reals defined by some properties of their continued fraction expansion (see for example [3,105]).

Infinite Iterated Function Systems and Parabolic Systems

In the previous section a system with infinitely many branches appeared. This can be regarded as an iterated function system with infinitely many maps F_i . This situation is quite general. If one considers a (one-dimensional) system with a parabolic (indifferent) fixed point, i.e. there is a fixed point where the derivative has absolute value equal to 1, one often uses an induced system. For this one chooses nearby the parabolic point a higher iterate of the map in order to achieve uniform expansion away from the parabolic point. This leads to infinitely many branches since the number of iterates has to be increased the closer the parabolic point is. The main difference to the finite iterated function system is that the setting is no longer compact and many properties of the pressure functional are lost.

Mauldin and Urbański and others (see for example [3,75,76,78]) developed a thermodynamic formalism adapted to the pressure functional for infinite iterated function systems. Besides noncompactness one of the main problems is the loss of analyticity (phase transitions) and convexity of the pressure functional for infinite iterated function systems.

Complex Dynamics

Let $T: \mathbb{C} \rightarrow \mathbb{C}$ be a polynomial and J its Julia set (repeller of this system). If this set is hyperbolic, i.e. the derivative at each point has absolute value larger than 1 the study of the dimension can be related to the study of a finite iterated function system. However, in the presence of a parabolic fixed point this leads to an infinite iterated function system.

If one considers the coefficients of the polynomial as parameters one often sees a qualitative change in the

asymptotic behavior. For example, the classical Mandelbrot set for polynomials $z^2 + c$ is the locus of values $c \in \mathbb{C}$ for which the orbit of the origin stays bounded. This set is well known to be fractal. However, its complete description is still not available.

Embedology and Computational Aspects of Dimension

Tools from dynamical systems are becoming increasingly important to study the time evolution of deterministic systems in engineering and the physical and biological sciences. One of the main ideas is to model a “real world” system by a smooth dynamical system which possesses a strange attractor with a natural ergodic invariant measure. When studying a complicated real world system, one can measure only a very small number of variables. The challenge is to reconstruct the underlying attractor from the time measurement of a scalar quantity. An idealized measurement is considered as a function $h: M^n \rightarrow \mathbb{R}$.

The main tool researchers currently use to *reconstruct* the model system is called attractor reconstruction (see papers and references in [86]). This method is based on embedding with time delays (see the influential paper [33], where the authors attribute the idea of delay coordinates to Ruelle), where one attempts to reconstruct the attractor for the model using a single long trajectory. Then one considers the points in \mathbb{R}^{p+1} defined by $(x_k, x_{k+\tau}, x_{k+2\tau}, \dots, x_{k+p\tau})$.

Takens [122] showed that for a smooth $T: M^n \rightarrow M^n$ and for typical smooth h , the mapping $\varphi: M^n \rightarrow \mathbb{R}^{2n+1}$ defined by $x \rightarrow (h(x), h(f^\tau(x)), \dots, h(f^{2n\tau}(x)))$ is an embedding. Since the box dimension of the attractor Λ may be much less than the dimension of the ambient manifold n , an interesting mathematical question is whether there exists $p < 2n + 1$ such that the mapping on the attractor $\varphi: \Lambda \rightarrow \mathbb{R}^p$ defined by $x \rightarrow (h(x), h(f^\tau(x)), \dots, h(f^{p\tau}(x)))$ is 1 – 1? It is known that for a typical smooth h the mapping φ is 1 – 1 for $p > 2 \dim_B(\Lambda)$ [29].

Denjoy Systems

This section will give some ideas indicating the principle difficulties that arise in systems with low complexity. Contrary to hyperbolic systems (each vector in the tangent space is either contracted or expanded) finer mechanisms determine the local behavior of the scaling of balls. While in hyperbolic systems the dynamical scaling of small balls is exponential in a low-complexity system this scaling is subexponential and hence the linearization error is of the same magnitude. Up to now there is no general dimension theory for low complexity systems.

A specific example presented here is considered in [67]. Poincaré showed that to each orientation preserving homeomorphism of the circle $S^1 = \mathbb{R}/\mathbb{Z}$ is associated a unique real parameter $\alpha \in [0, 1)$, called the rotation number, so that the ordered orbit structure of T is the same as that of the rigid rotation R_α , where $R_\alpha(t) = (t + \alpha) \bmod 1$, provided that α is irrational. Half a century later, Denjoy [35] constructed examples of C^1 diffeomorphisms that are not conjugate (via a homeomorphism) to rotations. This was improved later on by Herman [55]. In these examples, the minimal set of T is necessarily a Cantor set Ω .

The arithmetic properties of the rotation number have a strong effect on the properties of T . One area that has been well understood is the relation between the differentiability of T , the differentiability of the conjugation and the arithmetic properties of the rotation number. (See, for example, Herman [55]) Without stating any precise theorem, the results differ sharply for Diophantine and for Liouville rotation numbers (definition follows). Roughly speaking the conjugating map is always regular for Diophantine rotation numbers while it might be not smooth at all for Liouville rotation numbers.

Definition 2 An irrational number α is of Diophantine class $\nu = \nu(\alpha) \in \mathbb{R}^+$ if

$$\|q\alpha\| < \frac{1}{q^\mu}$$

has infinitely many solutions in integers q for $\mu < \nu$ and at most finitely many for $\mu > \nu$ where $\|\cdot\|$ denotes the distance to the nearest integer.

In [67] the effect of the rotation number on the dimension of Ω is studied. There the main result is

Theorem 6 Assume that $0 < \delta < 1$ and that $\alpha \in (0, 1)$ is of Diophantine class $\nu \in (0, \infty)$. Then an orientation preserving $C^{1+\delta}$ diffeomorphism of the circle with rotation number α and minimal set Ω_α^δ satisfies

$$\dim_H \Omega_\alpha^\delta \geq \frac{\delta}{\nu}.$$

Furthermore, these results are sharp, i. e. the standard Denjoy examples attain the minimum.

Return Times and Dimension

Recently an interesting connection between the pointwise dimensions, multifractal analysis, and recurrence behavior of trajectories was discovered [1, 11, 23]. Roughly speaking,

given an ergodic probability measure μ the return time asymptotics (as the neighborhood of the point shrinks) of μ -a.e point is determined by the pointwise dimension of μ at this point. The deeper understanding of this relation would help to get a unified approach to dimensions, exponents, entropies, recurrence times and correlation decay.

Dimension Theory of Low-Dimensional Dynamical Systems – Young’s Dimension Formula

In this section a remarkable extension of Ruelle’s dimension formula by Young [128] for the dimension of a measure is discussed.

Theorem 7 *Let $T: M^2 \rightarrow M^2$ be a C^2 surface diffeomorphism and let μ be an ergodic measure. Then*

$$\dim_H(\mu) = h_\mu(f) \left(\frac{1}{\lambda^1} - \frac{1}{\lambda^2} \right),$$

where $\lambda^1 \leq \lambda^2$ are the two Lyapunov exponents for μ .

In [79], Manning and McCluskey prove the following dimension formula for a basic set (horseshoe) of an Axiom-A surface diffeomorphism which is a set-version of Young’s formula.

Theorem 8 *Let Λ be a basic set for a C^2 Axiom-A surface diffeomorphism $T: M^2 \rightarrow M^2$. Then $\dim_H(\Lambda) = s_1 + s_2$, where s_1 and s_2 satisfy*

$$\begin{aligned} P(-s \log \|DT_x|E_x^u\|) &= 0 \\ P(s \log \|DT_x|E_x^s\|) &= 0. \end{aligned}$$

where E^s and E^u are the stable and unstable directions, respectively.

Some Remarks on Dimension Theory for Low-Dimensional versus High-Dimensional Dynamical Systems

Unlike lower dimensions (one, two, or conformal repellers), for higher-dimensional dynamical systems there are no general dimension formulas (besides the Ledrappier–Young formula), and in general dimension theory is much more difficult. This is due to several problems:

- 1) The geometry of the Bowen balls differs in a substantial way from round balls.
- 2) Number theoretic properties of some scaling rates [104, 106] enter into dimension calculations in ways they do

not in low dimensions (see Subsect. “Iterated Function Systems”).

- 3) The dimension theory of sets is often reduced to the theory of invariant measures. However, there is no invariant measure of full dimension in general and measure-theoretic considerations do not apply [79].
- 4) The stable and unstable foliations for higher dimensional systems are typically not C^1 [51, 109]. Hence, to split the system into an expanding and a contracting part is far more subtle.

Dimension Theory of Higher-Dimensional Dynamical Systems

Here an example of a hyperbolic attractor in dimension 3 is considered to highlight some of the difficulties. Let Δ denote the unit disc in \mathbb{R}^2 . Let $f: S^1 \times \Delta \rightarrow S^1 \times \Delta$ be of the form

$$T(t, x, y) = (\varphi(t), \psi^1(t, x), \psi^2(t, y)),$$

with

$$0 < \max_{S^1 \times \Delta} \frac{\partial}{\partial x} \psi^1(t, x) < \min_{S^1 \times \Delta} \frac{\partial}{\partial y} \psi^2(t, y) < \lambda < 1.$$

The limit set

$$\Lambda := \bigcap_{n \in \mathbb{N}} T^n(S^1 \times \Delta)$$

is called the attractor or the *solenoid*. It is an example of a structurally stable basic set and is one of the fundamental examples of a uniformly hyperbolic attractor.

The following result can be proved.

Theorem 9 ([24, 52]) *For all t , the thermodynamic pressure*

$$P \left(\dim_H \Lambda_t^s \log \left| \frac{\partial}{\partial y} \psi^2(t, y) \right| \right) = 0.$$

In particular, the stable dimension is independent of the stable section.

In this particular case the invariant axes for strong and weak contraction split the system smoothly and the difficulty is to show that the strong contraction is dominated by the weaker. In particular one has to ensure that effects as described in Subsect. “Iterated Function Systems” do not appear. In the general situation this is not the case and one lacks a similar theorem in the general situation. In particular, the unstable foliation is not better than Hölder and does not provide a “nice” coordinate.

Hyperbolic Measures

Given $x \in M$ and a vector $v \in T_x M$ the **Lyapunov exponent** is defined as

$$\chi(x, v) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \|D_x T^n v\|.$$

provided this limit exists. For fixed x , the numbers $\chi(x, \cdot)$ attain only finitely many values. By ergodicity they are μ -a.e. constant, and the corresponding values are denoted by

$$\chi_1 \leq \dots \leq \chi_p,$$

where $p = \dim M$. Denote by s (s stands for stable) the largest index such that $\chi_s < 0$.

Definition 3 The invariant ergodic measure μ is said to be **hyperbolic** if $\chi_i \neq 0$ for every $i = 1, \dots, p$.

The Kaplan–Yorke Conjecture

In this section a heuristic argument for the dimension of an invariant ergodic measure will be given. This argument uses a specific cover of a dynamical ball by “round” balls. These ideas are essentially developed by Kaplan and Yorke [61] for the estimation of the dimension of an invariant set. Their estimates always provide an upper bound for the dimension. Kaplan and Yorke conjectured that in typical situations these estimates also provide a lower bound for the dimension of an attractor. Ledrappier and Young [72,73] showed that in case this holds for an invariant measure, then this measure has to be very special: an SBR-measure (after Sinai, Ruelle, and Bowen) (an SRB-measure is a measure that describes the asymptotic behavior for a set of initial points of positive Lebesgue measure and has absolutely continuous conditional measures on unstable manifolds).

Consider a small ball B in the phase space. The image $T^n B$ is almost an ellipsoid with axes of length

$$e^{\chi_1 n}, \dots, e^{\chi_p n}.$$

For $1 \leq i \leq s$, cover $T^n B$ by balls of radius $e^{\chi_i n}$. Then approximately

$$\frac{\exp[\chi_{i+1} n]}{\exp[\chi_i n]} \cdots \frac{\exp[\chi_p n]}{\exp[\chi_i n]}.$$

balls are needed for covering. The dimension can be estimated from above by

$$\dim_B \Lambda \leq \frac{\sum_{j>i} \chi_j}{|\chi_i|} + (p - i) := \dim_L^i \Lambda. \quad (1)$$

This is the Kaplan–Yorke formula.

General Theory

In this section the dimension theory of higher-dimensional dynamical systems is investigated. Most developed is this theory for invariant measures. There is an important connection between Lyapunov exponents and the measure theoretic entropy and dimensions of measures that will be presented here.

Let μ be an ergodic invariant measure. The Oseledec and Pesin Theory guarantee that local stable manifolds exist at μ -a.e. point. As for the Kaplan–Yorke formula the idea is to consider the contributions to the entropy and to the dimension in the directions of χ_i .

Historically the first connections between exponents and entropy were discovered by Margulis and Ruelle. They proved that

$$h_\mu(f) \leq - \sum_{i=1}^s \chi_i$$

for a C^1 diffeomorphism T . Pesin [91] showed that this inequality is actually an equality if the measure μ is essentially the Riemannian volume on unstable manifolds. Ledrappier and Young [72] showed that this is indeed a necessary condition. They also provided an exact formula:

Theorem 10 (Ledrappier–Young [72,73]) With $d^0 = 0$ for a C^2 diffeomorphism holds

$$h_\mu(f) = - \sum_{i=1}^s \chi_i (d^i - d^{i-1})$$

where d^i are the dimensions of the (conditional) measure on the i th unstable leaves.

The proof of this theorem is difficult and uses the theory of *nonuniform hyperbolic systems* (Pesin theory).

In dimension 1 and 2 the above theorem resembles Ruelle’s and Young’s theorems.

The reader should note that the above theorem includes also the existence of the pointwise dimension along the stable and unstable direction. Here the question arises whether this implies the existence of the pointwise dimension itself.

The Existence of the Pointwise Dimension for Hyperbolic Measure – the Eckmann–Ruelle Conjecture

In [17], Barreira, Pesin, and Schmeling prove that every hyperbolic measure has an *almost* local product structure, i. e., the measure of a small ball can be approximated by the product of the stable conditional measure of the sta-

ble component and the unstable conditional measure of the unstable component, up to small exponential error. This was used to prove the existence of the pointwise dimension of every hyperbolic measure almost everywhere. Moreover, the pointwise dimension is the sum of the contributions from its stable and unstable part. This implies that most dimension-type characteristics of the measure (including the Hausdorff dimension, box dimension, and information dimension) coincide, and provides a rigorous mathematical justification of the concept of fractal dimension for hyperbolic measures. The existence of the pointwise dimension for hyperbolic measures had been conjectured long before by Eckmann and Ruelle.

The hypotheses of this theorem are sharp. Ledrappier and Misiurewicz [70] constructed an example of a non-hyperbolic measure for which the pointwise dimension is not constant a.e. In [101], Pesin and Weiss present an example of a Hölder homeomorphism with Hölder constant arbitrarily close to one, where the pointwise dimension for the unique measure of maximal entropy does not exist a.e. There is also a one-dimensional example by Cutler [34].

Endomorphisms

The previous section indicates that the dimensional properties of hyperbolic measures under invariant conditions for a diffeomorphism are understood. However, partial differential equations often generate only semi-flows and the corresponding dynamical system is noninvertible. Also, Poincaré sections sometimes introduce singularities. For such dynamical systems the theory of diffeomorphisms does not apply. However, the next theorem allows under some conditions application of this theory. It essentially rules out similar situations as considered in Subsect. “Iterated Function Systems”.

Definition 4 A system (possibly with singularities) is *almost surely invertible* if it is invertible on a set of full measure. This implies that a full measure set of points has unique forward and backward trajectories.

Theorem 11 (Schmeling–Troubetzkoy [114]) *A two-dimensional system with singularities is almost surely invertible (w.r.t. an SRB-measure) if and only if Young’s formula holds.*

Multifractal Analysis

A group of physicists [50] suggested the idea of a multifractal analysis.

The Dynamical Characteristic View of Multifractal Analysis

The aim of multifractal analysis is an attempt to understand the fine structure of the level sets of the fundamental asymptotic quantities in ergodic theory (e. g., Birkhoff averages, local entropy, Lyapunov exponents). For ergodic measures these quantities are a.e. constant, however may depend on the underlying ergodic measure. Important elements of multifractal analysis entail determining the range of values these characteristics attain, an analysis of the dimension of the level sets, and an understanding of the sets where the limits do not exist. A general concept of multifractal analysis was proposed by Barreira, Schmeling and Pesin [16]. An important field of applications of multifractal analysis is to describe sets of real numbers that have constraints on their digits or continued fraction expansion.

General Multifractal Formalism

In this section the abstract theory of multifractal analysis is described. Let X, Y be two measurable spaces and $g: X \setminus \mathcal{B} \rightarrow Y$ be any measurable function where \mathcal{B} is a measurable (possibly empty) subset of X (in the standard applications $Y = \mathbb{R}$ or $Y = \mathbb{C}$). The associated **multifractal decomposition** of X is defined as

$$X = \mathcal{B} \cup \bigcup_{\alpha \in Y} K_{\alpha}^g$$

where

$$K_{\alpha}^g := \{x \in X: g(x) = \alpha\}$$

For a given set function $G: 2^X \rightarrow \mathbb{R}$ the **multifractal spectrum** is defined by

$$\mathcal{F}(\alpha) := G(K_{\alpha}^g).$$

At this point some classical and concrete examples of this general framework are considered.

The Entropy Spectrum

Let μ be an ergodic invariant measure for $T: X \rightarrow X$. If one sets

$$g^E(x) := h_{\mu}(x) \in Y = \mathbb{R}$$

and

$$G^E(Z) = h_{\text{top}}(T|Z)$$

the associated multifractal spectrum $f_{\mu}^E(\alpha)$ is called the **entropy spectrum**.

The Dimension Spectrum

This is the **classical multifractal spectrum**.

Let μ be an invariant ergodic measure on a complete separable metric space X .

Set

$$g^D(x) := d_\mu(x) \in Y = \mathbb{R}$$

and

$$G^D(Z) = \dim_H Z.$$

The associated multifractal spectrum $f_\mu^D(\alpha)$ is called the **dimension spectrum**.

The Lyapunov Spectrum

Let

$$g^L(x) := \chi(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} \psi(\mathbb{T}^k x) \in Y = \mathbb{R}$$

where $\psi(x) = \log |D_x T|$ and

$$G^D(Z) = \dim_H Z \quad \text{or} \quad G^E(Z) = h_{\text{top}}(\sigma|Z).$$

The associated multifractal spectra $f_L^D(\alpha)$ and $f_L^E(\alpha)$ are called **Lyapunov spectra**.

It was observed by H. Weiss [126] and Barreira, Pesin and Schmeling [16] that for conformal repellers

$$f_L^D(\alpha) = f_{\mu_h}^D\left(\frac{\log 2}{\alpha}\right) \quad (2)$$

where the dimension spectrum on the right-hand side is with respect to the measure of maximal entropy, and

$$f_L^E(\alpha) = f_{\mu_D}^E(\dim_H \Lambda \cdot \alpha) \quad (3)$$

where the entropy spectrum on the right-hand side is with respect to the measure of maximal dimension.

The following list summarizes the state of the art for the dynamical characteristic multifractal analysis of dynamical systems. The precise statements can be found in the original papers.

- [102,126] For conformal repellers and Axiom-A surface diffeomorphisms, a complete multifractal analysis exists for the Lyapunov exponent.
- [16,102,126] For mixing a subshift of finite type, a complete multifractal analysis exists for the Birkhoff average for a Hölder continuous potential and for the local entropy for a Gibbs measure with Hölder continuous potential.

- [12,97] There is a complete multifractal analysis for hyperbolic flows.
- [123] There is a generalization of the multifractal analysis on subshifts with specification and continuous potentials.
- [13,19] There is an analysis of “mixed” spectra like the dimension spectrum of local entropies and also an analysis of joint level sets determined by more than one (measurable) function.
- [105] For the Gauss map (and a class of nonuniformly hyperbolic maps) a complete multifractal analysis exists for the Lyapunov exponent.
- [57] A general approach to multifractal analysis for repellers with countably many branches is developed. It shows in contrary to finitely many branches features of nonanalytic behavior.

In the first three statements the multifractal spectra are analytic concave functions that can be computed by means of the Legendre transform of the pressure functional with respect to a suitable chosen family of potentials. In the remaining items this is no longer the case. Analyticity and convexity properties of the pressure functional are lost. However, the authors succeeded to provide a satisfactory theory in these cases.

Multifractal Analysis and Large Deviation Theory

There are deep connections between large deviation theory and multifractal analysis. The variational formula for pressure is an important tool in the analysis, and can be viewed (and proven) as a large deviation result [41]. Some authors use large deviation theory as a tool to effect multifractal analysis.

Future Directions

The dimension theory is fast developing and of great importance in the theory of dynamical systems. In the most ideal situations (low dimensions and hyperbolicity) a generally far reaching and powerful theory has been developed. It uses ideas from statistical physics, fractal geometry, probability theory and other fields.

Unfortunately, the richness of this theory does not carry over to higher-dimensional systems. However, recent developments have shown that it is possible to obtain a general theory for the dimension of measures. Part of this theory is the development of the analytic tools of nonuniformly hyperbolic systems.

Therefore, the dimension theory of dynamical systems is far from complete. In particular, it is usually difficult to apply the general theory to concrete examples, for instance

if one really wants to *compute* the dimension. The general theory does not provide a way to compute the dimension but gives rather connections to other characteristics. Moreover, in the presence of neutral directions (zero Lyapunov exponents) one encounters all the difficulties arising in low-complexity systems.

Another important open problem is to understand the dimension theory of invariant sets in higher-dimensional spaces. One way would be to relate the dimension of sets to the dimension of measures. Such a connection is not clear. The reason is that most systems do not exhibit a measure whose dimension coincides with the dimension of its support (invariant set). But there are some reasons to conjecture that any compact invariant set of an expanding map in any dimension carries a measure of maximal dimension (see [48,63]). If this conjecture is true one obtains an invariant measure whose unstable dimension coincides with the unstable dimension of the invariant set. There is also a measure of maximal stable dimension. Combining these two measures one could establish an analogous theory for invariant sets as for invariant measures.

Last but not least one has to mention the impact of the dimension theory of dynamical systems on other fields. This new point of view makes in many cases the posed problems more tractable. This is illustrated in examples from number theory, geometric limit constructions and others. The applications of the dimension theory of dynamical systems to other questions seem to be unlimited.

Bibliography

Primary Literature

1. Afraimovich VS, Schmeling J, Ugalde J, Urias J (2000) Spectra of dimension for Poincaré recurrences. *Discret Cont Dyn Syst* 6(4):901–914
2. Afraimovich VS, Chernov NI, Sataev EA (1995) Statistical Properties of 2D Generalized Hyperbolic Attractors. *Chaos* 5:238–252
3. Aihua F, Yunping J, Jun W (2005) Asymptotic Hausdorff dimensions of Cantor sets associated with an asymptotically non-hyperbolic family. *Ergod Theory Dynam Syst* 25(6):1799–1808
4. Alexander J, Yorke J (1984) Fat Baker's transformations. *Erg Th Dyn Syst* 4:1–23
5. Ambroladze A, Schmeling J (2004) Lyapunov exponents are not stable with respect to arithmetic subsequences. In: *Fractal geometry and stochastics III*. Progr Probab 57. Birkhäuser, Basel, pp 109–116
6. Artin E (1965) Ein mechanisches System mit quasi-ergodischen Bahnen, *Collected papers*. Addison Wesley, pp 499–501
7. Barreira L () Variational properties of multifractal spectra. *IST preprint*
8. Barreira L (1995) Cantor sets with complicated geometry and modeled by general symbolic dynamics. *Random Comp Dyn* 3:213–239
9. Barreira L (1996) A non-additive thermodynamic formalism and applications to dimension theory of hyperbolic dynamical systems. *Erg Th Dyn Syst* 16:871–927
10. Barreira L (1996) A non-additive thermodynamic formalism and dimension theory of hyperbolic dynamical systems. *Math Res Lett* 3:499–509
11. Barreira L, Saussol B (2001) Hausdorff dimension of measure via Poincaré recurrence. *Comm Math Phys* 219(2):443–463
12. Barreira L, Saussol B (2001) Multifractal analysis of hyperbolic flows. *Comm Math Phys* 219(2):443–463
13. Barreira L, Saussol B (2001) Variational principles and mixed multifractal spectra. *Trans Amer Math Soc* 353(10):3919–3944 (electronic)
14. Barreira L, Schmeling J (2000) Sets of “Non-typical” Points Have Full Topological Entropy and Full Hausdorff Dimension. *Isr J Math* 116:29–70
15. Barreira L, Pesin Y, Schmeling J (1997) Multifractal spectra and multifractal rigidity for horseshoes. *J Dyn Contr Syst* 3:33–49
16. Barreira L, Pesin Y, Schmeling J (1997) On a General Concept of Multifractal Rigidity: Multifractal Spectra For Dimensions, Entropies, and Lyapunov Exponents. *Multifractal Rigidity*. *Chaos* 7:27–38
17. Barreira L, Pesin Y, Schmeling J (1999) Dimension and Product Structure of Hyperbolic Measures. *Annals Math* 149:755–783
18. Barreira L, Saussol B, Schmeling J (2002) Distribution of frequencies of digits via multifractal analysis. *J Number Theory* 97(2):410–438
19. Barreira L, Saussol B, Schmeling J (2002) Higher-dimensional multifractal analysis. *J Math Pures Appl* (9) 81(1):67–91
20. Belykh VP (1982) Models of discrete systems of phase synchronization. In: Shakhildyan VV, Belyushina LN (eds) *Systems of Phase Synchronization*. Radio i Svyaz, Moscow, pp 61–176
21. Billingsley P (1978) *Ergodic Theory and Information*. Krieger
22. Blinchevskaya M, Ilyashenko Y (1999) Estimate for the Entropy Dimension Of The Maximal Attractor For k —Constructing Systems In An Infinite-Dimensional Space. *Russ J Math Phys* 6(1):20–26
23. Boshernitzan M (1993) Quantitative recurrence results. *Invent Math* 113:617–631
24. Bothe H-G (1995) The Hausdorff dimension of certain solenoids. *Erg Th Dyn Syst* 15:449–474
25. Bousch T (2000) Le poisson n'a pas d'arêtes. *Ann IH Poincaré (Prob-Stat)* 36(4):489–508
26. Bowen R (1973) Topological entropy for noncompact sets. *Trans Amer Math Soc* 184:125–136
27. Bowen R (1979) Hausdorff Dimension Of Quasi-circles. *Publ Math IHES* 50:11–25
28. Bylov D, Vinograd R, Grobman D, Nemyckii V (1966) Theory of Lyapunov exponents and its application to problems of stability. *Izdat "Nauka", Moscow (in Russian)*
29. Casdagli M, Sauer T, Yorke J (1991) Embedology. *J Stat Phys* 65:589–616
30. Ciliberto S, Eckmann JP, Kamphorst S, Ruelle D (1971) Liapunov Exponents from Times. *Phys Rev A* 34
31. Collet P, Lebowitz JL, Porzio A (1987) The Dimension Spectrum of Some Dynamical Systems. *J Stat Phys* 47:609–644
32. Constantin P, Foias C (1988) *Navier-Stokes Equations*. Chicago U Press

33. Cruchfield J, Farmer D, Packard N, Shaw R (1980) Geometry from a Time Series. *Phys Rev Lett* 45:712–724
34. Cutler C (1990) Connecting Ergodicity and Dimension in Dynamical Systems. *Ergod Th Dynam Syst* 10:451–462
35. Denjoy A (1932) Sur les courbes définies par les équations différentielles à la surface du tore. *J Math Pures Appl* 2:333–375
36. Ding M, Grebogi C, Ott E, Yorke J (1993) Estimating correlation dimension from a chaotic times series: when does the plateau onset occur? *Phys D* 69:404–424
37. Dodson M, Rynne B, Vickers J (1990) Diophantine approximation and a lower bound for Hausdorff dimension. *Mathematika* 37:59–73
38. Douady A, Oesterle J (1980) Dimension de Hausdorff Des Attracteurs. *CRAS* 290:1135–1138
39. Eggleston HG (1952) Sets of Fractional Dimension Which Occur in Some Problems of Number Theory. *Proc Lond Math Soc* 54:42–93
40. Ellis R (1984) Large Deviations for a General Class of Random Vectors. *Ann Prob* 12:1–12
41. Ellis R (1985) *Entropy, Large Deviations, and Statistical Mechanics*. Springer
42. Falconer K (1990) *Fractal Geometry, Mathematical Foundations and Applications*. Cambridge U Press, Cambridge
43. Fan AH, Feng DJ, Wu J (2001) Recurrence, dimension and entropy. *J Lond Math Soc* 64(2):229–244
44. Frederickson P, Kaplan J, Yorke E, Yorke J (1983) The Liapunov Dimension Of Strange Attractors. *J Differ Equ* 49:185–207
45. Frostman O (1935) Potential d'équilibre Et Capacité des Ensembles Avec Quelques Applications à la Théorie des Fonctions. *Meddel Lunds Univ Math Sem* 3:1–118
46. Furstenberg H (1967) Disjointness in Ergodic Theory, Minimal Sets, and a Problem in Diophantine Approximation. *Math Syst Theory* 1:1–49
47. Furstenberg H (1970) Intersections of Cantor Sets and Transversality of Semigroups I. In: *Problems in Analysis. Sympos Salomon Bochner*, Princeton Univ. Princeton Univ Press, pp 41–59
48. Gatzouras D, Peres Y (1996) The variational principle for Hausdorff dimension: A survey, in *Ergodic theory of \mathbb{Z}^d actions*. In: Pollicott M et al (ed) *Proc of the Warwick symposium*. Cambridge University Press. *Lond Math Soc Lect Note Ser* 228:113–125
49. Grassberger P, Procaccia I, Hentschel H (1983) On the Characterization of Chaotic Motions, *Lect Notes. Physics* 179:212–221
50. Halsey T, Jensen M, Kadanoff L, Procaccia I, Shraiman B (1986) Fractal Measures and Their Singularities: The Characterization of Strange Sets. *Phys Rev A* 33(N2):1141–1151
51. Hasselblatt B (1994) Regularity of the Anosov splitting and of horospheric foliations. *Ergod Theory Dynam Syst* 14(4):645–666
52. Hasselblatt B, Schmeling J (2004) Dimension product structure of hyperbolic sets. In: *Modern dynamical systems and applications*. Cambridge Univ Press, Cambridge, pp 331–345
53. Henley D (1992) Continued Fraction Cantor Sets, Hausdorff Dimension, and Functional Analysis. *J Number Theory* 40:336–358
54. Hentschel HGE, Procaccia I (1983) The Infinite Number of Generalized Dimensions of Fractals and Strange Attractors. *Physica* 8D:435–444
55. Herman MR (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. *Publications de l'Institut de Mathématiques des Hautes Études Scientifiques* 49:5–234
56. Hunt B (1996) Maximal Local Lyapunov Dimension Bounds The Box Dimension Of Chaotic Attractors. *Nonlinearity* 9:845–852
57. Iommi G (2005) Multifractal analysis for countable Markov shifts. *Ergod Theory Dynam Syst* 25(6):1881–1907
58. Jarník V (1931) Über die simultanen diophantischen Approximationen. *Math Zeitschr* 33:505–543
59. Jenkinson O (2001) Rotation, entropy, and equilibrium states. *Trans Amer Math Soc* 353:3713–3739
60. Kalinin B, Sadovskaya V (2002) On pointwise dimension of non-hyperbolic measures. *Ergod Theory Dynam Syst* 22(6):1783–1801
61. Kaplan JL, Yorke JA (1979) Functional differential equations and approximation of fixed points. *Lecture Notes. In: Mathematics vol 730*. Springer, Berlin, pp 204–227
62. Katznelson Y, Weiss B (1982) A simple proof of some ergodic theorems. *Isr J of Math* 42:291–296
63. Kenyon R, Peres Y (1996) Measure of full dimension on affine invariant sets. *Erg Th Dyn Syst* 16:307–323
64. Kesseböhmer M (1999) *Multifraktale und Asymptotiken grosser Deviationen*. Thesis U Göttingen, Göttingen
65. Kingman JFC (1968) The ergodic theory of subadditive stochastic processes. *J Royal Stat Soc B30*:499–510
66. Kleinbock D, Margulis G (1998) Flows on Homogeneous Spaces and Diophantine Approximation on Manifold. *Ann Math* 148:339–360
67. Kra B, Schmeling J (2002) Diophantine classes, dimension and Denjoy maps. *Acta Arith* 105(4):323–340
68. Ledrappier F (1981) Some Relations Between Dimension And Lyapounov Exponents. *Comm Math Phys* 81:229–238
69. Ledrappier F (1986) Dimension of invariant measures. *Proceedings of the conference on ergodic theory and related topics II* (Georgenthal, 1986). *Teubner-texte Math* 94:137–173
70. Ledrappier F, Misiurewicz M (1985) Dimension of Invariant Measures for Maps with Exponent Zero. *Ergod Th Dynam Syst* 5:595–610
71. Ledrappier F, Strelcyn JM (1982) A proof of the estimate from below in Pesin's entropy formula. *Ergod Theory Dynam Syst* 2:203–219
72. Ledrappier F, Young LS (1985) The Metric Entropy Of Diffeomorphisms. I Characterization Of Measures Satisfying Pesin's Entropy Formula. *Ann Math* 122:509–539
73. Ledrappier F, Young LS (1985) The Metric Entropy Of Diffeomorphisms, II. Relations Between Entropy, Exponents and Dimension. *Ann Math* 122:540–574
74. Lopes A (1989) The Dimension Spectrum of the Maximal Measure. *SIAM J Math Analysis* 20:1243–1254
75. Mauldin RD, Urbański M (1996) Dimensions and measures in infinite iterated function systems. *Proc Lond Math Soc* 73(1):105–154
76. Mauldin RD, Urbański M (2002) Fractal measures for parabolic IFS. *Adv Math* 168(2):225–253
77. Mauldin DR, Urbański M (2003) *Graph directed Markov systems. Geometry and dynamics of limit sets* Cambridge Tracts in Mathematics, 148. Cambridge University Press, Cambridge

78. Mauldin RD; Urbański M (2000) Parabolic iterated function systems. *Ergod Theory Dynam Syst* 20(5):1423–1447
79. McCluskey H, Manning A (1983) Hausdorff Dimension For Horseshoes. *Erg Th Dyn Syst* 3:251–260
80. Moran P (1946) Additive Functions Of Intervals and Hausdorff Dimension. *Proceedings Of Cambridge Philosophical Society* 42:15–23
81. Moreira C, Yoccoz J (2001) Stable Intersections of Cantor Sets with Large Hausdorff Dimension. *Ann of Math* (2) 154(1):45–96
82. Mănă R (1981) On the Dimension of Compact Invariant Sets for Certain Nonlinear Maps. *Lecture Notes in Mathematics*, vol 898. Springer
83. Mănă R (1990) The Hausdorff Dimension of Horseshoes of Surfaces. *Bol Soc Bras Math* 20:1–24
84. Neunhäuserer J (1999) An analysis of dimensional theoretical properties of some affine dynamical systems. Thesis. Free University Berlin, Berlin
85. Oseledets V (1968) A multiplicative ergodic theorem. Liapunov characteristic numbers for dynamical systems. *Trans Moscow Math Soc* 19:197–221
86. Ott E, Sauer T, Yorke J (1994) Part I Background. In: *Coping with chaos*. Wiley Ser Nonlinear Sci, Wiley, New York, pp 1–62
87. Palis J, Takens F (1987) Hyperbolicity And The Creation Of Homoclinic Orbits. *Ann Math* 125:337–374
88. Palis J, Takens F (1993) Hyperbolicity And Sensitive Chaotic Dynamics At Homoclinic Bifurcations. Cambridge U Press, Cambridge
89. Palis J, Takens F (1994) Homoclinic Tangencies For Hyperbolic Sets Of Large Hausdorff Dimension. *Acta Math* 172:91–136
90. Palis J, Viana M (1988) On the continuity of Hausdorff dimension and limit capacity for horseshoes. *Lecture Notes in Math*, vol 1331. Springer
91. Pesin Y (1977) Characteristic exponents and smooth ergodic theory. *Russian Math Surveys* 32(4):55–114
92. Pesin Y (1992) Dynamical systems with generalized hyperbolic attractors: hyperbolic, ergodic and topological properties. *Erg Th Dyn Syst* 12:123–151
93. Pesin Y (1993) On Rigorous Mathematical Definition of Correlation Dimension and Generalized Spectrum for Dimensions. *J Statist Phys* 71(3–4):529–547
94. Pesin Y (1997) Dimension Theory In Dynamical Systems: Rigorous Results And Applications. Cambridge U Press, Cambridge
95. Pesin Y (1997) Dimension theory in dynamical systems: contemporary views and applications. In: *Chicago Lectures in Mathematics*. Chicago University Press, Chicago
96. Pesin Y, Pitskel' B (1984) Topological pressure and the variational principle for noncompact sets. *Funct Anal Appl* 18:307–318
97. Pesin Y, Sadovskaya V (2001) Multifractal analysis of conformal axiom A flows. *Comm Math Phys* 216(2):277–312
98. Pesin Y, Tempelman A (1995) Correlation Dimension of Measures Invariant Under Group Actions. *Random Comput Dyn* 3(3):137–156
99. Pesin Y, Weiss H (1994) On the Dimension of Deterministic and Random Cantor-like Sets. *Math Res Lett* 1:519–529
100. Pesin Y, Weiss H (1996) On The Dimension Of Deterministic and Random Cantor-like Sets, Symbolic Dynamics, And The Eckmann-Ruelle Conjecture. *Comm Math Phys* 182:105–153
101. Pesin Y, Weiss H (1997) A Multifractal Analysis of Equilibrium Measures For Conformal Expanding Maps and Moran-like Geometric Constructions. *J Stat Phys* 86:233–275
102. Pesin Y, Weiss H (1997) The Multifractal Analysis of Gibbs Measures: Motivation. *Mathematical Foundation and Examples*. *Chaos* 7:89–106
103. Petersen K (1983) *Ergodic theory*. Cambridge studies in advanced mathematics 2. Cambridge Univ Press, Cambridge
104. Pollicott M, Weiss H (1994) The Dimensions Of Some Self Affine Limit Sets In The Plane And Hyperbolic Sets. *J Stat Phys* 77:841–866
105. Pollicott M, Weiss H (1999) Multifractal Analysis for the Continued Fraction and Manneville-Pomeau Transformations and Applications to Diophantine Approximation. *Comm Math Phys* 207(1):145–171
106. Przytycki F, Urbański M (1989) On Hausdorff Dimension of Some Fractal Sets. *Studia Math* 93:155–167
107. Ruelle D (1978) *Thermodynamic Formalism*. Addison-Wesley
108. Ruelle D (1982) Repellers For Real Analytic Maps. *Erg Th Dyn Syst* 2:99–107
109. Schmeling J (1994) Hölder Continuity of the Holonomy Maps for Hyperbolic basic Sets II. *Math Nachr* 170:211–225
110. Schmeling J (1997) Symbolic Dynamics for Beta-shifts and Self-Normal Numbers. *Erg Th Dyn Syst* 17:675–694
111. Schmeling J (1998) A dimension formula for endomorphisms – the Belykh family. *Erg Th Dyn Syst* 18:1283–1309
112. Schmeling J (1999) On the Completeness of Multifractal Spectra. *Erg Th Dyn Syst* 19:1–22
113. Schmeling J (2001) Entropy Preservation under Markov Codings. *J Stat Phys* 104(3–4):799–815
114. Schmeling J, Troubetzkoy S (1998) Dimension and invertibility of hyperbolic endomorphisms with singularities. *Erg Th Dyn Syst* 18:1257–1282
115. Schmeling J, Weiss H (2001) Dimension theory and dynamics. *AMS Proceedings of Symposia in Pure Mathematics series* 69:429–488
116. Series C (1985) The Modular Surface and Continued Fractions. *J Lond Math Soc* 31:69–80
117. Simon K (1997) The Hausdorff dimension of the general Smale–Williams solenoid with different contraction coefficients. *Proc Am Math Soc* 125:1221–1228
118. Simpelaere D (1994) Dimension Spectrum of Axiom-A Diffeomorphisms, II. Gibbs Measures. *J Stat Phys* 76:1359–1375
119. Solomyak B (1995) On the random series $\sum \pm \lambda^n$ (an Erdős problem). *Ann of Math* (2) 142(3):611–625
120. Solomyak B (2004) Notes on Bernoulli convolutions. In: *Fractal geometry and applications: a jubilee of Benoît Mandelbrot*. *Proc Sympos Pure Math*, 72, Part 1. Amer Math Soc, Providence, pp 207–230
121. Stratmann B (1995) Fractal Dimensions for Jarnik Limit Sets of Geometrically Finite Kleinian Groups; The Semi-Classical Approach. *Ark Mat* 33:385–403
122. Takens F (1981) Detecting Strange Attractors in Turbulence. *Lecture Notes in Mathematics*, vol 898. Springer
123. Takens F, Verbitski E (1999) Multifractal analysis of local entropies for expansive homeomorphisms with specification. *Comm Math Phys* 203:593–612
124. Walters P (1982) *Introduction to Ergodic Theory*. Springer
125. Weiss H (1992) Some Variational Formulas for Hausdorff Dimension, Topological Entropy, and SRB Entropy for Hyperbolic Dynamical System. *J Stat Phys* 69:879–886

126. Weiss H (1999) The Lyapunov Spectrum Of Equilibrium Measures for Conformal Expanding Maps and Axiom-A Surface Diffeomorphisms. *J Statist Phys* 95(3–4):615–632
127. Young LS (1981) Capacity of Attractors. *Erg Th Dyn Syst* 1:381–388
128. Young LS (1982) Dimension, Entropy, and Lyapunov Exponents. *Erg Th Dyn Syst* 2:109–124

Books and Reviews

- Bowen R (1975) Equilibrium states and the ergodic theory of Anosov diffeomorphisms. *Lecture Notes in Mathematics*, vol 470. Springer
- Eckmann JP, Ruelle D (1985) Ergodic Theory Of Chaos And Strange Attractors. *Rev Mod Phys* 57:617–656
- Federer H (1969) Geometric measure theory. Springer
- Hasselblatt B, Katok A (2002) Handbook of Dynamical Systems, vol 1, Survey 1. Principal Structures. Elsevier
- Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst Hautes Études Sci Publ Math* 51:137–173
- Katok A, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ Press, Cambridge
- Keller G (1998) Equilibrium states in ergodic theory. In: London Mathematical Society Student Texts 42. Cambridge University Press, Cambridge
- Mañé R (1987) Ergodic theory and differentiable dynamics. In: *Ergebnisse der Mathematik und ihrer Grenzgebiete 3, Band 8*. Springer
- Mario R, Urbański M (2005) Regularity properties of Hausdorff dimension in infinite conformal iterated function systems. *Ergod Theory Dynam Syst* 25(6):1961–1983
- Mattila P (1995) Geometry of sets and measures in Euclidean spaces. In: *Fractals and rectifiability*. Cambridge University Press, Cambridge
- Pugh C, Shub M (1989) Ergodic attractors. *Trans Amer Math Soc* 312(1):1–54
- Takens F (1988) Limit capacity and Hausdorff dimension of dynamically defined Cantor sets. *Lecture Notes in Math*, vol 1331. Springer

Ergodic Theory on Homogeneous Spaces and Metric Number Theory

DMITRY KLEINBOCK
Department of Mathematics, Brandeis University,
Waltham, USA

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Basic Facts](#)
[Connection with Dynamics on the Space of Lattices](#)

[Diophantine Approximation](#)
[with Dependent Quantities: The Set-Up](#)
[Further Results](#)
[Future Directions](#)
[Acknowledgment](#)
[Bibliography](#)

Glossary

Diophantine approximation Diophantine approximation refers to approximation of real numbers by rational numbers, or more generally, finding integer points at which some (possibly vector-valued) functions attain values close to integers.

Metric number theory Metric number theory (or, specifically, metric Diophantine approximation) refers to the study of sets of real numbers or vectors with prescribed Diophantine approximation properties.

Homogeneous spaces A homogeneous space G/Γ of a group G by its subgroup Γ is the space of cosets $\{g\Gamma\}$. When G is a Lie group and Γ is a discrete subgroup, the space G/Γ is a smooth manifold and locally looks like G itself.

Lattice; unimodular lattice A lattice in a Lie group is a discrete subgroup of finite covolume; unimodular stands for covolume equal to 1.

Ergodic theory The study of statistical properties of orbits in abstract models of dynamical systems.

Hausdorff dimension A nonnegative number attached to a metric space and extending the notion of topological dimension of “sufficiently regular” sets, such as smooth submanifolds of real Euclidean spaces.

Definition of the Subject

The theory of Diophantine approximation, named after Diophantus of Alexandria, in its simplest set-up deals with the approximation of real numbers by rational numbers. Various higher-dimensional generalizations involve studying values of linear or polynomial maps at integer points. Often a certain “approximation property” is fixed, and one wants to characterize the set of numbers (vectors, matrices) which share this property, by means of certain measures (Lebesgue, or Hausdorff, or some other interesting measures). This is usually referred to as *metric* Diophantine approximation.

The starting point for the theory is an elementary fact that \mathbb{Q} , the set of rational numbers, is dense in \mathbb{R} , the reals. In other words, every real number can be approximated by rationals: for any $y \in \mathbb{R}$ and any $\varepsilon > 0$ there ex-

ists $p/q \in \mathbb{Q}$ with

$$|y - p/q| < \varepsilon. \quad (1)$$

To answer questions like “how well can various real numbers be approximated by rational numbers? i. e., how small can ε in (1) be chosen for varying $p/q \in \mathbb{Q}$?”, a natural approach has been to compare the accuracy of the approximation of y by p/q to the “complexity” of the latter, which can be measured by the size of its denominator q in its reduced form. This seemingly simple set-up has led to introducing many important Diophantine approximation properties of numbers/vectors/matrices, which show up in various fields of mathematics and physics, such as differential equations, KAM theory, transcendental number theory.

Introduction

As the first example of refining the statement about the density of \mathbb{Q} in \mathbb{R} , consider a theorem by Kronecker stating that for any $y \in \mathbb{R}$ and any $c > 0$, there exist infinitely many $q \in \mathbb{Z}$ such that

$$|y - p/q| < c/|q| \quad \text{i. e. } |qy - p| < c \quad (2)$$

for some $p \in \mathbb{Z}$. A comparison of (1) and (2) shows that it makes sense to multiply both sides of (1) by q , since in the right hand side of (2) one would still be able to get very small numbers. In other words, approximation of y by p/q translates into approximating integers by integer multiples of y .

Also, if y is irrational, (p, q) can be chosen to be relatively prime, i. e. one gets infinitely many different rational numbers p/q satisfying (2). However if $y \in \mathbb{Q}$ the latter is no longer true for small enough c . Thus it seems to be more convenient to talk about pairs (p, q) rather than $p/q \in \mathbb{Q}$, avoiding a necessity to consider the two cases separately.

At this point it is convenient to introduce the following central definition: if ψ is a function $\mathbb{N} \rightarrow \mathbb{R}_+$ and $y \in \mathbb{R}$, say that y is ψ -approximable (notation: $y \in \mathcal{W}(\psi)$) if there exist infinitely many $q \in \mathbb{N}$ such that

$$|qy - p| < \psi(q) \quad (3)$$

for some $p \in \mathbb{Z}$. Because of Kronecker’s Theorem, it is natural to assume that $\psi(x) \rightarrow 0$ as $x \rightarrow \infty$. Often ψ will be assumed non-increasing, although many results do not require monotonicity of ψ .

One can similarly consider a higher-dimensional version of the above set-up. Note that $y \in \mathbb{R}$ in the above formulas plays the role of a linear map from \mathbb{R} to another copy of \mathbb{R} , and one asks how close values of this map

at integers are from integers. It is natural to generalize it by taking a linear operator Y from \mathbb{R}^n to \mathbb{R}^m for fixed $m, n \in \mathbb{N}$, that is, an $m \times n$ -matrix (interpreted as a system of m linear forms Y_i on \mathbb{R}^n). We will denote by $M_{m,n}$ the space of $m \times n$ matrices with real coefficients. For ψ as above, one says that $Y \in M_{m,n}$ is ψ -approximable (notation: $Y \in \mathcal{W}_{m,n}(\psi)$) if there are infinitely many $\mathbf{q} \in \mathbb{Z}^n$ such that

$$\|Y\mathbf{q} + \mathbf{p}\| \leq \psi(\|\mathbf{q}\|) \quad (4)$$

for some $\mathbf{p} \in \mathbb{Z}^m$. Here $\|\cdot\|$ is the supremum norm on \mathbb{R}^k given by $\|\mathbf{y}\| = \max_{1 \leq i \leq k} |y_i|$. (This definition is slightly different from the one used in [58], where powers of norms were considered).

Traditionally, one of the main goals of metric Diophantine approximation has been to understand how big the sets $\mathcal{W}_{m,n}(\psi)$ are for fixed m, n and various functions ψ . Of course, (4) is not the only interesting condition that can be studied; various modifications of the approximation properties can also be considered. For example the Oppenheim Conjecture, now a theorem of Margulis [69] and a basis for many important recent developments [22,34,35], states that indefinite irrational quadratic forms can take arbitrary small values at integer points; Littlewood’s conjecture, see (18) below, deals with a similar statement about products of linear forms. See the article [► Ergodic Theory: Rigidity](#) by Nitica and surveys [32,70] for details.

We remark that the standard tool for studying Diophantine approximation properties of real numbers ($m = n = 1$) is the continued fraction expansion, or, equivalently, the Gauss map $x \mapsto 1/x \bmod 1$ of the unit interval, see [49]. However the emphasis of this survey lies in higher-dimensional theory, and the dynamical system described below can be thought of as a replacement for the continued fraction technique applicable in the one-dimensional case. Additional details about interactions between ergodic theory and number theory can be found in the article by Nitica mentioned above, in [► Ergodic Theory: Recurrence](#) by Frantzikinakis and McCutcheon and [► Ergodic Theory: Interactions with Combinatorics and Number Theory](#) by Ward, as well as in the survey papers [32,33,50,58,66,70,71].

Here is a brief outline of the rest of the article. In the next section we survey basic results, some classical, some obtained relatively recently, in metric Diophantine approximation. Sect. [“Connection with Dynamics on the Space of Lattices”](#) is devoted to a description of the connection between Diophantine approximation and dynamics, specifically flows on the space of

lattices. In Sect. “[Diophantine Approximation with Dependent Quantities: The Set-Up](#)” and Sect. “[Further Results](#)”, we specialize to the set-up of Diophantine approximation on manifolds, or, more generally, approximation properties of vectors with respect to measures satisfying some natural conditions, and show how applications of homogeneous dynamics contributed to important recent developments in the field. Sect. “[Future Directions](#)” mentions several open questions and directions for further investigation.

Basic Facts

General references for this section: [17,80].

The simplest choice for functions ψ happens to be the following: let us denote $\psi_{c,v}(x) = cx^{-v}$. It was shown by Dirichlet in 1842 that with the choice $c = 1$ and $v = n/m$, all $Y \in M_{m,n}$ are ψ -approximable. Moreover, Dirichlet’s Theorem states that for any $Y \in M_{m,n}$ and for any $t > 0$ there exist $\mathbf{q} = (q_1, \dots, q_n) \in \mathbb{Z}^n \setminus \{0\}$ and $\mathbf{p} = (p_1, \dots, p_m) \in \mathbb{Z}^m$ satisfying the following system of inequalities:

$$\|Y\mathbf{q} - \mathbf{p}\| < e^{-t/m} \quad \text{and} \quad \|\mathbf{q}\| \leq e^{t/n}. \quad (5)$$

From this it easily follows that $\mathcal{W}_{m,n}(\psi_{1,n/m}) = M_{m,n}$. In fact, it is this paper of Dirichlet which gave rise to his box principle. Later another proof of the same result was given by Minkowski. The constant $c = 1$ is not optimal: the smallest value of c for which $\mathcal{W}_{1,1}(\psi_{c,1}) = \mathbb{R}$ is $1/\sqrt{5}$, and the optimal constants are not known in higher dimensions, although some estimates can be given [80].

Systems of linear forms which do not belong to $\mathcal{W}_{m,n}(\psi_{c,n/m})$ for some positive c are called *badly approximable*; that is, we set

$$\text{BA}_{m,n} \stackrel{\text{def}}{=} M_{m,n} \setminus \bigcup_{c>0} \mathcal{W}_{m,n}(\psi_{c,n/m}).$$

Their existence in arbitrary dimensions was shown by Peron. Note that a real number y ($m = n = 1$) is badly approximable if and only if its continued fraction coefficients are uniformly bounded. It was proved by Jarník [46] in the case $m = n = 1$ and by Schmidt in the general case [78] that badly approximable matrices form a set of full Hausdorff dimension: that is, $\dim(\text{BA}_{m,n}) = mn$.

On the other hand, it can be shown that each of the sets $\mathcal{W}_{m,n}(\psi_{c,n/m})$ for any $c > 0$ has full Lebesgue measure, and hence the complement $\text{BA}_{m,n}$ to their intersection has measure zero. This is a special case of a theorem due to Khintchine [48] in the case $n = 1$ and to Groshev [42] in full generality, which gives the precise condition on the function ψ under which the set of ψ -approximable matrices has full measure. Namely, if ψ is non-increasing (this

assumption can be removed in higher dimensions but not for $n = 1$, see [29]), then λ -almost no (resp. λ -almost every) $Y \in M_{m,n}$ is ψ -approximable, provided the sum

$$\sum_{k=1}^{\infty} k^{n-1} \psi(k)^m \quad (6)$$

converges (resp. diverges). (Here and hereafter λ stands for Lebesgue measure). This statement is usually referred to as the Khintchine–Groshev Theorem. The convergence case of this theorem follows in a straightforward manner from the Borel–Cantelli Lemma, but the divergence case is harder. It was reproved and sharpened in 1960 by Schmidt [76], who showed that if the sum (6) diverges, then for almost all Y the number of solutions to (4) with $\|\mathbf{q}\| \leq N$ is asymptotic to the partial sum of the series (6) (up to a constant), and also gave an estimate for the error term.

A special case of the convergence part of the theorem shows that $\mathcal{W}_{m,n}(\psi_{1,v})$ has measure zero whenever $v > n/m$. Y is said to be *very well approximable* if it belongs to $\mathcal{W}_{m,n}(\psi_{1,v})$ for some $v > n/m$. That is,

$$\text{VWA}_{m,n} \stackrel{\text{def}}{=} \bigcup_{v>n/m} \mathcal{W}_{m,n}(\psi_{1,v}).$$

More specifically, let us define the *Diophantine exponent* $\omega(Y)$ of Y (sometimes called “the exact order” of Y) to be the supremum of $v > 0$ for which $Y \in \mathcal{W}_{m,n}(\psi_{1,v})$. Then $\omega(Y)$ is always not less than n/m , and is equal to n/m for Lebesgue-a.e. Y ; in fact, $\text{VWA}_{m,n} = \{Y \in M_{m,n} : \omega(Y) > n/m\}$.

The Hausdorff dimension of the null sets $\mathcal{W}_{m,n}(\psi_{1,v})$ was computed independently by Besicovitch [14] and Jarník [45] in the one-dimensional case and by Dodson [26] in general: when $v > n/m$, one has

$$\dim(\mathcal{W}_{m,n}(\psi_{1,v})) = (n-1)m + \frac{m+n}{v+1}. \quad (7)$$

See [27] for a nice exposition of ideas involved in the proof of both the aforementioned formula and the Khintchine–Groshev Theorem.

Note that it follows from (7) that the null set $\text{VWA}_{m,n}$ has full Hausdorff dimension. Matrices contained in the intersection

$$\bigcap_v \mathcal{W}_{m,n}(\psi_{1,v}) = \{Y \in M_{m,n} : \omega(Y) = \infty\}$$

are called *Liouville* and form a set of Hausdorff dimension $(n-1)m$, that is, to the dimension of Y for which $Y\mathbf{q} \in \mathbb{Z}$ for some $\mathbf{q} \in \mathbb{Z}^n \setminus \{0\}$ (the latter belong to $\mathcal{W}_{m,n}(\psi)$ for any positive ψ).

Note also that the aforementioned properties behave nicely with respect to transposition; this is described by the so-called Khintchine's Transference Principle (Chap. V in [17]). For example, $Y \in \text{BA}_{m,n}$ if and only if $Y^T \in \text{BA}_{n,m}$, and $Y \in \text{VWA}_{m,n}$ if and only if $Y^T \in \text{VWA}_{n,m}$. In particular, many problems related to approximation properties of vectors ($n = 1$) and linear forms ($m = 1$) reduce to one another.

We refer the readers to [43] and [8] for very detailed and comprehensive recent accounts of various further aspects of the theory.

Connection with Dynamics on the Space of Lattices

General references for this section: [4,86].

Interactions between Diophantine approximation and the theory of dynamical systems has a long history. Already in Kronecker's Theorem one can see a connection. Indeed, the statement of the theorem can be rephrased as follows: the points on the orbit of 0 under the rotation of the circle \mathbb{R}/\mathbb{Z} by y approach the initial point 0 arbitrarily closely. This is a special case of the Poincaré Recurrence Theorem in measurable dynamics. And, likewise, all the aforementioned properties of $Y \in M_{m,n}$ can be restated in terms of recurrence properties of the \mathbb{Z}^n -action on the m -dimensional torus $\mathbb{R}^m/\mathbb{Z}^m$ given by $x \mapsto Yx \bmod \mathbb{Z}^m$. In other words, fixing Y gives rise to a dynamical system in which approximation properties of Y show up.

However the theme of this section is a different dynamical system, whose phase space is (essentially) the space of parameters Y , and which can be used to read the properties of Y from the behavior of the associated trajectory.

It has been known for a long time (see [81] for a historical account) that Diophantine properties of real numbers can be coded by the behavior of geodesics on the quotient of the hyperbolic plane by $\text{SL}_2(\mathbb{Z})$. In fact, the latter flow can be viewed as the suspension flow of the Gauss map mentioned at the end of Sect. "Introduction". There have been many attempts to construct a higher-dimensional analogue of the Gauss map so that it captures all the features of simultaneous approximation, see [47,63,65] and references therein. On the other hand, it seems to be more natural and efficient to generalize the suspension flow itself, and this is where one needs higher rank homogeneous dynamics.

As was mentioned above, in the basic set-up of simultaneous Diophantine approximation one takes a system of m linear forms Y_1, \dots, Y_m on \mathbb{R}^n and looks at the values of $|Y_i(\mathbf{q}) + p_i|$, $p_i \in \mathbb{Z}$, when $\mathbf{q} = (q_1, \dots, q_n) \in \mathbb{Z}^n$

is far from 0. The trick is to put together

$$Y_1(\mathbf{q}) + p_1, \dots, Y_m(\mathbf{q}) + p_m \quad \text{and} \quad q_1, \dots, q_n,$$

and consider the collection of vectors

$$\left\{ \begin{pmatrix} Y\mathbf{q} + \mathbf{p} \\ \mathbf{q} \end{pmatrix} \mid \mathbf{p} \in \mathbb{Z}^m, \mathbf{q} \in \mathbb{Z}^n \right\} = L_Y \mathbb{Z}^k$$

where $k = m + n$ and

$$L_Y \stackrel{\text{def}}{=} \begin{pmatrix} I_m & Y \\ 0 & I_n \end{pmatrix}, \quad Y \in M_{m,n}. \quad (8)$$

This collection is a unimodular lattice in \mathbb{R}^k , that is, a discrete subgroup of \mathbb{R}^k with covolume 1. Our goal is to keep track of vectors in such a lattice having small projections onto the first m components of \mathbb{R}^k and big projections onto the last n components. This is where dynamics comes into the picture. Denote by g_t the one-parameter subgroup of $\text{SL}_k(\mathbb{R})$ given by

$$g_t = \text{diag}(\underbrace{e^{t/m}, \dots, e^{t/m}}_{m \text{ times}}, \underbrace{e^{-t/n}, \dots, e^{-t/n}}_{n \text{ times}}). \quad (9)$$

The vectors in the lattice $L_Y \mathbb{Z}^k$ are moved by the action of g_t , $t > 0$, and a special role is played by the moment t when the "small" and "big" projections equalize.

That is, one is led to consider a new dynamical system. Its phase space is the space of unimodular lattices in \mathbb{R}^k , which can be naturally identified with the homogeneous space

$$\Omega_k \stackrel{\text{def}}{=} G/\Gamma, \quad \text{where } G = \text{SL}_k(\mathbb{R}) \text{ and } \Gamma = \text{SL}_k(\mathbb{Z}), \quad (10)$$

and the action is given by left multiplication by elements of the subgroup (9) of G , or perhaps other subgroups $H \subset G$. Study of such systems has a rich history; for example, they are known to be ergodic and mixing whenever H is unbounded [74]. What is important in this particular case is that the space Ω_k happens to be noncompact, and its structure at infinity is described via Mahler's Compactness Criterion, see Chap. V in [4]: a sequence of lattices $g_i \mathbb{Z}^k$ goes to infinity in $\Omega_k \iff$ there exists a sequence $\{\mathbf{v}_i \in \mathbb{Z}^k \setminus \{0\}\}$ such that $g_i(\mathbf{v}_i) \rightarrow 0$ as $i \rightarrow \infty$. Equivalently, for $\varepsilon > 0$ consider a subset K_ε of Ω_k consisting of lattices with no nonzero vectors of norm less than ε ; then all the sets K_ε are compact, and every compact subset of Ω_k is contained in one of them. Moreover, one can choose a metric on Ω_k such that $\text{dist}(A, \mathbb{Z}^k)$ is, up to a uniform multiplicative constant,

equal to $-\log \min_{v \in \Lambda \setminus \{0\}} \|v\|$ (see [25]); then the length of the smallest nonzero vector in a lattice Λ will determine how far away is this lattice in the “cusp” of Ω_k .

Using Mahler’s Criterion, it is not hard to show that $Y \in \text{BA}_{m,n}$ if and only if the trajectory

$$\{g_t L_Y \mathbb{Z}^k : t \in \mathbb{R}_+\} \quad (11)$$

is bounded in Ω_k . This was proved by Dani [20] in 1985, and later generalized in [57] to produce a criterion for Y to be ψ -approximable for any non-increasing function ψ . An important special case is a criterion for a system of linear forms to be very well approximable: $Y \in \text{VWA}_{m,n}$ if and only if the trajectory (11) has linear growth, that is, there exists a positive γ such that $\text{dist}(g_t L_Y \mathbb{Z}^k, \mathbb{Z}^k) > \gamma t$ for an unbounded set of $t > 0$.

This correspondence allows one to link various Diophantine and dynamical phenomena. For example, from the results of [55] on abundance of bounded orbits on homogeneous spaces one can deduce the aforementioned theorem of Schmidt [78]: the set $\text{BA}_{m,n}$ has full Hausdorff dimension. And a dynamical Borel–Cantelli Lemma established in [57] can be used for an alternative proof of the Khintchine–Groshev Theorem; see also [87] for an earlier geometric approach. Note that both proofs are based on the following two properties of the g_t -action: mixing, which forces points to return to compact subsets and makes preimages of cusp neighborhoods quasi-independent, and hyperbolicity, which implies that the behavior of points on unstable leaves is generic. The latter is important since the orbits of the group $\{L_Y \mathbb{Z}^k : Y \in M_{m,n}\}$ are precisely the unstable leaves with respect to the g_t -action.

We note that other types of Diophantine problems, such as conjectures of Oppenheim and Littlewood mentioned in the previous section, can be reduced to statements involving Ω_k by means of the same principle: Mahler’s Criterion is used to relate small values of some function at integer points to excursions to infinity in Ω_k of orbit of the stabilizer of this function.

Other important and useful recent applications of homogeneous dynamics to metric Diophantine approximation are related to the circle of ideas roughly called “Diophantine approximation with dependent quantities” (terminology borrowed from [84]), to be surveyed in the next two sections.

Diophantine Approximation with Dependent Quantities: The Set-Up

General references for this section: [12,84].

Here we restrict ourselves to Diophantine properties of vectors in \mathbb{R}^n . In particular, we will look more closely

at the set of very well approximable vectors, which we will simply denote by VWA, dropping the subscripts. In many cases it does not matter whether one works with row or column vectors, in view of the duality remark made at the end of Sect. “Basic Facts”.

We begin with a non-example of an application of dynamics to Diophantine approximation: a celebrated and difficult theorem which currently, to the best of the author’s knowledge, has no dynamical proof. Suppose that $y = (y_1, \dots, y_n) \in \mathbb{R}^n$ is such that each y_i is algebraic and $1, y_1, \dots, y_n$ are linearly independent over \mathbb{Q} . It was established by Roth for $n = 1$ [75] and then generalized to arbitrary n by Schmidt [79], that y as above necessarily belongs to the complement of VWA. In other words, vectors with very special algebraic properties happen to follow the behavior of a generic vector in \mathbb{R}^n .

We would like to view the above example as a special case of a general class of problems. Namely, suppose we are given a Radon measure μ on \mathbb{R}^n . Let us say that μ is *extremal* [85] if μ -a.e. $y \in \mathbb{R}^n$ is not very well approximable. Further, define the *Diophantine exponent* $\omega(\mu)$ of μ to be the μ -essential supremum of the function $\omega(\cdot)$; in other words,

$$\omega(\mu) \stackrel{\text{def}}{=} \sup \{v | \mu(\mathcal{W}(\psi_{1,v})) > 0\}.$$

Clearly it only depends on the measure class of μ . If μ is naturally associated with a subset \mathcal{M} of \mathbb{R}^n supporting μ (for example, if \mathcal{M} is a smooth submanifold of \mathbb{R}^n and μ is the measure class of the Riemannian volume on \mathcal{M} , or, equivalently, the pushforward $\mathbf{f}_* \lambda$ of λ by a smooth map \mathbf{f} parametrizing \mathcal{M}), one defines the Diophantine exponent $\omega(\mathcal{M})$ of \mathcal{M} to be equal to that of μ , and says that \mathcal{M} is extremal if $\mathbf{f}(\mathbf{x})$ is not very well approximable for λ -a.e. \mathbf{x} .

Then $\omega(\mu) \geq n$ for any μ , and $\omega(\lambda) = \omega(\mathbb{R}^n)$ is equal to n . The latter justifies the use of the word “extremal”: μ is *extremal* if $\omega(\mu)$ is equal to n , i.e. attains the smallest possible value. The aforementioned results of Roth and Schmidt then can be interpreted as the extremality of atomic measures supported on algebraic vectors without rational dependence relations.

Historically, the first measure (other than λ) to be considered in the set-up described above was the pushforward of λ by the map

$$\mathbf{f}(x) = (x, x^2, \dots, x^n). \quad (12)$$

The extremality of $\mathbf{f}_* \lambda$ for \mathbf{f} as above was conjectured in 1932 by K. Mahler [67] and proved in 1964 by Sprindžuk [82,83]. It was important for Mahler’s study of transcendental numbers: this result, roughly speaking, says

that almost all transcendental numbers are “not very algebraic”. At about the same time Schmidt [77] proved the extremality of $\mathbf{f}_*\lambda$ when $\mathbf{f}: I \rightarrow \mathbb{R}^2$, $I \subset \mathbb{R}$, is C^3 and satisfies

$$\begin{vmatrix} f_1'(x) & f_2'(x) \\ f_1''(x) & f_2''(x) \end{vmatrix} \neq 0 \quad \text{for } \lambda\text{-a.e. } x \in I;$$

in other words, the curve parametrized by \mathbf{f} has nonzero curvature at almost all points. Since then, a lot of attention has been devoted to showing that measures $\mathbf{f}_*\lambda$ are extremal for other smooth maps \mathbf{f} .

To describe a broader class of examples, recall the following definition. Let $\mathbf{x} \in \mathbb{R}^d$ and let $\mathbf{f} = (f_1, \dots, f_n)$ be a C^k map from a neighborhood of \mathbf{x} to \mathbb{R}^n . Say that \mathbf{f} is *nondegenerate at \mathbf{x}* if \mathbb{R}^n is spanned by partial derivatives of \mathbf{f} at \mathbf{x} up to some order. Say that \mathbf{f} is *nondegenerate* if it is nondegenerate at λ -a.e. \mathbf{x} . It was conjectured by Sprindžuk [84] in 1980 that $\mathbf{f}_*\lambda$ for real analytic nondegenerate \mathbf{f} are extremal. Many special cases were established since then (see [12] for a detailed exposition of the theory and many related results), but the general case stood open until the mid-1990s [56], when Sprindžuk’s conjecture was proved using the dynamical approach (later Beresnevich [6] succeeded in establishing and extending this result without use of dynamics). The proof in [56] uses the correspondence outlined in the previous section plus a measure estimate for flows on the space of lattices which is described below.

In the subsequent work the method of [56] was adapted to a much broader class of measures. To define it we need to introduce some more notation and definitions. If $\mathbf{x} \in \mathbb{R}^d$ and $r > 0$, denote by $B(\mathbf{x}, r)$ the open ball of radius r centered at \mathbf{x} . If $B = B(\mathbf{x}, r)$ and $c > 0$, cB will denote the ball $B(\mathbf{x}, cr)$. For $B \subset \mathbb{R}^d$ and a real-valued function f on B , let

$$\|f\|_B \stackrel{\text{def}}{=} \sup_{\mathbf{x} \in B} |f(\mathbf{x})|.$$

If ν is a measure on \mathbb{R}^d such that $\nu(B) > 0$, define $\|f\|_{\nu, B} \stackrel{\text{def}}{=} \|f\|_{B \cap \text{supp } \nu}$; this is the same as the $L^\infty(\nu)$ -norm of $f|_B$ if f is continuous and B is open. If $D > 0$ and $U \subset \mathbb{R}^d$ is an open subset, let us say that ν is *D-Federer on U* if for any ball $B \subset U$ centered at $\text{supp } \nu$ one has $\frac{\nu(3B)}{\nu(B)} < D$ whenever $3B \subset U$. This condition is often called “doubling” in the literature. See [54, 72] for examples and references. ν is called *Federer* if for ν -a.e. $\mathbf{x} \in \mathbb{R}^d$ there exist a neighborhood U of \mathbf{x} and $D > 0$ such that ν is *D-Federer on U* .

Given $C, \alpha > 0$, open $U \subset \mathbb{R}^d$ and a measure ν on U , a function $f: U \rightarrow \mathbb{R}$ is called *(C, α)-good on U with respect to ν* if for any ball $B \subset U$ centered in $\text{supp } \nu$ and any

$\varepsilon > 0$ one has

$$\nu(\{\mathbf{x} \in B: |f(\mathbf{x})| < \varepsilon\}) \leq C \left(\frac{\varepsilon}{\|f\|_{\nu, B}} \right)^\alpha \nu(B). \quad (13)$$

This condition was formally introduced in [56] for ν being Lebesgue measure, and in [54] for arbitrary ν . A basic example is given by polynomials, and the upshot of the above definition is the formalization of a property needed for the proof of several basic facts [19, 21, 68] about polynomial maps into the space of lattices.

In [54] a strengthening of this property was considered: f was called *absolutely (C, α)-good on U with respect to ν* if for B and ε as above one has

$$\nu(\{\mathbf{x} \in B: |f(\mathbf{x})| < \varepsilon\}) \leq C \left(\frac{\varepsilon}{\|f\|_B} \right)^\alpha \nu(B). \quad (14)$$

There is no difference between (13) and (14) when ν has full support, but it turns out to be useful for describing measures supported on proper (e.g. fractal) subsets of \mathbb{R}^d .

Now suppose that we are given a measure ν on \mathbb{R}^d , an open $U \subset \mathbb{R}^d$ with $\nu(U) > 0$ and a map $\mathbf{f} = (f_1, \dots, f_n): \mathbb{R}^d \rightarrow \mathbb{R}^n$. Following [62], say that a pair (\mathbf{f}, ν) is *(absolutely) good on U* if any linear combination of $1, f_1, \dots, f_n$ is (absolutely) (C, α) -good on U with respect to ν . If for ν -a.e. \mathbf{X} there exists a neighborhood U of \mathbf{X} and $C, \alpha > 0$ such that ν is (absolutely) (C, α) -good on U , we will say that the pair (\mathbf{f}, ν) is *(absolutely) good*.

Another relevant notion is the nonplanarity of (\mathbf{f}, ν) . Namely, (\mathbf{f}, ν) is said to be *nonplanar* if whenever B is a ball with $\nu(B) > 0$, the restrictions of $1, f_1, \dots, f_n$ to $B \cap \text{supp } \nu$ are linearly independent over \mathbb{R} ; in other words, $\mathbf{f}(B \cap \text{supp } \nu)$ is not contained in any proper affine subspace of \mathbb{R}^n . Note that absolutely good implies both good and nonplanar, but the converse is in general not true.

Many examples of (absolutely) good and nonplanar pairs (\mathbf{f}, ν) can be found in the literature. Already the case $n = d$ and $\mathbf{f} = \text{Id}$ is very interesting. A measure μ on \mathbb{R}^n is said to be *friendly* (resp., *absolutely friendly*) if and only if it is Federer and the pair (Id, μ) is good and nonplanar (resp., absolutely good). See [54, 88, 89] for many examples. An important class of measures is given by limit measures of irreducible system of self-similar or self-conformal contractions satisfying the Open Set Condition [44]; those are shown to be absolutely friendly in [54]. The prime example is the middle-third Cantor set on the real line. The term “friendly” was cooked up as a loose abbreviation for “Federer, nonplanar and decaying”, and later proved to be particularly friendly in dealing with problems arising in metric number theory, see e.g. [36].

Also let us say that a pair (\mathbf{f}, ν) is *nondegenerate* if \mathbf{f} is nondegenerate at ν -a.e. \mathbf{X} . When ν is Lebesgue measure on \mathbb{R}^d , it is proved in Proposition 3.4 in [56], that a nondegenerate (\mathbf{f}, ν) is good and nonplanar. The same conclusion is derived in Proposition 7.3 in [54], assuming that ν is absolutely friendly. Thus volume measures on smooth nondegenerate manifolds are friendly, but not absolutely friendly.

It turns out that all the aforementioned examples of measures can be proved to be extremal by a generalization of the argument from [56]. Specifically, let ν be a Federer measure on \mathbb{R}^d , U an open subset of \mathbb{R}^d , and $\mathbf{f}: U \rightarrow \mathbb{R}^n$ a continuous map such that the pair (\mathbf{f}, ν) is good and nonplanar; then $\mathbf{f}_* \nu$ is extremal. This can be derived from the Borel–Cantelli Lemma, the correspondence described in the previous section, and the following measure estimate: if ν , U and \mathbf{f} are as above, then for ν -a.e. $\mathbf{x}_0 \in U$ there exists a ball $B \subset U$ centered at \mathbf{x}_0 and $\tilde{C}, \alpha > 0$ such that for any $t \in \mathbb{R}_+$ and any $\varepsilon > 0$,

$$\nu(\{\mathbf{x} \in B : g_t L_{\mathbf{f}(\mathbf{x})} \mathbb{Z}^{n+1} \notin K_\varepsilon\}) < \tilde{C} \varepsilon^\alpha. \quad (15)$$

Here g_t is as in (9) with $m = 1$ (assuming that the row vector viewpoint is adopted). This is a quantitative way of saying that for fixed t , the “flow” $\mathbf{x} \mapsto g_t L_{\mathbf{f}(\mathbf{x})} \mathbb{Z}^{n+1}$, $B \rightarrow \mathcal{O}_{n+1}$, cannot diverge, and in fact must spend a big (uniformly in t) proportion of time inside compact sets K_ε .

The inequality (15) is derived from a general “quantitative non-divergence” estimate, which can be thought of a substantial generalization of theorems of Margulis and Dani [19,21,68] on non-divergence of unipotent flows on homogeneous spaces. One of its most general versions [54] deals with a measure ν on \mathbb{R}^d , a continuous map $h: \tilde{B} \rightarrow G$, where \tilde{B} is a ball in \mathbb{R}^d centered at $\text{supp } \nu$ and G is as in (10). To describe the assumptions on h , one needs to employ the combinatorial structure of lattices in \mathbb{R}^k , and it will be convenient to use the following notation: if V is a nonzero rational subspace of \mathbb{R}^k and $g \in G$, define $\ell_V(g)$ to be the covolume of $g(V \cap \mathbb{Z}^k)$ in gV . Then, given positive constants C, D, α , there exists $C_1 = C_1(d, k, C, \alpha, D) > 0$ with the following property. Suppose ν is D -Federer on \tilde{B} , $0 < \rho \leq 1$, and h is such that for each rational $V \subset \mathbb{R}^k$

- (i) $\ell_V \circ h$ is (C, α) -good on \tilde{B} with respect to ν , and
- (ii) $\|\ell_V \circ h\|_{\nu, B} \geq \rho$, where $B = 3^{-(k-1)} \tilde{B}$. Then
- (iii) for any positive $\varepsilon \leq \rho$, one has

$$\nu(\{\mathbf{x} \in B : h(\mathbf{x}) \mathbb{Z}^k \notin K_\varepsilon\}) \leq C_1(\varepsilon \rho)^\alpha \nu(B). \quad (16)$$

Taking $h(\mathbf{x}) = g_t L_{\mathbf{f}(\mathbf{x})}$ and unwinding the definitions of good and nonplanar pairs, one can show that (i) and (ii)

can be verified for some balls B centered at ν -almost every point, and derive (15) from (16).

Further Results

The approach to metric Diophantine approximation using quantitative non-divergence, that is, the implication (i) + (ii) \Rightarrow (iii), is not omnipotent. In particular, it is difficult to use when more precise results are needed, such as for example computing/estimating the Hausdorff dimension of the set of $\psi_{1,\nu}$ -approximable vectors on a manifold. See [9,10] for such results. On the other hand, the dynamical approach can often treat much more general objects than its classical counterpart, and also can be perturbed in a lot of directions, producing many generalizations and modifications of the main theorems from the preceding section.

One of the most important of them is the so-called *multiplicative* version of the set-up of Sect. “Diophantine Approximation with Dependent Quantities: The Set-Up”. Namely, define functions $\Pi(\mathbf{x}) \stackrel{\text{def}}{=} \prod_i |x_i|$ and $\Pi_+(\mathbf{x}) \stackrel{\text{def}}{=} \prod_i \max(|x_i|, 1)$. Then, given a function $\psi: \mathbb{N} \rightarrow \mathbb{R}_+$, one says that $Y \in M_{m,n}$ is *multiplicatively ψ -approximable* (notation: $Y \in \mathcal{W}_{m,n}^\times(\psi)$) if there are infinitely many $\mathbf{q} \in \mathbb{Z}^n$ such that

$$\Pi(Y\mathbf{q} + \mathbf{p})^{1/m} \leq \psi(\Pi_+(\mathbf{q})^{1/n}) \quad (17)$$

for some $\mathbf{p} \in \mathbb{Z}^m$. Since $\Pi(\mathbf{x}) \leq \Pi_+(\mathbf{x}) \leq \|\mathbf{x}\|^k$ for $\mathbf{x} \in \mathbb{R}^k$, any ψ -approximable Y is multiplicatively ψ -approximable; but the converse is in general not true, see e.g. [37]. However if one, as before, considers the family $\{\psi_{1,\nu}\}$, the critical parameter for which the drop from full measure to measure zero occurs is again n/m . That is, if one defines the *multiplicative Diophantine exponent* $\omega^\times(Y)$ of Y by $\omega^\times(Y) \stackrel{\text{def}}{=} \sup\{\nu: Y \in \mathcal{W}_{m,n}^\times(\psi_{1,\nu})\}$, then clearly $\omega^\times(Y) \geq \omega(Y)$ for all Y , and yet $\omega^\times(Y) = n/m$ for λ -a.e. $Y \in M_{m,n}$.

Now specialize to \mathbb{R}^n (by the same duality principle as before, it does not matter whether to think in terms of row or column vectors, but we will adopt the row vector set-up), and define the *multiplicative exponent* $\omega^\times(\mu)$ of a measure μ on \mathbb{R}^n by $\omega^\times(\mu) \stackrel{\text{def}}{=} \sup\{\nu | \mu(\mathcal{W}^\times(\psi_{1,\nu})) > 0\}$; then $\omega^\times(\lambda) = n$. Following Sprindžuk [85], say that μ is *strongly extremal* if $\omega^\times(\mu) = n$. It turns out that all the results mentioned in the previous section have their multiplicative analogues; that is, the measures described there happen to be strongly extremal. This was conjectured by A. Baker [1] for the curve (12), and then by Sprindžuk in 1980 [85] for analytic nondegenerate manifolds. (We remark that only very few results in this set-up can be obtained by the standard methods, see e.g. [10]). The proof

of this stronger statement is based on using the multi-parameter action of

$$g_{\mathbf{t}} = \text{diag}(e^{t_1 + \dots + t_n}, e^{-t_1}, \dots, e^{-t_n}),$$

where $\mathbf{t} = (t_1, \dots, t_n)$

instead of g_t considered in the previous section. One can show that the choice $h(\mathbf{x}) = g_{\mathbf{t}} L_{\mathbf{f}(\mathbf{x})}$ allows one to verify (i) and (ii) uniformly in $\mathbf{t} \in \mathbb{R}_+^n$, and the proof is finished by applying a multi-parameter version of the correspondence described in Sect. “[Connection with Dynamics on the Space of Lattices](#)”. Namely, one can show that $\mathbf{y} \in \text{VWA}_{1,n}^\times$ if and only if the trajectory $\{g_{\mathbf{t}} L_{\mathbf{y}} \mathbb{Z}^k : \mathbf{t} \in \mathbb{R}_+^n\}$ grows linearly, that is, for some $\gamma > 0$ one has $\text{dist}(g_{\mathbf{t}} L_{\mathbf{y}} \mathbb{Z}^{n+1}, \mathbb{Z}^{n+1}) > \gamma \|\mathbf{t}\|$ for an unbounded set of $\mathbf{t} \in \mathbb{R}_+^n$. A similar correspondence was recently used in [30] to prove that the set of exceptions to Littlewood’s Conjecture, which, using the terminology introduced above, can be called *badly multiplicatively approximable* vectors:

$$\text{BA}_{n,1}^\times \stackrel{\text{def}}{=} \mathbb{R}^n \setminus \bigcup_{c>0} \mathcal{W}_{n,1}^\times(\psi_{c,1/n})$$

$$= \left\{ \mathbf{y} : \inf_{q \in \mathbb{Z} \setminus \{0\}, \mathbf{p} \in \mathbb{Z}^n} |q| \cdot \Pi(q\mathbf{y} - \mathbf{p}) > 0 \right\}, \quad (18)$$

has Hausdorff dimension zero. This was done using a measure rigidity result for the action of the group of diagonal matrices on the space of lattices. See [18] for an implicit description of this correspondence and [32,66,70] for more detail.

The dynamical approach also turned out to be fruitful in studying Diophantine properties of pairs (\mathbf{f}, ν) for which the nonplanarity condition fails. Note that obvious examples of non-extremal measures are provided by proper affine subspaces of \mathbb{R}^n whose coefficients are rational or are well enough approximable by rational numbers. On the other hand, it is clear from a Fubini argument that almost all translates of any given subspace are extremal. In [51] the method of [56] was pushed further to produce criteria for the extremality, as well as the strong extremality, of arbitrary affine subspaces \mathcal{L} of \mathbb{R}^n . Further, it was shown that if \mathcal{L} is extremal (resp. strongly extremal), then so is any smooth submanifold of \mathcal{L} which is nondegenerate in \mathcal{L} at a.e. point. (The latter property is a straightforward generalization of the definition of nondegeneracy in \mathbb{R}^n : a map \mathbf{f} is *nondegenerate in \mathcal{L} at \mathbf{x}* if the linear part of \mathcal{L} is spanned by partial derivatives of \mathbf{f} at \mathbf{x}). In other words, extremality and strong extremality pass from affine subspaces to their nondegenerate submanifolds.

A more precise analysis makes it possible to study Diophantine exponents of measures with supports contained

in arbitrary proper affine subspaces of \mathbb{R}^n . Namely, in [53] it is shown how to compute $\omega(\mathcal{L})$ for any \mathcal{L} , and furthermore proved that if ν is a Federer measure on \mathbb{R}^d , U an open subset of \mathbb{R}^d , and $\mathbf{f}: U \rightarrow \mathbb{R}^n$ a continuous map such that the pair (\mathbf{f}, ν) is good and nonplanar in \mathcal{L} , then $\omega(\mathbf{f}_* \nu) = \omega(\mathcal{L})$. Here we say, generalizing the definition from Sect. “[Diophantine Approximation with Dependent Quantities: The Set-Up](#)”, that (\mathbf{f}, ν) is *nonplanar in \mathcal{L}* if for any ball B with $\nu(B) > 0$, the \mathbf{f} -image of $B \cap \text{supp } \nu$ is not contained in any proper affine subspace of \mathcal{L} . (It is easy to see that for a smooth map $\mathbf{f}: U \rightarrow \mathcal{L}$, (\mathbf{f}, λ) is good and nonplanar in \mathcal{L} whenever \mathbf{f} is nondegenerate in \mathcal{L} at a.e. point). It is worthwhile to point out that these new applications require a strengthening of the measure estimate described at the end of Sect. “[Diophantine Approximation with Dependent Quantities: The Set-Up](#)”: it was shown in [53] that (i) and (ii) would still imply (iii) if ρ in (ii) is replaced by $\rho^{\dim V}$.

Another application concerns badly approximable vectors. Using the dynamical description of the set $\text{BA} \subset \mathbb{R}^n$ due to Dani [20], it turns out to be possible to find badly approximable vectors inside supports of certain measures on \mathbb{R}^n . Namely, if a subset K of \mathbb{R}^n supports an absolutely friendly measure, then $\text{BA} \cap K$ has Hausdorff dimension not less than the Hausdorff dimension of this measure. In particular, it proves that limit measures of irreducible system of self-similar/self-conformal contractions satisfying the Open Set Condition, such as e.g. the middle-third Cantor set on the real line, contain subsets of full Hausdorff dimension consisting of badly approximable vectors. This was established in [60] and later independently in [64] using a different approach. See also [36] for a stronger result.

The proof in [60] uses quantitative nondivergence estimates and an iterative procedure, which requires the measure in question to be absolutely friendly and not just friendly. A similar question for even the simplest not-absolutely friendly measures is completely open. For example, it is not known whether there exist uncountably many badly approximable pairs of the form (x, x^2) . An analogous problem for atomic measures supported on algebraic numbers, that is, a “badly approximable” version of Roth’s Theorem, is currently beyond reach as well — there are no known badly approximable (or, for that matter, well approximable) algebraic numbers of degree bigger than two.

It has been recently understood that the quantitative nondivergence method can be applied to the question of improvement to Dirichlet’s Theorem (see the beginning of Sect. “[Basic Facts](#)”). Given a positive $\varepsilon < 1$, let us say that Dirichlet’s Theorem *can be ε -improved* for $Y \in M_{m,n}$, writing $Y \in \text{DI}_{m,n}(\varepsilon)$, if for every sufficiently large t the

system

$$\|Y\mathbf{q} - \mathbf{p}\| < \varepsilon e^{-t/m} \quad \text{and} \quad \|\mathbf{q}\| < \varepsilon e^{t/n} \quad (19)$$

(that is, (5) with the right hand side terms multiplied by ε) has a nontrivial integer solution (\mathbf{p}, \mathbf{q}) . It is a theorem of Davenport and Schmidt [24] that $\lambda(\text{DI}_{m,n}(\varepsilon)) = 0$ for any $\varepsilon < 1$; in other words, Dirichlet's Theorem cannot be improved for Lebesgue-generic systems of linear forms. By a modification of the correspondence between dynamics and approximation, (19) is easily seen to be equivalent to $g_t L_Y \mathbb{Z}^k \in K_\varepsilon$, and since the complement to K_ε has nonempty interior for any $\varepsilon < 1$, the result of Davenport and Schmidt follows from the ergodicity of the g_t -action on Ω_k .

Similar questions with λ replaced by $\mathbf{f}_* \lambda$ for some specific smooth maps \mathbf{f} were considered in [2, 3, 15, 23]. For example, [15], Theorem 7, provides an explicitly computable constant $\varepsilon_0 = \varepsilon_0(n)$ such that for \mathbf{f} as in (12),

$$\mathbf{f}_* \lambda(\text{DI}_{1,n}(\varepsilon)) = 0 \quad \text{for } \varepsilon < \varepsilon_0.$$

This had been previously done in [23] for $n = 2$ and in [2] for $n = 3$. In [62] this is extended to a much broader class of measures using estimates described in Sect. “Diophantine Approximation with Dependent Quantities: The Set-Up”. In particular, almost every point of any nondegenerate smooth manifold is proved not to lie in $\text{DI}(\varepsilon)$ for small enough ε depending only on the manifold. Earlier this was done in [61] for the set of *singular* vectors, defined as the intersection of $\text{DI}(\varepsilon)$ over all positive ε ; those correspond to divergent g_t -trajectories. As before, the advantage of the method is allowing a multiplicative generalization of the Dirichlet-improvement set-up; see [62] for more detail.

It is also worthwhile to mention that a generalization of the measure estimate discussed in Sect. “Diophantine Approximation with Dependent Quantities: The Set-Up” was used in [13] to estimate the measure of the set of points \mathbf{x} in a ball $B \subset \mathbb{R}^d$ for which the system

$$\begin{cases} |\mathbf{f}(\mathbf{x}) \cdot \mathbf{q} + p| < \varepsilon \\ |\mathbf{f}'(\mathbf{x}) \cdot \mathbf{q}| < \delta \\ |q_i| < Q_i, \quad i = 1, \dots, n, \end{cases}$$

where \mathbf{f} is a smooth nondegenerate map $B \rightarrow \mathbb{R}^n$, has a nonzero integer solution. For that, $L_{\mathbf{f}(\mathbf{x})}$ as in (15) has to be replaced by the matrix

$$\begin{pmatrix} 1 & 0 & \mathbf{f}(\mathbf{x}) \\ 0 & 1 & \mathbf{f}'(\mathbf{x}) \\ 0 & 0 & I_n \end{pmatrix},$$

and therefore (i) and (ii) turn into more complicated conditions, which nevertheless can be checked when \mathbf{f} is smooth and nondegenerate and ν is Lebesgue measure. This has resulted in proving the convergence case of Khintchine–Groshev Theorem for nondegenerate manifolds [13], in both standard and multiplicative versions. The aforementioned estimate was also used in [7] for the proof of the divergence case, and in [38, 40] for establishing the convergence case of Khintchine–Groshev theorem for affine hyperplanes and their nondegenerate submanifolds. This generalized results obtained by standard methods for the curve (12) by Bernik and Beresnevich [6, 11].

Finally, let us note that in many of the problems mentioned above, the ground field \mathbb{R} can be replaced by \mathbb{Q}_p , and in fact several fields can be taken simultaneously, thus giving rise to the S -arithmetic setting where $S = \{p_1, \dots, p_s\}$ is a finite set of normalized valuations of \mathbb{Q} , which may or may not include the infinite valuation (cf. [83, 90]). The space of lattices in \mathbb{R}^{n+1} is replaced there by the space of lattices in \mathbb{Q}_S^{n+1} , where \mathbb{Q}_S is the product of the fields \mathbb{R} and $\mathbb{Q}_{p_1}, \dots, \mathbb{Q}_{p_s}$. This is the subject of the paper [59], where S -arithmetic analogues of many results reviewed in Sect. “Diophantine Approximation with Dependent Quantities: The Set-Up” have been established. Similarly one can consider versions of the above theorems over local fields of positive characteristic [39]. See also [52] where Sprindžuk's solution [83] of the complex case of Mahler's Conjecture has been generalized (the latter involves studying small values of linear forms with coefficients in \mathbb{C} at real integer points), and [31] which establishes a p -adic analogue of the result of [30] on the set of exceptions to Littlewood's Conjecture.

Future Directions

Interactions between ergodic theory and number theory have been rapidly expanding during the last two decades, and the author has no doubts that new applications of dynamics to Diophantine approximation will emerge in the near future. Specializing to the topics discussed in the present paper, it is fair to say that the list of “further results” contained in the previous section is by no means complete, and many even further results are currently in preparation. This includes: extending proofs of extremality and strong extremality of certain measures to the set-up of systems of linear forms (namely, with $\min(m, n) > 1$; this was mentioned as work in progress in [56]); proving Khintchine-type theorems (both convergence and divergence parts) for p -adic and S -arithmetic nondegenerate manifolds, see [73] for results in this direction; extending [38, 40] to establish Khintchine-type theo-

rems for submanifolds of arbitrary affine subspaces. The work of Druţu [28], who used ergodic theory on homogeneous spaces to compute the Hausdorff dimension of the intersection of $\mathcal{W}_{n,1}(\psi_{1,v})$, $v > 1$, with some rational quadratic hypersurfaces in \mathbb{R}^n deserves a special mention; it is plausible that using this method one can treat more general situations. Several other interesting open directions are listed in [41], Section 9, in the final sections of papers [7,54], in the book [15], and in surveys by Frantzikinakis–McCutcheon, Nitica and Ward in this volume.

Acknowledgment

The work on this paper was supported in part by NSF Grant DMS-0239463.

Bibliography

1. Baker A (1975) Transcendental number theory. Cambridge University Press, London
2. Baker RC (1976) Metric diophantine approximation on manifolds. *J Lond Math Soc* 14:43–48
3. Baker RC (1978) Dirichlet's theorem on diophantine approximation. *Math Proc Cambridge Phil Soc* 83:37–59
4. Bekka M, Mayer M (2000) Ergodic theory and topological dynamics of group actions on homogeneous spaces. Cambridge University Press, Cambridge
5. Beresnevich V (1999) On approximation of real numbers by real algebraic numbers. *Acta Arith* 90:97–112
6. Beresnevich V (2002) A Groshev type theorem for convergence on manifolds. *Acta Mathematica Hungarica* 94:99–130
7. Beresnevich V, Bernik VI, Kleinbock D, Margulis GA (2002) Metric Diophantine approximation: the Khintchine–Groshev theorem for nondegenerate manifolds. *Moscow Math J* 2:203–225
8. Beresnevich V, Dickinson H, Velani S (2006) Measure theoretic laws for lim sup sets. *Mem Amer Math Soc*:179:1–91
9. Beresnevich V, Dickinson H, Velani S (2007) Diophantine approximation on planar curves and the distribution of rational points. *Ann Math* 166:367–426
10. Beresnevich V, Velani S (2007) A note on simultaneous Diophantine approximation on planar curves. *Ann Math* 337: 769–796
11. Bernik VI (1984) A proof of Baker's conjecture in the metric theory of transcendental numbers. *Dokl Akad Nauk SSSR* 277:1036–1039
12. Bernik VI, Dodson MM (1999) Metric Diophantine approximation on manifolds. Cambridge University Press, Cambridge
13. Bernik VI, Kleinbock D, Margulis GA (2001) Khintchine-type theorems on manifolds: convergence case for standard and multiplicative versions. *Int Math Res Notices* 2001:453–486
14. Besicovitch AS (1929) On linear sets of points of fractional dimensions. *Ann Math* 101:161–193
15. Bugeaud Y (2002) Approximation by algebraic integers and Hausdorff dimension. *J London Math Soc* 65:547–559
16. Bugeaud Y (2004) Approximation by algebraic numbers. Cambridge University Press, Cambridge
17. Cassels JWS (1957) An introduction to Diophantine approximation. Cambridge University Press, New York
18. Cassels JWS, Swinnerton-Dyer H (1955) On the product of three homogeneous linear forms and the indefinite ternary quadratic forms. *Philos Trans Roy Soc London Ser A* 248:73–96
19. Dani SG (1979) On invariant measures, minimal sets and a lemma of Margulis. *Invent Math* 51:239–260
20. Dani SG (1985) Divergent trajectories of flows on homogeneous spaces and diophantine approximation. *J Reine Angew Math* 359:55–89
21. Dani SG (1986) On orbits of unipotent flows on homogeneous spaces. II. *Ergodic Theory Dynam Systems* 6:167–182
22. Dani SG, Margulis GA (1993) Limit distributions of orbits of unipotent flows and values of quadratic forms. In: IM Gelfand Seminar. American Mathematical Society, Providence, pp 91–137
23. Davenport H, Schmidt WM (1970) Dirichlet's theorem on diophantine approximation. In: *Symposia Mathematica*. INDAM, Rome, pp 113–132
24. Davenport H, Schmidt WM (1969/1970) Dirichlet's theorem on diophantine approximation. II. *Acta Arith* 16:413–424
25. Ding J (1994) A proof of a conjecture of C. L. Siegel. *J Number Theory* 46:1–11
26. Dodson MM (1992) Hausdorff dimension, lower order and Khintchine's theorem in metric Diophantine approximation. *J Reine Angew Math* 432:69–76
27. Dodson MM (1993) Geometric and probabilistic ideas in the metric theory of Diophantine approximations. *Uspekhi Mat Nauk* 48:77–106
28. Druţu C (2005) Diophantine approximation on rational quadrics. *Ann Math* 333:405–469
29. Duffin RJ, Schaeffer AC (1941) Khintchine's problem in metric Diophantine approximation. *Duke Math J* 8:243–255
30. Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood's conjecture. *Ann Math* 164:513–560
31. Einsiedler M, Kleinbock D (2007) Measure rigidity and p -adic Littlewood-type problems. *Compositio Math* 143:689–702
32. Einsiedler M, Lindenstrauss E (2006) Diagonalizable flows on locally homogeneous spaces and number theory. In: *Proceedings of the International Congress of Mathematicians*. Eur Math Soc, Zürich, pp 1731–1759
33. Eskin A (1998) Counting problems and semisimple groups. In: *Proceedings of the International Congress of Mathematicians*. Doc Math, Berlin, pp 539–552
34. Eskin A, Margulis GA, Mozes S (1998) Upper bounds and asymptotics in a quantitative version of the Oppenheim conjecture. *Ann Math* 147:93–141
35. Eskin A, Margulis GA, Mozes S (2005) Quadratic forms of signature (2,2) and eigenvalue spacings on rectangular 2-tori. *Ann Math* 161:679–725
36. Fishman L (2006) Schmidt's games on certain fractals. *Israel S Math* (to appear)
37. Gallagher P (1962) Metric simultaneous diophantine approximation. *J London Math Soc* 37:387–390
38. Ghosh A (2005) A Khintchine type theorem for hyperplanes. *J London Math Soc* 72:293–304
39. Ghosh A (2007) Metric Diophantine approximation over a local field of positive characteristic. *J Number Theor* 124:454–469
40. Ghosh A (2006) Dynamics on homogeneous spaces and Diophantine approximation on manifolds. Ph D Thesis, Brandeis University, Waltham

41. Gorodnik A (2007) Open problems in dynamics and related fields. *J Mod Dyn* 1:1–35
42. Groshev AV (1938) Une théorème sur les systèmes des formes linéaires. *Dokl Akad Nauk SSSR* 9:151–152
43. Harman G (1998) Metric number theory. Clarendon Press, Oxford University Press, New York
44. Hutchinson JE (1981) Fractals and self-similarity. *Indiana Univ Math J* 30:713–747
45. Jarník V (1928–9) Zur metrischen Theorie der diophantischen Approximationen. *Prace Mat-Fiz* 36:91–106
46. Jarník V (1929) Diophantischen Approximationen und Hausdorffsches Mass. *Mat Sb* 36:371–382
47. Khanin K, Lopes-Dias L, Marklof J (2007) Multidimensional continued fractions, dynamical renormalization and KAM theory. *Comm Math Phys* 270:197–231
48. Khintchine A (1924) Einige Sätze über Kettenbrüche, mit Anwendungen auf die Theorie der Diophantischen Approximationen. *Math Ann* 92:115–125
49. Khintchine A (1963) Continued fractions. P Noordhoff Ltd, Groningen
50. Kleinbock D (2001) Some applications of homogeneous dynamics to number theory. In: *Smooth ergodic theory and its applications*. American Mathematical Society, Providence, pp 639–660
51. Kleinbock D (2003) Extremal subspaces and their submanifolds. *Geom Funct Anal* 13:437–466
52. Kleinbock D (2004) Baker–Sprindžuk conjectures for complex analytic manifolds. In: *Algebraic groups and Arithmetic*. Tata Inst Fund Res, Mumbai, pp 539–553
53. Kleinbock D (2008) An extension of quantitative nondivergence and applications to Diophantine exponents. *Trans AMS*, to appear
54. Kleinbock D, Lindenstrauss E, Weiss B (2004) On fractal measures and diophantine approximation. *Selecta Math* 10: 479–523
55. Kleinbock D, Margulis GA (1996) Bounded orbits of nonquasi-unipotent flows on homogeneous spaces. In: *Sinai's Moscow Seminar on Dynamical Systems*. American Mathematical Society, Providence, pp 141–172
56. Kleinbock D, Margulis GA (1998) Flows on homogeneous spaces and Diophantine approximation on manifolds. *Ann Math* 148:339–360
57. Kleinbock D, Margulis GA (1999) Logarithm laws for flows on homogeneous spaces. *Invent Math* 138:451–494
58. Kleinbock D, Shah N, Starkov A (2002) Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In: *Handbook on Dynamical Systems*, vol 1A. Elsevier Science, North Holland, pp 813–930
59. Kleinbock D, Tomanov G (2007) Flows on S -arithmetic homogeneous spaces and applications to metric Diophantine approximation. *Comm Math Helv* 82:519–581
60. Kleinbock D, Weiss B (2005) Badly approximable vectors on fractals. *Israel J Math* 149:137–170
61. Kleinbock D, Weiss B (2005) Friendly measures, homogeneous flows and singular vectors. In: *Algebraic and Topological Dynamics*. American Mathematical Society, Providence, pp 281–292
62. Kleinbock D, Weiss B (2008) Dirichlet's theorem on diophantine approximation and homogeneous flows. *J Mod Dyn* 2:43–62
63. Kontsevich M, Suhov Y (1999) Statistics of Klein polyhedra and multidimensional continued fractions. In: *Pseudoperiodic topology*. American Mathematical Society, Providence, pp 9–27
64. Kristensen S, Thorn R, Velani S (2006) Diophantine approximation and badly approximable sets. *Adv Math* 203:132–169
65. Lagarias JC (1994) Geodesic multidimensional continued fractions. *Proc London Math Soc* 69:464–488
66. Lindenstrauss E (2007) Some examples how to use measure classification in number theory. In: *Equidistribution in number theory, an introduction*. Springer, Dordrecht, pp 261–303
67. Mahler K (1932) Über das Mass der Menge aller S -Zahlen. *Math Ann* 106:131–139
68. Margulis GA (1975) On the action of unipotent groups in the space of lattices. In: *Lie groups and their representations* (Budapest, 1971). Halsted, New York, pp 365–370
69. Margulis GA (1989) Discrete subgroups and ergodic theory. In: *Number theory, trace formulas and discrete groups* (Oslo, 1987). Academic Press, Boston, pp 377–398
70. Margulis GA (1997) Oppenheim conjecture. In: *Fields Medalists' lectures*. World Sci Publishing, River Edge, pp 272–327
71. Margulis GA (2002) Diophantine approximation, lattices and flows on homogeneous spaces. In: *A panorama of number theory or the view from Baker's garden*. Cambridge University Press, Cambridge, pp 280–310
72. Mauldin D, Urbański M (1996) Dimensions and measures in infinite iterated function systems. *Proc London Math Soc* 73:105–154
73. Mohammadi A, Salehi Golsefidy A (2008) S -Arithmetic Khintchine-Type Theorem. Preprint
74. Moore CC (1966) Ergodicity of flows on homogeneous spaces. *Amer J Math* 88:154–178
75. Roth KF (1955) Rational Approximations to Algebraic Numbers. *Mathematika* 2:1–20
76. Schmidt WM (1960) A metrical theorem in diophantine approximation. *Canad J Math* 12:619–631
77. Schmidt WM (1964) Metrische Sätze über simultane Approximation abhängiger Größen. *Monatsh Math* 63:154–166
78. Schmidt WM (1969) Badly approximable systems of linear forms. *J Number Theory* 1:139–154
79. Schmidt WM (1972) Norm form equations. *Ann Math* 96: 526–551
80. Schmidt WM (1980) Diophantine approximation. Springer, Berlin
81. Sheingorn M (1993) Continued fractions and congruence subgroup geodesics. In: *Number theory with an emphasis on the Markoff spectrum* (Provo, UT, 1991). Dekker, New York, pp 239–254
82. Sprindžuk VG (1964) More on Mahler's conjecture. *Dokl Akad Nauk SSSR* 155:54–56
83. Sprindžuk VG (1969) Mahler's problem in metric number theory. American Mathematical Society, Providence
84. Sprindžuk VG (1979) Metric theory of Diophantine approximations. *VH Winston & Sons*, Washington DC
85. Sprindžuk VG (1980) Achievements and problems of the theory of Diophantine approximations. *Uspekhi Mat Nauk* 35: 3–68
86. Starkov A (2000) Dynamical systems on homogeneous spaces. American Mathematical Society, Providence
87. Sullivan D (1982) Disjoint spheres, approximation by imaginary quadratic numbers, and the logarithm law for geodesics. *Acta Math* 149:215–237

88. Stratmann B, Urbanski M (2006) Diophantine extremality of the Patterson measure. *Math Proc Cambridge Phil Soc* 140:297–304
89. Urbanski M (2005) Diophantine approximation of self-conformal measures. *J Number Theory* 110:219–235
90. Želudevič F (1986) Simultane diophantische Approximationen abhängiger Größen in mehreren Metriken. *Acta Arith* 46: 285–296

Ergodic Theory: Interactions with Combinatorics and Number Theory

TOM WARD

School of Mathematics, University of East Anglia,
Norwich, UK

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Ergodic Theory](#)
[Frequency of Returns](#)
[Ergodic Ramsey Theory and Recurrence](#)
[Orbit-Counting as an Analogous Development](#)
[Diophantine Analysis as a Toolbox](#)
[Future Directions](#)
[Bibliography](#)

Glossary

Almost everywhere (abbreviated a. e.) A property that makes sense for each point x in a measure space (X, \mathcal{B}, μ) is said to hold almost everywhere (or a. e.) if the set $N \subset X$ on which it does not hold satisfies $N \in \mathcal{B}$ and $\mu(N) = 0$.

Čech–Stone compactification of \mathbb{N} , $\beta\mathbb{N}$

A compact Hausdorff space that contains \mathbb{N} as a dense subset with the property that any map from \mathbb{N} to a compact Hausdorff space K extends uniquely to a continuous map $\beta\mathbb{N} \rightarrow K$. This property and the fact that $\beta\mathbb{N}$ is a compact Hausdorff space containing \mathbb{N} characterizes $\beta\mathbb{N}$ up to homeomorphism.

Curvature An intrinsic measure of the curvature of a Riemannian manifold depending only on the Riemannian metric; in the case of a surface it determines whether the surface is locally convex (positive curvature), locally saddle-shaped (negative) or locally flat (zero).

Diophantine approximation Theory of the approximation of real numbers by rational numbers: how small

can the distance from a given irrational real number to a rational number be made in terms of the denominator of the rational?

Equidistributed A sequence is equidistributed if the asymptotic proportion of time it spends in an interval is proportional to the length of the interval.

Ergodic A measure-preserving transformation is ergodic if the only invariant functions are equal to a constant a. e.; equivalently if the transformation exhibits the convergence in the quasi-ergodic hypothesis.

Ergodic theory The study of statistical properties of orbits in abstract models of dynamical systems; more generally properties of measure-preserving (semi-) group actions on measure spaces.

Geodesic (flow) The shortest path between two points on a Riemannian manifold; such a geodesic path is uniquely determined by a starting point and the initial tangent vector to the path (that is, a point in the unit tangent bundle). The transformation on the unit tangent bundle defined by flowing along the geodesic defines the geodesic flow.

Haar measure (on a compact group) If G is a compact topological group, the unique measure μ defined on the Borel sets of G with the property that $\mu(A + g) = \mu(A)$ for all $g \in G$ and $\mu(G) = 1$.

Measure-theoretic entropy A numerical invariant of measure-preserving systems that reflects the asymptotic growth in complexity of measurable partitions refined under iteration of the map.

Mixing A measure-preserving system is mixing if measurable sets (events) become asymptotically independent as they are moved apart in time (under iteration).

(Quasi) Ergodic hypothesis The assumption that, in a dynamical system evolving in time and preserving a natural measure, there are some reasonable conditions under which the ‘time average’ along orbits of an observable (that is, the average value of a function defined on the phase space) will converge to the ‘space average’ (that is, the integral of the function with respect to the preserved measure).

Recurrence Return of an orbit in a dynamical system close to its starting point infinitely often.

S-Unit theorems A circle of results stating that linear equations in fields of zero characteristic have only finitely many solutions taken from finitely-generated multiplicative subgroups of the multiplicative group of the field (apart from infinite families of solutions arising from vanishing sub-sums).

Topological entropy A numerical invariant of topological dynamical systems that measures the asymptotic

growth in the complexity of orbits under iteration. The *variational principle* states that the topological entropy of a topological dynamical system is the supremum over all invariant measures of the measure-theoretic entropies of the dynamical systems viewed as measurable dynamical systems.

Definition of the Subject

Number theory is a branch of pure mathematics concerned with the properties of numbers in general, and integers in particular. The areas of most relevance to this article are *Diophantine analysis* (the study of how real numbers may be approximated by rational numbers, and the consequences for solutions of equations in integers); *analytic number theory*, and in particular asymptotic estimates for the number of primes smaller than X as a function of X ; *equidistribution*, and questions about how the digits of real numbers are distributed. Combinatorics is concerned with identifying structures in discrete objects; of most interest here is that part of combinatorics connected with Ramsey theory, asserting that large subsets of highly structured objects must automatically contain large replicas of that structure. Ergodic theory is the study of asymptotic behavior of group actions preserving a probability measure; it has proved to be a powerful part of dynamical systems with wide applications.

Introduction

Ergodic theory, part of the mathematical study of dynamical systems, has pervasive connections with number theory and combinatorics. This article briefly surveys how these arise through a small sample of results. Unsurprisingly, many details are suppressed, and of course the selection of topics reflects the author's interests far more than it does the full extent of the flow of ideas between ergodic theory and number theory. In addition the selection of topics has been chosen in part to be complementary to those in related articles in the Encyclopedia. A particularly enormous lacuna is the theory of arithmetic dynamical systems itself – the recent monograph by Silverman [94] gives a comprehensive overview.

More sophisticated aspects of this connection – in particular the connections between ergodic theory on homogeneous spaces and Diophantine analysis – are covered in the articles ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#) and ► [Ergodic Theory: Rigidity](#); more sophisticated overviews of the connections with combinatorics may be found in the article ► [Ergodic Theory: Recurrence](#).

Ergodic Theory

While the early origins of ergodic theory lie in the quasi-ergodic hypothesis of classical Hamiltonian dynamics, the mathematical study of ergodic theory concerns various properties of group actions on measure spaces, including but not limited to several special branches:

1. The classical study of single measure-preserving transformations.
2. Measure-preserving actions of \mathbb{Z}^d ; more generally of countable amenable groups.
3. Measure-preserving actions of \mathbb{R}^d and more general amenable groups, called flows.
4. Measure-preserving actions of lattices in Lie groups.
5. Measure-preserving actions of Lie groups.

The ideas and conditions surrounding the quasi-ergodic hypothesis were eventually placed on a firm mathematical footing by developments starting in 1890. For a single measure-preserving transformation $T: X \rightarrow X$ of a probability space (X, \mathcal{B}, μ) , Poincaré [74] showed a *recurrence theorem*: if $E \in \mathcal{B}$ is any measurable set, then for a. e. $x \in E$ there is an infinite set of return times, $0 < n_1 < n_2 < \dots$ with $T^{n_i}(x) \in E$ (of course Poincaré noted this in a specific setting, concerned with a natural invariant measure for the “three-body” problem in planetary motion).

Poincaré's qualitative result was made quantitative in the 1930s, when von Neumann [105] used the approach of Koopman [52] to show the *mean ergodic theorem*: if $f \in L^2(\mu)$ then there is some $\bar{f} \in L^2(\mu)$ for which

$$\left\| \frac{1}{N} \sum_{n=0}^{N-1} f \circ T^n - \bar{f} \right\|_2 \longrightarrow 0 \quad \text{as } N \longrightarrow \infty;$$

clearly \bar{f} then has the property that $\|\bar{f} - \bar{f} \circ T\|_2 = 0$ and $\int \bar{f} d\mu = \int f d\mu$. Around the same time, Birkhoff [13] showed the more delicate *pointwise ergodic theorem*: for any $g \in L^1(\mu)$ there is some $\bar{g} \in L^1(\mu)$ for which

$$\frac{1}{N} \sum_{n=0}^{N-1} g(T^n x) \rightarrow \bar{g}(x) \text{ a. e.};$$

again it is then clear that $\bar{g}(Tx) = \bar{g}(x)$ a. e. and $\int \bar{g} d\mu = \int g d\mu$.

The map T is called *ergodic* if the invariance condition forces the function (f or g) to be equal to a constant a. e. Thus an ergodic map has the property that the *time or ergodic average* $(1/N) \sum_{n=0}^{N-1} f \circ T^n$ converges to the *space average* $\int f d\mu$. An overview of ergodic theorems and their

many extensions may be found in the article ► [Ergodic Theorems](#).

Thus ergodic theory at its most basic level makes strong statements about the asymptotic behavior of orbits of a dynamical system as seen by *observables* (measurable functions on the space X). Applying the ergodic theorem to the indicator function of a measurable set A shows that ergodicity guarantees that a.e. orbit spends an asymptotic proportion of time in A equal to the volume $\mu(A)$ of that set (as measured by the invariant measure). This points to the start of the pervasive connections between ergodic theory and number theory – but as this and other articles relate, the connections extend far beyond this.

Frequency of Returns

In this section we illustrate the way in which a dynamical point of view may unify, explain and extend quite disparate results from number theory.

Normal Numbers

Borel [15] showed (as a consequence of what became the Borel–Cantelli Lemma in probability) that a.e. real number (with respect to Lebesgue measure) is *normal* to every base: that is, has the property that any block of k digits in the base- r expansion appears with asymptotic frequency r^{-k} .

Continued Fraction Digits

Analogs of normality results for the continued fraction expansion of real numbers were found by Khinchin, Kuz'min, Lévy and others. Any irrational $x \in [0, 1]$ has a unique expansion as a continued fraction

$$x = \frac{1}{a_1(x) + \frac{1}{a_2(x) + \frac{1}{a_3(x) + \dots}}}$$

and, just as in the case of the familiar base- r expansion, it turns out that the digits $(a_n(x))$ obey precise statistical rules for a.e. x . Gauss conjectured that the appearance of individual digits would obey the law

$$\frac{1}{N} |\{k: 1 \leq k \leq N, a_k(x) = j\}| \rightarrow \frac{2 \log(1+j) - \log j - \log(2+j)}{\log 2}. \quad (1)$$

This was eventually proved by Kuz'min [54] and Lévy [59], and the probability distribution of the digits is the *Gauss–*

Kuz'min law. Khinchin [50] developed this further, showing for example that

$$\begin{aligned} \lim_{n \rightarrow \infty} (a_1(x) a_2(x) \dots a_n(x))^{1/n} \\ = \prod_{n=1}^{\infty} \left(\frac{(n+1)^2}{n(n+2)} \right)^{\log n / \log 2} \\ = 2.68545 \dots \quad \text{for a.e. } x. \end{aligned}$$

Lévy [60] showed that the denominator $q_n(x)$ of the n th convergent $(p_n(x))/(q_n(x))$ (the rational obtained by truncating the continued fraction expansion of x at the n th term) grows at a specific exponential rate,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log q_n(x) = \frac{\pi^2}{12 \log 2} \quad \text{for a.e. } x.$$

First Digits

The astronomer Newcomb [67] noted that the first digits of large collections of numerical data that are not dimensionless have a specific and non-uniform distribution:

“The law of probability of the occurrence of numbers is such that all mantissæ of their logarithms are equally probable.”

This is now known as Benford's Law, following his popularization and possible rediscovery of the phenomenon [6]. In both cases, this was an empirical observation eventually made rigorous by Hill [42]. Arnold (App. 12 in [3]), pointed out the dynamics behind this phenomena in certain cases, best illustrated by the statistical behavior of the sequence 1, 2, 4, 8, 1, 3, 6, 1, ... of first digits of powers of 2. Empirically, the digit 1 appears about 30% of the time, while the digit 9 appears about 5% of the time.

Equidistribution

Weyl [109] (and, separately, Bohl [14] and Sierpiński [91]) found an important instance of *equidistribution*. Writing $\{\cdot\}$ for the fractional part, a sequence (a_n) of real numbers is said to be equidistributed modulo 1 if, for any interval $[a, b] \subset [0, 1)$,

$$\frac{1}{N} |\{k: 1 \leq k \leq N, \{a_k\} \in [a, b]\}| \rightarrow (b - a) \quad \text{as } N \rightarrow \infty;$$

equivalently if

$$\frac{1}{N} \sum_{k=1}^N f(a_k) \rightarrow \int_0^1 f(t) dt \quad \text{as } N \rightarrow \infty$$

for all continuous functions f . Weyl showed that the sequence $\{n\alpha\}$ is equidistributed if and only if α is irrational. This result was refined and extended in many directions; for example, Hlawka [44] and others found rates for the convergence in terms of the discrepancy of the sequence, Weyl [110] proved equidistribution for $\{n^2\alpha\}$, and Vinogradov for $\{p_n\alpha\}$ where p_n is the n th prime.

The Ergodic Context

All the results of this section are manifestations of various kinds of convergence of ergodic averages. Borel's theorem on normal numbers is an immediate consequence of the fact that Lebesgue measure on $[0, 1]$ is invariant and ergodic for the map $x \mapsto bx$ modulo 1 with $b \geq 2$. The asymptotic properties of continued fraction digits are all a consequence of the fact that the *Gauss measure* defined by

$$\mu(A) = \frac{1}{\log 2} \int_A \frac{dx}{1+x} \quad \text{for } A \subset [0, 1]$$

is invariant and ergodic for the Gauss map $x \mapsto \{\frac{1}{x}\}$, and the orbit of an irrational number under the Gauss map determine the digits appearing in the continued fraction expansion much as the orbit under the map $x \mapsto bx \pmod{1}$ determines the digits in the base b expansion.

The results on equidistribution and the frequency of first digits are related to ergodic averaging of a different sort. For example, writing $R_\alpha(t) = t + \alpha$ modulo 1 for the circle rotation by α , the first digit of 2^n is the digit j if and only if

$$\log_{10} j \leq R_{\log_{10}(2)}(0) < \log_{10}(j+1).$$

Thus the asymptotic frequency of appearance concerns the orbit of a *specific point*. In order to see what this means, consider a continuous map $T: X \rightarrow X$ of a compact metric space (X, d) . The space $\mathcal{M}(T)$ of Borel probability measures on the Borel σ -algebra of (X, d) is a non-empty compact convex set in the weak*-topology, each extreme point is an ergodic measure for T , and these ergodic measures are mutually singular. If $\mathcal{M}(T)$ is not a singleton and $\mu_1, \mu_2 \in \mathcal{M}(T)$ are distinct ergodic measures, then for a continuous function f with $\int_X f d\mu_1 \neq \int_X f d\mu_2$ it is clear that the ergodic averages $\frac{1}{N} \sum_{n=0}^{N-1} f(T^n x)$ must converge to $\int_X f d\mu_1$ a. e. with respect to μ_1 and to $\int_X f d\mu_2$ a. e. with respect to μ_2 . Thus the presence of many invariant measures for a continuous map means that ergodic averages along the orbits of specific points need not converge to the space average with respect to a chosen invariant measure.

In the extreme situation of *unique ergodicity* (a single invariant measure, which is necessarily an extreme point of $\mathcal{M}(T)$ and hence ergodic) the convergence of ergodic averages is much more uniform. Indeed, if T is uniquely ergodic with $\mathcal{M}(T) = \{\mu\}$ then, for any continuous function $f: X \rightarrow \mathbb{R}$,

$$\frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) \longrightarrow \int_X f d\mu \quad \text{uniformly in } x$$

(see Oxtoby [69]). The circle rotation R_α is uniquely ergodic for irrational α , leading to the equidistribution results.

The ergodic viewpoint on equidistribution also places equidistribution results in a wider context. Weyl's result that $\{n^2\alpha\}$ (indeed, the fractional part of any polynomial with at least one irrational coefficient) is equidistributed for irrational α was given another proof by Furstenberg [29] using the notion of unique ergodicity. These methods were then used in the study of nilsystems (translations on quotients of nilpotent Lie groups) by Auslander, Green and Hahn [4] and Parry [70], and these nilsystems play an essential role for polynomial (and other non-conventional) ergodic averaging (see Host and Kra [45] and Leibman [58] in the polynomial case; Host and Kra [46] in the multiple linear case). Remarkably, nilsystems are starting to play a role within combinatorics – an example is the work on the asymptotic number of 4-step arithmetic progressions in the primes by Green and Tao [40]. Pointwise ergodic theorems have also been found along sequences other than \mathbb{N} ; notably for integer-valued polynomials and along the primes for L^2 functions by Bourgain [16,17]. For more details, see the survey paper of del Junco on ergodic theorems.

Ergodic Ramsey Theory and Recurrence

In 1927 van der Waerden proved a conjecture attributed to Baudet: if the natural numbers are written as a disjoint union of finitely many sets,

$$\mathbb{N} = C_1 \sqcup C_2 \sqcup \cdots \sqcup C_r, \quad (2)$$

then there must be one set C_j that contains arbitrarily long arithmetic progressions. That is, there is some $j \in \{1, \dots, r\}$ such that for any $k \geq 1$ there are $a \geq 1$ and $n \geq 1$ with

$$a, a+n, a+2n, \dots, a+(k-1)n \in C_j.$$

The original proof appears in van der Waerden's paper [103], and there is a discussion of how he found the proof in [104].

Work of Furstenberg and Weiss [34] and others placed the theorem of van der Waerden in the context of topological dynamics, giving alternative proofs. Specifically, van der Waerden's theorem is a consequence of *topological multiple recurrence*: the return of points under iteration in a topological dynamical system close to their starting point along finite sequences of times. The same approach readily gives dynamical proofs of Rado's extension [76] of van der Waerden's theorem, and of Hindman's theorem [43]. The theorems of Rado and Hindman introduce a new theme: given a set $A = \{n_1, n_2, \dots\}$ of natural numbers, write $FS(A)$ for the set of numbers obtained as finite sums $n_{i_1} + \dots + n_{i_j}$ with $i_1 < i_2 < \dots < i_j$. Rado showed that for any large n there is some C_s containing some $FS(A)$ for a set A of cardinality n . Hindman showed that there is some C_s containing some $FS(A)$ for an *infinite* set A .

In the theorem of van der Waerden, it is clear that for any reasonable notion of "proportion" or "density" one of the sets C_j must occupy a positive proportion of \mathbb{N} . A set $A \subset \mathbb{N}$ is said to have *positive upper density* if there are sequences (M_i) and (N_i) with $N_i - M_i \rightarrow \infty$ as $i \rightarrow \infty$ such that

$$\lim_{i \rightarrow \infty} \frac{1}{N_i - M_i} |\{a \in A : M_i < a < N_i\}| > 0.$$

Erdős and Turán [26] conjectured the stronger statement that any subset of \mathbb{N} with positive upper density must contain arbitrary long arithmetic progressions. This statement was shown for arithmetic progressions of length 3 by Roth [79] in 1952, then for length 4 by Szemerédi [96] in 1969. The general result was eventually proved by Szemerédi [97] in 1975 in a lengthy and extremely difficult argument.

Furstenberg saw that Szemerédi's Theorem would follow from a deep extension of the Poincaré recurrence phenomena described in Sect. "Ergodic Theory" and proved that extension [30] (see also the survey article by Furstenberg, Katznelson and Ornstein [33]). The *multiple recurrence* result of Furstenberg says that for any measure-preserving system (X, \mathcal{B}, μ, T) and set $A \in \mathcal{B}$ with $\mu(A) > 0$, and for any $k \in \mathbb{N}$,

$$\liminf_{N-M \rightarrow \infty} \frac{1}{N-M+1} \cdot \sum_{n=M}^N \mu(A \cap T^{-n}A \cap T^{-2n}A \cap \dots \cap T^{-kn}A) > 0.$$

An immediate consequence is that in the same setting there must be some $n \geq 1$ for which

$$\mu(A \cap T^{-n}A \cap T^{-2n}A \cap \dots \cap T^{-kn}A) > 0. \quad (3)$$

A general *correspondence principle*, due to Furstenberg, shows that statements in combinatorics like Szemerédi's Theorem are equivalent to statements in ergodic theory like (3).

This opened up a significant new field of *ergodic Ramsey theory*, in which methods from dynamical systems and ergodic theory are used to produce new results in infinite combinatorics. For an overview, see the articles ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#), ► [Ergodic Theory: Rigidity](#), ► [Ergodic Theory: Recurrence](#) and the survey articles of Bergelson [7,8,10]. The field is too large to give an overview here, but a few examples will give a flavor of some of the themes.

Call a set $R \subset \mathbb{Z}$ a *set of recurrence* if, for any finite measure-preserving invertible transformation T of a finite measure space (X, \mathcal{B}, μ) and any set $A \in \mathcal{B}$ with $\mu(A) > 0$, there are infinitely many $n \in R$ for which $\mu(A \cap T^{-n}A) > 0$. Thus Poincaré recurrence is the statement that \mathbb{N} is a set of recurrence. Furstenberg and Katznelson [31] showed that if T_1, \dots, T_k form a family of commuting measure-preserving transformations and A is a set of positive measure, then

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} \mu(T_1^{-n}A \cap \dots \cap T_k^{-n}A) > 0.$$

This remarkable multiple recurrence implies a multi-dimensional form of Szemerédi's theorem. Recently, Gowers has found a non-ergodic proof of this [38].

Furstenberg also gave an ergodic proof of Sárközy's theorem [80]: if $p \in \mathbb{Q}[t]$ is a polynomial with $p(\mathbb{Z}) \subset \mathbb{Z}$ and $p(0) = 0$, then $\{p(n)\}_{n>0}$ is a set of recurrence. This was extended to multiple polynomial recurrence by Bergelson and Leibman [11].

Topology and Coloring Theorems

The existence of idempotent ultrafilters in the Čech–Stone compactification $\beta\mathbb{N}$ gives rise to an algebraic approach to many questions in topological dynamics (this notion has its origins in the work of Ellis [24]). Using these methods, results like Hindman's finite sums theorem find elegant proofs, and many new results in combinatorics have been found. For example, in the partition (2) there must be one set C_j containing a triple x, y, z solving $x - y = z^2$.

A deeper application is to improve a strengthening of Kronecker's theorem. To explain this, recall that a set S is called *IP* if there is a sequence (n_i) of natural numbers (which do not need to be distinct) with the property that S contains all the terms of the sequence and all finite sums of terms of the sequence with distinct indices.

A set S is called IP^* if it has non-empty intersection with every IP set, and a set S is called IP_+^* if there is some $t \in \mathbb{Z}$ for which $S - t$ is IP^* . Thus being IP^* (or IP_+^*) is an extreme form of ‘fatness’ for a set. Now let $1, \alpha_1, \dots, \alpha_k$ be numbers that are linearly independent over the rationals, and for any $d \in \mathbb{N}$ and kd non-empty intervals $I_{ij} \subset [0, 1]$ ($1 \leq i \leq d, 1 \leq j \leq k$), let

$$D = \{n \in \mathbb{N} : \{n^i \alpha_j\} \in I_{ij} \text{ for all } i, j\}.$$

Kronecker showed that if $d = 1$ then D is non-empty; Hardy and Littlewood showed that D is infinite, and Weyl showed that D has positive density. Bergelson [9] uses these algebraic methods to improve the result by showing that D is an IP_+^* set.

Polynomialization and IP -sets

As mentioned above, Bergelson and Leibman [11] extended multiple recurrence to a polynomial setting. For example, let

$$\{p_{i,j} : 1 \leq i \leq k, 1 \leq j \leq t\}$$

be a collection of polynomials with rational coefficients and $p_{i,j}(\mathbb{Z}) \subset \mathbb{Z}$, $p_{i,j}(0) = 0$. Then if $\mu(A) > 0$, we have

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \mu \left(\bigcap_{i=1}^k \left(\prod_{j=1}^t T_j^{p_{i,j}(n)} \right)^{-1} A \right) > 0.$$

Using the Furstenberg correspondence principle, this gives a multi-dimensional polynomial Szemerédi theorem: If $P: \mathbb{Z}^r \rightarrow \mathbb{Z}^\ell$ is a polynomial mapping with the property that $P(0) = 0$, and $F \subset \mathbb{Z}^r$ is a finite configuration, then any set $S \subset \mathbb{Z}^\ell$ of positive upper Banach density contains a set of the form $u + P(nF)$ for some $u \in \mathbb{Z}^\ell$ and $n \in \mathbb{N}$.

In a different direction, motivated in part by Hindman’s theorem, the multiple recurrence results generalize to IP -sets. Furstenberg and Katznelson [32] proved a linear IP -multiple recurrence theorem in which the recurrence is guaranteed to occur along an IP -set. A combinatorial proof of this result has been found by Nagle, Rödl and Schacht [66]. Bergelson and McCutcheon [12] extended these results by proving a polynomial IP -multiple recurrence theorem. To formulate this, make the following definitions. Write \mathcal{F} for the family of non-empty finite subsets of \mathbb{N} , so that a sequence indexed by \mathcal{F} is an IP -set. More generally, an \mathcal{F} -sequence $(n_\alpha)_{\alpha \in \mathcal{F}}$ taking values in an abelian group is called an IP -sequence if $n_{\alpha \cup \beta} = n_\alpha + n_\beta$ whenever $\alpha \cap \beta = \emptyset$. An IP -ring is a set of the form $\mathcal{F}^{(1)} = \{\bigcup_{i \in \beta} \alpha_i : \beta \in \mathcal{F}\}$ where $\alpha_1 < \alpha_2 < \dots$ is a sequence in \mathcal{F} , and $\alpha < \beta$ means $a < b$ for all $a \in \alpha, b \in \beta$;

write $\mathcal{F}_{<}^m$ for the set of m -tuples $(\alpha_1, \dots, \alpha_m)$ from \mathcal{F}^m with $\alpha_i < \alpha_j$ for $i < j$. Write $PE(m, d)$ for the collection of all expressions of the form $T(\alpha_1, \dots, \alpha_m) = \prod_{i=1}^r T_i^{p_i((n_{\alpha_j}^{(b)}))_{1 \leq b \leq k, 1 \leq j \leq m}}, (\alpha_1, \dots, \alpha_m) \in (F \cup \emptyset)_{<}^m$, where each p_i is a polynomial in a $k \times m$ matrix of variables with integer coefficients and zero constant term with degree $\leq d$. Then for every $m, t \in \mathbb{N}$, there is an IP -ring $\mathcal{F}^{(1)}$, and an $a = a(A, m, t, d) > 0$, such that, for every set of polynomial expressions $\{S_0, \dots, S_t\} \subset PE(m, d)$,

$$IP - \lim_{(\alpha_1, \dots, \alpha_m) \in (\mathcal{F}^{(1)})_{<}^m} \mu \left(\bigcap_{i=0}^t S_i(\alpha_1, \dots, \alpha_m)^{-1} A \right) > 0.$$

There are a large number of deep combinatorial consequences of this result, not all of which seem accessible by other means.

Sets of Primes

In a remarkable development, Szemerédi’s theorem and some of the ideas behind ergodic Ramsey theory joined results of Goldston and Yıldırım [37] in playing a part in Green and Tao’s proof [39] that the set of primes contains arbitrarily long arithmetic progressions. This profound result is surveyed from an ergodic point of view in the article of Kra [53]. As with Szemerédi’s theorem itself, this result has been extended to a polynomial setting by Tao and Ziegler [101]. Given integer-valued polynomials $f_1, \dots, f_k \in \mathbb{Z}[t]$ with

$$f_1(0) = \dots = f_k(0) = 0$$

and any $\varepsilon > 0$, Tao and Ziegler proved that there are infinitely many integers x, m with $1 \leq m \leq x^\varepsilon$ for which $x + f_1(m), \dots, x + f_k(m)$ are primes.

Orbit-Counting as an Analogous Development

Some of the connections between number theory and ergodic theory arise through developments that are analogous but not directly related. A remarkable instance of this concerns the long history of attempts to count prime numbers laid alongside the problem of counting closed orbits in dynamical systems.

Counting Orbits and Geodesics

Consider first the fundamental arithmetic function $\pi(X) = |\{p \leq X : p \text{ is prime}\}|$. Tables of primes prepared by Felkel and Vega in the 18th century led Legendre to suggest that $\pi(X)$ is approximately $x/(\log(X) - 1.08)$. Gauss,

using both computational evidence and a heuristic argument, suggested that $\pi(X)$ is approximated by

$$\text{li}(X) = \int_2^X \frac{dt}{\log t}.$$

Both of these suggestions imply the well-known asymptotic formula

$$\pi(X) \sim \frac{X}{\log X}. \quad (4)$$

Riemann brought the analytic ideas of Dirichlet and Chebyshev (who used the zeta function to find a weaker version of (4) with upper and lower bounds for the quantity $\frac{\pi(X) \log(X)}{X}$) to bear by proposing that the zeta function

$$\zeta(s) = \sum_{n=1}^{\infty} \frac{1}{n^s} = \prod_p (1 - p^{-s})^{-1}, \quad (5)$$

already studied by Euler, would connect properties of the primes to analytic methods. An essential step in these developments, due to Riemann, is the meromorphic extension of ζ from the region $\Re(s) > 1$ in (5) to the whole complex plane and a functional equation relating the value of the extension at s to the value at $1 - s$. Moreover, Riemann showed that the extension has readily understood real zeros, and that all the other zeros he could find were symmetric about $\Re(s) = \frac{1}{2}$. The *Riemann hypothesis* asserts that zeros in the region $0 < \Re(s) < 1$ all lie on the line $\Re(s) = \frac{1}{2}$, and this remains open.

Analytic properties of the Riemann zeta function were used by Hadamard and de la Vallée Poussin to prove (4), the *Prime Number Theorem*, in 1896. Tauberian methods developed by Wiener and Ikehara [11] later gave different approaches to the Prime Number Theorem.

These ideas initiated the widespread use of zeta functions in several parts of mathematics, but it was not until the middle of the 20th century that Selberg [88] introduced a zeta function dealing directly with quantities arising in dynamical systems: the lengths of closed geodesics on surfaces of constant curvature -1 . The geodesic flow acts on the unit tangent bundle to the surface by moving a point and unit tangent vector at that point along the unique geodesic they define at unit speed. Closed geodesics are then in one-to-one correspondence with periodic orbits of the associated geodesic flow on the unit tangent bundle, and it is in this sense that the quantities are dynamical. The function defined by Selberg takes the form

$$Z(s) = \prod_{\tau} \prod_{k=0}^{\infty} (1 - e^{-(s+k)|\tau|}),$$

in which τ runs over all the closed geodesics, and $|\tau|$ denotes the length of the geodesic. In a direct echo of the Riemann zeta function, Selberg found an analytic continuation to the complex plane, and showed that the zeros of Z lie on the real axis or on the line $\Re(s) = \frac{1}{2}$ (the analogue of the Riemann hypothesis for Z ; see also the paper of Hejhal [41]). The zeros of Z are closely connected to the eigenvalues for the Laplace–Beltrami operator, and thus give information about the lengths of closed geodesics via Selberg’s trace formula in the same paper. Huber [47] and others used this approach to give an analogue of the prime number theorem for closed geodesics – a *prime orbit theorem*.

Sinai [92] considered closed geodesics on a manifold M with negative curvature bounded between $-R^2$ and $-r^2$, and found the bounds

$$(\dim(M) - 1)r \leq \liminf_{T \rightarrow \infty} \frac{\log \pi(T)}{T} \leq \limsup_{T \rightarrow \infty} \frac{\log \pi(T)}{T} \leq (\dim(M) - 1)R$$

for the number $\pi(T)$ of closed geodesics of multiplicity one with length less than T , analogous to Chebyshev’s result.

The essential dynamical feature behind the geodesic flow on a manifold of negative curvature is that it is an example of an *Anosov flow* [2]. These are smooth dynamical \mathbb{R} -actions (equivalently, first-order differential equations on Riemannian manifolds) with the property that the tangent bundle has a continuously varying splitting into a direct sum $E^u \oplus E^s \oplus E^o$ and the action of the differential of the flow acts on E^u as an exponential expansion, on E^s as an exponential contraction, E^o is the one-dimensional bundle of vectors that are tangent to orbits, and the expansion and contraction factors are bounded. In the setting of Anosov flows, the natural orbit counting function is $\pi(X) = |\{\tau : \tau \text{ a closed orbit of length } |\tau| \leq X\}|$. Margulis [62,63] generalized the picture to weak-mixing Anosov flows by showing a prime orbit theorem of the form

$$\pi(X) \sim \frac{e^{h_{\text{top}} X}}{h_{\text{top}} X} \quad (6)$$

for the counting function $\pi(X) = |\{\tau : \tau \text{ a closed orbit of length } |\tau| \leq X\}|$ where as before h_{top} denotes the topological entropy of the flow. Integral to Margulis’ work is a result on the spatial distribution of the closed geodesics reflected in a flow-invariant probability measure, now called the Margulis measure.

Anosov also studied discrete dynamical systems with similar properties: diffeomorphisms of compact manifolds

with a similar splitting of the tangent space (though in this setting E^0 disappears). The archetypal Anosov diffeomorphism is a hyperbolic toral automorphism of the sort considered in Subsect. “[Orbit Growth and Convergence](#)”; for such automorphisms of the 2-torus Adler and Weiss [1] constructed Markov partitions, allowing the dynamics of the toral automorphism to be modeled by a topological Markov shift, and used this to determine when two such automorphisms are measurably isomorphic. Sinai [93], Ratner [77], Bowen [18,20] and others developed the construction of Markov partitions in general for Anosov diffeomorphisms and flows.

Around the same time, Smale [95] introduced a more permissive hyperbolicity axiom for diffeomorphisms, Axiom A. Maps satisfying Axiom A are diffeomorphisms satisfying the same hypothesis as that of Anosov diffeomorphisms, but only on the set of points that return arbitrarily close under the action of the flow (or iteration of the map).

Thus Markov partitions, and with them associated transfer operators became a substitute for the geometrical Laplace–Beltrami operators of the setting considered by Selberg. Bowen [19] extended the uniform distribution result of Margulis to this setting and found an analogue of Chebychev’s theorem for closed orbits. Parry [71] (in a restricted case) and Parry and Pollicott [73] went on to prove the prime orbit theorem in this more general setting. The methods are an adaptation of the Ikehara–Wiener–Tauberian approach to the prime number theorem.

Thus many facets of the prime number theorem story find their echoes in the study of closed orbits for hyperbolic flows: the role played by meromorphic extensions of suitable zeta functions, Tauberian methods, and so on. Moreover, related results from number theory have analogues in dynamics, for example Mertens’ theorem [65] in the work of Sharp [89] and Noorani [68] and Dirichlet’s theorem in work of Parry [72].

The “elementary” proof (not using analytic methods) of the prime number theorem by Erdős [25] and Selberg [87] (see the survey by Goldfeld [36] for the background to the results and the unfortunate priority dispute) has an echo in some approaches to orbit-counting problems from an elementary (non-Tauberian) perspective, including work of Lalley [56] on special flows and Everest, Miles, Stevens and the author [27] in the algebraic setting.

In a different direction Lalley [55] found orbit asymptotics for closed orbits satisfying constraints in the Axiom A setting without using Tauberian theorems. His more direct approach is still analytic, using complex transfer operators (the same objects used to by Parry and Pol-

licott to study the dynamical zeta function at complex values) and indeed somewhat parallels a Tauberian argument.

Further resonances with number theory arise here. For example, there are results on the distribution of closed orbits for group extensions (analogous to Chebotarev’s theorem) and for orbits with homological constraints (see Sharp [90], Katsuda and Sunada [49]).

Of course the great diversity of dynamical systems subsumed in the phrase “prime orbit theorem” creates new problems and challenges, and in particular if there is not much geometry to work with then the reliance on Markov partitions and transfer operators makes it difficult to find higher-order asymptotics.

Dolgopyat [22] has nonetheless managed to push the Markov methods to obtain uniform bounds on iterates of the associated transfer operators to the region $\Re(s) > \sigma_0$ with $\sigma_0 < 1$. This result has wide implications; an example most relevant to the analogy with number theory is the work of Pollicott and Sharp [75] in which Dolgopyat’s result is used to show that for certain geodesic flows there is a two-term prime orbit theorem of the form

$$\pi(X) = \text{li} \left(e^{h_{\text{top}} X} \right) + O \left(e^{cX} \right)$$

for some $c < h_{\text{top}}$.

For non-positive curvature manifolds less is known: Knieper [51] finds upper and lower bounds for the function counting closed geodesics on rank-1 manifolds of non-positive curvature of the form

$$A \frac{e^{hX}}{X} \leq \pi(X) \leq B e^{hX}$$

for constants $A, B > 0$.

Counting Orbits for Group Endomorphisms

A prism through which to view some of the deeper issues that arise in Subsect. “[Counting Orbits and Geodesics](#)” is provided by group endomorphisms. The price paid for having simple closed formulas for all the quantities involved is of course a severe loss of generality, but the diversity of examples illustrates many of the phenomena that may be expected in more general settings when hyperbolicity is lost.

Consider an endomorphism $T: X \rightarrow X$ of a compact group with the property that $F_n(T) < \infty$ for all $n \geq 1$. The number of closed orbits of length n under T is then

$$O_n(T) = \frac{1}{n} \sum_{d|n} \mu(n/d) F_d(T). \quad (7)$$

In simple situations (hyperbolic toral automorphisms for example) it is straightforward to show that

$$\pi_T(X) = |\{\tau: \tau \text{ a closed orbit under } T \text{ of length } \leq X\}| \\ \sim \frac{e^{(X+1)h_{\text{top}}(T)}}{X}. \quad (8)$$

Waddington [106] considered quasihyperbolic toral automorphisms, showing that the asymptotic (8) in this case is multiplied by an explicit almost-periodic function bounded away from zero and infinity.

This result has been extended further into non-hyperbolic territory, which is most easily seen via the so-called connected S -integer dynamical systems introduced by Chothi, Everest and the author [21]. Fix an algebraic number field \mathbb{K} with set of places $P(\mathbb{K})$ and set of infinite places $P_\infty(\mathbb{K})$, an element of infinite multiplicative order $\xi \in \mathbb{K}^*$, and a finite set $S \subset P(\mathbb{K}) \setminus P_\infty(\mathbb{K})$ with the property that $|\xi|_w \leq 1$ for all $w \notin S \cup P_\infty(\mathbb{K})$. The associated ring of S -integers is

$$R_S = \{x \in \mathbb{K}: |x|_w \leq 1 \text{ for all } w \notin S \cup P_\infty(\mathbb{K})\}.$$

Let X be the compact character group of R_S , and define the endomorphism $T: X \rightarrow X$ to be the dual of the map $x \mapsto \xi x$ on R_S . Following Weil [108], write \mathbb{K}_w for the completion at w , and for w finite, write r_w for the maximal compact subring of \mathbb{K}_w . Notice that if $S = P$ then $R_S = \mathbb{K}$ and $F_n(T) = 1$ for all $n \geq 1$ by the product formula for \mathbb{A} -fields. As the set S shrinks, more and more periodic orbits come into being, and if S is as small as possible (given ξ) then the resulting system is (more or less) hyperbolic or quasi-hyperbolic.

For S finite, it turns out that there are still sufficiently many periodic orbits to have the growth rate result (10), but the asymptotic (8) is modified in much the same way as Waddington observed for quasi-hyperbolic toral automorphisms:

$$\liminf_{X \rightarrow \infty} \frac{X \pi_T(X)}{e^{(X+1)h_{\text{top}}(T)}} > 0 \quad (9)$$

and there is an associated pair (X^*, a_T) , where X^* is a compact group and $a_T \in X^*$, with the property that if $a_T^{N_j}$ converges in X^* as $j \rightarrow \infty$, then there is convergence in (9).

A simple special case will illustrate this. Taking $\mathbb{K} = \mathbb{Q}$, $\xi = 2$ and $S = \{3\}$ gives a compact group endomorphism T with

$$F_n(T) = (2^n - 1)|2^n - 1|_3.$$

For this example the results of [21] are sharper: The expression in (9) converges along (X_j) if and only if 2^{X_j} converges in the ring of 3-adic integers \mathbb{Z}_3 , the expression has uncountably many limit points, and the upper and lower limits are transcendental.

Similarly, the dynamical analogue of Mertens' theorem found by Sharp may be found for S -integer systems with S finite. Writing

$$\mathcal{M}_T(N) = \sum_{|\tau| \leq N} \frac{1}{e^{h(T)|\tau|}},$$

it is shown in [21] that for an ergodic S -integer map T with $\mathbb{K} = \mathbb{Q}$ and S finite, there are constants $k_T \in \mathbb{Q}$ and C_T such that

$$\mathcal{M}_T(N) = k_T \log N + C_T + O(1/N).$$

Without the restriction that $\mathbb{K} = \mathbb{Q}$, it is shown that there are constants $k_T \in \mathbb{Q}$, C_T and $\delta > 0$ with

$$\mathcal{M}_T(N) = k_T \log N + C_T + O(N^{-\delta}).$$

Diophantine Analysis as a Toolbox

Many problems in ergodic theory and dynamical system exploit ideas and results from number theory in a direct way; we illustrate this by describing a selection of dynamical problems that call on particular parts of number theory in an essential way. The example of mixing in Subsect. "Mixing and Additive Relations in Fields" is particularly striking for two reasons: the results needed from number theory are relatively recent, and the ergodic application directly motivated a further development in number theory.

Orbit Growth and Convergence

The analysis of periodic orbits – how their number grows as the length grows and how they spread out through space – is of central importance in dynamics (see Katok [48] for example). An instance of this is that for many simple kinds of dynamical systems $T: X \rightarrow X$ (where T is a continuous map of a compact metric space (X, d)) the logarithmic growth rate of the number of periodic points exists and coincides with the topological entropy $h(T)$ (an invariant giving a quantitative measure of the average rate of growth in orbit complexity under T). That is, writing

$$F_n(T) = |\{x \in X: T^n x = x\}|,$$

we find

$$\frac{1}{n} \log F_n(T) \longrightarrow h_{\text{top}}(T) \quad (10)$$

for many of the simplest dynamical systems. For example, if $X = \mathbb{T}^r$ is the r -torus and $T = T_A$ is the automorphism of the torus corresponding to a matrix A in $GL_r(\mathbb{Z})$, then T_A is ergodic with respect to Lebesgue measure if and only if no eigenvalue of A is a root of unity. Under this assumption, we have

$$F_n(T_A) = \prod_{i=1}^r |\lambda_i^n - 1|$$

and

$$h_{\text{top}}(T_A) = \sum_{i=1}^r \log \max\{1, |\lambda_i|\} \quad (11)$$

where $\lambda_1, \dots, \lambda_r$ are the eigenvalues of A . It follows that the convergence in (10) is clear under the assumption that T_A is *hyperbolic* (that is, no eigenvalue has modulus one). Without this assumption the convergence is less clear: for $r \geq 4$ the automorphism T_A may be ergodic without being hyperbolic. That is, while no eigenvalues are unit roots some may have unit modulus. As pointed out by Lind [61] in his study of these *quasihyperbolic* automorphisms, the convergence (10) does still hold for these systems, but this requires a significant Diophantine result (the theorem of Gel'fond [35] suffices; one may also use Baker's theorem [5]). Even further from hyperbolicity lie the family of S -integer systems [21,107]; their orbit-growth properties are intimately tied up with Artin's conjecture on primitive roots and prime divisors of linear recurrence sequences.

Mixing and Additive Relations in Fields

The problem of higher-order mixing for commuting group automorphisms provides a striking example of the dialogue between ergodic theory and number theory, in which deep results from number theory have been used to solve problems in ergodic theory, and questions arising in ergodic theory have motivated further developments in number theory.

An action T of a countable group Γ on a probability space (X, \mathcal{B}, μ) is called *k-fold mixing* or *mixing on $(k+1)$ sets* if

$$\mu(A_0 \cap T^{-g_1} A_1 \cap \dots \cap T^{-g_k} A_k) \longrightarrow \mu(A_0) \dots \mu(A_k) \quad (12)$$

as

$$g_i g_j^{-1} \longrightarrow \infty \quad \text{for } i \neq j$$

with the convention that $g_0 = 1_\Gamma$, for any sets $A_0, \dots, A_k \in \mathcal{B}$; $g_n \rightarrow \infty$ in Γ means that for any finite set $F \subset \Gamma$ there is an N with $n > N \implies g_n \notin F$. For $k = 1$

the property is called simply *mixing*. This notion for single transformations goes back to the foundational work of Rohlin [78], where he showed that ergodic group endomorphisms are mixing of all orders (and so the notion is not useful for distinguishing between group endomorphisms as measurable dynamical systems). He raised the (still open) question of whether any measure-preserving transformation can be mixing without being mixing of all orders.

A class of group actions that are particularly easy to understand are the *algebraic dynamical systems* studied systematically by Schmidt [83]: here X is a compact abelian group, each T^g is a continuous automorphism of X , and μ is the Haar measure on X . Schmidt [82] related mixing properties of algebraic dynamical systems with $\Gamma = \mathbb{Z}^d$ to statements in arithmetic, and showed that a mixing action on a connected group could only fail to mix in a certain way. Later Schmidt and the author [85] showed that for X connected, mixing implies mixing of all orders. The proof proceeds by showing that the result is exactly equivalent to the following statement: if \mathbb{K} is a field of characteristic zero, and G is a finitely generated subgroup of the multiplicative group \mathbb{K}^\times , then the equation

$$a_1 x_1 + \dots + a_n x_n = 1 \quad (13)$$

for fixed $a_1, \dots, a_n \in \mathbb{K}^\times$ has a finite number of solutions $x_1, \dots, x_n \in G$ for which no subsum $\sum_{i \in I} a_i x_i$ with $I \subsetneq \{1, \dots, n\}$ vanishes. The bound on the number of solutions to (13) follows from the profound extensions to W. Schmidt's subspace theorem in Diophantine geometry [86] by Evertse and Schlickewei (see [28,81,102] for the details).

The argument in [85] may be cast as follows: failure of k -fold mixing in a connected algebraic dynamical system implies (via duality) an infinite set of solutions to an equation of the shape (13) in some field of characteristic zero. The S -unit theorem means that this can only happen if there is some proper subsum that vanishes infinitely often. This infinite family of solutions to a homogeneous form of (13) with fewer terms can then be translated back via duality to show that the system fails to mix for some strictly lower order, proving that mixing implies mixing of all orders by induction.

Mixing properties for algebraic dynamical systems without the assumption of connectedness are quite different, and in particular it is possible to have mixing actions that are not mixing of all orders. This is a simple consequence of the fact that the constituents of a disconnected algebraic dynamical system are associated with fields of positive characteristic, where the presence of the

Frobenius automorphism can prevent higher-order mixing. Ledrappier [57] pointed this out via examples of the following shape. Let

$$X = \left\{ x \in \mathbb{F}_2^{\mathbb{Z}^2} : x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod{2} \right\}$$

and define the \mathbb{Z}^2 -action T to be the natural shift action,

$$(T^{(n,m)}x)_{(a,b)} = x_{(a+n,b+m)}.$$

It is readily seen that this action is mixing with respect to the Haar measure. The condition $x_{(a+1,b)} + x_{(a,b)} + x_{(a,b+1)} = 0 \pmod{2}$ implies that, for any $k \geq 1$,

$$x_{(0,2^k)} = \sum_{j=0}^{2^k} \binom{2^k}{j} x_{(j,0)} = x_{(0,0)} + x_{(2^k,0)} \pmod{2} \quad (14)$$

since every entry in the 2^k th row of Pascal's triangle is even apart from the first and the last. Now let $A = \{x \in X : x_{(0,0)} = 0\}$ and let $x_* \in X$ be any element with $x_{(0,0)} = 1$. Then X is the disjoint union of A and $A + x_*$, so

$$\mu(A) = \mu(A + x_*) = \frac{1}{2}.$$

However, (14) shows that

$$x \in A \cap T_{-(2^k,0)}A \implies x \in T_{-(0,2^k)}A,$$

so

$$A \cap T_{-(2^k,0)}A \cap T_{-(0,2^k)}(A + x_*) = \emptyset$$

for all $k \geq 1$, which shows that T cannot be mixing on three sets.

The full picture of higher-order mixing properties on disconnected groups is rather involved; see Schmidt's monograph [83]. A simple illustration is the construction by Einsiedler and the author [23] of systems with any prescribed order of mixing. When such systems fail to be mixing of all orders, they fail in a very specific way – along dilates of a specific *shape* (a finite subset of \mathbb{Z}^d). In the example above, the shape that fails to mix is $\{(0,0), (1,0), (0,1)\}$. This gives an order of mixing as detected by shapes; computing this is in principle an algebraic problem. On the other hand, there is a more natural definition of the order of mixing, namely the largest k for which (12) holds; computing this is in principle a Diophantine problem. A conjecture emerged (formulated explicitly by Schmidt [84]) that for any algebraic dynamical system, if every set of cardinality $r \geq 2$ is a mixing shape, then the system is mixing on r sets.

This question motivated Masser [64] to prove an appropriate analogue of the S -unit theorem on the number of solutions to (13) in positive characteristic as follows. Let H be a multiplicative group and fix $n \in \mathbb{N}$. An infinite subset $A \subset H^n$ is called *broad* if it has both of the following properties:

- if $h \in H$ and $1 \leq j \leq n$, then there are at most finitely many (a_1, \dots, a_n) in A with $a_j = h$;
- if $n \geq 2$, $h \in H$ and $1 \leq i < j \leq n$ then there are at most finitely many $(a_1, \dots, a_n) \in H$ with $a_i a_j^{-1} = h$.

Then Masser's theorem says the following. Let \mathbb{K} be a field of characteristic $p > 0$, let G be a finitely-generated subgroup of \mathbb{K}^\times and suppose that the equation

$$a_1 x_1 + \dots + a_n x_n = 1$$

has a broad set of solutions $(x_1, \dots, x_n) \in G^n$ for some constants $a_1, \dots, a_n \in \mathbb{K}^\times$. Then there is an $m \leq n$, constants $b_1, \dots, b_m \in \mathbb{K}^\times$ and some $(g_1, \dots, g_m) \in G^m$ with the following properties:

- $g_j \neq 1$ for $1 \leq j \leq m$;
- $g_i g_j^{-1} \neq 1$ for $1 \leq i < j \leq m$;
- there are infinitely many k for which

$$b_1 g_1^k + b_2 g_2^k + \dots + b_m g_m^k = 1.$$

The proof that shapes detect the order of mixing in algebraic dynamics then proceeds much as in the connected case.

Future Directions

The interaction between ergodic theory, number theory and combinatorics continues to expand rapidly, and many future directions of research are discussed in the articles ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#), ► [Ergodic Theory: Rigidity](#) and ► [Ergodic Theory: Recurrence](#). Some of the directions most relevant to the examples discussed in this article include the following.

The recent developments mentioned in Subsect. “[Sets of Primes](#)” clearly open many exciting prospects involving finding new structures in arithmetically significant sets (like the primes). The original conjecture of Erdős and Turán [26] asked if $\sum_{a \in A \subset \mathbb{N}} \frac{1}{a} = \infty$ is sufficient to force the set A to contain arbitrary long arithmetic progressions, and remains open. This would of course imply both Szemerédi's theorem [97] and the result of Green and Tao [39] on arithmetic progressions in the primes. More generally, it is clear that there is still much to come from the dialogue subsuming the four parallel proofs of

Szemerédi's: one by purely combinatorial methods, one by ergodic theory, one by hypergraph theory, and one by Fourier analysis and additive combinatorics. For an overview, see the survey papers of Tao [98,99,100].

In the context of the orbit-counting results in Sect. "Orbit-Counting as an Analogous Development", a natural problem is to on the one hand obtain finer asymptotics with better control of the error terms, and on the other to extend the situations that can be handled. In particular, relaxing the hypotheses related to hyperbolicity (or negative curvature) is a constant challenge.

Bibliography

Primary Literature

- Adler RL, Weiss B (1970) Similarity of automorphisms of the torus. *Memoirs of the American Mathematical Society*, No. 98. American Mathematical Society, Providence
- Anosov DV (1967) Geodesic flows on closed Riemannian manifolds of negative curvature. *Trudy Mat Inst Steklov* 90:209
- Arnol'd VI, Avez A (1968) *Ergodic problems of classical mechanics*. Translated from the French by A Avez. Benjamin Inc, New York
- Auslander L, Green L, Hahn F (1963) Flows on homogeneous spaces. *Annals of Mathematics Studies*, No 53. Princeton University Press, Princeton
- Baker A (1990) *Transcendental number theory*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2nd edn
- Benford F (1938) The law of anomalous numbers. *Proc Amer Philos Soc* 78:551–572
- Bergelson V (1996) Ergodic Ramsey theory – an update. In: *Ergodic theory of \mathbf{Z}^d actions* (Warwick, 1993–1994), vol 228. London Math Soc Lecture Note Ser. Cambridge University Press, Cambridge, pp 1–61
- Bergelson V (2000) Ergodic theory and Diophantine problems. In: *Topics in symbolic dynamics and applications* (Temuco, 1997), vol 279. London Math Soc Lecture Note Ser. Cambridge University Press, Cambridge, pp 167–205
- Bergelson V (2003) Minimal idempotents and ergodic Ramsey theory. In: *Topics in dynamics and ergodic theory*, vol 310. London Math Soc, Lecture Note Series, Cambridge University Press, Cambridge, pp 8–39
- Bergelson V (2006) Combinatorial and Diophantine applications of ergodic theory. In: *Handbook of dynamical systems*, vol 1B. Elsevier, Amsterdam, pp 745–869
- Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J Amer Math Soc* 9(3):725–753
- Bergelson V, McCutcheon R (2000) An ergodic IP polynomial Szemerédi theorem. *Mem Amer Math Soc* 146(695):viii–106
- Birkhoff GD (1931) Proof of the ergodic theorem. *Proc Natl Acad Sci USA* 17:656–660
- Bohl P (1909) Über ein in der Theorie der säkularen Störungen vorkommendes Problem. *J Math* 135:189–283
- Borel E (1909) Les probabilités denombrables et leurs applications arithmetiques. *Rend Circ Math Palermo* 27:247–271
- Bourgain J (1988) An approach to pointwise ergodic theorems. In: *Geometric aspects of functional analysis* (1986/87), vol 1317. Lecture Notes in Math, pp 204–223. Springer, Berlin
- Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the integers. *Israel J Math* 61(1):39–72
- Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. *Amer J Math* 92:725–747
- Bowen R (1972) The equidistribution of closed geodesics. *Amer J Math* 94:413–423
- Bowen R (1973) Symbolic dynamics for hyperbolic flows. *Amer J Math* 95:429–460
- Chothi V, Everest G, Ward T (1997) S-integer dynamical systems: periodic points. *J Reine Angew Math* 489:99–132
- Dolgopyat D (1998) On decay of correlations in Anosov flows. *Ann Math* 147(2):357–390
- Einsiedler M, Ward T (2003) Asymptotic geometry of non-mixing sequences. *Ergodic Theory Dyn Syst* 23(1):75–85
- Ellis R (1969) *Lectures on topological dynamics*. Benjamin Inc, New York
- Erdős P (1949) On a new method in elementary number theory which leads to an elementary proof of the prime number theorem. *Proc Natl Acad Sci USA* 35:374–384
- Erdős P, Turán P (1936) On some sequences of integers. *J London Math Soc* 11:261–264
- Everest G, Miles R, Stevens S, Ward T (2007) Orbit-counting in non-hyperbolic dynamical systems. *J Reine Angew Math* 608:155–182
- Evertse JH, Schlickewei HP (1999) The absolute subspace theorem and linear equations with unknowns from a multiplicative group. In: *Number theory in progress*, vol 1 (Zakopane-Kościelisko, 1997). de Gruyter, Berlin, pp 121–142
- Furstenberg H (1961) Strict ergodicity and transformation of the torus. *Amer J Math* 83:573–601
- Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J Analyse Math* 31:204–256
- Furstenberg H, Katznelson Y (1979) An ergodic Szemerédi theorem for commuting transformations. *J Analyse Math* 34:275–291
- Furstenberg H, Katznelson Y (1985) An ergodic Szemerédi theorem for IP-systems and combinatorial theory. *J Analyse Math* 45:117–168
- Furstenberg H, Katznelson Y, Ornstein D (1982) The ergodic theoretical proof of Szemerédi's theorem. *Bull Amer Math Soc* 7(3):527–552
- Furstenberg H, Weiss B (1979) Topological dynamics and combinatorial number theory. *J Analyse Math* 34:61–85
- Gel'fond AO (1960) *Transcendental and algebraic numbers*. Translated from the first Russian edition by Leo F Boron. Dover Publications, New York
- Goldfeld D (2004) The elementary proof of the prime number theorem: an historical perspective. In: *Number theory* (2003). Springer, New York, pp 179–192
- Goldston DA, Yıldırım CY (2005) Small gaps between primes I. *arXiv:math.NT/0504336*
- Gowers WT (2007) Hypergraph regularity and the multidimensional Szemerédi Theorem. *Ann of Math* 166:897–946
- Green B, Tao T (2004) The primes contain arbitrarily long arithmetic progressions. *arXiv:math.NT/0404188*
- Green B, Tao T (2006) Linear equations in primes. *arXiv:math.NT/0606088*

41. Hejhal DA (1976) The Selberg trace formula and the Riemann zeta function. *Duke Math J* 43(3):441–482
42. Hill TP (1995) Base-invariance implies Benford's law. *Proc Amer Math Soc* 123(3):887–895
43. Hindman N (1974) Finite sums from sequences within cells of a partition of \mathbb{N} . *J Comb Theory Ser A* 17:1–11
44. Hlawka E (1964) Discrepancy and uniform distribution of sequences. *Compositio Math* 16:83–91
45. Host B, Kra B (2005) Convergence of polynomial ergodic averages. *Israel J Math* 149:1–19, *Probability in mathematics*
46. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. *Ann Math* 161(1):397–488
47. Huber H (1959) Zur analytischen Theorie hyperbolischen Raumformen und Bewegungsgruppen. *Math Ann* 138:1–26
48. Katok A (1980) Lyapunov exponents, entropy and periodic orbits for diffeomorphisms. *Inst Hautes Études Sci Publ Math* (51):137–173
49. Katsuda A, Sunada T (1990) Closed orbits in homology classes. *Inst Hautes Études Sci Publ Math* (71):5–32
50. Khinchin AI (1964) Continued fractions. The University of Chicago Press, Chicago
51. Knieper G (1997) On the asymptotic geometry of nonpositively curved manifolds. *Geom Funct Anal* 7(4):755–782
52. Koopman B (1931) Hamiltonian systems and transformations in Hilbert spaces. *Proc Natl Acad Sci USA* 17:315–318
53. Kra B (2006) The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view. *Bull Amer Math Soc* 43(1):3–23
54. Kuz'min RO (1928) A problem of Gauss. *Dokl Akad Nauk*, pp 375–380
55. Lalley SP (1987) Distribution of periodic orbits of symbolic and Axiom A flows. *Adv Appl Math* 8(2):154–193
56. Lalley SP (1988) The “prime number theorem” for the periodic orbits of a Bernoulli flow. *Amer Math Monthly* 95(5):385–398
57. Ledrappier F (1978) Un champ markovien peut être d'entropie nulle et mélangeant. *C R Acad Sci Paris Sér A-B* 287(7):A561–A563
58. Leibman A (2005) Convergence of multiple ergodic averages along polynomials of several variables. *Israel J Math* 146:303–315
59. Levy P (1929) Sur les lois de probabilité dont dependent les quotients complets et incomplets d'une fraction continue. *Bull Soc Math France* 57:178–194
60. Levy P (1936) Sur quelques points de la théorie des probabilités dénombrables. *Ann Inst H Poincaré* 6(2):153–184
61. Lind DA (1982) Dynamical properties of quasihyperbolic toral automorphisms. *Ergodic Theory Dyn Syst* 2(1):49–68
62. Margulis GA (1969) Certain applications of ergodic theory to the investigation of manifolds of negative curvature. *Funkcional Anal i Prilozhen* 3(4):89–90
63. Margulis GA (2004) On some aspects of the theory of Anosov systems. *Springer Monographs in Mathematics*. Springer, Berlin.
64. Masser DW (2004) Mixing and linear equations over groups in positive characteristic. *Israel J Math* 142:189–204
65. Mertens F (1874) Ein Beitrag zur analytischen Zahlentheorie. *J Reine Angew Math* 78:46–62
66. Nagle B, Rödl V, Schacht M (2006) The counting lemma for regular k -uniform hypergraphs. *Random Structures Algorithms* 28(2):113–179
67. Newcomb S (1881) Note on the frequency of the use of digits in natural numbers. *Amer J Math* 4(1):39–40
68. Noorani MS (1999) Mertens' theorem and closed orbits of ergodic toral automorphisms. *Bull Malaysian Math Soc* 22(2):127–133
69. Oxtoby JC (1952) Ergodic sets. *Bull Amer Math Soc* 58:116–136
70. Parry W (1969) Ergodic properties of affine transformations and flows on nilmanifolds. *Amer J Math* 91:757–771
71. Parry W (1983) An analogue of the prime number theorem for closed orbits of shifts of finite type and their suspensions. *Israel J Math* 45(1):41–52
72. Parry W (1984) Bowen's equidistribution theory and the Dirichlet density theorem. *Ergodic Theory Dyn Syst* 4(1):117–134
73. Parry W, Pollicott M (1983) An analogue of the prime number theorem for closed orbits of Axiom A flows. *Ann Math* 118(3):573–591
74. Poincaré H (1890) Sur le problème des trois corps et les équations de la Dynamique. *Acta Math* 13:1–270
75. Pollicott M, Sharp R (1998) Exponential error terms for growth functions on negatively curved surfaces. *Amer J Math* 120(5):1019–1042
76. Rado R (1933) Studien zur Kombinatorik. *Math Z* 36(1):424–470
77. Ratner M (1973) Markov partitions for Anosov flows on n -dimensional manifolds. *Israel J Math* 15:92–114
78. Rohlin VA (1949) On endomorphisms of compact commutative groups. *Izvestiya Akad Nauk SSSR Ser Mat* 13:329–340
79. Roth K (1952) Sur quelques ensembles d'entiers. *C R Acad Sci Paris* 234:388–390
80. Sárközy A (1978) On difference sets of sequences of integers. III. *Acta Math Acad Sci Hungar* 31(3–4):355–386
81. Schlickewei HP (1990) S -unit equations over number fields. *Invent Math* 102(1):95–107
82. Schmidt K (1989) Mixing automorphisms of compact groups and a theorem by Kurt Mahler. *Pacific J Math* 137(2):371–385
83. Schmidt K (1995) Dynamical systems of algebraic origin, vol 128. *Progress in Mathematics*. Birkhäuser, Basel
84. Schmidt K (2001) The dynamics of algebraic \mathbb{Z}^d -actions. In: *European Congress of Mathematics*, vol I (Barcelona, 2000), vol 201. *Progress in Math*. Birkhäuser, Basel, pp 543–553
85. Schmidt K, Ward T (1993) Mixing automorphisms of compact groups and a theorem of Schlickewei. *Invent Math* 111(1):69–76
86. Schmidt WM (1972) Norm form equations. *Ann Math* 96(2):526–551
87. Selberg A (1949) An elementary proof of the prime-number theorem. *Ann Math* 50(2):305–313
88. Selberg A (1956) Harmonic analysis and discontinuous groups in weakly symmetric Riemannian spaces with applications to Dirichlet series. *J Indian Math Soc* 20:47–87
89. Sharp R (1991) An analogue of Mertens' theorem for closed orbits of Axiom A flows. *Bol Soc Brasil Mat* 21(2):205–229
90. Sharp R (1993) Closed orbits in homology classes for Anosov flows. *Ergodic Theory Dyn Syst* 13(2):387–408
91. Sierpiński W (1910) Sur la valeur asymptotique d'une certaine somme. *Bull Intl Acad Polonaise des Sci et des Lettres (Cracovie)*, pp 9–11
92. Sinai JG (1966) Asymptotic behavior of closed geodesics on compact manifolds with negative curvature. *Izv Akad Nauk SSSR Ser Mat* 30:1275–1296

93. Sinai JG (1968) Construction of Markov partitionings. *Funkcional Anal i Priložen* 2(3):70–80
94. Silverman JH (2007) *The Arithmetic of Dynamical Systems*, vol 241. Graduate Texts in Mathematics. Springer, New York
95. Smale S (1967) Differentiable dynamical systems. *Bull Amer Math Soc* 73:747–817
96. Szemerédi E (1969) On sets of integers containing no four elements in arithmetic progression. *Acta Math Acad Sci Hungar* 20:89–104
97. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. *Acta Arith* 27:199–245
98. Tao T (2005) The dichotomy between structure and randomness, arithmetic progressions, and the primes. [arXiv:math/0512114v2](https://arxiv.org/abs/math/0512114v2)
99. Tao T (2006) Arithmetic progressions and the primes. *Collect Math*, vol extra, Barcelona, pp 37–88
100. Tao T (2007) What is good mathematics? [arXiv:math/0702396v1](https://arxiv.org/abs/math/0702396v1)
101. Tao T, Ziegler T (2006) The primes contain arbitrarily long polynomial progressions. [arXiv:math.NT/0610050](https://arxiv.org/abs/math.NT/0610050)
102. van der Poorten AJ, Schlickewei HP (1991) Additive relations in fields. *J Austral Math Soc Ser A* 51(1):154–170
103. van der Waerden BL (1927) Beweis einer Baudet'schen Vermutung. *Nieuw Arch Wisk* 15:212–216
104. van der Waerden BL (1971) How the proof of Baudet's conjecture was found. In: *Studies in Pure Mathematics* (Presented to Richard Rado). Academic Press, London, pp 251–260
105. von Neumann J (1932) Proof of the quasi-ergodic hypothesis. *Proc Natl Acad Sci USA* 18:70–82
106. Waddington S (1991) The prime orbit theorem for quasi-hyperbolic toral automorphisms. *Monatsh Math* 112(3):235–248
107. Ward T (1998) Almost all S -integer dynamical systems have many periodic points. *Ergodic Theory Dyn Syst* 18(2):471–486
108. Weil A (1967) *Basic number theory*. Die Grundlehren der mathematischen Wissenschaften, Band 144. Springer, New York
109. Weyl H (1910) Über die Gibbssche Erscheinung und verwandte Konvergenzphänomene. *Rendiconti del Circolo Matematico di Palermo* 30:377–407
110. Weyl H (1916) Über die Gleichverteilung von Zahlen mod Eins. *Math Ann* 77:313–352
111. Wiener N (1932) Tauberian theorems. *Ann Math* 33(1):1–100
- Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems, vol 54. In: *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge
- Krengel U (1985) *Ergodic theorems*, vol 6. de Gruyter Studies in Mathematics. de Gruyter, Berlin
- McCutcheon R (1999) *Elemental methods in ergodic Ramsey theory*, vol 1722. Lecture Notes in Mathematics. Springer, Berlin
- Petersen K (1989) *Ergodic theory*, vol 2. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge
- Schweiger F (1995) *Ergodic theory of fibred systems and metric number theory*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York
- Totoki H (1969) *Ergodic theory*. Lecture Notes Series, No 14. Matematisk Institut, Aarhus Universitet, Aarhus
- Walters P (1982) *An introduction to ergodic theory*, vol 79. Graduate Texts in Mathematics. Springer, New York

Ergodic Theory, Introduction to

BRYNA KRA

Northwestern University, Evanston, USA

Ergodic theory lies at the intersection of many areas of mathematics, including smooth dynamics, statistical mechanics, probability, harmonic analysis, and group actions. Problems, techniques, and results are related to many other areas of mathematics, and ergodic theory has had applications both within mathematics and to numerous other branches of science. Ergodic theory has particularly strong overlap with other branches of dynamical systems; to clarify what distinguishes it from other areas of dynamics, we start with a quick overview of dynamical systems.

Dynamical systems is the study of systems that evolve with time. The evolution of a dynamical system is given by some fixed rule that determines the states of the system a short time into the future, given only the present states. Reflecting the origins of the subject in celestial mechanics, the set of states through which the system evolves with time is called an orbit. Many important concepts in dynamical systems are related to understanding the orbits in the system: Do the orbits fill out the entire space? Do orbits collapse? Do orbits return to themselves? What are statistical properties of the orbits? Are orbits stable under perturbation? For simple dynamical systems, knowing the individual orbits is often sufficient to answer such questions. However, in most dynamical systems it is impossible to write down explicit formulae for orbits, and even when one can, many systems are too complicated to be understood just in terms of individual orbits. The orbits may only be known approximately, some orbits may appear to be random while others exhibit regular behavior, and vary-

Books and Reviews

- Cornfeld IP, Fomin SV, Sinai YG (1982) *Ergodic theory*. Springer, New York
- Dajani K, Kraaikamp C (2002) *Ergodic theory of numbers*, vol 29. In: *Carus Mathematical Monographs*. Mathematical Association of America, Washington DC
- Denker M, Grillenberger C, Sigmund K (1976) *Ergodic theory on compact spaces*. Lecture Notes in Mathematics, vol 527. Springer, Berlin
- Furstenberg H (1981) *Recurrence in ergodic theory and combinatorial number theory*. Princeton University Press, Princeton
- Glasner E (2003) *Ergodic theory via joinings*, vol 101. *Mathematical Surveys and Monographs*. American Mathematical Society, Providence
- Iosifescu M, Kraaikamp C (2002) *Metrical theory of continued fractions*, vol 547. *Mathematics and its Applications*. Kluwer, Dordrecht

ing the parameters defining the system may give rise to qualitatively different behaviors. The various branches of dynamical systems have been developed for understanding long term properties of the orbits.

To make the notion of a dynamical system more precise, let X denote the collection of all states of the system. The evolution of these states is given by some fixed rule $T: X \rightarrow X$, dictating where each state $x \in X$ is mapped. An application of the transformation $T: X \rightarrow X$ corresponds to the passage of a unit of time and for a positive integer n , the map $T^n = T \circ T \circ \dots \circ T$ denotes the composition of T with itself taken n times. Given a state $x \in X$, the orbit of the point x under the transformation T is the collection of iterates x, Tx, T^2x, \dots of the state x . Thus the single transformation T generates a semigroup of transformations acting on X , by considering the powers T^n . More generally, one can consider a family of transformations $\{T_t: t \in \mathbb{R}\}$ with each $T_t: X \rightarrow X$. Assuming that $T_0(x) = x$ and that $T_{t+s}(x) = T_t(T_s(x))$ for all states $x \in X$ and all real t and s , this models the evolution of continuous time in a system. Autonomous differential equations are examples of such continuous time systems.

In almost all cases of interest, the space X has some underlying structure which is preserved by the transformation T . Different underlying structures X and different properties of the transformation T give rise to different branches of dynamical systems. When X is a smooth manifold and $T: X \rightarrow X$ is a differentiable mapping, one is in the framework of differentiable dynamics. When X is a topological space and $T: X \rightarrow X$ is a continuous map, one is in the framework of topological dynamics. When X is a measure space and $T: X \rightarrow X$ is a measure preserving map, one is in the framework of ergodic theory. These categories are not mutually exclusive, and the relations among them are deep and interesting. Some of these relations are explored in the articles [► Topological Dynamics](#), [► Symbolic Dynamics](#), and [► Smooth Ergodic Theory](#).

To further explain the role of ergodic theory, a few definitions are needed. The state space X is assumed to be a measure space, endowed with a σ -algebra \mathcal{B} of measurable sets and a measure μ . The measure μ assigns each set $B \in \mathcal{B}$ a non-negative number (its measure), and usually one assumes that $\mu(X) = 1$ (thus μ is a probability measure). The transformation $T: X \rightarrow X$ is assumed to be a measurable and measure preserving map: For all $B \in \mathcal{B}$, $\mu(T^{-1}B) = \mu(B)$. The quadruple (X, \mathcal{B}, μ, T) is called a measure preserving system. For the precise definitions and background on measure preserving transformations, see the article [► Measure Preserving Systems](#). An extensive discussion of examples of measure preserving systems and basic constructions used in ergodic theory is

given in [► Ergodic Theory: Basic Examples and Constructions](#).

The origins of ergodic theory are in the nineteenth century work of Boltzmann on the foundations of statistical mechanics. Boltzmann hypothesized that for large systems of interacting particles in equilibrium, the “time average” is equal to the “space average”. This question can be reformulated in the context of modern terminology. Assume that (X, \mathcal{B}, μ, T) is a measure preserving system and that $f: X \rightarrow \mathbb{R}$ is some measurement taken on the system. Thus if $x \in X$ is a state, evaluating the sequence

$$f(x), f(Tx), f(T^2x), \dots$$

can be viewed as successive values of this measurement. Boltzmann’s question can be phrased as: Under what conditions is the time mean equal to the space mean? In short, when does

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x) = \int_X f \, d\mu ?$$

Boltzmann hypothesized that if orbits went “everywhere” in the space, then such a conclusion would hold.

The study of the equality of space and time averages has been a major direction of research in ergodic theory. The long term behavior of the average

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=0}^{N-1} f(T^n x),$$

and especially the existence of this limit, is a basic question. Roughly speaking, the ergodic theorem states that starting at almost any initial point, the distribution of its iterates obeys some asymptotic law. This, and more general convergence questions, are addressed in the article [► Ergodic Theorems](#).

Perhaps the earliest result in ergodic theory is the Poincaré Recurrence Theorem: In a finite measure space, some iterate of any set with positive measure intersects the set itself in a set of positive measure. More generally, the qualitative behavior of orbits is used to understand conditions under which the time average is equal to the space average of the system (see the article [► Ergodic Theory: Recurrence](#)). If the time average is equal almost everywhere to the space average, then the system is said to be ergodic. Ergodicity is a key notion, giving a simple expression for the time average of an arbitrary function. Moreover, using the Ergodic Decomposition Theorem, the study of arbitrary measure preserving systems can be reduced to ergodic ones. Ergodicity and related properties of a system

are discussed in the article ► [Ergodicity and Mixing Properties](#).

Another central problem in ergodic theory is the classification of measure preserving systems. There are various notions of equivalence, and a classical approach to checking if systems are equivalent is finding invariants that are preserved under the equivalence. This subject, including an introduction to Ornstein Theory, is covered in the article ► [Isomorphism Theory in Ergodic Theory](#). A map $T: X \rightarrow X$ determines an associated unitary operator $U = U_T$ defined on $L^2(X)$ by

$$U_T f(x) = f(Tx).$$

There are also numerical invariants that can be assigned to a system, for example entropy, and this is discussed in the article ► [Entropy in Ergodic Theory](#).

When two systems are not equivalent, one would like to understand what properties they do have in common. An essential tool in such a classification is the notion of joinings (see the article ► [Joinings in Ergodic Theory](#)). Roughly speaking, a joining is a way of embedding two systems in the same space. When this can be done in a nontrivial manner, one obtains information on properties shared by the systems.

Some systems have predictable behavior and can be classified according to the behavior of individual points and their iterates. Others have behavior that is too complex or unpredictable to be understood on the level of orbits. Ergodic theory provides a statistical understanding of such systems and this is discussed in the article ► [Chaos and Ergodic Theory](#). A prominent role in chaotic dynamical systems is played by one dimensional Gibbs measure and by equilibrium states and (see the article ► [Pressure and Equilibrium States in Ergodic Theory](#)). Rigidity theory addresses the opposite case, studying what kinds of properties in a system are obstructions to general chaotic behavior. The role of ergodic theory in this area is discussed in the article ► [Ergodic Theory: Rigidity](#).

Another important class of systems arises when one relaxes the condition that the transformation T preserves the measure of sets on X , only requiring that the transformation preserve the negligible sets. Such systems are discussed in ► [Joinings in Ergodic Theory](#).

Ergodic theory has seen a burst of recent activity, and most of this activity comes from interaction with other fields. Historically, ergodic theory has interacted with numerous fields, including other areas of dynamics, probability, statistical mechanics, and harmonic analysis. More recently, ergodic theory and its techniques have been imported into number theory and combinatorics, proving new results that have yet to be proved by other methods,

and in turn, combinatorial problems have given rise to new areas of research within ergodic theory itself. Problems related to Diophantine approximation are discussed in ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#) and ► [Ergodic Theory: Rigidity](#) and ones related to combinatorial problems are addressed in ► [Ergodic Theory: Interactions with Combinatorics and Number Theory](#) and in ► [Ergodic Theory: Recurrence](#). Interaction with problems that are geometric in nature, in particular dimension theory, is discussed in ► [Ergodic Theory: Fractal Geometry](#).

Ergodic Theory: Non-singular Transformations

ALEXANDRE I. DANILENKO¹, CESAR E. SILVA²

¹ Institute for Low Temperature Physics & Engineering,
Ukrainian National Academy of Sciences, Kharkov,
Ukraine

² Department of Mathematics, Williams College,
Williamstown, USA

Article Outline

Glossary
Definition of the Subject
Basic Results
Panorama of Examples
Mixing Notions and multiple recurrence
Topological Group $\text{Aut}(X, \mu)$
Orbit Theory
Smooth Nonsingular Transformations
Spectral Theory for Nonsingular Systems
Entropy and Other Invariants
Nonsingular Joinings and Factors
Applications. Connections with Other Fields
Concluding Remarks
Bibliography

Glossary

Nonsingular dynamical system Let (X, \mathcal{B}, μ) be a standard Borel space equipped with a σ -finite measure. A Borel map $T: X \rightarrow X$ is a *nonsingular transformation* of X if for any $N \in \mathcal{B}$, $\mu(T^{-1}N) = 0$ if and only if $\mu(N) = 0$. In this case the measure μ is called *quasi-invariant* for T ; and the quadruple (X, \mathcal{B}, μ, T) is called a *nonsingular dynamical system*. If $\mu(A) = \mu(T^{-1}A)$ for all $A \in \mathcal{B}$ then μ is said to be

invariant under T or, equivalently, T is *measure-preserving*.

Conservativeness T is *conservative* if for all sets A of positive measure there exists an integer $n > 0$ such that $\mu(A \cap T^{-n}A) > 0$.

Ergodicity T is *ergodic* if every measurable subset A of X that is invariant under T (i. e., $T^{-1}A = A$) is either μ -null or μ -conull. Equivalently, every Borel function $f: X \rightarrow \mathbb{R}$ such that $f \circ T = f$ is constant a. e.

Types II, II₁, II_∞ and III Suppose that μ is non-atomic and T ergodic (and hence conservative). If there exists a σ -finite measure ν on \mathcal{B} which is equivalent to μ and invariant under T then T is said to be of *type II*. It is easy to see that ν is unique up to scaling. If ν is finite then T is of *type II₁*. If ν is infinite then T is of *type II_∞*. If T is not of type II then T is said to be of *type III*.

Definition of the Subject

An abstract measurable dynamical system consists of a set X (phase space) with a transformation $T: X \rightarrow X$ (evolution law or time) and a finite or σ -finite measure μ on X that specifies a class of negligible subsets. Nonsingular ergodic theory studies systems where T respects μ in a weak sense: the transformation preserves only the class of negligible subsets but it may not preserve μ . This survey is about dynamics and invariants of nonsingular systems. Such systems model ‘non-equilibrium’ situations in which events that are impossible at some time remain impossible at any other time. Of course, the first question that arises is whether it is possible to find an equivalent invariant measure, i. e. pass to a hidden equilibrium without changing the negligible subsets? It turns out that there exist systems which do not admit an equivalent invariant finite or even σ -finite measure. They are of our primary interest here. In a way (Baire category) most of systems are like that.

Nonsingular dynamical systems arise naturally in various fields of mathematics: topological and smooth dynamics, probability theory, random walks, theory of numbers, von Neumann algebras, unitary representations of groups, mathematical physics and so on. They also can appear in the study of probability preserving systems: some criteria of mild mixing and distality, a problem of Furstenberg on disjointness, etc. We briefly discuss this in Sect. “Applications. Connections with Other Fields”. Nonsingular ergodic theory studies all of them from a general point of view:

- What is the qualitative nature of the dynamics?
- What are the orbits?
- Which properties are typical withing a class of systems?

- How do we find computable invariants to compare or distinguish various systems?

Typically there are two kinds of results: some are extensions to nonsingular systems of theorems for finite measure-preserving transformations (for instance, the entire Sect. “Basic Results”) and the other are about new properly ‘nonsingular’ phenomena (see Sect. “Mixing Notions and Multiple Recurrence” or Sect. “Orbit Theory”). Philosophically speaking, the dynamics of nonsingular systems is more diverse comparatively with their finite measure-preserving counterparts. That is why it is usually easier to construct counterexamples than to develop a general theory. Because of shortage of space we concentrate only on invertible transformations, and we have not included as many references as we had wished. Nonsingular endomorphisms and general group or semigroup actions are practically not considered here (with some exceptions in Sect. “Applications. Connections with Other Fields” devoted to applications). A number of open problems are scattered through the entire text.

We thank J. Aaronson, J.R. Choksi, V.Ya. Golodets, M. Lemańczyk, F. Parreau, E. Roy for useful remarks.

Basic Results

This section includes the basic results involving conservativeness and ergodicity as well as some direct nonsingular counterparts of the basic machinery from classic ergodic theory: mean and pointwise ergodic theorems, Rokhlin lemma, ergodic decomposition, generators, Glimm–Effros theorem and special representation of nonsingular flows. The historically first example of a transformation of type III (due to Ornstein) is also given here with full proof.

Nonsingular Transformations

In this paper we will consider only *invertible* nonsingular transformations, i. e. those which are bijections when restricted to an invariant Borel subset of full measure. Thus when we refer to a nonsingular dynamical system (X, \mathcal{B}, μ, T) we shall assume that T is an invertible nonsingular transformation. Of course, each measure ν on \mathcal{B} which is *equivalent* to μ , i. e. μ and ν have the same null sets, is also quasi-invariant under T . In particular, since μ is σ -finite, T admits an equivalent quasi-invariant probability measure. For each $i \in \mathbb{Z}$, we denote by ω_i^μ or ω_i the Radon–Nikodym derivative $d(\mu \circ T^i)/d\mu \in L^1(X, \mu)$. The derivatives satisfy the cocycle equation $\omega_{i+j}(x) = \omega_i(x) \omega_j(T^i x)$ for a. e. x and all $i, j \in \mathbb{Z}$.

Basic Properties of Conservativeness and Ergodicity

A measurable set W is said to be *wandering* if for all $i, j \geq 0$ with $i \neq j$, $T^{-i}W \cap T^{-j}W = \emptyset$. Clearly, if T has a wandering set of positive measure then it cannot be conservative. A nonsingular transformation T is *incompressible* if whenever $T^{-1}C \subset C$, then $\mu(C \setminus T^{-1}C) = 0$. A set W of positive measure is said to be *weakly wandering* if there is a sequence $n_i \rightarrow \infty$ such that $T^{n_i}W \cap T^{n_j}W = \emptyset$ for all $i \neq j$. Clearly, a finite measure-preserving transformation cannot have a weakly wandering set. Hajian and Kakutani [83] showed that a nonsingular transformation T admits an equivalent finite invariant measure if and only if T does not have a weakly wandering set.

Proposition 1 (see e. g. [123]) *Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system. The following are equivalent:*

- (i) T is conservative.
- (ii) For every measurable set A , $\mu(A \setminus \bigcup_{n=1}^{\infty} T^{-n}A) = 0$.
- (iii) T is incompressible.
- (iv) Every wandering set for T is null.

Since any finite measure-preserving transformation is incompressible, we deduce that it is conservative. This is the statement of the classical Poincaré recurrence lemma. If T is a conservative nonsingular transformation of (X, \mathcal{B}, μ) and $A \in \mathcal{B}$ a subset of positive measure, we can define an *induced transformation* T_A of the space $(A, \mathcal{B} \cap A, \mu \upharpoonright A)$ by setting $T_A x := T^n x$ if $n = n(x)$ is the smallest natural number such that $T^n x \in A$. T_A is also conservative. As shown in [179], if $\mu(X) = 1$ and T is conservative and ergodic, $\int_A \sum_{i=0}^{n(x)-1} \omega(x) d\mu(x) = 1$, which is a nonsingular version of the well-known Kac's formula.

Theorem 2 (Hopf Decomposition, see e. g. [3]) *Let T be a nonsingular transformation. Then there exist disjoint invariant sets $C, D \in \mathcal{B}$ such that $X = C \sqcup D$, T restricted to C is conservative, and $D = \bigsqcup_{n=-\infty}^{\infty} T^n W$, where W is a wandering set. If $f \in L^1(X, \mu)$, $f > 0$, then $C = \{x: \sum_{i=0}^{n(x)-1} f(T^i x) \omega_i(x) = \infty \text{ a. e.}\}$ and $D = \{x: \sum_{i=0}^{n(x)-1} f(T^i x) \omega_i(x) < \infty \text{ a. e.}\}$.*

The set C is called the *conservative part* of T and D is called the *dissipative part* of T .

If T is ergodic and μ is non-atomic then T is automatically conservative. The translation by 1 on the group \mathbb{Z} furnished with the counting measure is an example of an ergodic non-conservative (infinite measure-preserving) transformation.

Proposition 4 *Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system. The following are equivalent:*

- (i) T is conservative and ergodic.

- (ii) For every set A of positive measure, $\mu(X \setminus \bigcup_{n=1}^{\infty} T^{-n}A) = 0$. (In this case we will say A sweeps out.)
- (iii) For every measurable set A of positive measure and for a. e. $x \in X$ there exists an integer $n > 0$ such that $T^n x \in A$.
- (iv) For all sets A and B of positive measure there exists an integer $n > 0$ such that $\mu(T^{-n}A \cap B) > 0$.
- (v) If A is such that $T^{-1}A \subset A$, then $\mu(A) = 0$ or $\mu(A^c) = 0$.

This survey is mainly about systems of type III. For some time it was not quite obvious whether such systems exist at all. The historically first example was constructed by Ornstein in 1960.

Example 5 (Ornstein [149]) Let $A_n = \{0, 1, \dots, n\}$, $v_n(0) = 0.5$ and $v_n(i) = 1/(2n)$ for $0 < i \leq n$ and all $n \in \mathbb{N}$. Denote by (X, μ) the infinite product probability space $\bigotimes_{n=1}^{\infty} (A_n, v_n)$. Of course, μ is non-atomic. A point of X is an infinite sequence $x = (x_n)_{n=1}^{\infty}$ with $x_n \in A_n$ for all n . Given $a_1 \in A_1, \dots, a_n \in A_n$, we denote the cylinder $\{x = (x_i)_{i=1}^{\infty} \in X: x_1 = a_1, \dots, x_n = a_n\}$ by $[a_1, \dots, a_n]$. Define a Borel map $T: X \rightarrow X$ by setting

$$(Tx)_i = \begin{cases} 0, & \text{if } i < l(x) \\ x_i + 1, & \text{if } i = l(x) \\ x_i, & \text{if } i > l(x), \end{cases} \quad (1)$$

where $l(x)$ is the smallest number l such that $x_l \neq l$. It is easy to verify that T is a nonsingular transformation of (X, μ) and

$$\begin{aligned} \frac{d\mu \circ T}{d\mu}(x) &= \prod_{n=1}^{\infty} \frac{v_n((Tx)_n)}{v_n(x_n)} \\ &= \begin{cases} (l(x)-1)!, & \text{if } x_{l(x)} = 0 \\ l(x), & \text{if } x_{l(x)} \neq 0. \end{cases} \end{aligned}$$

We prove that T is of type III by contradiction. Suppose that there exists a T -invariant σ -finite measure ν equivalent to μ . Let $\varphi := d\mu/d\nu$. Then

$$\omega_i^\mu(x) = \varphi(x) \varphi(T^i x)^{-1} \text{ for a. a. } x \in X \text{ and all } i \in \mathbb{Z}. \quad (2)$$

Fix a real $C > 1$ such that the set $E_C := \varphi^{-1}([C^{-1}, C]) \subset X$ is of positive measure. By a standard approximation argument, for each sufficiently large n , there is a cylinder $[a_1, \dots, a_n]$ such that $\mu(E_C \cap [a_1, \dots, a_n]) > 0.9\mu([a_1, \dots, a_n])$. Since $v_{n+1}(0) = 0.5$, it follows that $\mu(E_C \cap [a_1, \dots, a_n, 0]) > 0.8\mu([a_1, \dots, a_n, 0])$. Moreover, by the pigeon hole principle there is $0 < i \leq n+1$ with $\mu(E_C \cap$

$[a_1, \dots, a_n, i] > 0.8\mu([a_1, \dots, a_n, i])$. Find $N_n > 0$ such that $T^{N_n}[a_1, \dots, a_n, 0] = [a_1, \dots, a_n, i]$. Since $\omega_{N_n}^\mu$ is constant on $[a_1, \dots, a_n, 0]$, there is a subset $E_0 \subset E_C \cap [a_1, \dots, a_n, 0]$ of positive measure such that $T^{N_n}E_0 \subset E_C \cap [a_1, \dots, a_n, i]$. Moreover, $\omega_{N_n}^\mu(x) = v_{n+1}(i)/v_{n+1}(0) = (n+1)^{-1}$ for a.a. $x \in [a_1, \dots, a_n, 0]$. On the other hand, we deduce from (2) that $\omega_{N_n}^\mu(x) \geq C^{-2}$ for all $x \in E_0$, a contradiction.

Mean and Pointwise Ergodic Theorems.

Rokhlin Lemma

Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system. Define a unitary operator U_T of $L^2(X, \mu)$ by setting

$$U_T f := \sqrt{\frac{d(\mu \circ T)}{d\mu}} \cdot f \circ T. \quad (3)$$

We note that U_T preserves the cone of positive functions $L^2_+(X, \mu)$. Conversely, every positive unitary operator in $L^2(X, \mu)$ that preserves $L^2_+(X, \mu)$ equals U_T for a μ -nonsingular transformation T .

Theorem 6 (von Neumann mean Ergodic Theorem, see e.g. [3]) *If T has no μ -absolutely continuous T -invariant probability, then $n^{-1} \sum_{i=0}^{n-1} U_T^i \rightarrow 0$ in the strong operator topology.*

Denote by \mathcal{I} the sub- σ -algebra of T -invariant sets. Let $\mathbb{E}_\mu[\cdot|\mathcal{I}]$ stand for the conditional expectation with respect to \mathcal{I} . Note that if T is ergodic, then $\mathbb{E}_\mu[f|\mathcal{I}] = \int f d\mu$. Now we state a nonsingular analogue of Birkhoff's pointwise ergodic theorem, due to Hurewicz [105] and in the form stated by Halmos [84].

Theorem 7 (Hurewicz pointwise Ergodic Theorem) *If T is conservative, $\mu(X) = 1$, $f, g \in L^1(X, \mu)$ and $g > 0$, then*

$$\frac{\sum_{i=0}^{n-1} f(T^i x) \omega_i(x)}{\sum_{i=0}^{n-1} g(T^i x) \omega_i(x)} \rightarrow \frac{\mathbb{E}_\mu[f|\mathcal{I}]}{\mathbb{E}_\mu[g|\mathcal{I}]} \text{ as } n \rightarrow \infty \text{ for a.e. } x.$$

A transformation T is *aperiodic* if the T -orbit of a.e. point from X is infinite. The following classical statement can be deduced easily from Proposition 1.

Lemma 8 (Rokhlin's lemma [161]) *Let T be an aperiodic nonsingular transformation. For each $\varepsilon > 0$ and integer $N > 1$ there exists a measurable set A such that the sets $A, TA, \dots, T^{N-1}A$ are disjoint and $\mu(A \cup TA \cup \dots \cup T^{N-1}A) > 1 - \varepsilon$.*

This lemma was refined later (for ergodic transformations) by Lehrer and Weiss as follows.

Theorem 9 (ε -free Rokhlin lemma [132]) *Let T be ergodic and μ non-atomic. Then for a subset $B \subset X$ and any N for which $\bigcup_{k=0}^{\infty} T^{-kN}(X \setminus B) = X$, there is a set A such that the sets $A, TA, \dots, T^{N-1}A$ are disjoint and $A \cup TA \cup \dots \cup T^{N-1}A \supset B$.*

The condition $\bigcup_{k=0}^{\infty} T^{-kN}(X \setminus B) = X$ holds of course for each $B \neq X$ if T is *totally ergodic*, i.e. T^p is ergodic for any p , or if N is prime.

Ergodic Decomposition

A proof of the following theorem may be found in [3].

Theorem 10 (Ergodic Decomposition Theorem) *Let T be a conservative nonsingular transformation on a standard probability space (X, \mathcal{B}, μ) . There there exists a standard probability space (Y, ν, \mathcal{A}) and a family of probability measures μ_y on (X, \mathcal{B}) , for $y \in Y$, such that*

- (i) *For each $A \in \mathcal{B}$ the map $y \mapsto \mu_y(A)$ is Borel and for each $A \in \mathcal{B}$*

$$\mu(A) = \int \mu_y(A) d\nu(y).$$

- (ii) *For $y, y' \in Y$ the measures μ_y and $\mu_{y'}$ are mutually singular.*
 (iii) *For each $y \in Y$ the transformation T is nonsingular and conservative, ergodic on (X, \mathcal{B}, μ_y) .*
 (iv) *For each $y \in Y$*

$$\frac{d\mu \circ T}{d\mu} = \frac{d\mu_y \circ T}{d\mu_y} \mu_{y-a.e.}.$$

- (v) *(Uniqueness) If there exists another probability space (Y', ν', \mathcal{A}') and a family of probability measures $\mu'_{y'}$ on (X, \mathcal{B}) , for $y' \in Y'$, satisfying (i)–(iv), then there exists a measure-preserving isomorphism $\theta: Y \rightarrow Y'$ such that $\mu_y = \mu'_{\theta y}$ for ν -a.e. y .*

It follows that if T preserves an equivalent σ -finite measure then the system $(X, \mathcal{B}, \mu_y, T)$ is of type II for a.a. y . The space (Y, ν, \mathcal{A}) is called *the space of T -ergodic components*.

Generators

It was shown in [157, 162] that a nonsingular transformation T on a standard probability space (X, \mathcal{B}, μ) has a *countable generator*, i.e. a countable partition \mathcal{P} so that $\bigvee_{n=-\infty}^{\infty} T^n \mathcal{P}$ generates the measurable sets. It was refined by Krengel [126]: if T is of

type II_∞ or III then there exists a generator P consisting of two sets only. Moreover, given a sub- σ -algebra $\mathcal{F} \subset \mathcal{B}$ such that $\mathcal{F} \subset T\mathcal{F}$ and $\bigcup_{k>0} T^k\mathcal{F} = \mathcal{B}$, the set $\{A \in \mathcal{F} \mid (A, X \setminus A) \text{ is a generator of } T\}$ is dense in \mathcal{F} . It follows, in particular, that T is isomorphic to the shift on $\{0, 1\}^{\mathbb{Z}}$ equipped with a quasi-invariant probability measure.

The Glimm–Effros Theorem

The classical Bogoliouboff–Krylov theorem states that each homeomorphism of a compact space admits an ergodic invariant probability measure [33]. The following statement by Glimm [76] and Effros [61] is a “nonsingular” analogue of that theorem. (We consider here only a particular case of \mathbb{Z} -actions.)

Theorem 11 *Let X be a Polish space and $T: X \rightarrow X$ an aperiodic homeomorphism. Then the following are equivalent:*

- (i) *T has a recurrent point x , i. e. $x = \lim_{n \rightarrow \infty} T^{n_i}x$ for a sequence $n_1 < n_2 < \dots$.*
- (ii) *There is an orbit of T which is not locally closed.*
- (iii) *There is no a Borel set which intersects each orbit of T exactly once.*
- (iv) *There is a continuous probability Borel measure μ on X such that (X, μ, T) is an ergodic nonsingular system.*

A natural question arises: under the conditions of the theorem how many such μ can exist? It turns out that there is a wealth of such measures. To state a corresponding result we first write an important definition.

Definition 12 Two nonsingular systems (X, \mathcal{B}, μ, T) and $(X, \mathcal{B}', \mu', T')$ are called *orbit equivalent* if there is a one-to-one bi-measurable map $\varphi: X \rightarrow X$ with $\mu' \circ \varphi \sim \mu$ and such that φ maps the T -orbit of x onto the T' -orbit of $\varphi(x)$ for a. a. $x \in X$.

The following theorem was proved in [116, 128] and [174].

Theorem 13 *Let (X, T) be as in Theorem 11. Then for each ergodic dynamical system (Y, \mathcal{C}, ν, S) of type II_∞ or III , there exist uncountably many mutually disjoint Borel measures μ on X such that (X, T, \mathcal{B}, μ) is orbit equivalent to (Y, \mathcal{C}, ν, S) .*

On the other hand, T may not have any finite invariant measure. Indeed, let T be an irrational rotation on the circle \mathbb{T} and X a non-empty T -invariant G_δ subset of \mathbb{T} of full Lebesgue measure. Let (X, T) contain a recurrent point. Then the unique ergodicity of (\mathbb{T}, T) implies that (X, T) has no finite invariant measures.

Let T be an aperiodic Borel transformation of a standard Borel space X . Denote by $\mathcal{M}(T)$ the set of all ergodic T -nonsingular continuous measures on X . Given $\mu \in \mathcal{M}(T)$, let $N(\mu)$ denote the family of all Borel μ -null subsets. Shelah and Weiss showed [178] that $\bigcap_{\mu \in \mathcal{M}(T)} N(\mu)$ coincides with the collection of all Borel T -wandering sets.

Special Representations of Ergodic Flows

Nonsingular flows ($= \mathbb{R}$ -actions) appear naturally in the study of orbit equivalence for systems of type III (see Sect. “Orbit Theory”). Here we record some basic notions related to nonsingular flows. Let (X, \mathcal{B}, μ) be a standard Borel space with a σ -finite measure μ on \mathcal{B} . A nonsingular flow on (X, μ) is a Borel map $S: X \times \mathbb{R} \ni (x, t) \mapsto S_t x \in X$ such that $S_t S_s = S_{t+s}$ for all $s, t \in \mathbb{R}$ and each S_t is a nonsingular transformation of (X, μ) . Conservativeness and ergodicity for flows are defined in a similar way as for transformations.

A very useful example of a flow is a flow built under a function. Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system and f a positive Borel function on X such that $\sum_{i=0}^{\infty} f(T^i x) = \sum_{i=0}^{\infty} f(T^{-i} x) = \infty$ for all $x \in X$. Set $X^f := \{(x, s): x \in X, 0 \leq s < f(x)\}$. Define μ^f to be the restriction of the product measure $\mu \times \text{Leb}$ on $X \times \mathbb{R}$ to X^f and define, for $t \geq 0$,

$$S_t^f(x, s) := \left(T^n x, s + t - \sum_{i=0}^{n-1} f(T^i x) \right),$$

where n is the unique integer that satisfies

$$\sum_{i=0}^{n-1} f(T^i x) < s + t \leq \sum_{i=0}^n f(T^i x).$$

A similar definition applies when $t < 0$. In particular, when $0 < s + t < \varphi(x)$, $S_t^f(x, s) = (x, s + t)$, so that the flow moves the point (x, s) up t units, and when it reaches $(x, \varphi(x))$ it is sent to $(Tx, 0)$. It can be shown that $S^f = (S_t^f)_{t \in \mathbb{R}}$ is a free μ^f -nonsingular flow and that it preserves μ^f if and only if T preserves μ [148]. It is called the *flow built under the function φ with the base transformation T* . Of course, S^f is conservative or ergodic if and only if so is T .

Two flows $S = (S_t)_{t \in \mathbb{R}}$ on (X, \mathcal{B}, μ) and $V = (V_t)_{t \in \mathbb{R}}$ on (Y, \mathcal{C}, ν) are said to be *isomorphic* if there exist invariant co-null sets $X' \subset X$ and $Y' \subset Y$ and an invertible nonsingular map $\rho: X' \rightarrow Y'$ that intertwines the actions of the flows: $\rho \circ S_t = V_t \circ \rho$ on X' for all t . The following nonsingular version of Ambrose–Kakutani representation theorem was proved by Krengel [120] and Kubo [130].

Theorem 14 Let S be a free nonsingular flow. Then it is isomorphic to a flow built under a function.

Rudolph showed that in the Ambrose–Kakutani theorem one can choose the function φ to take two values. Krenge [122] showed that this can also be assumed in the nonsingular case.

Panorama of Examples

This section is devoted entirely to examples of nonsingular systems. We describe here the most popular (and simple) constructions of nonsingular systems: odometers, nonsingular Markov odometers, tower transformations, rank-one and finite rank systems and nonsingular Bernoulli shifts.

Nonsingular Odometers

Given a sequence m_n of natural numbers, we let $A_n := \{0, 1, \dots, m_n - 1\}$. Let ν_n be a probability on A_n and $\nu_n(a) > 0$ for all $a \in A_n$. Consider now the infinite product probability space $(X, \mu) := \bigotimes_{n=1}^{\infty} (A_n, \nu_n)$. Assume that $\prod_{n=1}^{\infty} \max\{\nu_n(a) \mid a \in A_n\} = 0$. Then μ is non-atomic. Given $a_1 \in A_1, \dots, a_n \in A_n$, we denote by $[a_1, \dots, a_n]$ the cylinder $x = (x_i)_{i>0} \mid x_1 = a_1, \dots, x_n = a_n$. If $x \neq (0, 0, \dots)$, we let $l(x)$ be the smallest number l such that the l th coordinate of x is not $m_l - 1$. We define a Borel map $T: X \rightarrow X$ by (1) if $x \neq (m_1, m_2, \dots)$ and put $Tx := (0, 0, \dots)$ if $x = (m_1, m_2, \dots)$. Of course, T is isomorphic to a rotation on a compact monothetic totally disconnected Abelian group. It is easy to check that T is μ -nonsingular and

$$\begin{aligned} \frac{d\mu \circ T}{d\mu}(x) &= \prod_{n=1}^{\infty} \frac{\nu_n((Tx)_n)}{\nu_n(x_n)} \\ &= \frac{\nu_{l(x)}(x_{l(x)} + 1)}{\nu_{l(x)}(x_{l(x)})} \prod_{n=1}^{l(x)-1} \frac{\nu_n(0)}{\nu_n(m_n - 1)} \end{aligned}$$

for a. a. $x = (x_n)_{n>0} \in X$. It is also easy to verify that T is ergodic. It is called the *nonsingular odometer* associated to $(m_n, \nu_n)_{n=1}^{\infty}$. We note that Ornstein's transformation (Example 5) is a nonsingular odometer.

Markov Odometers

We define Markov odometers as in [54]. An ordered Bratteli diagram B [102] consists of

- (i) a vertex set V which is a disjoint union of finite sets $V^{(n)}$, $n \geq 0$, V_0 is a singleton;
- (ii) an edge set E which is a disjoint union of finite sets $E^{(n)}$, $n > 0$;

- (iii) source mappings $s_n: E^{(n)} \rightarrow V^{(n-1)}$ and range mappings $r_n: E^{(n)} \rightarrow V^{(n)}$ such that $s_n^{-1}(v) \neq \emptyset$ for all $v \in V^{(n-1)}$ and $r_n^{-1}(v) \neq \emptyset$ for all $v \in V^{(n)}$, $n > 0$;
- (iv) a partial order on E so that $e, e' \in E$ are comparable if and only if $e, e' \in E^{(n)}$ for some n and $r_n(e) = r_n(e')$.

A Bratteli compactum X_B of the diagram B is the space of infinite paths

$$\{x = (x_n)_{n>0} \mid x_n \in E^{(n)} \text{ and } r(x_n) = s(x_{n+1})\}$$

on B . X_B is equipped with the natural topology induced by the product topology on $\prod_{n>0} E^{(n)}$. We will assume always that the diagram is *essentially simple*, i. e. there is only one infinite path $x_{\max} = (x_n)_{n>0}$ with x_n maximal for all n and only one $x_{\min} = (x_n)_{n>0}$ with x_n minimal for all n . The Bratteli–Vershik map $T_B: X_B \rightarrow X_B$ is defined as follows: $Tx_{\max} = x_{\min}$. If $x = (x_n)_{n>0} \neq x_{\max}$ then let k be the smallest number such that x_k is not maximal. Let y_k be a successor of x_k . Let (y_1, \dots, y_k) be the unique path such that y_1, \dots, y_{k-1} are all minimal. Then we let $T_B x := (y_1, \dots, y_k, x_{k+1}, x_{k+2}, \dots)$. It is easy to see that T_B is a homeomorphism of X_B . Suppose that we are given a sequence $P^{(n)} = (P_{(v,e) \in V^{n-1} \times E^{(n)}}^{(n)})$ of stochastic matrices, i. e.

- (i) $P_{v,e}^{(n)} > 0$ if and only if $v = s_n(e)$ and
- (ii) $\sum_{\{e \in E^{(n)} \mid s_n(e)=v\}} P_{v,e}^{(n)} = 1$ for each $v \in V^{(n-1)}$.

For $e_1 \in E^{(1)}, \dots, e_n \in E^{(n)}$, let $[e_1, \dots, e_n]$ denote the cylinder $\{x = (x_j)_{j>0} \mid x_1 = e_1, \dots, x_n = e_n\}$. Then we define a *Markov measure* on X_B by setting

$$\mu_P([e_1, \dots, e_n]) = P_{s_1(e_1), e_1}^1 P_{s_2(e_2), e_2}^2 \cdots P_{s_n(e_n), e_n}^n$$

for each cylinder $[e_1, \dots, e_n]$. The dynamical system (X_B, μ_P, T_B) is called a *Markov odometer*. It is easy to see that every nonsingular odometer is a Markov odometer where the corresponding $V^{(n)}$ are all singletons.

Tower Transformations

This construction is a discrete analogue of flow under a function. Given a nonsingular dynamical system (X, μ, T) and a measurable map $f: X \rightarrow \mathbb{N}$, we define a new dynamical system (X^f, μ^f, T^f) by setting

$$\begin{aligned} X^f &:= \{(x, i) \in X \times \mathbb{Z}_+ \mid 0 \leq i < f(x)\}, \\ d\mu^f(x, i) &:= d\mu(x) \quad \text{and} \\ T^f(x, i) &:= \begin{cases} (x, i+1), & \text{if } i+1 < f(x) \\ (Tx, 0), & \text{otherwise.} \end{cases} \end{aligned}$$

Then T^f is μ^f -nonsingular and $(d\mu^f \circ T^f / d\mu^f)(x, i) = (d\mu \circ T / d\mu)(x)$ for a. a. $(x, i) \in X^f$. This transformation is called the (Kakutani) *tower over T with height function f* . It is easy to check that T^f is conservative if and only if T is conservative; T^f is ergodic if and only if T is ergodic; T^f is of type III if and only if T is of type III. Moreover, the induced transformation $(T^f)_{X \times \{0\}}$ is isomorphic to T . Given a subset $A \subset X$ of positive measure, T is the tower over the induced transformation T_A with the first return time to A as the height function.

Rank-One Transformations. Chacón Maps. Finite Rank

The definition uses the process of “cutting and stacking”. We construct by induction a sequence of columns C_n . A column C_n consists of a finite sequence of bounded intervals (left-closed, right-open) $C_n = \{I_{n,0}, \dots, I_{n,h_n-1}\}$ of height h_n . A column C_n determines a *column map* T_{C_n} that sends each interval $I_{n,i}$ to the interval above it $I_{n,i+1}$ by the unique orientation-preserving affine map between the intervals. T_{C_n} remains undefined on the top interval I_{n,h_n-1} . Set $C_0 = \{[0, 1]\}$ and let $\{r_n > 2\}$ be a sequence of positive integers, let $\{s_n\}$ be a sequence of functions $s_n: \{0, \dots, r_n - 1\} \rightarrow \mathbb{N}_0$, and let $\{w_n\}$ be a sequence of probability vectors on $\{0, \dots, r_n - 1\}$. If C_n has been defined, column C_{n+1} is defined as follows. First “cut” (i. e., subdivide) each interval $I_{n,i}$ in C_n into r_n subintervals $I_{n,i}[j]$, $j = 0, \dots, r_n - 1$, whose lengths are in the proportions $w_n(0): w_n(1): \dots: w_n(r_n - 1)$. Next place, for each $j = 0, \dots, r_n - 1$, $s_n(j)$ new subintervals above $I_{n,h_n-1}[j]$, all of the same length as $I_{n,h_n-1}[j]$. Denote these intervals, called *spacers*, by $S_{n,0}[j], \dots, S_{n,s_n(j)-1}[j]$. This yields, for each $j \in \{0, \dots, r_n - 1\}$, r_n subcolumns each consisting of the subintervals

$$I_{n,0}[j], \dots, I_{n,h_n-1}[j]$$

followed by the spacers $S_{n,0}[j], \dots, S_{n,s_n(j)-1}[j]$.

Finally each subcolumn is stacked from left to right so that the top subinterval in subcolumn j is sent to the bottom subinterval in subcolumn $j + 1$, for $j = 0, \dots, r_n - 2$ (by the unique orientation-preserving affine map between the intervals). For example, $S_{n,s_n(0)-1}[0]$ is sent to $I_{n,0}[1]$. This defines a new column C_{n+1} and new column map $T_{C_{n+1}}$, which remains undefined on its top subinterval. Let X be the union of all intervals in all columns and let μ be Lebesgue measure restricted to X . We assume that as $n \rightarrow \infty$ the maximal length of the intervals in C_n converges to 0, so we may define a transformation T of (X, μ) by $Tx := \lim_{n \rightarrow \infty} T_{C_n}x$. One can verify that T is well-defined a. e. and that it is nonsingular and ergodic. T is

said to be the *rank-one* transformation associated with $(r_n, w_n, s_n)_{n=1}^\infty$. If all the probability vectors w_n are uniform the resulting transformation is measure-preserving. The measure is infinite (σ -finite) if and only if the total mass of the spacers is infinite. In the case $r_n = 3$ and $s_n(0) = s_n(2) = 0$, $s_n(1) = 1$ for all $n \geq 0$, the associated rank-one transformation is called a *nonsingular Chacón map*.

It is easy to see that every nonsingular odometer is of rank-one (the corresponding maps s_n are all trivial). Each rank-one map T is a tower over a nonsingular odometer (to obtain such an odometer reduce T to a column C_n).

A rank N transformation is defined in a similar way. A nonsingular transformation T is said to be of *rank N or less* if at each stage of its construction there exists N disjoint columns, the levels of the columns generate the σ -algebra and the Radon–Nikodym derivative of T is constant on each non-top level of every column. T is said to be of *rank N* if it is of rank N or less and not of rank $N - 1$ or less. A rank N transformation, $N \geq 2$, need not be ergodic.

Nonsingular Bernoulli Transformations – Hamachi’s Example

A *nonsingular Bernoulli* transformation is a transformation T such that there exists a countable generator \mathcal{P} (see Subsect. “[Generators](#)”) such that the partitions $T^n \mathcal{P}$, $n \in \mathbb{Z}$, are mutually independent and such that the Radon–Nikodym derivative ω_1 is measurable with respect to the sub- σ -algebra $\bigvee_{n=-\infty}^0 T^n \mathcal{P}$.

In [87], Hamachi constructed examples of conservative nonsingular Bernoulli transformations, hence ergodic (see Subsect. “[Weak Mixing](#), [Mixing](#), [K-Property](#)”), with a 2-set generating partition that are of type III. Krenge [121] asked if there are of type II_∞ examples of nonsingular Bernoulli automorphisms and the question remains open. Hamachi’s construction is the left-shift on the space $X = \prod_{n=-\infty}^\infty \{0, 1\}$. The measure is a product $\mu = \prod_{n=-\infty}^\infty \mu_n$ where $\mu_n = (1/2, 1/2)$ for $n \geq 0$ and for $n < 0$ μ_n is chosen carefully alternating on large blocks between the uniform measure and different non-uniform measures. Kakutani’s criterion for equivalence of infinite product measures is used to verify that μ is nonsingular.

Mixing Notions and multiple recurrence

The study of mixing and multiple recurrence are central topics in classical ergodic theory [33,70]. Unfortunately, these notions are considerably less ‘smooth’ in the world of nonsingular systems. The very concepts of any kind of mixing and multiple recurrence are not well understood in view of their ambiguity. Below we discuss nonsingular

systems possessing a surprising diversity of such properties that seem equivalent but are different indeed.

Weak Mixing, Mixing, K -Property

Let T be an ergodic conservative nonsingular transformation. A number $\lambda \in \mathbb{C}$ is an L^∞ -eigenvalue for T if there exists a nonzero $f \in L^\infty$ so that $f \circ T = \lambda f$ a. e. It follows that $|\lambda| = 1$ and f has constant modulus, which we assume to be 1. Denote by $e(T)$ the set of all L^∞ -eigenvalues of T . T is said to be *weakly mixing* if $e(T) = \{1\}$. We refer to Theorem 2.7.1 in [3] for proof of the following Keane's ergodic multiplier theorem: given an ergodic probability preserving transformation S , the product transformation $T \times S$ is ergodic if and only if $\sigma_S(e(T)) = 0$, where σ_S denotes the measure of (reduced) maximal spectral type of the unitary U_S (see (3)). It follows that T is weakly mixing if and only if $T \times S$ is ergodic for every ergodic probability preserving S . While in the finite measure-preserving case this implies that $T \times T$ is ergodic, it was shown in [5] that there exists a weakly mixing nonsingular T with $T \times T$ not conservative, hence not ergodic. In [11], a weakly mixing T was constructed with $T \times T$ conservative but not ergodic. A nonsingular transformation T is said to be *doubly ergodic* if for all sets of positive measure A and B there exists an integer $n > 0$ such that $\mu(A \cap T^{-n}A) > 0$ and $\mu(A \cap T^{-n}B) > 0$. Furstenberg [70] showed that for finite measure-preserving transformations double ergodicity is equivalent to weak mixing. In [20] it is shown that for nonsingular transformations weak mixing does not imply double ergodicity and double ergodicity does not imply that $T \times T$ is ergodic.

T is said to have *ergodic index* k if the Cartesian product of k copies of T is ergodic but the product of $k + 1$ copies of T is not ergodic. If all finite Cartesian products of T are ergodic then T is said to have *infinite ergodic index*. Parry and Kakutani [113] constructed for each $k \in \mathbb{N} \cup \{\infty\}$, an infinite Markov shift of ergodic index k . A stronger property is *power weak mixing*, which requires that for all nonzero integers k_1, \dots, k_r the product $T^{k_1} \times \dots \times T^{k_r}$ is ergodic [47]. The following examples were constructed in [12,36,38]:

- (i) power weakly mixing rank-one transformations,
- (ii) non-power weakly mixing rank-one transformations with infinite ergodic index,
- (iii) non-power weakly mixing rank-one transformations with infinite ergodic index and such that $T^{k_1} \times \dots \times T^{k_r}$ are all conservative, $k_1, \dots, k_r \in \mathbb{Z}$,

of types II_∞ and III (and various subtypes of III, see Sect. "Orbit Theory"). Thus we have the following scale

of properties (equivalent to weak mixing in the probability preserving case), where every next property is strictly stronger than the previous ones:

$$\begin{aligned}
 T \text{ is weakly mixing} &\Leftarrow T \text{ is doubly ergodic} \\
 &\Leftarrow T \times T \text{ is ergodic} \\
 &\Leftarrow T \times T \times T \text{ is ergodic} \\
 &\Leftarrow \dots \\
 &\Leftarrow T \text{ has infinite ergodic index} \\
 &\Leftarrow T \text{ is power weakly mixing.}
 \end{aligned}$$

We also mention a recent example of a power weakly mixing transformation of type II_∞ which embeds into a flow [46].

We now consider several attempts to generalize the notion of (strong) mixing. Given a sequence of measurable sets $\{A_n\}$ let $\sigma_k(\{A_n\})$ denote the σ -algebra generated by A_k, A_{k+1}, \dots . A sequence $\{A_n\}$ is said to be *remotely trivial* if $\bigcap_{k=0}^\infty \sigma_k(\{A_n\}) = \{\emptyset, X\} \bmod \mu$, and it is *semi-remotely trivial* if every subsequence contains a subsequence that is remotely trivial. Krengel and Sucheston [124] define a nonsingular transformation T of a σ -finite measure space to be *mixing* if for every set A of finite measure the sequence $\{T^{-n}A\}$ is semi-remotely trivial, and *completely mixing* if $\{T^{-n}A\}$ is semi-remotely trivial for all measurable sets A . They show that T is completely mixing if and only if it is type II_1 and mixing for the equivalent finite invariant measure. Thus there are no type III and II_∞ completely mixing nonsingular transformations on probability spaces. We note that this definition of mixing in infinite measure spaces depends on the choice of measure inside the equivalence class (but it is independent if we replace the measure by an equivalent measure with the same collection of sets of finite measure).

Hajian and Kakutani showed [83] that an ergodic infinite measure-preserving transformation T is either of *zero type*: $\lim_{n \rightarrow \infty} \mu(T^{-n}A \cap A) = 0$ for all sets A of finite measure, or of *positive type*: $\limsup_{n \rightarrow \infty} \mu(T^{-n}A \cap A) > 0$ for all sets A of finite positive measure. T is *mixing* if and only if it is of zero type [124]. For $0 \leq \alpha \leq 1$ Kakutani suggested a related definition of α -type: an infinite measure preserving transformation is of α -type if $\limsup_{n \rightarrow \infty} \mu(A \cap T^n A) = \alpha \mu(A)$ for every subset A of finite measure. In [153] examples of ergodic transformations of any α -type and a transformation of not any type were constructed.

It may seem that mixing is stronger than any kind of nonsingular weak mixing considered above. However, it is not the case: if T is a weakly mixing infinite measure preserving transformation of zero type and S is an ergodic probability preserving transformation then $T \times S$ is er-

godic and of zero type. On the other hand, the L^∞ -spectrum $e(T \times S)$ is nontrivial, i. e. $T \times S$ is not weakly mixing, whenever S is not weakly mixing. We also note that there exist rank-one infinite measure-preserving transformations T of zero type such that $T \times T$ is not conservative (hence not ergodic) [11]. In contrast to that, if T is of positive type all of its finite Cartesian products are conservative [7]. Another result that suggests that there is no good definition of mixing in the nonsingular case was proved recently in [110]. It is shown there that while the mixing finite measure-preserving transformations are measurably sensitive, there exists no infinite measure-preserving system that is measurably sensitive. (Measurable sensitivity is a measurable version of the strong sensitive dependence on initial conditions – a concept from topological theory of chaos.)

A nonsingular transformation T of (X, \mathcal{B}, μ) is called *K-automorphism* [180] if there exists a sub- σ -algebra $\mathcal{F} \subset \mathcal{B}$ such that $T^{-1}\mathcal{F} \subset \mathcal{F}$, $\bigcap_{k \geq 0} T^{-k}\mathcal{F} = \{\emptyset, X\}$, $\bigvee_{k=0}^{+\infty} T^k\mathcal{F} = \mathcal{B}$ and the Radon-Nikodym derivative $d\mu \circ T/d\mu$ is \mathcal{F} -measurable (see also [156] for the case when T is of type II_∞ ; the authors in [180] required T to be conservative). Evidently, a nonsingular Bernoulli transformation (see Subject. “[Nonsingular Bernoulli Transformations – Hamachi’s Example](#)”) is a *K-automorphism*. Parry [156] showed that a type II_∞ *K-automorphism* is either dissipative or ergodic. Krengel [121] proved the same for a class of Bernoulli nonsingular transformations, and finally Silva and Thieullen extended this result to nonsingular *K-automorphisms* [180]. It is also shown in [180] that if T is a nonsingular *K-automorphism*, for any ergodic nonsingular transformation S , if $S \times T$ is conservative, then it is ergodic. It follows that a conservative nonsingular *K-automorphism* is weakly mixing. However, it does not necessarily have infinite ergodic index [113]. Krengel and Sucheston [124] showed that an infinite measure-preserving conservative *K-automorphism* is mixing.

Multiple and Polynomial Recurrence

Let p be a positive integer. A nonsingular transformation T is called *p-recurrent* if for every subset B of positive measure there exists a positive integer k such that

$$\mu(B \cap T^{-k}B \cap \dots \cap T^{-kp}B) > 0.$$

If T is *p-recurrent* for any $p > 0$, then it is called *multiply recurrent*. It is easy to see that T is 1-recurrent if and only if it is conservative. T is called *rigid* if $T^{n_k} \rightarrow \text{Id}$ for a sequence $n_k \rightarrow \infty$. Clearly, if T is rigid then it is multiply recurrent. Furstenberg showed [70] that every finite measure-preserving transformation is multiply recurrent.

In contrast to that Eigen, Hajian and Halverson [64] constructed for any $p \in \mathbb{N} \cup \{\infty\}$, a nonsingular odometer of type II_∞ which is *p-recurrent* but not $(p+1)$ -recurrent. Aaronson and Nakada showed in [7] that an infinite measure preserving Markov shift T is *p-recurrent* if and only if the product $T \times \dots \times T$ (p times) is conservative. It follows from this and [5] that in the class of ergodic Markov shifts infinite ergodic index implies multiple recurrence. However, in general this is not true. It was shown in [12,45] and [82] that for each $p \in \mathbb{N} \cup \{\infty\}$ there exist

- (i) power weakly mixing rank-one transformations and
- (ii) non-power weakly mixing rank-one transformations with infinite ergodic index

which are *p-recurrent* but not $(p+1)$ -recurrent (the latter holds when $p \neq \infty$, of course).

A subset A is called *p-wandering* if $\mu(A \cap T^k A \cap \dots \cap T^{pk} A) = 0$ for each k . Aaronson and Nakada established in [7] a *p*-analogue of Hopf decomposition (see Theorem 2).

Proposition 15 *If (X, \mathcal{B}, μ, T) is conservative aperiodic nonsingular dynamical system and $p \in \mathbb{N}$ then $X = C_p \cup D_p$, where C_p and D_p are T -invariant disjoint subsets, D_p is a countable union of *p-wandering* sets, $T \upharpoonright C_p$ is *p-recurrent* and $\sum_{k=1}^{\infty} \mu(B \cap T^{-k}B \cap \dots \cap T^{-dk}B) = \infty$ for every $B \subset C_p$.*

Let T be an infinite measure-preserving transformation and let \mathcal{F} be a σ -finite factor (i. e., invariant subalgebra) of T . Inoue [106] showed that for each $p > 0$, if $T \upharpoonright \mathcal{F}$ is *p-recurrent* then so is T provided that the extension $T \rightarrow T \upharpoonright \mathcal{F}$ is isometric. It is unknown yet whether the latter assumption can be dropped. However, partial progress was recently achieved in [140]: if $T \upharpoonright \mathcal{F}$ is multiply recurrent then so is T .

Let $\mathcal{P} := \{q \in \mathbb{Q}[t] \mid q(\mathbb{Z}) \subset \mathbb{Z} \text{ and } q(0) = 0\}$. An ergodic conservative nonsingular transformation T is called *p-polynomially recurrent* if for every $q_1, \dots, q_p \in \mathcal{P}$ and every subset B of positive measure there exists $k \in \mathbb{N}$ with

$$\mu(B \cap T^{q_1(k)}B \cap \dots \cap T^{q_p(k)}B) > 0.$$

If T is *p-polynomially recurrent* for every $p \in \mathbb{N}$ then it is called *polynomially recurrent*. Furstenberg’s theorem on multiple recurrence was significantly strengthened in [17], where it was shown that every finite measure-preserving transformation is polynomially recurrent. However, Danilenko and Silva [45] constructed

- (i) nonsingular transformations T which are *p-polynomially recurrent* but not $(p+1)$ -polynomially recurrent (for each fixed $p \in \mathbb{N}$),

- (ii) polynomially recurrent transformations T of type Π_∞ ,
- (iii) rigid (and hence multiply recurrent) transformations T which are not polynomially recurrent.

Moreover, such T can be chosen inside the class of rank-one transformations with infinite ergodic index.

Topological Group $\text{Aut}(X, \mu)$

Let (X, \mathcal{B}, μ) be a standard probability space and let $\text{Aut}(X, \mu)$ denote the group of all nonsingular transformations of X . Let ν be a finite or σ -finite measure equivalent to μ ; the subgroup of the ν -preserving transformations is denoted by $\text{Aut}_0(X, \nu)$. Then $\text{Aut}(X, \mu)$ is a simple group [62] and it has no outer automorphisms [63]. Ryzhikov showed [169] that every element of this group is a product of three involutions (i. e. transformations of order 2). Moreover, a nonsingular transformation is a product of two involutions if and only if it is conjugate to its inverse by an involution.

Inspired by [85], Ionescu Tulcea [107] and Chacon and Friedman [21] introduced the *weak* and the *uniform* topologies respectively on $\text{Aut}(X, \mu)$. The weak one – we denote it by d_w – is induced from the weak operator topology on the group of unitary operators in $L^2(X, \mu)$ by the embedding $T \mapsto U_T$ (see Subsect. “[Mean and Pointwise Ergodic Theorems. Rokhlin Lemma](#)”). Then $(\text{Aut}(X, \mu), d_w)$ is a Polish topological group and $\text{Aut}_0(X, \nu)$ is a closed subgroup of $\text{Aut}(X, \mu)$. This topology will not be affected if we replace μ with any equivalent measure. We note that T_n weakly converges to T if and only if $\mu(T_n^{-1}A \Delta T^{-1}A) \rightarrow 0$ for each $A \in \mathcal{B}$ and $d(\mu \circ T_n)/d\mu \rightarrow d(\mu \circ T)/d\mu$ in $L^1(X, \mu)$. Danilenko showed in [34] that $(\text{Aut}(X, \mu), d_w)$ is contractible. It follows easily from the Rokhlin lemma that periodic transformations are dense in $\text{Aut}(X, \mu)$.

For each $p \geq 1$, one can also embed $\text{Aut}(X, \mu)$ into the isometry group of $L^p(X, \mu)$ via a formula similar to (3) but with another power of the Radon–Nikodym derivative in it. The strong operator topology on the isometry group induces the very same weak topology on $\text{Aut}(X, \mu)$ for all $p \geq 1$ [24].

It is natural to ask which properties of nonsingular transformations are typical in the sense of Baire category. The following technical lemma (see [24,68]) is an indispensable tool when considering such problems.

Lemma 16 *The conjugacy class of each aperiodic transformation T is dense in $\text{Aut}(X, \mu)$ endowed with the weak topology.*

Using this lemma and the Hurewicz ergodic theorem Choksi and Kakutani [24] proved that the ergodic transformations form a dense G_δ in $\text{Aut}(X, \mu)$. The same holds for the subgroup $\text{Aut}_0(X, \nu)$ [24,170]. Combined with [107] the above implies that the ergodic transformations of type III is a dense G_δ in $\text{Aut}(X, \mu)$. For further refinement of this statement we refer to Sect. “[Orbit Theory](#)”.

Since the map $T \mapsto T \times \cdots \times T$ (p times) from $\text{Aut}(X, \mu)$ to $\text{Aut}(X^p, \mu^{\otimes p})$ is continuous for each $p > 0$, we deduce that the set \mathcal{E}_∞ of transformations with infinite ergodic index is a G_δ in $\text{Aut}(X, \mu)$. It is non-empty by [113]. Since this \mathcal{E}_∞ is invariant under conjugacy, it is dense in $\text{Aut}(X, \mu)$ by Lemma 16. Thus we obtain that \mathcal{E}_∞ is a dense G_δ . In a similar way one can show that $\mathcal{E}_\infty \cap \text{Aut}_0(X, \nu)$ is a dense G_δ in $\text{Aut}_0(X, \nu)$ (see also [24,26,170] for original proofs of these claims).

The rigid transformations form a dense G_δ in $\text{Aut}(X, \mu)$. It follows that the set of multiply recurrent nonsingular transformations is residual [13]. A finer result was established in [45]: the set of polynomially recurrent transformations is also residual.

Given $T \in \text{Aut}(X, \mu)$, we denote the *centralizer* $\{S \in \text{Aut}(X, \mu) \mid ST = TS\}$ of T by $C(T)$. Of course, $C(T)$ is a closed subgroup of $\text{Aut}(X, \mu)$ and $C(T) \supset \{T^n \mid n \in \mathbb{Z}\}$. The following problems solved recently (by the efforts of many authors) for probability preserving systems are still open for the nonsingular case. Are the properties:

- (i) T has square root;
- (ii) T embeds into a flow;
- (iii) T has non-trivial invariant sub- σ -algebra;
- (iv) $C(T)$ contains a torus of arbitrary dimension

typical (residual) in $\text{Aut}(X, \mu)$?

The *uniform* topology on $\text{Aut}(X, \mu)$, finer than d_w , is defined by the metric

$$d_u(T, S) = \mu(\{x: Tx \neq Sx\}) + \mu(\{x: T^{-1}x \neq S^{-1}x\}).$$

This topology is also complete metric. It depends only on the measure class of μ . However the uniform topology is not separable and that is why it is of less importance in ergodic theory. We refer to [21,24,27] and [68] for the properties of d_u .

Orbit Theory

Orbit theory is, in a sense, the most complete part of nonsingular ergodic theory. We present here the seminal Krieger’s theorem on orbit classification of ergodic nonsingular transformations in terms of ratio sets and associated flows. Examples of transformations of various types

III_λ , $0 \leq \lambda \leq 1$ are also given here. Next, we consider the outer conjugacy problem for automorphisms of the orbit equivalence relations. This problem is solved in terms of a simple complete system of invariants. We discuss also a general theory of cocycles (of nonsingular systems) taking values in locally compact Polish groups and present an important orbit classification theorem for cocycles. This theorem is an analogue of the aforementioned result of Krieger. We complete the section by considering ITPFI-systems and their relation to AT-flows.

Full Groups. Ratio Set and Types III_λ , $0 \leq \lambda \leq 1$

Let T be a nonsingular transformation of a standard probability space (X, \mathcal{B}, μ) . Denote by $\text{Orb}_T(x)$ the T -orbit of x , i. e. $\text{Orb}_T(x) = \{T^n x \mid n \in \mathbb{Z}\}$. The *full group* $[T]$ of T consists of all transformations $S \in \text{Aut}(X, \mu)$ such that $Sx \in \text{Orb}_T(x)$ for a. a. x . If T is ergodic then $[T]$ is topologically simple (or even algebraically simple if T is not of type II_∞) [62]. It is easy to see that $[T]$ endowed with the uniform topology d_u is a Polish group. If T is ergodic then $([T], d_u)$ is contractible [34].

The *ratio set* $r(T)$ of T was defined by Krieger [126] and as we shall see below it is the key concept in the orbit classification (see Definition 1). The ratio set is a subset of $[0, +\infty)$ defined as follows: $t \in r(T)$ if and only if for every $A \in \mathcal{B}$ of positive measure and each $\epsilon > 0$ there is a subset $B \subset A$ of positive measure and an integer $k \neq 0$ such that $T^k B \subset A$ and $|\omega_k^\mu(x) - t| < \epsilon$ for all $x \in B$. It is easy to verify that $r(T)$ depends only on the equivalence class of μ and not on μ itself. A basic fact is that $1 \in r(T)$ if and only if T is conservative. Assume now T to be conservative and ergodic. Then $r(T) \cap (0, +\infty)$ is a closed subgroup of the multiplicative group $(0, +\infty)$. Hence $r(T)$ is one of the following sets:

- (i) $\{1\}$;
- (ii) $\{0, 1\}$; in this case we say that T is of type III_0 ,
- (iii) $\{\lambda^n \mid n \in \mathbb{Z}\} \cup \{0\}$ for $0 < \lambda < 1$; then we say that T is of type III_λ ,
- (iv) $[0, +\infty)$; then we say that T is of type III_1 .

Krieger showed that $r(T) = \{1\}$ if and only if T is of type II. Hence we obtain a further subdivision of type III into subtypes III_0 , III_λ , or III_1 .

Example 17 (i) Fix $\lambda \in (0, 1)$. Let $v_n(0) := 1/(1 + \lambda)$ and $v_n(1) := \lambda/(1 + \lambda)$ for all $n = 1, 2, \dots$. Let T be the nonsingular odometer associated with the sequence $(2, v_n)_{n=1}^\infty$ (see Subsect. “Nonsingular Odometers”). We claim that T is of type III_λ . Indeed, the group Σ of finite permutations of \mathbb{N} acts on X by $(\sigma x)_n = x_{\sigma^{-1}(n)}$, for all $n \in \mathbb{N}$, $\sigma \in \Sigma$ and $x = (x_n)_{n=1}^\infty \in X$. This action preserves μ . Moreover,

it is ergodic by the Hewitt–Savage 0–1 law. It remains to notice that $(d\mu \circ T/d\mu)(x) = \lambda$ on the cylinder $[0]$ which is of positive measure.

(ii) Fix positive reals ρ_1 and ρ_2 such that $\log \rho_1$ and $\log \rho_2$ are rationally independent. Let $v_n(0) := 1/(1 + \rho_1 + \rho_2)$, $v_n(1) := \rho_1/(1 + \rho_1 + \rho_2)$ and $v_n(2) := \rho_2/(1 + \rho_1 + \rho_2)$ for all $n = 1, 2, \dots$. Then the nonsingular odometer associated with the sequence $(3, v_n)_{n=1}^\infty$ is of type III_1 . This can be shown in a similar way as (i).

Non-singular odometer of type III_0 will be constructed in Example 19 below.

Maharam Extension, Associated Flow and Orbit Classification of Type III Systems

On $X \times \mathbb{R}$ with the σ -finite measure $\mu \times \kappa$, where $d\kappa(y) = \exp(y)dy$, consider the transformation

$$\tilde{T}(x, y) := \left(Tx, y - \log \frac{d\mu \circ T}{d\mu}(x) \right).$$

We call it the *Maharam extension* of T (see [136], where these transformations were introduced). It is measure-preserving and it commutes with the flow $S_t(x, y) := (x, y + t)$, $t \in \mathbb{R}$. It is conservative if and only if T is conservative [136]. However \tilde{T} is not necessarily ergodic. Let (Z, ν) denote the space of \tilde{T} -ergodic components. Then $(S_t)_{t \in \mathbb{R}}$ acts nonsingularly on this space. The restriction of $(S_t)_{t \in \mathbb{R}}$ to (Z, ν) is called the *associated flow* of T . The associated flow is ergodic whenever T is ergodic. It is easy to verify that the isomorphism class of the associated flow is an invariant of the orbit equivalence of the underlying system.

Proposition 18 ([90])

- (i) T is of type II if and only if its associated flow is the translation on \mathbb{R} , i. e. $x \mapsto x + t$, $x, t \in \mathbb{R}$,
- (ii) T is of type III_λ , $0 \leq \lambda < 1$ if and only if its associated flow is the periodic flow on the interval $[0, -\log \lambda)$, i. e. $x \mapsto x + t \pmod{-\log \lambda}$,
- (iii) T is of type III_1 if and only if its associated flow is the trivial flow on a singleton or, equivalently, \tilde{T} is ergodic,
- (iv) T is of type III_0 if and only if its associated flow is non-transitive.

Example 19 Let $A_n = \{0, 1, \dots, 2^{2^n}\}$ and $v_n(0) = 0.5$ and $v_n(i) = 0.5 \cdot 2^{-2^n}$ for all $0 < i \leq 2^n$. Let T be the nonsingular odometer associated with $(2^{2^n} + 1, v_n)_{n=0}^\infty$. It is straightforward that the associated flow of T is the flow built under the constant function 1 with the probability preserving 2-adic odometer (associated with $(2, \kappa_n)_{n=1}^\infty$,

$\kappa_n(0) = \kappa_n(1) = 0.5$) as the base transformation. In particular, T is of type III_0 .

A natural problem arises: to compute Krieger's type (or the ratio set) for the nonsingular odometers – the simplest class of nonsingular systems. Some partial progress was achieved in [56,141,152], etc. However in the general setting this problem remains open.

The map $\Psi: \text{Aut}(X, \mu) \ni T \mapsto \tilde{T} \in \text{Aut}(X \times \mathbb{R}, \mu \times \kappa)$ is a continuous group homomorphism. Since the set \mathcal{E} of ergodic transformations on $X \times \mathbb{R}$ is a G_δ in $\text{Aut}(X \times \mathbb{R}, \mu \times \kappa)$ (see Sect. “Topological Group $\text{Aut}(X, \mu)$ ”), the subset $\Psi^{-1}(\mathcal{E})$ of type III_1 ergodic transformations on X is also G_δ . The latter subset is non-empty in view of Example 17(ii). Since it is invariant under conjugacy, we deduce from Lemma 16 that the set of ergodic transformations of type III_1 is a dense G_δ in $(\text{Aut}(X, \mu), d_w)$ [23,159].

Now we state the main result of this section – Krieger's theorem on orbit classification for ergodic transformations of type III . It is a far reaching generalization of the basic result by H. Dye: any two ergodic probability preserving transformations are orbit equivalent [60].

Theorem 20 (Orbit equivalence for type III systems [125]–[129]) *Two ergodic transformations of type III are orbit equivalent if and only if their associated flows are isomorphic. In particular, for a fixed $0 < \lambda \leq 1$, any two ergodic transformations of type III_λ are orbit equivalent.*

The original proof of this theorem is rather complicated. Simpler treatment of it can be found in [90] and [117].

We also note that every nontransitive ergodic flow can be realized as the associated flow of a type III_0 transformation. However it is somewhat easier to construct a \mathbb{Z}^2 -action of type III_0 whose associated flow is the given one. For this, we take an ergodic nonsingular transformation Q on a probability space $(Z, \mathcal{B}, \lambda)$ and a measure-preserving transformation R of an infinite σ -finite measure space (Y, \mathcal{F}, ν) such that there is a continuous homomorphism $\pi: \mathbb{R} \rightarrow C(R)$ with $(d\nu \circ \pi(t)/d\nu)(y) = \exp(t)$ for a. a. y (for instance, take a type III_1 transformation T and put $R := \tilde{T}$ and $\pi(t) := S_t$). Let $\varphi: Z \rightarrow \mathbb{R}$ be a Borel map with $\inf_Z \varphi > 0$. Define two transformations R_0 and Q_0 of $(Z \times Y, \lambda \times \nu)$ by setting:

$$R_0(x, y) := (x, Ry), \quad Q_0(x, y) = (Qx, U_x y),$$

where $U_x = \pi(\varphi(x) - \log(d\mu \circ Q/d\mu)(x))$. Notice that R_0 and Q_0 commute. The corresponding \mathbb{Z}^2 -action generated by these transformations is ergodic. Take any transformation $V \in \text{Aut}(Z \times Y, \lambda \times \nu)$ whose orbits coincide

with the orbits of the \mathbb{Z}^2 -action. (According to [29], any ergodic nonsingular action of any countable amenable group is orbit equivalent to a single transformation.) Then V is of type III_0 . It is now easy to verify that the associated flow of V is the special flow built under $\varphi \circ Q^{-1}$ with the base transformation Q^{-1} . Since Q and φ are arbitrary, we deduce the following from Theorem 14.

Theorem 21 *Every nontransitive ergodic flow is an associated flow of an ergodic transformation of type III_0 .*

In [129] Krieger introduced a map Φ as follows. Let T be an ergodic transformation of type III_0 . Then the associated flow of T is a flow built under function with a base transformation $\Phi(T)$. We note that the orbit equivalence class of $\Phi(T)$ is well defined by the orbit equivalent class of T . If $\Phi^n(T)$ fails to be of type III_0 for some $1 \leq n < \infty$ then T is said to *belong to Krieger's hierarchy*. For instance, the transformation constructed in Example 19 belongs to Krieger's hierarchy. Connes gave in [28] an example of T such that $\Phi(T)$ is orbit equivalent to T (see also [73] and [90]). Hence T is not in Krieger's hierarchy.

Normalizer of the Full Group.

Outer Conjugacy Problem

Let

$$N[T] = \{R \in \text{Aut}(X, \mu) \mid R[T]R^{-1} = [T]\},$$

i. e. $N[T]$ is the *normalizer* of the full group $[T]$ in $\text{Aut}(X, \mu)$. We note that a transformation R belongs to $N[T]$ if and only if $R(\text{Orb}_T(x)) = \text{Orb}_T(Rx)$ for a. a. x . To define a topology on $N[T]$ consider the T -orbit equivalence relation $\mathcal{R}_T \subset X \times X$ and a σ -finite measure $\mu_{\mathcal{R}_T}$ on \mathcal{R}_T given by $\mu_{\mathcal{R}_T} = \int_X \sum_{y \in \text{Orb}_T(x)} \delta_{(x,y)} d\mu(x)$. For $R \in N[T]$, we define a transformation $i(R) \in \text{Aut}(\mathcal{R}_T, \mu_{\mathcal{R}_T})$ by setting $i(R)(x, y) := (Rx, Ry)$. Then the map $R \mapsto i(R)$ is an embedding of $N[T]$ into $\text{Aut}(\mathcal{R}_T, \mu_{\mathcal{R}_T})$. Denote by τ the topology on $N[T]$ induced by the weak topology on $\text{Aut}(\mathcal{R}_T, \mu_{\mathcal{R}_T})$ via i [34]. Then $(N[T], \tau)$ is a Polish group. A sequence R_n converges to R in $N[T]$ if $R_n \rightarrow R$ weakly (in $\text{Aut}(X, \mu)$) and $R_n T R_n^{-1} \rightarrow R T R^{-1}$ uniformly (in $[T]$).

Given $R \in N[T]$, denote by \tilde{R} the Maharam extension of R . Then $\tilde{R} \in N[\tilde{T}]$ and it commutes with $(S_t)_{t \in \mathbb{R}}$. Hence it defines a nonsingular transformation mod R on the space (Z, ν) of the associated flow $W = (W_t)_{t \in \mathbb{R}}$ of T . Moreover, mod R belongs to the centralizer $C(W)$ of W in $\text{Aut}(Z, \nu)$. Note that $C(W)$ is a closed subgroup of $(\text{Aut}(Z, \nu), d_w)$.

Let T be of type II_∞ and let μ' be the invariant σ -finite measure equivalent to μ . If $R \in N[T]$ then it is easy

to see that the Radon–Nikodym derivative $d\mu' \circ R/d\mu'$ is invariant under T . Hence it is constant, say c . Then $\text{mod } R = \log c$.

Theorem 22 ([86,90]) *If T is of type III then the map $\text{mod}: N[T] \rightarrow C(W)$ is a continuous onto homomorphism. The kernel of this homomorphism is the τ -closure of $[T]$. Hence the quotient group $N[T]/\overline{[T]}^\tau$ is (topologically) isomorphic to $C(W)$. In particular, $\overline{[T]}^\tau$ is co-compact in $N[T]$ if and only if W is a finite measure-preserving flow with a pure point spectrum.*

The following theorem describes the homotopical structure of normalizers.

Theorem 23 ([34]) *Let T be of type II or III_λ , $0 \leq \lambda < 1$. The group $\overline{[T]}^\tau$ is contractible. $N[T]$ is homotopically equivalent to $C(W)$. In particular, $N[T]$ is contractible if T is of type II. If T is of type III_λ with $0 < \lambda < 1$ then $\pi_1(N[T]) = \mathbb{Z}$.*

The outer period $p(R)$ of $R \in N[T]$ is the smallest positive integer n such that $R^n \in [T]$. We write $p(R) = 0$ if no such n exists.

Two transformations R and R' in $N[T]$ are called *outer conjugate* if there are transformations $V \in N[T]$ and $S \in [T]$ such that $VRV^{-1} = R'S$. The following theorem provides convenient (for verification) necessary and sufficient conditions for the outer conjugacy.

Theorem 24 ([30] for type II and [18] for type III) *Transformations $R, R' \in N[T]$ are outer conjugate if and only if $p(R) = p(R')$ and $\text{mod } R$ is conjugate to $\text{mod } R'$ in the centralizer of the associated flow for T .*

We note that in the case T is of type II, the second condition in the theorem is just $\text{mod } R = \text{mod } R'$. It is always satisfied when T is of type II_1 .

Cocycles of Dynamical Systems. Weak Equivalence of Cocycles

Let G be a locally compact Polish group and λ_G a left Haar measure on G . A Borel map $\varphi: X \rightarrow G$ is called a *cocycle* of T . Two cocycles φ and φ' are *cohomologous* if there is a Borel map $b: X \rightarrow G$ such that

$$\varphi'(x) = b(Tx)^{-1}\varphi(x)b(x)$$

for a.a. $x \in X$. A cocycle cohomologous to the trivial one is called a *coboundary*. Given a dense subgroup $G' \subset G$, then every cocycle is cohomologous to a cocycle with values in G' [81]. Each cocycle φ extends to a (unique) map $\alpha_\varphi: \mathcal{R}_T \rightarrow G$ such that $\alpha_\varphi(Tx, x) = \varphi(x)$ for a.a. x and $\alpha_\varphi(x, y)\alpha_\varphi(y, z) = \alpha_\varphi(x, z)$ for a.a. $(x, y), (y, z) \in \mathcal{R}_T$. α_φ is called the *cocycle of \mathcal{R}_T generated by φ* .

Moreover, φ and φ' are cohomologous via b as above if and only if α_φ and $\alpha_{\varphi'}$ are cohomologous via b , i.e. $\alpha_\varphi(x, y) = b(x)^{-1}\alpha_{\varphi'}(x, y)b(y)$ for $\mu_{\mathcal{R}_T}$ -a.a. $(x, y) \in \mathcal{R}_T$. The following notion was introduced by Golodets and Sinelshchikov [78,81]: two cocycles φ and φ' are *weakly equivalent* if there is a transformation $R \in N[T]$ such that the cocycles α_φ and $\alpha_{\varphi'} \circ (R \times R)$ of \mathcal{R}_T are cohomologous. Let $\mathcal{M}(X, G)$ denote the set of Borel maps from X to G . It is a Polish group when endowed with the topology of convergence in measure. Since T is ergodic, it is easy to deduce from Rokhlin's lemma that the cohomology class of any cocycle is dense in $\mathcal{M}(X, G)$. Given $\varphi \in \mathcal{M}(X, G)$, we define the φ -skew product extension T_φ of T acting on $(X \times G, \mu \times \lambda_G)$ by setting $T_\varphi(x, g) := (Tx, \varphi(x)g)$. Thus Maharam extension is (isomorphic to) the Radon–Nikodym cocycle-skew product extension. We now specify some basic classes of cocycles [19,35,81,173]:

- (i) φ is called *transient* if T_φ is of type I.
- (ii) φ is called *recurrent* if T_φ is conservative (equivalently, T_φ is not transient).
- (iii) φ has *dense range* in G if T_φ is ergodic.
- (iv) φ is called *regular* if φ cobounds with dense range into a closed subgroup H of G (then H is defined up to conjugacy).

These properties are invariant under the cohomology and the weak equivalence. The Radon–Nikodym cocycle ω_1 is a coboundary if and only if T is of type II. It is regular if and only if T is of type II or III_λ , $0 < \lambda \leq 1$. It has dense range (in the multiplicative group \mathbb{R}_+^*) if and only if T is of type III_1 . Notice that ω_1 is never transient (since T is conservative).

Schmidt introduced in [176] an invariant $R(\varphi) := \{g \in G \mid \varphi - g \text{ is recurrent}\}$. He showed in particular that

- (i) $R(\varphi)$ is a cohomology invariant,
- (ii) $R(\varphi)$ is a Borel set in G ,
- (iii) $R(\log \omega_1) = \{0\}$ for each aperiodic conservative T ,
- (iv) there are cocycles φ such that $R(\varphi)$ and $G \setminus R(\varphi)$ are dense in G ,
- (v) if $\mu(X) = 1$, $\mu \circ T = \mu$ and $\varphi: X \rightarrow \mathbb{R}$ is integrable then $R(\varphi) = \{\int \varphi d\mu\}$.

We note that (v) follows from Atkinson theorem [15]. A nonsingular version of this theorem was established in [183]: if T is ergodic and μ -nonsingular and $f \in L^1(\mu)$ then

$$\liminf_{n \rightarrow \infty} \left| \sum_{j=0}^{n-1} f(T^j x) \omega_j(x) \right| = 0 \quad \text{for a.a. } x$$

if and only if $\int f d\mu = 0$.

Since T_φ commutes with the action of G on $X \times G$ by inverted right translations along the second coordinate, this action induces an ergodic G -action $W_\varphi = (W_\varphi(g))_{g \in G}$ on the space (Z, ν) of T_φ -ergodic components. It is called the *Mackey range (or Poincaré flow)* of φ [66,135,173,188]. We note that φ is regular (and cobounds with dense range into $H \subset G$) if and only if W_φ is transitive (and H is the stabilizer of a point $z \in Z$, i. e. $H = \{g \in G \mid W_\varphi(g)z = z\}$). Hence every cocycle taking values in a compact group is regular.

It is often useful to consider the *double cocycle* $\varphi_0 := \varphi \times \omega_1$ instead of φ . It takes values in the group $G \times \mathbb{R}_+^*$. Since T_{φ_0} is exactly the Maharam extension of T_φ , it follows from [136] that φ_0 is transient or recurrent if and only if φ is transient or recurrent respectively.

Theorem 25 (Orbit classification of cocycles [81]) *Let $\varphi, \varphi': X \rightarrow G$ be two recurrent cocycles of an ergodic transformation T . They are weakly equivalent if and only if their Mackey ranges W_{φ_0} and $W_{\varphi'_0}$ are isomorphic.*

Another proof of this theorem was presented in [65].

Theorem 26 *Let T be an ergodic nonsingular transformation. Then there is a cocycle of T with dense range in G if and only if G is amenable.*

It follows that if G is amenable then the subset of cocycles of T with dense range in G is a dense G_δ in $\mathcal{M}(X, G)$ (just adapt the argument following Example 19). The ‘only if’ part of Theorem 26 was established in [187]. The ‘if’ part was considered by many authors in particular cases: G is compact [186], G is solvable or amenable almost connected [79], G is amenable unimodular [108], etc. The general case was proved in [78] and [100] (see also a recent treatment in [9]).

Theorem 21 is a particular case of the following result.

Theorem 27 ([10,65,80]) *Let G be amenable. Let V be an ergodic nonsingular action of $G \times \mathbb{R}_+^*$. Then there is an ergodic nonsingular transformation T and a recurrent cocycle φ of T with values in G such that V is isomorphic to the Mackey range of the double cocycle φ_0 .*

Given a cocycle $\varphi \in \mathcal{M}(X, G)$ of T , we say that a transformation $R \in N[T]$ is *compatible with φ* if the cocycles α_φ and $\alpha_\varphi \circ (R \times R)$ of \mathcal{R}_T are cohomologous. Denote by $D(T, \varphi)$ the group of all such R . It has a natural Polish topology which is stronger than τ [41]. Since $[T]$ is a normal subgroup in $D(T, \varphi)$, one can consider the outer conjugacy equivalence relation inside $D(T, \varphi)$. It is called *φ -outer conjugacy*. Suppose that G is Abelian. Then an analogue of Theorem 24 for the φ -outer conjugacy is established in [41]. Also, the cocycles φ with $D(T, \varphi) = N[T]$ are described there.

ITPFI Transformations and AT-Flows

A nonsingular transformation T is called *ITPFI*¹ if it is orbit equivalent to a nonsingular odometer (associated to a sequence $(m_n, \nu_n)_{n=1}^\infty$, see Subsect. “Nonsingular Odometers”). If the sequence m_n can be chosen bounded then T is called ITPFI of bounded type. If $m_n = 2$ for all n then T is called ITPFI₂. By [74], every ITPFI-transformation of bounded type is ITPFI₂. A remarkable characterization of ITPFI transformations in terms of their associated flows was obtained by Connes and Woods [31]. We first single out a class of ergodic flows. A nonsingular flow $V = (V_t)_{t \in \mathbb{R}}$ on a space (Ω, ν) is called *approximate transitive (AT)* if given $\epsilon > 0$ and $f_1, \dots, f_n \in L_+^1(X, \mu)$, there exists $f \in L_+^1(X, \mu)$ and $\lambda_1, \dots, \lambda_n \in L_+^1(\mathbb{R}, dt)$ such that

$$\left\| f_j - \int_{\mathbb{R}} f \circ V_t \frac{d\nu \circ V_t}{d\nu} \lambda_j(t) dt \right\|_1 < \epsilon$$

for all $1 \leq j \leq n$. A flow built under a constant ceiling function with a funny rank-one [67] probability preserving base transformation is AT [31]. In particular, each ergodic finite measure-preserving flow with a pure point spectrum is AT.

Theorem 28 ([31]) *An ergodic nonsingular transformation is ITPFI if and only if its associated flow is AT.*

The original proof of this theorem was given in the framework of von Neumann algebras theory. A simpler, purely measure theoretical proof was given later in [96] (the ‘only if’ part) and [88] (the ‘if’ part). It follows from Theorem 28 that every ergodic flow with pure point spectrum is the associated flow of an ITPFI transformation. If the point spectrum of V is $\theta\Gamma$, where Γ is a subgroup of \mathbb{Q} and $\theta \in \mathbb{R}$, then V is the associated flow of an ITPFI₂ transformation [91].

Theorem 29 ([54]) *Each ergodic nonsingular transformation is orbit equivalent to a Markov odometer (see Subsect. “Markov Odometers”).*

The existence of non-ITPFI transformations and ITPFI transformations of unbounded type was shown in [127]. In [55], an explicit example of a non-ITPFI Markov odometer was constructed.

Smooth Nonsingular Transformations

Diffeomorphisms of smooth manifolds equipped with smooth measures are commonly considered as physically natural examples of dynamical systems. Therefore

¹This abbreviates ‘infinite tensor product of factors of type I’ (came from the theory of von Neumann algebras).

the construction of smooth models for various dynamical properties is a well established problem of the modern (probability preserving) ergodic theory. Unfortunately, the corresponding ‘nonsingular’ counterpart of this problem is almost unexplored. We survey here several interesting facts related to the topic.

For $r \in \mathbb{N} \cup \{\infty\}$, denote by $\text{Diff}_+^r(\mathbb{T})$ the group of orientation preserving C^r -diffeomorphisms of the circle \mathbb{T} . Endow this set with the natural Polish topology. Fix $T \in \text{Diff}_+^r(\mathbb{T})$. Since $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, there exists a C^1 -function $f: \mathbb{R} \rightarrow \mathbb{R}$ such that $T(x + \mathbb{Z}) = f(x) + \mathbb{Z}$ for all $x \in \mathbb{R}$. The *rotation number* $\rho(T)$ of T is the limit

$$\lim_{n \rightarrow \infty} \underbrace{(f \circ \dots \circ f)(x)}_{n \text{ times}} \pmod{1}.$$

The limit exists and does not depend on the choice of x and f . It is obvious that T is nonsingular with respect to Lebesgue measure $\lambda_{\mathbb{T}}$. Moreover, if $T \in \text{Diff}_+^r(\mathbb{T})$ and $\rho(T)$ is irrational then the dynamical system $(\mathbb{T}, \lambda_{\mathbb{T}}, T)$ is ergodic [33]. It is interesting to ask: which Krieger’s type can such systems have?

Katznelson showed in [114] that the subset of type III C^∞ -diffeomorphisms and the subset of type II $_\infty$ C^∞ -diffeomorphisms are dense in $\text{Diff}_+^\infty(\mathbb{T})$. Hawkins and Schmidt refined the idea of Katznelson from [114] to construct, for every irrational number $\alpha \in [0, 1)$ which is not of constant type (i. e. in whose continued fraction expansion the denominators are not bounded) a transformation $T \in \text{Diff}_+^2(\mathbb{T})$ which is of type III $_\alpha$ and $\rho(T) = \alpha$ [97]. It should be mentioned that class C^2 in the construction is essential, since it follows from a remarkable result of Herman that if $T \in \text{Diff}_+^3(\mathbb{T})$ then under some condition on α (which determines a set of full Lebesgue measure), T is measure theoretically (and topologically) conjugate to a rotation by $\rho(T)$ [101]. Hence T is of type II $_\alpha$.

In [94], Hawkins shows that every smooth paracompact manifold of dimension ≥ 3 admits a type III $_\lambda$ diffeomorphism for every $\lambda \in [0, 1]$. This extends a result of Herman [100] on the existence of type III $_\alpha$ diffeomorphisms in the same circumstances.

It is also of interest to ask: which free ergodic flows are associated with smooth dynamical systems of type III $_\alpha$? Hawkins proved that any free ergodic C^∞ -flow on a smooth, connected, paracompact manifold is the associated flow for a C^∞ -diffeomorphism on another manifold (of higher dimension) [95].

A nice result was obtained in [115]: if $T \in \text{Diff}_+^2(\mathbb{T})$ and the rotation number of T has unbounded continued fraction coefficients then $(\mathbb{T}, \lambda_{\mathbb{T}}, T)$ is ITPFI. Moreover, a converse also holds: given a nonsingular odometer R , the set of orientation-preserving C^∞ -diffeomorphisms of

the circle which are orbit equivalent to R is C^∞ -dense in the Polish set of all C^∞ -orientation-preserving diffeomorphisms with irrational rotation numbers. In contrast to that, Hawkins constructs in [93] a type III $_\alpha$ C^∞ -diffeomorphism of the 4-dimensional torus which is not ITPFI.

Spectral Theory for Nonsingular Systems

While the spectral theory for probability preserving systems is developed in depth, the spectral theory of nonsingular systems is still in its infancy. We discuss below some problems related to L^∞ -spectrum which may be regarded as an analogue of the discrete spectrum. We also include results on computation of the maximal spectral type of the ‘nonsingular’ Koopman operator for rank-one nonsingular transformations.

L^∞ -Spectrum and Groups of Quasi-Invariance

Let T be an ergodic nonsingular transformation of (X, \mathcal{B}, μ) . A number $\lambda \in \mathbb{T}$ belongs to the L^∞ -spectrum $e(T)$ of T if there is a function $f \in L^\infty(X, \mu)$ with $f \circ T = \lambda f$. f is called an L^∞ -eigenfunction of T corresponding to λ . Denote by $\mathcal{E}(T)$ the group of all L^∞ -eigenfunctions of absolute value 1. It is a Polish group when endowed with the topology of converges in measure. If T is of type II $_\alpha$ then the L^∞ -eigenfunctions are $L^2(\mu')$ -eigenfunctions of T , where μ' is an equivalent invariant probability measure. Hence $e(T)$ is countable. Osikawa constructed in [151] the first examples of ergodic nonsingular transformations with uncountable $e(T)$.

We state now a nonsingular version of the von Neumann–Halmos discrete spectrum theorem. Let $Q \subset \mathbb{T}$ be a countable infinite subgroup. Let K be a compact dual of Q , where Q_d denotes Q with the discrete topology. Let $k_0 \in K$ be the element defined by $k_0(q) = q$ for all $q \in Q$. Let $R: K \rightarrow K$ be defined by $Rk = k + k_0$. The system (K, R) is called a *compact group rotation*. The following theorem was proved in [6].

Theorem 30 Assume that the L^∞ -eigenfunctions of T generate the entire σ -algebra \mathcal{B} . Then T is isomorphic to a compact group rotation equipped with an ergodic quasi-invariant measure.

A natural question arises: which subgroups of \mathbb{T} can appear as $e(T)$ for an ergodic T ?

Theorem 31 ([1, 143]) $e(T)$ is a Borel subset of \mathbb{T} and carries a unique Polish topology which is stronger than the usual topology on \mathbb{T} . The Borel structure of $e(T)$ under this topology agrees with the Borel structure inherited from \mathbb{T} . There is a Borel map $\psi: e(T) \ni \lambda \mapsto \psi_\lambda \in \mathcal{E}(T)$ such that $\psi_\lambda \circ T = \lambda \psi_\lambda$ for each λ . Moreover, $e(T)$ is of Lebesgue

measure 0 and it can have an arbitrary Hausdorff dimension.

A proper Borel subgroup E of \mathbb{T} is called

- (i) *weak Dirichlet* if $\limsup_{n \rightarrow \infty} \widehat{\lambda}(n) = 1$ for each finite complex measure λ supported on E ;
- (ii) *saturated* if $\limsup_{n \rightarrow \infty} |\widehat{\lambda}(n)| \geq |\lambda(E)|$ for each finite complex measure λ on \mathbb{T} , where $\widehat{\lambda}(n)$ denote the n th Fourier coefficient of λ .

Every countable subgroup of \mathbb{T} is saturated.

Theorem 32 $e(T)$ is σ -compact in the usual topology on \mathbb{T} [104] and saturated [104,139].

It follows that $e(T)$ is weak Dirichlet (this fact was established earlier in [175]).

It is not known if every Polish group continuously embedded in \mathbb{T} as a σ -compact saturated group is the eigenvalue group of some ergodic nonsingular transformation. This is the case for the so-called H_2 -groups and the groups of quasi-invariance of measures on \mathbb{T} (see below). Given a sequence n_j of positive integers and a sequence $a_j \geq 0$, the set of all $z \in \mathbb{T}$ such that $\sum_{j=1}^{\infty} a_j |1 - z^{n_j}|^2 < \infty$ is a group. It is called an H_2 -group. Every H_2 -group is Polish in an intrinsic topology stronger than the usual circle topology.

Theorem 33 ([104])

- (i) Every H_2 -group is a saturated (and hence weak Dirichlet) σ -compact subset of \mathbb{T} .
- (ii) If $\sum_{j=0}^{\infty} a_j = +\infty$ then the corresponding H_2 -group is a proper subgroup of \mathbb{T} .
- (iii) If $\sum_{j=0}^{\infty} a_j (n_j/n_{j+1})^2 < \infty$ then the corresponding H_2 -group is uncountable.
- (iv) Any H_2 -group is $e(T)$ for an ergodic nonsingular compact group rotation T .

It is an open problem whether every eigenvalue group $e(T)$ is an H_2 -group. It is known however that $e(T)$ is close 'to be an H_2 -group': if a compact subset $L \subset \mathbb{T}$ is disjoint from $e(T)$ then there is an H_2 -group containing $e(T)$ and disjoint from L .

Example 34 ([6], see also [151]) Let (X, μ, T) be the nonsingular odometer associated to a sequence $(2, v_j)_{j=1}^{\infty}$. Let n_j be a sequence of positive integers such that $n_j > \sum_{i < j} n_i$ for all j . For $x \in X$, we put $h(x) := n_{1(x)} - \sum_{j < 1(x)} n_j$. Then h is a Borel map from X to the positive integers. Let S be the tower over T with height function h (see Subsect. "Tower Transformations"). Then $e(S)$ is the H_2 -group of all $z \in \mathbb{T}$ with $\sum_{j=1}^{\infty} v_j(0)v_j(1)|1 - z^{n_j}|^2 < \infty$.

It was later shown in [104] that if $\sum_{j=1}^{\infty} v_j(0)v_j(1)(n_j/n_{j+1})^2 < \infty$ then the L^{∞} -eigenfunctions of S generate the entire σ -algebra, i. e. S is isomorphic (measure theoretically) to a nonsingular compact group rotation.

Let μ be a finite measure on \mathbb{T} . Let $H(\mu) := \{z \in \mathbb{Z} \mid \delta_z * \mu \sim \mu\}$, where $*$ means the convolution of measures. Then H_{μ} is a group called the *group of quasi-invariance of μ* . It has a Polish topology whose Borel sets agree with the Borel sets which $H(\mu)$ inherits from \mathbb{T} and the injection map of $H(\mu)$ into \mathbb{T} is continuous. This topology is induced by the weak operator topology on the unitary group in the Hilbert space $L^2(\mathbb{T}, \mu)$ via the map $H(\mu) \ni z \mapsto U_z$, $(U_z f)(x) = \sqrt{(d(\delta_z * \mu)/d\mu)(x)} f(xz)$ for $f \in L^2(\mathbb{T}, \mu)$. Moreover, $H(\mu)$ is saturated [104]. If $\mu(H(\mu)) > 0$ then either $H(\mu)$ is countable or μ is equivalent to $\lambda_{\mathbb{T}}$ [137].

Theorem 35 ([6]) Let μ be an ergodic with respect to the $H(\mu)$ -action by translations on \mathbb{T} . Then there is a compact group rotation (K, R) and a finite measure on K quasi-invariant and ergodic under R such that $e(R) = H(\mu)$. Moreover, there is a continuous one-to-one homomorphism $\psi: e(R) \rightarrow E(R)$ such that $\psi_{\lambda} \circ R = \lambda \psi_{\lambda}$ for all $\lambda \in e(R)$.

It was shown by Aaronson and Nadkarni [6] that if $n_1 = 1$ and $n_j = a_j a_{j-1} \cdots a_1$ for positive integers $a_j \geq 2$ with $\sum_{j=1}^{\infty} a_j^{-1} < \infty$ then the transformation S from Example 34 does not admit a continuous homomorphism $\psi: e(S) \rightarrow E(S)$ with $\psi_{\lambda} \circ T = \lambda \psi_{\lambda}$ for all $\lambda \in e(S)$. Hence $e(S) \neq H(\mu)$ for any measure μ satisfying the conditions of Theorem 35.

Assume that T is an ergodic nonsingular compact group rotation. Let \mathcal{B}_0 be the σ -algebra generated by a subcollection of eigenfunctions. Then \mathcal{B}_0 is invariant under T and hence a factor (see Sect. "Nonsingular Joinings and Factors") of T . It is not known if every factor of T is of this form. It is not even known whether every factor of T must have non-trivial eigenvalues.

Unitary Operator Associated with a Nonsingular System

Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system. In this subsection we consider spectral properties of the unitary operator U_T defined by (3). First, we note that the spectrum of T is the entire circle \mathbb{T} [147]. Next, if U_T has an eigenvector then T is of type II_1 . Indeed, if there are $\lambda \in \mathbb{T}$ and $0 \neq f \in L^2(X, \mu)$ with $U_T f = \lambda f$ then the measure ν , $d\nu(x) := |f(x)|^2 d\mu(x)$, is finite, T -invariant and equivalent to μ . Hence if T is of type III or II_{∞} then the maximal spectral type σ_T of U_T is continuous. Another 'restriction' on σ_T was recently found in [166]: no

Foiaş–Strătilă measure is absolutely continuous with respect to σ_T if T is of type Π_∞ . We recall that a symmetric measure on \mathbb{T} possesses *Foiaş–Strătilă property* if for each ergodic probability preserving system (Y, ν, S) and $f \in L^2(Y, \nu)$, if σ is the spectral measure of f then f is a Gaussian random variable [134]. For instance, measures supported on Kronecker sets possess this property.

Mixing is an L^2 -spectral property for type Π_∞ transformations: T is mixing if and only if σ_T is a Rajchman measure, i.e. $\widehat{\sigma}_T(n) := \int z^n d\sigma_T(z) \rightarrow 0$ as $|n| \rightarrow \infty$. Also, T is mixing if and only if $n^{-1} \sum_{i=0}^{n-1} U_T^{k_i} \rightarrow 0$ in the strong operator topology for each strictly increasing sequence $k_1 < k_2 < \dots$ [124]. This generalizes a well known theorem of Blum and Hanson for probability preserving maps. For comparison, we note that ergodicity is not an L^2 -spectral property of infinite measure preserving systems.

Now let T be a rank-one nonsingular transformation associated with a sequence $(r_n, w_n, s_n)_{n=1}^\infty$ as in Subsect. “Rank-One Transformations. Chacón Maps. Finite Rank”.

Theorem 36 ([25,104]) *The spectral multiplicity of U_T is 1 and the maximal spectral type σ_T of U_T (up to a discrete measure in the case T is of type Π_1) is the weak limit of the measures ρ_k defined as follows:*

$$d\rho_k(z) = \prod_{j=1}^k w_j(0) |P_j(z)|^2 dz,$$

where $P_j(z) := 1 + \sqrt{w_j(1)/w_j(0)} z^{-R_{1,j}} + \dots + \sqrt{w_j(m_j-1)/w_j(0)} z^{-R_{j-1,j}}$, $z \in \mathbb{T}$, $R_{i,j} := ih_{j-1} + s_j(0) + \dots + s_j(i)$, $1 \leq i \leq r_k - 1$ and h_j is the height of the j th column.

Thus the maximal spectral type of U_T is given by a so-called *generalized Riesz product*. We refer the reader to [25,103,104,148] for a detailed study of Riesz products: their convergence, mutual singularity, singularity to $\lambda_{\mathbb{T}}$, etc.

It was shown in [6] that $H(\sigma_T) \supset e(T)$ for any ergodic nonsingular transformation T . Moreover, σ_T is ergodic under the action of $e(T)$ by translations if T is isomorphic to an ergodic nonsingular compact group rotation. However it is not known:

- (i) Whether $H(\sigma_T) = e(T)$ for all ergodic T .
- (ii) Whether ergodicity of σ_T under $e(T)$ implies that T is an ergodic compact group rotation.

The first claim of Theorem 36 extends to the rank N nonsingular systems as follows: if T is an ergodic nonsingular transformation of rank N then the spectral multiplicity of U_T is bounded by N (as in the finite measure-preserving case). It is not known whether this claim is true for a more general class of transformations which are defined as rank N but without the assumption that the Radon–Nikodym cocycle is constant on the tower levels.

Entropy and Other Invariants

Let T be an ergodic conservative nonsingular transformation of a standard probability space (X, \mathcal{B}, μ) . If \mathcal{P} is a finite partition of X , we define the entropy $H(\mathcal{P})$ of \mathcal{P} as $H(\mathcal{P}) = -\sum_{P \in \mathcal{P}} \mu(P) \log \mu(P)$. In the study of measure-preserving systems the classical (Kolmogorov–Sinai) entropy proved to be a very useful invariant for isomorphism [33]. The key fact of the theory is that if $\mu \circ T = \mu$ then the limit $\lim_{n \rightarrow \infty} n^{-1} H(\bigvee_{i=1}^n T^{-i} \mathcal{P})$ exists for every \mathcal{P} . However if T does not preserve μ , the limit may no longer exist. Some efforts have been made to extend the use of entropy and similar invariants to the nonsingular domain. These include Krengel’s entropy of conservative measure-preserving maps and its extension to nonsingular maps, Parry’s entropy and Parry’s nonsingular version of Shannon–McMillan–Breiman theorem, critical dimension by Mortiss and Dooley, etc. Unfortunately, these invariants are less informative than their classical counterparts and they are more difficult to compute.

Krengel’s and Parry’s Entropies

Let S be a conservative measure-preserving transformation of a σ -finite measure space (Y, \mathcal{E}, ν) . The *Krengel entropy* [119] of S is defined by

$$h_{\text{Kr}}(S) = \sup \{ \nu(E) h(S_E) \mid 0 < \nu(E) < +\infty \},$$

where $h(S_E)$ is the finite measure-preserving entropy of S_E . It follows from Abramov’s formula for the entropy of induced transformation that $h_{\text{Kr}}(S) = \mu(E) h(S_E)$ whenever E sweeps out, i.e. $\bigcup_{i \geq 0} S^{-i} E = X$. A generic transformation from $\text{Aut}_0(X, \mu)$ has entropy 0. Krengel raised a question in [119]: does there exist a zero entropy infinite measure-preserving S and a zero entropy finite measure-preserving R such that $h_{\text{Kr}}(S \times R) > 0$? This problem was recently solved in [44] (a special case was announced by Silva and Thieullen in an October 1995 AMS conference (unpublished)):

- (i) if $h_{\text{Kr}}(S) = 0$ and R is distal then $h_{\text{Kr}}(S \times R) = 0$;
- (ii) if R is not distal then there is a rank-one transformation S with $h_{\text{Kr}}(S \times R) = \infty$.

We also note that if a conservative $S \in \text{Aut}_0(X, \mu)$ commutes with another transformation R such that

$\nu \circ R = c\nu$ for a constant $c \neq 1$ then $h_{Kr}(S)$ is either 0 or ∞ [180].

Now let T be a type III ergodic transformation of (X, \mathcal{B}, μ) . Silva and Thieullen define an entropy $h^*(T)$ of T by setting $h^*(T) := h_{Kr}(\tilde{T})$, where \tilde{T} is the Maharam extension of T (see Subsect. “Maharam Extension, Associated Flow and Orbit Classification of Type III Systems”). Since \tilde{T} commutes with transformations which ‘multiply’ \tilde{T} -invariant measure, it follows that $h^*(T)$ is either 0 or ∞ .

Let T be the standard III_λ -odometer from Example 17(i). Then $h^*(T) = 0$. The same is true for a so-called ternary odometer associated with the sequence $(3, \nu_n)_{n=1}^\infty$, where $\nu_n(0) = \nu_n(2) = \lambda/(1+2\lambda)$ and $\nu_n(1) = \lambda/(1+\lambda)$ [180]. It is not known however whether every ergodic nonsingular odometer has zero entropy. On the other hand, it was shown in [180] that $h^*(T) = \infty$ for every K -automorphism.

The Parry entropy [158] of S is defined by

$$h_{Pa}(S) := \left\{ H(S^{-1}\mathfrak{F}|\mathfrak{F}) \mid \mathfrak{F} \text{ is a } \sigma\text{-finite subalgebra of } \mathfrak{B} \text{ such that } \mathfrak{F} \subset S^{-1}\mathfrak{F} \right\}.$$

Parry showed [158] that $h_{Pa}(S) \leq h_{Kr}(S)$. It is still an open question whether the two entropies coincide. This is the case when S is of rank one (since $h_{Kr}(S) = 0$) and when S is quasi-finite [158]. The transformation S is called *quasi-finite* if there exists a subset of finite measure $A \subset Y$ such that the first return time partition $(A_n)_{n>0}$ of A has finite entropy. We recall that $x \in A_n \iff n$ is the smallest positive integer such that $T^n x \in A$. An example of non-quasi-finite ergodic infinite measure preserving transformation was constructed recently in [8].

Parry’s Generalization of Shannon–MacMillan–Breiman Theorem

Let T be an ergodic transformation of a standard nonatomic probability space (X, \mathcal{B}, μ) . Suppose that $f \circ T \in L^1(X, \mu)$ if and only if $f \in L^1(X, \mu)$. This means that there is $K > 0$ such that $K^{-1} < (d\mu \circ T)/(d\mu)(x) < K$ for a. a. x . Let \mathcal{P} be a finite partition of X . Denote by $C_n(x)$ the atom of $\bigvee_{i=0}^n T^{-i}\mathcal{P}$ which contains x . We put $\omega_{-1} = 0$. Parry shows in [155] that

$$\frac{\sum_{j=0}^n \log \mu(C_{n-j}(T^j x))(\omega_j(x) - \omega_{j-1}(x))}{\sum_{i=0}^n \omega_j(x)} \rightarrow$$

$$H\left(\mathcal{P} \mid \bigvee_{i=1}^\infty T^{-i}\mathcal{P}\right) - \int_X \log E\left(\frac{d\mu \circ T}{d\mu} \mid \bigvee_{i=0}^\infty T^{-i}\mathcal{P}\right) d\mu$$

for a. a. x . Parry also shows that under the aforementioned conditions on T ,

$$\begin{aligned} \frac{1}{n} \left(\sum_{j=0}^n H\left(\bigvee_{i=0}^j T^{-i}\mathcal{P}\right) - \sum_{j=0}^{n-1} H\left(\bigvee_{i=1}^{j+1} T^{-i}\mathcal{P}\right) \right) \\ \rightarrow H\left(\mathcal{P} \mid \bigvee_{i=1}^\infty T^{-i}\mathcal{P}\right). \end{aligned}$$

Critical Dimension

The critical dimension introduced by Mortiss [146] measures the order of growth for sums of Radon–Nikodym derivatives. Let (X, \mathcal{B}, μ, T) be an ergodic nonsingular dynamical system. Given $\delta > 0$, let

$$X_\delta := \left\{ x \in X \mid \liminf_{n \rightarrow \infty} \frac{\sum_{i=0}^{n-1} \omega_i(x)}{n^\delta} > 0 \right\} \text{ and } \quad (4)$$

$$X^\delta := \left\{ x \in X \mid \liminf_{n \rightarrow \infty} \frac{\sum_{i=0}^{n-1} \omega_i(x)}{n^\delta} = 0 \right\}. \quad (5)$$

Then X_δ and X^δ are T -invariant subsets.

Definition 37 ([57, 146]) The *lower critical dimension* $\alpha(T)$ of T is $\sup \{\delta \mid \mu(X_\delta) = 1\}$. The *upper critical dimension* $\beta(T)$ of T is $\inf \{\delta \mid \mu(X^\delta) = 1\}$.

It was shown in [57] that the lower and upper critical dimensions are invariants for isomorphism of nonsingular systems. Notice also that

$$\begin{aligned} \alpha(T) &= \liminf_{n \rightarrow \infty} \frac{\log \left(\sum_{i=1}^n \omega_i(x) \right)}{\log n} \text{ and} \\ \beta(T) &= \limsup_{n \rightarrow \infty} \frac{\log \left(\sum_{i=1}^n \omega_i(x) \right)}{\log n}. \end{aligned}$$

Moreover, $0 \leq \alpha(T) \leq \beta(T) \leq 1$. If T is of type II_1 then $\alpha(T) = \beta(T) = 1$. If T is the standard III_λ -odometer from Example 17 then $\alpha(T) = \beta(T) = \log(1+\lambda) - \lambda/(1+\lambda) \log \lambda$.

Theorem 38

(i) For every $\lambda \in [0, 1]$ and every $c \in [0, 1]$ there exists a nonsingular odometer of type III_λ with critical dimension equal to c [145].

(ii) For every $c \in [0, 1]$ there exists a nonsingular odometer of type Π_∞ with critical dimension equal to c [57].

Let T be the nonsingular odometer associated with a sequence $(m_n, v_n)_{n=1}^\infty$. Let $s(n) = m_1 \cdots m_n$ and let $H(\mathcal{P}_n)$ denote the entropy of the partition of the first n coordinates with respect to μ . We now state a nonsingular version of Shannon–MacMillan–Breiman theorem for T from [57].

Theorem 39 *Let m_i be bounded from above. Then*

(i)

$$\alpha(T) = \liminf_{n \rightarrow \infty} \inf \frac{-\sum_{i=1}^n \log m_i(x_i)}{\log s(n)} \\ = \liminf_{n \rightarrow \infty} \frac{H(\mathcal{P}_n)}{\log s(n)}$$

and

(ii)

$$\beta(T) = \limsup_{n \rightarrow \infty} \inf \frac{-\sum_{i=1}^n \log m_i(x_i)}{\log s(n)} \\ = \limsup_{n \rightarrow \infty} \frac{H(\mathcal{P}_n)}{\log s(n)}$$

for a. a. $x = (x_i)_{i \geq 1} \in X$.

It follows that in the case when $\alpha(T) = \beta(T)$, the critical dimension coincides with $\lim_{n \rightarrow \infty} H(\mathcal{P}_n)/(\log s(n))$. In [145] this expression (when it exists) was called *AC-entropy* (average coordinate). It also follows from Theorem 39 that if T is an odometer of bounded type then $\alpha(T^{-1}) = \alpha(T)$ and $\beta(T^{-1}) = \beta(T)$. In [58], Theorem 39 was extended to a subclass of Markov odometers. The critical dimensions for Hamachi shifts (see Subsect. “Nonsingular Bernoulli Transformations – Hamachi’s Example”) were investigated in [59]:

Theorem 40 *For any $\epsilon > 0$, there exists a Hamachi shift S with $\alpha(S) < \epsilon$ and $\beta(S) > 1 - \epsilon$.*

Nonsingular Restricted Orbit Equivalence

In [144] Mortiss initiated study of a nonsingular version of Rudolph’s restricted orbit equivalence [167]. This work is still in its early stages and does not yet deal with any form of entropy. However she introduced nonsingular orderings of orbits, defined sizes and showed that much of the basic machinery still works in the nonsingular setting.

Nonsingular Joinings and Factors

The theory of joinings is a powerful tool to study probability preserving systems and to construct striking counterexamples. It is interesting to study what part of this machinery can be extended to the nonsingular case. However, the definition of nonsingular joining is far from being obvious. Some progress was achieved in understanding 2-fold joinings and constructing prime systems of any Krieger type. As far as we know the higher-fold nonsingular joinings have not been considered so far. It turned out however that an alternative coding technique, predating joinings in studying the centralizer and factors of the classical measure-preserving Chacón maps, can be used as well to classify factors of Cartesian products of some nonsingular Chacón maps.

Joinings, Nonsingular MSJ and Simplicity

In this section all measures are probability measures. A *nonsingular joining* of two nonsingular systems $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ is a measure $\hat{\mu}$ on the product $\mathcal{B}_1 \times \mathcal{B}_2$ that is nonsingular for $T_1 \times T_2$ and satisfies: $\hat{\mu}(A \times X_2) = \mu_1(A)$ and $\hat{\mu}(X_1 \times B) = \mu_2(B)$ for all $A \in \mathcal{B}_1$ and $B \in \mathcal{B}_2$. Clearly, the product $\mu_1 \times \mu_2$ is a nonsingular joining. Given a transformation $S \in C(T)$, the measure μ_S given by $\mu_S(A \times B) := \mu(A \cap S^{-1}B)$ is a nonsingular joining of (X, μ, T) and $(X, \mu \circ S^{-1}, T)$. It is called a *graph-joining* since it is supported on the graph of S . Another important kind of joinings that we are going to define now is related to factors of dynamical systems. Recall that given a nonsingular system (X, \mathcal{B}, μ, T) , a sub- σ -algebra \mathcal{A} of \mathcal{B} such that $T^{-1}(\mathcal{A}) = \mathcal{A} \bmod \mu$ is called a *factor* of T . There is another, equivalent, definition. A nonsingular dynamical system (Y, \mathcal{C}, ν, S) is called a *factor* of T if there exists a measure-preserving map $\varphi: X \rightarrow Y$, called a *factor map*, with $\varphi T = S\varphi$ a.e. (If φ is only nonsingular, ν may be replaced with the equivalent measure $\mu \circ \varphi^{-1}$, for which φ is measure-preserving.) Indeed, the sub- σ -algebra $\varphi^{-1}(\mathcal{C}) \subset \mathcal{B}$ is T -invariant and, conversely, any T -invariant sub- σ -algebra of \mathcal{B} defines a factor map by immanent properties of standard probability spaces, see e.g. [3]. If φ is a factor map as above, then μ has a disintegration with respect to φ , i.e., $\mu = \int \mu_y d\nu(y)$ for a measurable map $y \mapsto \mu_y$ from Y to the probability measures on X so that $\mu_y(\varphi^{-1}(y)) = 1$, the measure $\mu_{S\varphi(x)} \circ T$ is equivalent to $\mu_{\varphi(x)}$ and

$$\frac{d\mu \circ T}{d\mu}(x) = \frac{d\nu \circ S}{d\nu}(\varphi(x)) \frac{d\mu_{S\varphi(x)} \circ T}{d\mu_{\varphi(x)}}(x) \quad (6)$$

for a.e. $x \in X$. Define now the *relative product* $\hat{\mu} = \mu \times_\varphi \mu$ on $X \times X$ by setting $\hat{\mu} = \int \mu_y \times \mu_y d\nu(y)$. Then

it is easy to deduce from (6) that $\hat{\mu}$ is a nonsingular self-joining of T .

We note however that the above definition of joining is not satisfactory since it does not reduce to the classical definition when we consider probability preserving systems. Indeed, the following result was proved in [168].

Theorem 41 *Let $(X_1, \mathcal{B}_1, \mu_1, T_1)$ and $(X_2, \mathcal{B}_2, \mu_2, T_2)$ be two finite measure-preserving systems such that $T_1 \times T_2$ is ergodic. Then for every $\lambda, 0 < \lambda < 1$, there exists a nonsingular joining $\hat{\mu}$ of μ_1 and μ_2 such that $(T_1 \times T_2, \hat{\mu})$ is ergodic and of type III_λ .*

It is not known however if the nonsingular joining $\hat{\mu}$ can be chosen in every orbit equivalence class. In view of the above, Rudolph and Silva [168] isolate an important subclass of joining. It is used in the definition of a nonsingular version of minimal self-joinings.

Definition 42

- (i) A nonsingular joining $\hat{\mu}$ of (X_1, μ_1, T_1) and (X_2, μ_2, T_2) is *rational* if there exist measurable functions $c^1: X_1 \rightarrow \mathbb{R}_+$ and $c^2: X_2 \rightarrow \mathbb{R}_+$ such that

$$\begin{aligned}\hat{\omega}_1^{\hat{\mu}}(x_1, x_2) &= \omega_1^{\mu_1}(x_1)\omega_1^{\mu_2}(x_2)c^1(x_1) \\ &= \omega_1^{\mu_1}(x_1)\omega_1^{\mu_2}(x_2)c^2(x_2) \quad \hat{\mu} \text{ a. e.}\end{aligned}$$

- (ii) A nonsingular dynamical system (X, \mathcal{B}, μ, T) has *minimal self-joinings (MSJ)* over a class \mathcal{M} of probability measures equivalent to μ , if for every $\mu_1, \mu_2 \in \mathcal{M}$, for every rational joining $\hat{\mu}$ of μ_1, μ_2 , a.e. ergodic component of $\hat{\mu}$ is either the product of its marginals or is the graph-joining supported on T^j for some $j \in \mathbb{Z}$.

Clearly, product measure, graph-joinings and the relative products are all rational joinings. Moreover, a rational joining of finite measure-preserving systems is measure-preserving and a rational joining of type II_1 's is of type II_1 [168]. Thus we obtain the finite measure-preserving theory as a special case. As for the definition of MSJ, it depends on a class \mathcal{M} of equivalent measures. In the finite measure-preserving case $\mathcal{M} = \{\mu\}$. However, in the nonsingular case no particular measure is distinguished. We note also that Definition 42(ii) involves some restrictions on all rational joinings and not only ergodic ones as in the finite measure-preserving case. The reason is that an ergodic component of a nonsingular joining needs not be a joining of measures equivalent to the original ones [2]. For finite measure-preserving transformations, MSJ over $\{\mu\}$ is the same as the usual 2-fold MSJ [49].

A nonsingular transformation T on (X, \mathcal{B}, μ) is called *prime* if its only factors are \mathcal{B} and $\{X, \emptyset\} \bmod \mu$. A (non-empty) class \mathcal{M} of probability measures equivalent to μ is said to be *centralizer stable* if for each $S \in C(T)$ and $\mu_1 \in \mathcal{M}$, the measure $\mu_1 \circ S$ is in \mathcal{M} .

Theorem 43 ([168]) *Let (X, \mathcal{B}, μ, T) be a ergodic nonatomic dynamical system such that T has MSJ over a class \mathcal{M} that is centralizer stable. Then T is prime and the centralizer of T consists of the powers of T .*

A question that arises is whether if such nonsingular dynamical system (not of type II_1) exist. Expanding on Ornstein's original construction from [150], Rudolph and Silva construct in [168], for each $0 \leq \lambda \leq 1$, a nonsingular rank-one transformation T_λ that is of type III_λ and that has MSJ over a class \mathcal{M} that is centralizer stable. Type II_∞ examples with analogous properties were also constructed there. In this connection it is worth to mention the example by Aaronson and Nadkarni [6] of II_∞ ergodic transformations that have no factor algebras on which the invariant measure is σ -finite (except for the trivial and the entire ones); however these transformations are not prime.

A more general notion than MSJ called *graph self-joinings (GSJ)*, was introduced [181]: just replace the words “on T^j for some $j \in \mathbb{Z}$ ” in Definition 3(ii) with “on S for some element $S \in C(T)$ ”. For finite measure-preserving transformations, GSJ over $\{\mu\}$ is the same as the usual 2-fold simplicity [49]. The famous Veech theorem on factors of 2-fold simple maps (see [49]) was extended to nonsingular systems in [181] as follows: if a system (X, \mathcal{B}, μ, T) has GSJ then for every non-trivial factor \mathcal{A} of T there exists a locally compact subgroup H in $C(T)$ (equipped with the weak topology) which acts smoothly (i.e. the partition into H -orbits is measurable) and such that $\mathcal{A} = \{B \in \mathcal{B} \mid \mu(hB\Delta B) = 0 \text{ for all } h \in H\}$. It follows that there is a cocycle φ from $(X, \mathcal{A}, \mu \upharpoonright \mathcal{A})$ to H such that T is isomorphic to the φ -skew product extension $(T \upharpoonright \mathcal{A})_\varphi$ (see Subsect. “Cocycles of Dynamical Systems. Weak Equivalence of Cocycles”). Of course, the ergodic nonsingular odometers and, more generally, ergodic nonsingular compact group rotation (see Subsect. “ L^∞ -Spectrum and Groups of Quasi-Invariance”) have GSJ. However, except for this trivial case (the Cartesian square is non-ergodic) plus the systems with MSJ from [168], no examples of type III systems with GSJ are known. In particular, no smooth examples have been constructed so far. This is in sharp contrast with the finite measure preserving case where abundance of simple (or close to simple) systems are known (see [39,40,49,182]).

Nonsingular Coding and Factors of Cartesian Products of Nonsingular Maps

As we have already noticed above, the nonsingular MSJ theory was developed in [168] only for 2-fold self-joinings. The reasons for this were technical problems with extending the notion of rational joinings from 2-fold to n -fold self-joinings. However while the 2-fold nonsingular MSJ or GSJ properties of T are sufficient to control the centralizer and the factors of T , it is not clear whether it implies anything about the factors or centralizer of $T \times T$. Indeed, to control them one needs to know the 4-fold joinings of T . However even in the finite measure-preserving case it is a long standing open question whether 2-fold MSJ implies n -fold MSJ. That is why del Junco and Silva [51] apply an alternative – nonsingular coding – techniques to classify the factors of Cartesian products of nonsingular Chacón maps. The techniques were originally used in [48] to show that the classical Chacón map is prime and has trivial centralizer. They were extended to nonsingular systems in [50].

For each $0 < \lambda < 1$ we denote by T_λ the Chacón map (see Subsect. “Rank-One Transformations. Chacón Maps. Finite Rank”) corresponding the sequence of probability vectors $w_n = (\lambda/(1 + 2\lambda), 1/(1 + 2\lambda), \lambda/(1 + 2\lambda))$ for all $n > 0$. One can verify that the maps T_λ are of type III_λ . (The classical Chacón map corresponds to $\lambda = 1$.) All of these transformations are defined on the same standard Borel space (X, \mathcal{B}) . These transformations were shown to be power weakly mixing in [12]. The centralizer of any finite Cartesian product of nonsingular Chacón maps is computed in the following theorem.

Theorem 44 ([51]) *Let $0 < \lambda_1 < \dots < \lambda_k \leq 1$ and n_1, \dots, n_k be positive integers. Then the centralizer of the Cartesian product $T_{\lambda_1}^{\otimes n_1} \times \dots \times T_{\lambda_k}^{\otimes n_k}$ is generated by maps of the form $U_1 \times \dots \times U_k$, where each U_i , acting on the n_i -dimensional product space X^{n_i} , is a Cartesian product of powers of T_{λ_i} or a co-ordinate permutation on X^{n_i} .*

Let π denote the permutation on $X \times X$ defined by $\pi(x, y) = (y, x)$ and let $\mathcal{B}^{2\odot}$ denote the symmetric factor, i.e. $\mathcal{B}^{2\odot} = \{A \in \mathcal{B} \otimes \mathcal{B} \mid \pi(A) = A\}$. The following theorem classifies the factors of the Cartesian product of any two nonsingular type III_λ , $0 < \lambda < 1$, or type II_1 Chacón maps.

Theorem 45 ([51]) *Let T_{λ_1} and T_{λ_2} be two nonsingular Chacón systems. Let \mathcal{F} be a factor algebra of $T_{\lambda_1} \times T_{\lambda_2}$.*

- (i) *If $\lambda_1 \neq \lambda_2$ then \mathcal{F} is equal mod 0 to one of the four algebras $\mathcal{B} \otimes \mathcal{B}$, $\mathcal{B} \otimes \mathcal{N}$, $\mathcal{N} \otimes \mathcal{B}$, or $\mathcal{N} \otimes \mathcal{N}$, where $\mathcal{N} = \{\emptyset, X\}$.*

- (ii) *If $\lambda_1 = \lambda_2$ then \mathcal{F} is equal mod 0 to one of the following algebras $\mathcal{B} \otimes \mathcal{C}$, $\mathcal{B} \otimes \mathcal{N}$, $\mathcal{N} \otimes \mathcal{C}$, $\mathcal{N} \otimes \mathcal{N}$, or $(T^m \times \text{Id})\mathcal{B}^{2\odot}$ for some integer m .*

It is not hard to obtain type III_1 examples of Chacón maps for which the previous two theorems hold. However the construction of type II_∞ and type III_0 nonsingular Chacón transformations is more subtle as it needs the choice of ω_n to vary with n . In [92], Hamachi and Silva construct type III_0 and type II_∞ examples, however the only property proved for these maps is ergodicity of their Cartesian square. More recently, Danilenko [38] has shown that all of them (in fact, a wider class of nonsingular Chacón maps of all types) are power weakly mixing.

In [22], Choksi, Eigen and Prasad asked whether there exists a zero entropy, finite measure-preserving mixing automorphism S , and a nonsingular type III automorphism T , such that $T \times S$ has no Bernoulli factors. Theorem 45 provides a partial answer (with a mildly mixing only instead of mixing) to this question: if S is the finite measure-preserving Chacón map and T is a nonsingular Chacón map as above, the factors of $T \times S$ are only the trivial ones, so $T \times S$ has no Bernoulli factors.

Applications. Connections with Other Fields

In this – final – section we shed light on numerous mathematical sources of nonsingular systems. They come from the theory of stochastic processes, random walks, locally compact Cantor systems, horocycle flows on hyperbolic surfaces, von Neumann algebras, statistical mechanics, representation theory for groups and anticommutation relations, etc. We also note that such systems sometimes appear in the context of probability preserving dynamics (see also a criterium of distality in Subsect. “Krengel’s and Parry’s Entropies”).

Mild Mixing

An ergodic finite measure-preserving dynamical system (X, \mathcal{B}, μ, T) is called *mildly mixing* if for each non-trivial factor algebra $\mathcal{A} \subset \mathcal{B}$, the restriction $T \upharpoonright \mathcal{A}$ is not rigid. For equivalent definitions and extensions to actions of locally compact groups we refer to [3] and [177]. There is an interesting criterium of the mild mixing that involves nonsingular systems: T is mildly mixing if and only if for each ergodic nonsingular transformation S , the product $T \times S$ is ergodic [71]. Furthermore, $T \times S$ is then orbit equivalent to S [98]. Moreover, if R is a nonsingular transformation such that $R \times S$ is ergodic for any ergodic nonsingular S then R is of type II_1 (and mildly mixing) [177].

Disjointness and Furstenberg's Class \mathcal{W}^\perp

Two probability preserving systems (X, μ, T) and (Y, ν, S) are called *disjoint* if $\mu \times \nu$ is the only $T \times S$ -invariant probability measure on $X \times Y$ whose coordinate projections are μ and ν respectively. Furstenberg in [69] initiated studying the class \mathcal{W}^\perp of transformations disjoint from all weakly mixing ones. Let \mathcal{D} denote the class of distal transformations and $\mathcal{M}(\mathcal{W}^\perp)$ the class of multipliers of \mathcal{W}^\perp (for the definitions see [75]). Then $\mathcal{D} \subset \mathcal{M}(\mathcal{W}^\perp) \subset \mathcal{W}^\perp$. In [43] and [133] it was shown by constructing explicit examples that these inclusions are strict. We record this fact here because nonsingular ergodic theory was the key ingredient of the arguments in the two papers pertaining to the theory of probability preserving systems. The examples are of the form $T_{\varphi, S}(x, y) = (Tx, S_{\varphi(x)}y)$, where T is an ergodic rotation on (X, μ) , $(S_g)_{g \in G}$ a mildly mixing action of a locally compact group G on Y and $\varphi: X \rightarrow G$ a measurable map. Let W_φ denote the Mackey action of G associated with φ and let (Z, κ) be the space of this action. The key observation is that there exists an affine isomorphism of the simplex of $T_{\varphi, S}$ -invariant probability measures whose pullback on X is μ and the simplex of $W_\varphi \times S$ quasi-invariant probability measures whose pullback on Z is κ and whose Radon–Nikodym cocycle is measurable with respect to Z . This is a far reaching generalization of Furstenberg theorem on relative unique ergodicity of ergodic compact group extensions.

Symmetric Stable and Infinitely Divisible Stationary Processes

Rosinsky in [163] established a remarkable connection between structural studies of stationary stochastic processes and ergodic theory of nonsingular transformations (and flows). For simplicity we consider only real processes in discrete time. Let $X = (X_n)_{n \in \mathbb{Z}}$ be a measurable stationary symmetric α -stable (S α S) process, $0 < \alpha < 2$. This means that any linear combination $\sum_{k=1}^n a_k X_{j_k}$, $j_k \in \mathbb{Z}$, $a_k \in \mathbb{R}$ has an S α S-distribution. (The case $\alpha = 2$ corresponds to Gaussian processes.) Then the process admits a spectral representation

$$X_n = \int_Y f_n(y) M(dy), \quad n \in \mathbb{Z}, \quad (7)$$

where $f_n \in L^\alpha(Y, \mu)$ for a standard σ -finite measure space (Y, \mathcal{B}, μ) and M is an independently scattered random measure on \mathcal{B} such that $E \exp(iuM(A)) = \exp(-|u|^\alpha \mu(A))$ for every $A \in \mathcal{B}$ of finite measure. By [163], one can choose the kernel $(f_n)_{n \in \mathbb{Z}}$ in a special way: there are a μ -nonsingular transformation T and measurable maps $\varphi: X \rightarrow \{-1, 1\}$ and $f \in L^\alpha(Y, \mu)$ such that

$f_n = U^n f$, $n \in \mathbb{Z}$, where U is the isometry of $L^\alpha(X, \mu)$ given by $Ug = \varphi \cdot (d\mu \circ T/d\mu)^{1/\alpha} \cdot g \circ T$. If, in addition, the smallest T -invariant σ -algebra containing $f^{-1}(\mathcal{B}_\mathbb{R})$ coincides with \mathcal{B} and $\text{Supp}\{f \circ T^n: n \in \mathbb{Z}\} = Y$ then the pair (T, φ) is called minimal. It turns out that minimal pairs always exist. Moreover, two minimal pairs (T, φ) and (T', φ') representing the same S α S process are equivalent in some natural sense [163]. Then one can relate ergodic-theoretical properties of (T, φ) to probabilistic properties of $(X_n)_{n \in \mathbb{Z}}$. For instance, let $Y = C \sqcup D$ be the Hopf decomposition of Y (see Theorem 2). We let $X_n^D := \int_D f_n(y) M(dy)$ and $X_n^C := \int_C f_n(y) M(dy)$. Then we obtain a unique (in distribution) decomposition of X into the sum $X^D + X^C$ of two independent stationary S α S-processes.

Another kind of decomposition was considered in [171]. Let P be the largest invariant subset of Y such that $T \upharpoonright P$ has a finite invariant measure. Partitioning Y into P and $N := Y \setminus P$ and restricting the integration in (7) to P and N we obtain a unique (in distribution) decomposition of X into the sum $X^P + X^N$ of two independent stationary S α S-processes. Notice that the process X is ergodic if and only if $\mu(P) = 0$.

Recently, Roy considered a more general class of *infinitely divisible (ID)* stationary processes [165]. Using Maruyama's representation of the characteristic function of an ID process X without Gaussian part he singled out the Lévy measure Q of X . Then Q is a shift invariant σ -finite measure on $\mathbb{R}^\mathbb{Z}$. Decomposing the dynamical system $(\mathbb{R}^\mathbb{Z}, \tau, Q)$ in various natural ways (Hopf decomposition, 0-type and positive type, so-called 'rigidity free' part and its complement) he obtains corresponding decompositions for the process X . Here τ stands for the shift on $\mathbb{R}^\mathbb{Z}$.

Poisson Suspensions

Poisson suspensions are widely used in statistical mechanics to model ideal gas, Lorentz gas, etc (see [33]). Let (X, \mathcal{B}, μ) be a standard σ -finite non-atomic measure space and $\mu(X) = \infty$. Denote by \tilde{X} the space of unordered countable subsets of X . It is called the space of *configurations*. Fix $t > 0$. Let $A \in \mathcal{B}$ have positive finite measure and let $j \in \mathbb{Z}_+$. Denote by $[A, j]$ the subset of all configurations $\tilde{x} \in \tilde{X}$ such that $\#(\tilde{x} \cap A) = j$. Let $\tilde{\mathcal{B}}$ be the σ -algebra generated by all $[A, j]$. We define a probability measure $\tilde{\mu}_t$ on $\tilde{\mathcal{B}}$ by two conditions:

- (i) $\tilde{\mu}_t([A, j]) = \frac{(t\mu(A))^j}{j!} \exp(-t\mu(A))$;
- (ii) if A_1, \dots, A_p are pairwise disjoint then $\tilde{\mu}_t(\bigcap_{k=1}^p [A_k, j_k]) = \prod_{k=1}^p \tilde{\mu}_t([A_k, j_k])$.

If T is a μ -preserving transformation of X and $\tilde{x} = (x_1, x_2, \dots)$ is a configuration then we set $\tilde{T}\omega := (Tx_1, Tx_2, \dots)$. It is easy to verify that \tilde{T} is a $\tilde{\mu}$ -preserving transformation of \tilde{X} . The dynamical system $(\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu}, \tilde{T})$ is called the *Poisson suspension* above (X, \mathcal{B}, μ, T) . It is ergodic if and only if T has no invariant sets of finite positive measure. There is a canonical representation of $L^2(\tilde{X}, \tilde{\mu})$ as the Fock space over $L^2(X, \mu)$ such that the unitary operator $U_{\tilde{T}}$ is the ‘exponent’ of U_T . Thus, the maximal spectral type of $U_{\tilde{T}}$ is $\sum_{n \geq 0} (n!)^{-1} \sigma^{*n}$, where σ is a measure of the maximal spectral type of U_T . It is easy to see that a σ -finite factor of T corresponds to a factor (called Poissonian) of \tilde{T} . Moreover, a σ -finite measure-preserving joining (with σ -finite projections) of two infinite measure-preserving transformations T_1 and T_2 generates a joining (called Poissonian) of \tilde{T}_1 and \tilde{T}_2 [52, 164]. Thus we see a similarity with the well studied theory of Gaussian dynamical systems [134]. However, the Poissonian case is less understood. There was a recent progress in this field. Parreau and Roy constructed Poisson suspensions whose ergodic self-joinings are all Poissonian [154]. In [111] partial solutions of the following (still open) problems are found:

- (i) whether the Pinsker factor of \tilde{T} is Poissonian,
- (ii) what is the relationship between Krengel’s entropy of T , Parry’s entropy of T and Kolmogorov–Sinai entropy of \tilde{T} .

Recurrence of Random Walks with Non-stationary Increments

Using nonsingular ergodic theory one can introduce the notion of recurrence for random walks obtained from certain non-stationary processes. Let T be an ergodic nonsingular transformation of a standard probability space (X, \mathcal{B}, μ) and let $f: X \rightarrow \mathbb{R}^n$ a measurable function. Define for $m \geq 1$, $Y_m: X \rightarrow \mathbb{R}^n$ by $Y_m := \sum_{n=0}^{m-1} f \circ T^n$. In other words, $(Y_m)_{m \geq 1}$ is the random walk associated with the (non-stationary) process $(f \circ T^n)_{n \geq 0}$. Let us call this random walk *recurrent* if the cocycle f of T is recurrent (see Subsect. “Cocycles of Dynamical Systems. Weak Equivalence of Cocycles”). It was shown in [176] that in the case $\mu \circ T = \mu$, i. e. the process is stationary, this definition is equivalent to the standard one.

Boundaries of Random Walks

Boundaries of random walks on groups retain valuable information on the underlying groups (amenability, entropy, etc.) and enable one to obtain integral representation for harmonic functions of the random

walk [112, 186, 187]. Let G be a locally compact group and ν a probability measure on G . Let T denote the (one-sided) shift on the probability space $(X, \mathcal{B}_X, \mu) := (G, \mathcal{B}_G, \nu)^{\mathbb{Z}^+}$ and $\varphi: X \rightarrow G$ a measurable map defined by $(y_0, y_1, \dots) \mapsto y_0$. Let T_φ be the φ -skew product extension of T acting on the space $(X \times G, \mu \times \lambda_G)$ (for non-invertible transformations the skew product extension is defined in the very same way as for invertible ones, see Subsect. “Cocycles of Dynamical Systems. Weak Equivalence of Cocycles”). Then T_φ is isomorphic to the *homogeneous random walk* on G with jump probability ν . Let $\mathcal{I}(T_\varphi)$ denote the sub- σ -algebra of T_φ -invariant sets and let $\mathcal{F}(T_\varphi) := \bigcap_{n \geq 0} T_\varphi^{-n}(\mathcal{B}_X \otimes \mathcal{B}_G)$. The former is called the *Poisson boundary* of T_φ and the latter one is called the *tail boundary* of T_φ . Notice that a nonsingular action of G by inverted right translations along the second coordinate is well defined on each of the two boundaries. The two boundaries (or, more precisely, the G -actions on them) are ergodic. The Poisson boundary is the Mackey range of φ (as a cocycle of T). Hence the Poisson boundary is amenable [187]. If the support of ν generates a dense subgroup of G then the corresponding Poisson boundary is weakly mixing [4]. As for the tail boundary, we first note that it can be defined for a wider family of *non-homogeneous* random walks. This means that the jump probability ν is no longer fixed and a sequence $(\nu_n)_{n \geq 0}$ of probability measures on G is considered instead. Now let $(X, \mathcal{B}_X, \mu) := \prod_{n \geq 0} (G, \mathcal{B}_G, \nu_n)$. The one-sided shift on X may not be nonsingular now. Instead of it, we consider the tail equivalence relation \mathcal{R} on X and a cocycle $\alpha: \mathcal{R} \rightarrow G$ given by $\alpha(x, y) = x_1 \cdots x_n y_n^{-1} \cdots y_1$, where $x = (x_i)_{i \geq 0}$ and $y = (y_i)_{i \geq 0}$ are \mathcal{R} -equivalent and n is the smallest integer such that $x_i = y_i$ for all $i > n$. The tail boundary of the random walk on G with time dependent jump probabilities $(\nu_n)_{n \geq 0}$ is the Mackey G -action associated with α . In the case of homogeneous random walks this definition is equivalent to the initial one. Connes and Woods showed [32] that the tail boundary is always amenable and AT. It is unknown whether the converse holds for general G . However it is true for $G = \mathbb{R}$ and $G = \mathbb{Z}$: the class of AT-flows coincides with the class of tail boundaries of the random walks on \mathbb{R} and a similar statement holds for \mathbb{Z} [32]. Jaworsky showed [109] that if G is countable and a random walk is homogeneous then the tail boundary of the random walk possesses a so-called SAT-property (which is stronger than AT).

Classifying σ -Finite Ergodic Invariant Measures

The description of ergodic finite invariant measures for topological (or, more generally, standard Borel) systems

is a well established problem in the classical ergodic theory [33]. On the other hand, it seems impossible to obtain any useful information about the system by analyzing the set of all ergodic quasi-invariant (or just σ -finite invariant) measures because this set is wildly huge (see Subsect. “The Glimm–Effros Theorem”). The situation changes if we impose some restrictions on the measures. For instance, if the system under question is a homeomorphism (or a topological flow) defined on a locally compact Polish space then it is natural to consider the class of (σ -finite) invariant Radon measures, i. e. measures taking finite values on the compact subsets. We give two examples.

First, the seminal results of Giordano, Putnam and Skau on the topological orbit equivalence of compact Cantor minimal systems were extended to locally compact Cantor minimal (l.c.c.m.) systems in [37] and [138]. Given a l.c.c.m. system X , we denote by $\mathcal{M}(X)$ and $\mathcal{M}_1(X)$ the set of invariant Radon measures and the set of invariant probability measures on X . Notice that $\mathcal{M}_1(X)$ may be empty [37]. It was shown in [138] that two systems X and X' are topologically orbit equivalent if and only if there is a homeomorphism of X onto X' which maps bijectively $\mathcal{M}(X)$ onto $\mathcal{M}(X')$ and $\mathcal{M}_1(X)$ onto $\mathcal{M}_1(X')$. Thus $\mathcal{M}(X)$ retains an important information on the system – it is ‘responsible’ for the topological orbit equivalence of the underlying systems. Uniquely ergodic l.c.c.m. systems (with unique up to scaling infinite invariant Radon measure) were constructed in [37].

The second example is related to study of the smooth horocycle flows on tangent bundles of hyperbolic surfaces. Let \mathbb{D} be the open disk equipped with the hyperbolic metric $|dz|/(1 - |z|^2)$ and let $\text{Möb}(\mathbb{D})$ denote the group of Möbius transformations of \mathbb{D} . A hyperbolic surface can be written in the form $M := \Gamma \backslash \text{Möb}(\mathbb{D})$, where Γ is a torsion free discrete subgroup of $\text{Möb}(\mathbb{D})$. Suppose that Γ is a nontrivial normal subgroup of a lattice Γ_0 in $\text{Möb}(\mathbb{D})$. Then M is a regular cover of the finite volume surface $M_0 := \Gamma_0 \backslash \text{Möb}(\mathbb{D})$. The group of deck transformations $G = \Gamma_0 / \Gamma$ is finitely generated. The horocycle flow $(h_t)_{t \in \mathbb{R}}$ and the geodesic flow $(g_t)_{t \in \mathbb{R}}$ defined on the unit tangent bundle $T^1(\mathbb{D})$ descend naturally to flows, say h and g , on $T^1(M)$. We consider the problem of classification of the h -invariant Radon measures on M . According to Ratner, h has no finite invariant measures on M if G is infinite (except for measures supported on closed orbits). However there are infinite invariant Radon measures, for instance the volume measure. In the case when G is free Abelian and Γ_0 is co-compact, every homomorphism $\varphi: G \rightarrow \mathbb{R}$ determines a unique up to scaling ergodic invariant Radon measure (e.i.r.m.) m on $T^1(M)$ such that $m \circ dD = \exp(\varphi(D))m$ for all $D \in G$ [16] and ev-

ery e.i.r.m. arises this way [172]. Moreover all these measures are quasi-invariant under g . In the general case, an interesting bijection is established in [131] between the e.i.r.m. which are quasi-invariant under g and the ‘non-trivial minimal’ positive eigenfunctions of the hyperbolic Laplacian on M .

Von Neumann Algebras

There is a fascinating and productive interplay between nonsingular ergodic theory and von Neumann algebras. The two theories alternately influenced development of each other. Let (X, \mathcal{B}, μ, T) be a nonsingular dynamical system. Given $\varphi \in L^\infty(X, \mu)$ and $j \in \mathbb{Z}$, we define operators A_φ and U_j on the Hilbert space $L^2(X \times \mathbb{Z}, \mu \times \nu)$ by setting

$$\begin{aligned}(A_\varphi f)(x, i) &:= \varphi(T^i x) f(x, i), \\ (U_j f)(x, i) &:= f(x, i - j).\end{aligned}$$

Then $U_j A_\varphi U_j^* = A_{\varphi \circ T^j}$. Denote by \mathcal{M} the von Neumann algebra (i. e. the weak closure of the $*$ -algebra) generated by A_φ , $\varphi \in L^\infty(X, \mu)$ and U_j , $j \in \mathbb{Z}$. If T is ergodic and aperiodic then \mathcal{M} is a factor, i. e. $\mathcal{M} \cap \mathcal{M}' = \mathbb{C}1$, where \mathcal{M}' denotes the algebra of bounded operators commuting with \mathcal{M} . It is called a *Krieger’s factor*. Murray–von Neumann–Connes’ type of \mathcal{M} is exactly the Krieger’s type of T . The flow of weights of \mathcal{M} is isomorphic to the associated flow of T . Two Krieger’s factors are isomorphic if and only if the underlying dynamical systems are orbit equivalent [129]. Moreover, a number of important problems in the theory of von Neumann algebras such as classification of subfactors, computation of the flow of weights and Connes’ invariants, outer conjugacy for automorphisms, etc. are intimately related to the corresponding problems in nonsingular orbit theory. We refer to [42,66,73,74,89,142] for details.

Representations of CAR

Representations of canonical anticommutation relations (CAR) is one of the most elegant and useful chapters of mathematical physics, providing a natural language for many body quantum physics and quantum field theory. By a representation of CAR we mean a sequence of bounded linear operators a_1, a_2, \dots in a separable Hilbert space \mathcal{K} such that $a_j a_k + a_k a_j = 0$ and $a_j a_k^* + a_k^* a_j = \delta_{j,k}$.

Consider $\{0, 1\}$ as a group with addition mod 2. Then $X = \{0, 1\}^{\mathbb{N}}$ is a compact Abelian group. Let $\Gamma := \{x = (x_1, x_2, \dots) : \lim_{n \rightarrow \infty} x_n = 0\}$. Then Γ is a dense countable subgroup of X . It is generated by elements γ_k whose k -coordinate is 1 and the other ones

are 0. Γ acts on X by translations. Let μ be an ergodic Γ -quasi-invariant measure on X . Let $(C_k)_{k \geq 1}$ be Borel maps from X to the group of unitary operators in a Hilbert space \mathcal{H} satisfying $C_k^*(x) = C_k(x + \delta_k)$, $C_k(x)C_l(x + \delta_l) = C_l(x)C_k(x + \delta_k)$, $k \neq l$ for a. a. x . In other words, $(C_k)_{k \geq 1}$ defines a cocycle of the Γ -action. We now put $\widetilde{\mathcal{H}} := L^2(X, \mu) \otimes \mathcal{H}$ and define operators a_k in $\widetilde{\mathcal{H}}$ by setting

$$(a_k f)(x) = (-1)^{x_1 + \dots + x_{k-1}} (1 - x_k) \\ \times C_k(x) \sqrt{\frac{d\mu \circ \delta_k}{d\mu}}(x) f(x + \delta_k),$$

where $f: X \rightarrow \mathcal{H}$ is an element of $\widetilde{\mathcal{H}}$ and $x = (x_1, x_2, \dots) \in X$. It is easy to verify that a_1, a_2, \dots is a representation of CAR. The converse was established in [72] and [77]: every factor-representation (this means that the von Neumann algebra generated by all a_k is a factor) of CAR can be represented as above for some ergodic measure μ , Hilbert space \mathcal{H} and a Γ -cocycle $(C_k)_{k \geq 1}$. Moreover, using nonsingular ergodic theory Golodets [77] constructed for each $k = 2, 3, \dots, \infty$, an irreducible representation of CAR such that $\dim \mathcal{H} = k$. This answered a question of Gårding and Wightman [72] who considered only the case $k = 1$.

Unitary Representations of Locally Compact Groups

Nonsingular actions appear in a systematic way in the theory of unitary representations of groups. Let G be a locally compact second countable group and H a closed normal subgroup of G . Suppose that H is commutative (or, more generally, of type I, see [53]). Then the natural action of G by conjugation on H induces a Borel G -action, say α , on the dual space \widehat{H} – the set of unitarily equivalent classes of irreducible unitary representations of H . If now $U = (U_g)_{g \in G}$ is a unitary representation of G in a separable Hilbert space then by applying Stone decomposition theorem to $U \upharpoonright H$ one can deduce that α is nonsingular with respect to a measure μ of the ‘maximal spectral type’ for $U \upharpoonright H$ on \widehat{H} . Moreover, if U is irreducible then α is ergodic. Whenever μ is fixed, we obtain a one-to-one correspondence between the set of cohomology classes of irreducible cocycles for α with values in the unitary group on a Hilbert space \mathcal{H} and the subset of \widehat{G} consisting of classes of those unitary representations V for which the measure associated to $V \upharpoonright H$ is equivalent to μ . This correspondence is used in both directions. From information about cocycles we can deduce facts about representations and vice versa [53, 118].

Concluding Remarks

While some of the results that we have cited for nonsingular \mathbb{Z} -actions extend to actions of locally compact Polish groups (or subclasses of Abelian or amenable ones), many natural questions remain open in the general setting. For instance: what is Rokhlin lemma, or the pointwise ergodic theorem, or the definition of entropy for nonsingular actions of general countable amenable groups? The theory of abstract nonsingular equivalence relations [66] or, more generally, nonsingular groupoids [160] and polymorphisms [184] is also a beautiful part of nonsingular ergodic theory that has nice applications: description of semifinite traces of AF-algebras, classification of factor representations of the infinite symmetric group [185], path groups [14], etc. Nonsingular ergodic theory is getting even more sophisticated when we pass from \mathbb{Z} -actions to noninvertible endomorphisms or, more generally, semigroup actions (see [3] and references therein). However, due to restrictions of space we do not consider these issues in our survey.

Bibliography

1. Aaronson J (1983) The eigenvalues of nonsingular transformations. *Isr J Math* 45:297–312
2. Aaronson J (1987) The intrinsic normalizing constants of transformations preserving infinite measures. *J Analyse Math* 49:239–270
3. Aaronson J (1997) *An Introduction to Infinite Ergodic Theory*. Amer Math Soc, Providence
4. Aaronson J, Lemańczyk M (2005) Exactness of Rokhlin endomorphisms and weak mixing of Poisson boundaries, *Algebraic and Topological Dynamics*. Contemporary Mathematics, vol 385. Amer Math Soc, Providence 77–88
5. Aaronson J, Lin M, Weiss B (1979) Mixing properties of Markov operators and ergodic transformations, and ergodicity of cartesian products. *Isr J Math* 33:198–224
6. Aaronson J, Nadkarni M (1987) L_∞ eigenvalues and L_2 spectra of nonsingular transformations. *Proc Lond Math Soc* 55(3):538–570
7. Aaronson J, Nakada H (2000) Multiple recurrence of Markov shifts and other infinite measure preserving transformations. *Isr J Math* 117:285–310
8. Aaronson J, Park KK (2008) Predictability, entropy and information of infinite transformations, preprint. ArXiv:0705.2148v3
9. Aaronson J, Weiss B (2004) On Herman’s theorem for ergodic, amenable group extensions of endomorphisms. *Ergod Theory Dynam Syst* 5:1283–1293
10. Adams S, Elliott GA, Giordano T (1994) Amenable actions of groups. *Trans Amer Math Soc* 344:803–822
11. Adams T, Friedman N, Silva CE (1997) Rank-One Weak Mixing for Nonsingular Transformations. *Isr J Math* 102:269–281
12. Adams T, Friedman N, Silva CE (2001) Rank one power weak mixing for nonsingular transformations. *Ergod Theory Dynam Systems* 21:1321–1332

13. Ageev ON, Silva CE (2002) Genericity of rigid and multiply recurrent infinite measure-preserving and nonsingular transformations. In: Proceedings of the 16th Summer Conference on General Topology and its Applications. *Topology Proc* 26(2):357–365
14. Albeverio S, Hoegh-Krohn R, Testard D, Vershik AM (1983) Factorial representations of Path groups. *J Funct Anal* 51: 115–231
15. Atkinson G (1976) Recurrence of co-cycles and random walks. *J Lond Math Soc* 13:486–488
16. Babillot M, Ledrappier F (1998) Geodesic paths and horocycle flow on abelian covers. In: Lie groups and ergodic theory. *Tata Inst Fund Res Stud Math* 14, Tata Inst Fund Res, Bombay, pp 1–32
17. Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Semerédi's theorems. *J Amer Math Soc* 9:725–753
18. Bezuglyi SI, Golodets VY (1985) Groups of measure space transformations and invariants of outer conjugation for automorphisms from normalizers of type III full groups. *J Funct Anal* 60(3):341–369
19. Bezuglyi SI, Golodets VY (1991) Weak equivalence and the structures of cocycles of an ergodic automorphism. *Publ Res Inst Math Sci* 27(4):577–625
20. Bowles A, Fidkowski L, Marinello A, Silva CE (2001) Double ergodicity of nonsingular transformations and infinite measure-preserving staircase transformations. *Ill J Math* 45(3):999–1019
21. Chacon RV, Friedman NA (1965) Approximation and invariant measures. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 3: 286–295
22. Choksi JR, Eigen S, Prasad V (1989) Ergodic theory on homogeneous measure algebras revisited. In: Mauldin RD, Shortt RM, Silva CE (eds) *Measure and measurable dynamics*. *Contemp Math* 94, Amer Math Soc, Providence, pp 73–85
23. Choksi JR, Hawkins JM, Prasad VS (1987) Abelian cocycles for nonsingular ergodic transformations and the genericity of type III₁ transformations. *Monat fur Math* 103:187–205
24. Choksi JR, Kakutani S (1979) Residuality of ergodic measurable transformations and of ergodic transformations which preserve an infinite measure. *Ind Univ Math J* 28:453–469
25. Choksi JR, Nadkarni MG (1994) The maximal spectral type of a rank one transformation. *Canad Math Bull* 37(1):29–36
26. Choksi JR, Nadkarni MG (2000) Genericity of nonsingular transformations with infinite ergodic index. *Colloq Math* 84/85:195–201
27. Choksi JR, Prasad VS (1983) Approximation and Baire category theorems in ergodic theory. In: Belley JM, Dubois J, Morales P (eds) *Measure theory and its applications*. *Lect Notes Math* 1033. Springer, Berlin, pp 94–113
28. Connes A (1975) On the hierarchy of W Krieger. *Ill J Math* 19:428–432
29. Connes A, Feldman J, Weiss B (1981) An amenable equivalence relation is generated by a single transformation. *Ergod Theory Dynam Systems* 1:431–450
30. Connes A, Krieger W (1977) Measure space automorphisms, the normalizers of their full groups, and approximate finiteness. *J Funct Anal* 24(4):336–352
31. Connes A, Woods EJ (1985) Approximately transitive flows and ITPFI factors. *Ergod Theory Dynam Syst* 5(2):203–236
32. Connes A, Woods EJ (1989) Hyperfinite von Neumann algebras and Poisson boundaries of time dependent random walks. *Pac J Math* 37:225–243
33. Cornfeld IP, Fomin VS, Sinai YG (1982) *Ergodic theory*. *Grundlehren der Mathematischen Wissenschaften*, vol 245. Springer, New York
34. Danilenko AI (1995) The topological structure of Polish groups and groupoids of measure space transformations. *Publ Res Inst Math Sci* 31(5):913–940
35. Danilenko AI (1998) Quasinormal subrelations of ergodic equivalence relations. *Proc Amer Math Soc* 126(11): 3361–3370
36. Danilenko AI (2001) Funny rank one weak mixing for nonsingular Abelian actions. *Isr J Math* 121:29–54
37. Danilenko AI (2001) Strong orbit equivalence of locally compact Cantor minimal systems. *Int J Math* 12:113–123
38. Danilenko AI (2004) Infinite rank one actions and nonsingular Chacon transformations. *Ill J Math* 48(3):769–786
39. Danilenko AI (2007) On simplicity concepts for ergodic actions. *J d'Anal Math* 102:77–118
40. Danilenko AI (2007) (C, F)-actions in ergodic theory. In: Kapranov M, Kolyada S, Manin YI, Moree P, Potyagailo L (eds) *Geometry and Dynamics of Groups and Spaces*. *Progr Math* 265:325–351
41. Danilenko AI, Golodets VY (1996) On extension of cocycles to normalizer elements, outer conjugacy, and related problems. *Trans Amer Math Soc* 348(12):4857–4882
42. Danilenko AI, Hamachi T (2000) On measure theoretical analogues of the Takesaki structure theorem for type III factors. *Colloq Math* 84/85:485–493
43. Danilenko AI, Lemańczyk M (2005) A class of multipliers for \mathcal{W}^\perp . *Isr J Math* 148:137–168
44. Danilenko AI, Rudolph DJ: Conditional entropy theory in infinite measure and a question of Krengel. *Isr J Math*, to appear
45. Danilenko AI, Silva CE (2004) Multiple and polynomial recurrence for Abelian actions in infinite measure. *J Lond Math Soc* 2 69(1):183–200
46. Danilenko AI, Solomko AV: Infinite measure preserving flows with infinite ergodic index. *Colloq Math*, to appear
47. Day S, Grivna B, McCartney E, Silva CE (1999) Power Weakly Mixing Infinite Transformations. *N Y J Math* 5:17–24
48. del Junco A (1978) A simple measure-preserving transformation with trivial centralizer. *Pac J Math* 79:357–362
49. del Junco A, Rudolph DJ (1987) On ergodic actions whose self-joinings are graphs. *Ergod Theory Dynam Syst* 7:531–557
50. del Junco A, Silva CE (1995) Prime type III_λ automorphisms: An instance of coding techniques applied to nonsingular maps. In: Takahashi Y (ed) *Fractals and Dynamics*. Plenum, New York, pp 101–115
51. del Junco A, Silva CE (2003) On factors of nonsingular Cartesian products. *Ergod Theory Dynam Syst* 23(5):1445–1465
52. Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2008) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. *Colloq Math* 110:81–115
53. Dixmier J (1969) *Les C*-algèbres et leurs représentations*. Gauthier-Villars Editeur, Paris
54. Dooley AH, Hamachi T (2003) Nonsingular dynamical systems, Bratteli diagrams and Markov odometers. *Isr J Math* 138:93–123

55. Dooley AH, Hamachi T (2003) Markov odometer actions not of product type. *Ergod Theory Dynam Syst* 23(3):813–829
56. Dooley AH, Klemes I, Quas AN (1998) Product and Markov measures of type III. *J Aust Math Soc Ser A* 65(1):84–110
57. Dooley AH, Mortiss G: On the critical dimension of product odometers, preprint
58. Dooley AH, Mortiss G (2006) On the critical dimension and AC entropy for Markov odometers. *Monatsh Math* 149:193–213
59. Dooley AH, Mortiss G (2007) The critical dimensions of Hamachi shifts. *Tohoku Math J* 59(2):57–66
60. Dye H (1963) On groups of measure-preserving transformations I. *Amer J Math* 81:119–159, and II, *Amer J Math* 85: 551–576
61. Effros EG (1965) Transformation groups and C^* -algebras. *Ann Math* 81(2):38–55
62. Eigen SJ (1981) On the simplicity of the full group of ergodic transformations. *Isr J Math* 40(3–4):345–349
63. Eigen SJ (1982) The group of measure preserving transformations of $[0,1]$ has no outer automorphisms. *Math Ann* 259:259–270
64. Eigen S, Hajian A, Halverson K (1998) Multiple recurrence and infinite measure preserving odometers. *Isr J Math* 108:37–44
65. Fedorov A (1985) Krieger's theorem for cocycles, preprint
66. Feldman J, Moore CC (1977) Ergodic equivalence relations, cohomology, and von Neumann algebras. I. *Trans Amer Math Soc* 234:289–324
67. Ferenczi S (1985) Systèmes de rang un gauche. *Ann Inst H Poincaré Probab Statist* 21(2):177–186
68. Friedman NA (1970) Introduction to Ergodic Theory. Van Nostrand Reinhold Mathematical Studies, No 29. Van Nostrand Reinhold Co., New York
69. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets and diophantine approximation. *Math Syst Theory* 1: 1–49
70. Furstenberg H (1981) Recurrence in Ergodic Theory and Combinatorial Number Theory. Princeton University Press, Princeton
71. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations, The structure of attractors in dynamical systems. In: Markley NG, Martin JC, Perrizo W (eds) *Lecture Notes in Math* 668. Springer, Berlin, pp 127–132
72. Gårding L, Wightman AS (1954) Representation of anticommutation relations. *Proc Nat Acad Sci USA* 40:617–621
73. Giordano T, Skandalis G (1985) Krieger factors isomorphic to their tensor square and pure point spectrum flows. *J Funct Anal* 64(2):209–226
74. Giordano T, Skandalis G (1985) On infinite tensor products of factors of type I_2 . *Ergod Theory Dynam Syst* 5:565–586
75. Glasner E (1994) On the multipliers of \mathcal{W}^\perp . *Ergod Theory Dynam Syst* 14:129–140
76. Glimm J (1961) Locally compact transformation groups. *Trans Amer Math Soc* 101:124–138
77. Golodets YV (1969) A description of the representations of anticommutation relations. *Uspehi Matemat Nauk* 24(4):43–64
78. Golodets YV, Sinel'shchikov SD (1983) Existence and uniqueness of cocycles of ergodic automorphism with dense range in amenable groups. Preprint FTINT AN USSR, pp 19–83
79. Golodets YV, Sinel'shchikov SD (1985) Locally compact groups appearing as ranges of cocycles of ergodic Z -actions. *Ergod Theory Dynam Syst* 5:47–57
80. Golodets YV, Sinel'shchikov SD (1990) Amenable ergodic actions of groups and images of cocycles. *Dokl Akad Nauk SSSR* 312(6):1296–1299, in Russian
81. Golodets YV, Sinel'shchikov SD (1994) Classification and structure of cocycles of amenable ergodic equivalence relations. *J Funct Anal* 121(2):455–485
82. Gruher K, Hines F, Patel D, Silva CE, Waelder R (2003) Power weak mixing does not imply multiple recurrence in infinite measure and other counterexamples. *N Y J Math* 9:1–22
83. Hajian AB, Kakutani S (1964) Weakly wandering sets and invariant measures. *Trans Amer Math Soc* 110:136–151
84. Halmos PR (1946) An ergodic theorem. *Proc Nat Acad Sci USA* 32:156–161
85. Halmos PR (1956) Lectures on ergodic theory. *Publ Math Soc Jpn* 3
86. Hamachi T (1981) The normalizer group of an ergodic automorphism of type III and the commutant of an ergodic flow. *J Funct Anal* 40:387–403
87. Hamachi T (1981) On a Bernoulli shift with nonidentical factor measures. *Ergod Theory Dynam Syst* 1:273–283
88. Hamachi T (1992) A measure theoretical proof of the Connes–Woods theorem on AT-flows. *Pac J Math* 154:67–85
89. Hamachi T, Kosaki H (1993) Orbital factor map. *Ergod Theory Dynam Syst* 13:515–532
90. Hamachi T, Osikawa M (1981) Ergodic groups of automorphisms and Krieger's theorems. *Semin Math Sci* 3, Keio Univ
91. Hamachi T, Osikawa M (1986) Computation of the associated flows of $ITPFI_2$ factors of type III_0 . In: *Geometric methods in operator algebras*. Pitman Res Notes Math Ser 123, Longman Sci Tech, Harlow, pp 196–210
92. Hamachi T, Silva CE (2000) On nonsingular Chacon transformations. *Ill J Math* 44:868–883
93. Hawkins JM (1982) Non- $ITPFI$ diffeomorphisms. *Isr J Math* 42:117–131
94. Hawkins JM (1983) Smooth type III diffeomorphisms of manifolds. *Trans Amer Math Soc* 276:625–643
95. Hawkins JM (1990) Diffeomorphisms of manifolds with nonsingular Poincaré flows. *J Math Anal Appl* 145(2):419–430
96. Hawkins JM (1990) Properties of ergodic flows associated to product odometers. *Pac J Math* 141:287–294
97. Hawkins J, Schmidt K (1982) On C^2 -diffeomorphisms of the circle which are of type III_1 . *Invent Math* 66(3):511–518
98. Hawkins J, Silva CE (1997) Characterizing mildly mixing actions by orbit equivalence of products. In: *Proceedings of the New York Journal of Mathematics Conference*, June 9–13 1997. *N Y J Math* 3A:99–115
99. Hawkins J, Woods EJ (1984) Approximately transitive diffeomorphisms of the circle. *Proc Amer Math Soc* 90(2):258–262
100. Herman M (1979) Construction de difféomorphismes ergodiques, preprint
101. Herman M-R (1979) Sur la conjugaison différentiable des difféomorphismes du cercle a des rotations. *Inst Hautes Etudes Sci Publ Math* 49:5–233, in French
102. Herman RH, Putnam IF, Skau CF (1992) Ordered Bratteli diagrams, dimension groups and topological dynamics. *Int J Math* 3(6):827–864
103. Host B, Méla J-F, Parreau F (1986) Analyse harmonique des mesures. *Astérisque* 135–136:1–261
104. Host B, Méla J-F, Parreau F (1991) Nonsingular transformations and spectral analysis of measures. *Bull Soc Math France* 119:33–90

105. Hurewicz W (1944) Ergodic theorem without invariant measure. *Ann Math* 45:192–206
106. Inoue K (2004) Isometric extensions and multiple recurrence of infinite measure preserving systems. *Isr J Math* 140: 245–252
107. Ionescu Tulcea A (1965) On the category of certain classes of transformations in ergodic theory. *Trans Amer Math Soc* 114:261–279
108. Ismagilov RS (1987) Application of a group algebra to problems on the tail σ -algebra of a random walk on a group and to problems on the ergodicity of a skew action. *Izv Akad Nauk SSSR Ser Mat* 51(4):893–907
109. Jaworsky W (1994) Strongly approximately transitive actions, the Choquet–Deny theorem, and polynomial growth. *Pac J Math* 165:115–129
110. James J, Koberda T, Lindsey K, Silva CE, Speh P (2008) Measurable sensitivity. *Proc Amer Math Soc* 136(10):3549–3559
111. Janvresse E, Meyerovitch T, de la Rue T, Roy E: Poisson suspensions and entropy of infinite transformations, preprint
112. Kaimanovich VA, Vershik AM (1983) Random walks on groups: boundary and entropy. *Ann Probab* 11:457–490
113. Kakutani S, Parry W (1963) Infinite measure preserving transformations with “mixing”. *Bull Amer Math Soc* 69:752–756
114. Katznelson Y (1977) Sigma-finite invariant measures for smooth mappings of the circle. *J Anal Math* 31:1–18
115. Katznelson Y (1979) The action of diffeomorphism of the circle on the Lebesgue measure. *J Anal Math* 36:156–166
116. Katznelson Y, Weiss B (1972) The construction of quasi-invariant measures. *Isr J Math* 12:1–4
117. Katznelson Y, Weiss B (1991) The classification of nonsingular actions, revisited. *Ergod Theory Dynam Syst* 11:333–348
118. Kirillov AA (1978) Elements of the theory of representations. Nauka, Moscow
119. Krengel U (1967) Entropy of conservative transformations. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 7:161–181
120. Krengel U (1969) Darstellungssätze für Strömungen und Halbströmungen, vol II. *Math Ann* 182:1–39
121. Krengel U (1970) Transformations without finite invariant measure have finite strong generators. In: *Contributions to Ergodic Theory and Probability*. Proc Conf Ohio State Univ, Columbus, Ohio. Springer, Berlin, pp 133–157
122. Krengel U (1976) On Rudolph’s representation of aperiodic flows. *Ann Inst H Poincaré Sect B (NS)* 12(4):319–338
123. Krengel U (1985) Ergodic Theorems. De Gruyter Studies in Mathematics, Berlin
124. Krengel U, Sucheston L (1969) On mixing in infinite measure spaces. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 13: 150–164
125. Krieger W (1969) On nonsingular transformations of a measure space, vol I, II. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 11:83–119
126. Krieger W (1970) On the Araki–Woods asymptotic ratio set and nonsingular transformations of a measure space. In: *Contributions to Ergodic Theory and Probability*. Proc Conf Ohio State Univ, Columbus, Ohio. In: *Lecture Notes in Math*, vol 160. Springer, Berlin, pp 158–177
127. Krieger W (1972) On the infinite product construction of nonsingular transformations of a measure space. *Invent Math* 15:144–163; Erratum in 26:323–328
128. Krieger W (1976) On Borel automorphisms and their quasi-invariant measures. *Math Z* 151:19–24
129. Krieger W (1976) On ergodic flows and isomorphism of factors. *Math Ann* 223:19–70
130. Kubo I (1969) Quasi-flows. *Nagoya Math J* 35:1–30
131. Ledrappier F, Sarig O (2007) Invariant measures for the horocycle flow on periodic hyperbolic surfaces. *Isr J Math* 160:281–315
132. Lehrer E, Weiss B (1982) An ε -free Rokhlin lemma. *Ergod Theory Dynam Syst* 2:45–48
133. Lemańczyk M, Parreau F (2003) Rokhlin extensions and lifting disjointness. *Ergod Theory Dynam Syst* 23:1525–1550
134. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self-joinings are Gaussian. *Fund Math* 164:253–293
135. Mackey GW (1966) Ergodic theory and virtual group. *Math Ann* 166:187–207
136. Maharam D (1964) Incompressible transformations. *Fund Math* LVI:35–50
137. Mandrekar V, Nadkarni M (1969) On ergodic quasi-invariant measures on the circle group. *J Funct Anal* 3:157–163
138. Matui H (2002) Topological orbit equivalence of locally compact Cantor minimal systems. *Ergod Theory Dynam Syst* 22:1871–1903
139. Méla J-F (1983) Groupes de valeurs propres des systèmes dynamiques et sous-groupes saturés du cercle. *CR Acad Sci Paris Sér I Math* 296(10):419–422
140. Meyerovitch T (2007) Extensions and Multiple Recurrence of infinite measure preserving systems, preprint. ArXiv: <http://arxiv.org/abs/math/0703914>
141. Moore CC (1967) Invariant measures on product spaces. *Proc Fifth Berkeley Symp*. University of California Press, Berkeley, pp 447–459
142. Moore CC (1982) Ergodic theory and von Neumann algebras. *Proc Symp Pure Math* 38:179–226
143. Moore CC, Schmidt K (1980) Coboundaries and homomorphisms for nonsingular actions and a problem of H Helson. *Proc Lond Math Soc* 3 40:443–475
144. Mortiss G (2000) A non-singular inverse Vitali lemma with applications. *Ergod Theory Dynam Syst* 20:1215–1229
145. Mortiss G (2002) Average co-ordinate entropy. *J Aust Math Soc* 73:171–186
146. Mortiss G (2003) An invariant for nonsingular isomorphism. *Ergod Theory Dynam Syst* 23:885–893
147. Nadkarni MG (1979) On spectra of nonsingular transformations and flows. *Sankhya Ser A* 41(1–2):59–66
148. Nadkarni MG (1998) Spectral theory of dynamical systems. In: *Birkhäuser Advanced Texts: Basler Lehrbücher*. Birkhäuser, Basel
149. Ornstein D (1960) On invariant measures. *Bull Amer Math Soc* 66:297–300
150. Ornstein D (1972) On the Root Problem in Ergodic Theory. In: *Proc Sixth Berkeley Symp Math Stat Probab*. University of California Press, Berkeley, pp 347–356
151. Osikawa M (1977/78) Point spectra of nonsingular flows. *Publ Res Inst Math Sci* 13:167–172
152. Osikawa M (1988) Ergodic properties of product type odometers. *Springer Lect Notes Math* 1299:404–414
153. Osikawa M, Hamachi T (1971) On zero type and positive type transformations with infinite invariant measures. *Mem Fac Sci Kyushu Univ* 25:280–295
154. Parreau F, Roy E: Poisson joinings of Poisson suspensions, preprint

155. Parry W (1963) An ergodic theorem of information theory without invariant measure. *Proc Lond Math Soc* 3 13:605–612
156. Parry W (1965) Ergodic and spectral analysis of certain infinite measure preserving transformations. *Proc Amer Math Soc* 16:960–966
157. Parry W (1966) Generators and strong generators in ergodic theory. *Bull Amer Math Soc* 72:294–296
158. Parry W (1969) *Entropy and generators in ergodic theory*. WA Benjamin, New York, Amsterdam
159. Parthasarathy KR, Schmidt K (1977) On the cohomology of a hyperfinite action. *Monatsh Math* 84(1):37–48
160. Ramsay A (1971) Virtual groups and group actions. *Adv Math* 6:243–322
161. Rokhlin VA (1949) Selected topics from the metric theory of dynamical systems. *Uspekhi Mat Nauk* 4:57–125
162. Rokhlin VA (1965) Generators in ergodic theory, vol II. *Vestnik Leningrad Univ* 20(13):68–72, in Russian, English summary
163. Rosinsky J (1995) On the structure of stationary stable processes. *Ann Probab* 23:1163–1187
164. Roy E (2005) Mesures de Poisson, infinie divisibilité et propriétés ergodiques. Thèse de doctorat de l'Université Paris 6
165. Roy E (2007) Ergodic properties of Poissonian ID processes. *Ann Probab* 35:551–576
166. Roy E: Poisson suspensions and infinite ergodic theory, preprint
167. Rudolph DJ (1985) Restricted orbit equivalence. *Mem Amer Math Soc* 54(323)
168. Rudolph D, Silva CE (1989) Minimal self-joinings for nonsingular transformations. *Ergod Theory Dynam Syst* 9:759–800
169. Ryzhikov VV (1994) Factorization of an automorphism of a full Boolean algebra into the product of three involutions. *Mat Zametki* 54(2):79–84, 159; in Russian. Translation in: *Math Notes* 54(1–2):821–824
170. Sachdeva U (1971) On category of mixing in infinite measure spaces. *Math Syst Theory* 5:319–330
171. Samorodnitsky G (2005) Null flows, positive flows and the structure of stationary symmetric stable processes. *Ann Probab* 33:1782–1803
172. Sarig O (2004) Invariant measures for the horocycle flows on Abelian covers. *Invent Math* 157:519–551
173. Schmidt K (1977) Cocycles on ergodic transformation groups. *Macmillan Lectures in Mathematics*, vol 1. Macmillan Company of India, Delhi
174. Schmidt K (1977) Infinite invariant measures in the circle. *Symp Math* 21:37–43
175. Schmidt K (1982) Spectra of ergodic group actions. *Isr J Math* 41(1–2):151–153
176. Schmidt K (1984) On recurrence. *Z Wahrscheinlichkeitstheorie Verw Gebiete* 68:75–95
177. Schmidt K, Walters P (1982) Mildly mixing actions of locally compact groups. *Proc Lond Math Soc* 45:506–518
178. Shelah S, Weiss B (1982) Measurable recurrence and quasi-invariant measures. *Isr Math J* 43:154–160
179. Silva CE, Thieullen P (1991) The subadditive ergodic theorem and recurrence properties of Markovian transformations. *J Math Anal Appl* 154(1):83–99
180. Silva CE, Thieullen P (1995) A skew product entropy for nonsingular transformations. *J Lond Math Soc* 2 52:497–516
181. Silva CE, Witte D (1992) On quotients of nonsingular actions whose self-joinings are graphs. *Int J Math* 5:219–237
182. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. In: *Ergodic theory and its connections with harmonic analysis* (Alexandria, 1993), pp 207–235. *Lond Math Soc Lect Notes Ser* 205. Cambridge Univ Press, Cambridge
183. Ullman D (1987) A generalization of a theorem of Atkinson to non-invariant measures. *Pac J Math* 130:187–193
184. Vershik AM (1983) Manyvalued mappings with invariant measure (polymorphisms) and Markov processes. *J Sov Math* 23:2243–2266
185. Vershik AM, Kerov SV (1985) Locally semisimple algebras. In: *Combinatorial theory and K_0 -functor*. *Mod Probl Math* 26:3–56
186. Zimmer RJ (1977) Random walks on compact groups and the existence of cocycles. *Isr J Math* 26:84–90
187. Zimmer RJ (1978) Amenable ergodic group actions and an application to Poisson boundaries of random walks. *J Funct Anal* 27:350–372
188. Zimmer RJ (1984) *Ergodic theory and semisimple Lie groups*. Birkhäuser, Basel, Boston

Ergodic Theory: Recurrence

NIKOS FRANTZIKINAKIS, RANDALL MCCUTCHEON
Department of Mathematics, University of Memphis,
Memphis, USA

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)
[Quantitative Poincaré Recurrence](#)
[Subsequence Recurrence](#)
[Multiple Recurrence](#)
[Connections with Combinatorics and Number Theory](#)
[Future Directions](#)
[Bibliography](#)

Glossary

Almost every, essentially Given a Lebesgue measure space (X, \mathcal{B}, μ) , a property $P(x)$ predicated of elements of X is said to hold for almost every $x \in X$, if the set $X \setminus \{x: P(x) \text{ holds}\}$ has zero measure. Two sets $A, B \in \mathcal{B}$ are essentially disjoint if $\mu(A \cap B) = 0$.

Conservative system Is an infinite measure preserving system such that for no set $A \in \mathcal{B}$ with positive measure are $A, T^{-1}A, T^{-2}A, \dots$ pairwise essentially disjoint.

(c_n) -Conservative system If $(c_n)_{n \in \mathbb{N}}$ is a decreasing sequence of positive real numbers, a conservative

ergodic measure preserving transformation T is (c_n) -conservative if for some non-negative function $f \in L^1(\mu)$, $\sum_{n=1}^{\infty} c_n f(T^n x) = \infty$ a.e.

Doubling map If \mathbb{T} is the interval $[0, 1]$ with its end-points identified and addition performed modulo 1, the (non-invertible) transformation $T: \mathbb{T} \rightarrow \mathbb{T}$, defined by $Tx = 2x \bmod 1$, preserves Lebesgue measure, hence induces a measure preserving system on \mathbb{T} .

Ergodic system Is a measure preserving system (X, \mathcal{B}, μ, T) (finite or infinite) such that every $A \in \mathcal{B}$ that is T -invariant (i.e. $T^{-1}A = A$) satisfies either $\mu(A) = 0$ or $\mu(X \setminus A) = 0$. (One can check that the rotation R_α is ergodic if and only if α is irrational, and that the doubling map is ergodic).

Ergodic decomposition Every measure preserving system (X, \mathcal{X}, μ, T) can be expressed as an integral of ergodic systems; for example, one can write $\mu = \int \mu_t d\lambda(t)$, where λ is a probability measure on $[0, 1]$ and μ_t are T -invariant probability measures on (X, \mathcal{X}) such that the systems $(X, \mathcal{X}, \mu_t, T)$ are ergodic for $t \in [0, 1]$.

Ergodic theorem States that if (X, \mathcal{B}, μ, T) is a measure preserving system and $f \in L^2(\mu)$, then $\lim_{N \rightarrow \infty} \left\| \frac{1}{N} \sum_{n=1}^N T^n f - P_f \right\|_{L^2(\mu)} = 0$, where P_f denotes the orthogonal projection of the function f onto the subspace $\{f \in L^2(\mu): Tf = f\}$.

Hausdorff a -measure Let (X, \mathcal{B}, μ, T) be a measure preserving system endowed with a μ -compatible metric d . The Hausdorff a -measure $\mathcal{H}_a(X)$ of X is an outer measure defined for all subsets of X as follows: First, for $A \subset X$ and $\varepsilon > 0$ let $\mathcal{H}_{a,\varepsilon}(A) = \inf \left\{ \sum_{i=1}^{\infty} r_i^a \right\}$, where the infimum is taken over all countable coverings of A by sets $U_i \subset X$ with diameter $r_i < \varepsilon$. Then define $\mathcal{H}_a(A) = \limsup_{\varepsilon \rightarrow 0} \mathcal{H}_{a,\varepsilon}(A)$.

Infinite measure preserving system Same as measure preserving system, but $\mu(X) = \infty$.

Invertible system Is a measure preserving system (X, \mathcal{B}, μ, T) (finite or infinite), with the property that there exists $X_0 \in \mathcal{B}$, with $\mu(X \setminus X_0) = 0$, and such that the transformation $T: X_0 \rightarrow X_0$ is bijective, with T^{-1} measurable.

Measure preserving system Is a quadruple (X, \mathcal{B}, μ, T) , where X is a set, \mathcal{B} is a σ -algebra of subsets of X (i.e. \mathcal{B} is closed under countable unions and complementation), μ is a probability measure (i.e. a countably additive function from \mathcal{B} to $[0, 1]$ with $\mu(X) = 1$), and $T: X \rightarrow X$ is measurable (i.e. $T^{-1}A = \{x \in X: Tx \in A\} \in \mathcal{B}$ for $A \in \mathcal{B}$), and μ -preserving (i.e. $\mu(T^{-1}A) = \mu(A)$). Moreover, throughout the discussion we assume that the measure space (X, \mathcal{B}, μ) is Lebesgue (see Sect. 1.0 in [2]).

μ -Compatible metric Is a separable metric on X , where (X, \mathcal{B}, μ) is a probability space, having the property that open sets are measurable.

Positive definite sequence Is a complex-valued sequence $(a_n)_{n \in \mathbb{Z}}$ such that for any $n_1, \dots, n_k \in \mathbb{Z}$ and $z_1, \dots, z_k \in \mathbb{C}$, $\sum_{i,j=1}^k a_{n_i-n_j} z_i \bar{z}_j \geq 0$.

Rotations on \mathbb{T} If \mathbb{T} is the interval $[0, 1]$ with its end-points identified and addition performed modulo 1, then for every $\alpha \in \mathbb{R}$ the transformation $R_\alpha: \mathbb{T} \rightarrow \mathbb{T}$, defined by $R_\alpha x = x + \alpha \bmod 1$, preserves Lebesgue measure on \mathbb{T} and hence induces a measure preserving system on \mathbb{T} .

Syndetic set Is a subset $E \subset \mathbb{Z}$ having bounded gaps. If G is a general discrete group, a set $E \subset G$ is syndetic if $G = FE$ for some finite set $F \subset G$.

Upper density Is the number $\bar{d}(A) = \limsup_{N \rightarrow \infty} (|A \cap \{-N, \dots, N\}|) / (2N + 1)$, where $A \subset \mathbb{Z}$ (assuming the limit to exist). Alternatively for measurable $E \subset \mathbb{R}^m$, $\bar{D}(E) = \limsup_{l(S) \rightarrow \infty} (m(S \cap E)) / (m(S))$, where S ranges over all cubes in \mathbb{R}^m , and $l(S)$ denotes the length of the shortest edge of S .

Notation The following notation will be used throughout the article: $Tf = f \circ T$, $\{x\} = x - [x]$, $D\text{-}\lim_{n \rightarrow \infty} (a_n) = a \leftrightarrow \bar{d}(\{n: |a_n - a| > \varepsilon\}) = 0$ for every $\varepsilon > 0$.

Definition of the Subject

The basic principle that lies behind several recurrence phenomena is that the typical trajectory of a system with finite volume comes back infinitely often to any neighborhood of its initial point. This principle was first exploited by Poincaré in his 1890 King Oscar prize-winning memoir that studied planetary motion. Using the prototype of an ergodic-theoretic argument, he showed that in any system of point masses having fixed total energy that restricts its dynamics to bounded subsets of its phase space, the typical state of motion (characterized by configurations and velocities) must recur to an arbitrary degree of approximation.

Among the recurrence principle's more spectacularly counterintuitive ramifications is that isolated ideal gas systems that do not lose energy will return arbitrarily closely to their initial states, even when such a return entails a decrease in entropy from equilibrium, in apparent contradiction to the second law of thermodynamics. Such concerns, previously canvassed by Poincaré himself, were more infamously expounded by Zermelo [74] in 1896. Subsequent clarifications by Boltzmann, Maxwell and others led to an improved understanding of the second law's primarily statistical nature. (For an interesting historical/philosophical

discussion, see [68]; also [10]. For a probabilistic analysis of the likelihood of observing second law violations in small systems over short time intervals, see [28]).

These discoveries had a profound impact in dynamics, and the theory of measure preserving transformations (ergodic theory) evolved from these developments. Since then, the Poincaré recurrence principle has been applied to a variety of different fields in mathematics, physics, and information theory. In this article we survey the impact it has had in ergodic theory, especially as pertains to the field of *ergodic Ramsey theory*. (The heavy emphasis herein on the latter reflects authorial interest, and is not intended to transmit a proportionate image of the broader landscape of research relating to recurrence in ergodic theory.) Background information we assume in this article can be found in the books [35,63,71] (► [Measure Preserving Systems](#)).

Introduction

In this section we shall give several formulations of the Poincaré recurrence principle using the language of ergodic theory. Roughly speaking, the principle states that in a finite (or conservative) measure preserving system, every set of positive measure (or almost every point) comes back to itself infinitely many times under iteration. Despite the profound importance of these results, their proofs are extremely simple.

Theorem 1 (Poincaré Recurrence for Sets) *Let (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then $\mu(A \cap T^{-n}A) > 0$ for infinitely many $n \in \mathbb{N}$.*

Proof Since T is measure preserving, the sets $A, T^{-1}A, T^{-2}A, \dots$ have the same measure. These sets cannot be pairwise essentially disjoint, since then the union of finitely many of them would have measure greater than $\mu(X) = 1$. Therefore, there exist $m, n \in \mathbb{N}$, with $n > m$, such that $\mu(T^{-m}A \cap T^{-n}A) > 0$. Again since T is measure preserving, we conclude that $\mu(A \cap T^{-k}A) > 0$, where $k = n - m > 0$. Repeating this argument for the iterates $A, T^{-m}A, T^{-2m}A, \dots$, for all $m \in \mathbb{N}$, we easily deduce that $\mu(A \cap T^{-n}A) > 0$ for infinitely many $n \in \mathbb{N}$. \square

We remark that the above argument actually shows that $\mu(A \cap T^{-n}A) > 0$ for some $n \leq \lceil \frac{1}{\mu(A)} \rceil + 1$.

Theorem 2 (Poincaré Recurrence for Points) *Let (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$. Then for almost every $x \in A$ we have that $T^n x \in A$ for infinitely many $n \in \mathbb{N}$.*

Proof Let B be the set of $x \in A$ such that $T^n x \notin A$ for all $n \in \mathbb{N}$. Notice that $B = A \setminus \bigcup_{n \in \mathbb{N}} T^{-n}A$; in particular, B is measurable. Since the iterates $B, T^{-1}B, T^{-2}B, \dots$ are

pairwise essentially disjoint, we conclude (as in the proof of Theorem 1) that $\mu(B) = 0$. This shows that for almost every $x \in A$ we have that $T^n x \in A$ for some $n \in \mathbb{N}$. Repeating this argument for the transformation T^m in place of T for all $m \in \mathbb{N}$, we easily deduce the advertised statement. \square

Next we give a variation of Poincaré recurrence for measure preserving systems endowed with a compatible metric:

Theorem 3 (Poincaré Recurrence for Metric Systems)

Let (X, \mathcal{B}, μ, T) be a measure preserving system, and suppose that X is endowed with a μ -compatible metric. Then for almost every $x \in X$ we have

$$\liminf_{n \rightarrow \infty} d(x, T^n x) = 0.$$

The proof of this result is similar to the proof of Theorem 2 (see p. 61 in [35]). Applying this result to the doubling map $Tx = 2x$ on \mathbb{T} , we get that for almost every $x \in X$, every string of zeros and ones in the dyadic expansion of x occurs infinitely often.

We remark that all three formulations of the Poincaré Recurrence Theorem that we have given hold for conservative systems as well. See, e.g., [2] for details.

This article is structured as follows. In Sect. “[Quantitative Poincaré Recurrence](#)” we give a few quantitative versions of the previously mentioned qualitative results. In Sects. “[Subsequence Recurrence](#)” and “[Multiple Recurrence](#)” we give several refinements of the Poincaré recurrence theorem, by restricting the scope of the return time n , and by considering multiple intersections (for simplicity we focus on \mathbb{Z} -actions). In Sect. “[Connections with Combinatorics and Number Theory](#)” we give various implications of the recurrence results in combinatorics and number theory (► [Ergodic Theory: Interactions with Combinatorics and Number Theory](#)). Lastly, in Sect. “[Future Directions](#)” we give several open problems related to the material presented in Sects. “[Subsequence Recurrence](#)” to “[Connections with Combinatorics and Number Theory](#)”.

Quantitative Poincaré Recurrence

Early Results

For applications it is desirable to have quantitative versions of the results mentioned in the previous section. For example one would like to know how large $\mu(A \cap T^{-n}A)$ can be made and for how many n .

Theorem 4 (Khinchine [55]) *Let (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$. Then for every $\varepsilon > 0$ we*

have $\mu(A \cap T^{-n}A) > \mu(A)^2 - \varepsilon$ for a set of $n \in \mathbb{N}$ that has bounded gaps.

By considering the doubling map $Tx = 2x$ on \mathbb{T} and letting $A = \mathbf{1}_{[0, 1/2]}$, it is easy to check that the lower bound of the previous result cannot be improved. We also remark that it is not possible to estimate the size of the gap by a function of $\mu(A)$ alone. One can see this by considering the rotations $R_k x = x + 1/k$ for $k \in \mathbb{N}$, defined on \mathbb{T} , and letting $A = \mathbf{1}_{[0, 1/3]}$.

Concerning the second version of the Poincaré recurrence theorem, it is natural to ask whether for almost every $x \in X$ the set of return times $S_x = \{n \in \mathbb{N} : T^n x \in A\}$ has bounded gaps. This is not the case, as one can see by considering the doubling map $Tx = 2x$ on \mathbb{T} with the Lebesgue measure, and letting $A = \mathbf{1}_{[0, 1/2]}$. Since Lebesgue almost every $x \in \mathbb{T}$ contains arbitrarily large blocks of ones in its dyadic expansion, the set S_x has unbounded gaps. Nevertheless, as an easy consequence of the Birkhoff ergodic theorem [19], one has the following:

Theorem 5 *Let (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for almost every $x \in X$ the set $S_x = \{n \in \mathbb{N} : T^n x \in A\}$ has well defined density and $\int d(S_x) d\mu(x) = \mu(A)$. Furthermore, for ergodic measure preserving systems we have $d(S_x) = \mu(A)$ a. e.*

Another question that arises naturally is, given a set A with positive measure and an $x \in A$, how long should one wait until some iterate $T^n x$ of x hits A ? By considering an irrational rotation R_α on \mathbb{T} , where α is very near to, but not less than, $\frac{1}{100}$, and letting $A = \mathbf{1}_{[0, 1/2]}$, one can see that the first return time is a member of the set $\{1, 50, 51\}$. So it may come as a surprise that the average first return time does not depend on the system (as long as it is ergodic), but only on the measure of the set A .

Theorem 6 (Kac [51]) *Let (X, \mathcal{B}, μ, T) be an ergodic measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. For $x \in X$ define $R_A(x) = \min\{n \in \mathbb{N} : T^n x \in A\}$. Then for $x \in A$ the expected value of $R_A(x)$ is $1/\mu(A)$, i. e. $\int_A R_A(x) d\mu = 1$.*

More Recent Results

As we mentioned in the previous section, if the space X is endowed with a μ -compatible metric d , then for almost every $x \in X$ we have that $\liminf_{n \rightarrow \infty} d(x, T^n x) = 0$. A natural question is, how much iteration is needed to come back within a small distance of a given typical point? Under some additional hypothesis on the metric d we have the following answer:

Theorem 7 (Boshernitzan [20]) *Let (X, \mathcal{B}, μ, T) be a measure preserving system endowed with a μ -compatible metric d . Assume that the Hausdorff a -measure $\mathcal{H}_a(X)$ of X is σ -finite (i. e., X is a countable union of sets X_i with $\mathcal{H}_a(X_i) < \infty$). Then for almost every $x \in X$,*

$$\liminf_{n \rightarrow \infty} \{n^{\frac{1}{a}} \cdot d(x, T^n x)\} < \infty.$$

Furthermore, if $\mathcal{H}_a(X) = 0$, then for almost every $x \in X$,

$$\liminf_{n \rightarrow \infty} \{n^{\frac{1}{a}} \cdot d(x, T^n x)\} = 0.$$

One can see from rotations by “badly approximable” vectors $\alpha \in \mathbb{T}^k$ that the exponent $1/a$ in the previous theorem cannot be improved. Several applications of Theorem 7 to billiard flows, dyadic transformations, symbolic flows and interval exchange transformations are given in [20]. For a related result dealing with mean values of the limits in Theorem 7 see [67].

An interesting connection between rates of recurrence and entropy of an ergodic measure preserving system was established by Ornstein and Weiss [62], following earlier work of Wyner and Ziv [73]:

Theorem 8 (Ornstein and Weiss [62]) *Let (X, \mathcal{B}, μ, T) be an ergodic measure preserving system and \mathcal{P} be a finite partition of X . Let $P_n(x)$ be the element of the partition $\bigvee_{i=0}^{n-1} T^{-i} \mathcal{P} = \{\bigcap_{i=0}^{n-1} T^{-i} P^{(i)} : P^{(i)} \in \mathcal{P}, 0 \leq i < n\}$ that contains x . Then for almost every $x \in X$, the first return time $R_n(x)$ of x to $P_n(x)$ is asymptotically equivalent to $e^{h(T, \mathcal{P})n}$, where $h(T, \mathcal{P})$ denotes the entropy of the system with respect to the partition \mathcal{P} . More precisely,*

$$\lim_{n \rightarrow \infty} \frac{\log R_n(x)}{n} = h(T, \mathcal{P}).$$

An extension of the above result to some classes of infinite measure preserving systems was given in [42].

Another connection of recurrence rates, this time with the local dimension of an invariant measure, is given by the next result:

Theorem 9 (Barreira [4]) *Let (X, \mathcal{B}, μ, T) be an ergodic measure preserving system. Define the upper and lower recurrence rates*

$$\underline{R}(x) = \liminf_{r \rightarrow 0} \frac{\log \tau_r(x)}{-\log r} \quad \text{and} \quad \overline{R}(x) = \limsup_{r \rightarrow 0} \frac{\log \tau_r(x)}{-\log r},$$

where $\tau_r(x)$ is the first return time of $T^k x$ to $B(x, r)$, and the upper and lower pointwise dimensions

$$\underline{d}_\mu(x) = \liminf_{r \rightarrow 0} \frac{\log \mu(B(x, r))}{\log r} \quad \text{and} \\ \bar{d}_\mu(x) = \limsup_{r \rightarrow 0} \frac{\log \mu(B(x, r))}{\log r}.$$

Then for almost every $x \in X$, we have

$$\underline{R}(x) \leq \underline{d}_\mu(x) \quad \text{and} \quad \bar{R}(x) \leq \bar{d}_\mu(x).$$

Roughly speaking, this theorem asserts that for typical $x \in X$ and for small r , the first return time of x to $B(x, r)$ is at most $r^{-d_\mu(x)}$. Since $\underline{d}_\mu(x) \leq \mathcal{H}_a(X)$ for almost every $x \in X$, we can conclude the first part of Theorem 7 from Theorem 9. For related results the interested reader should consult the survey [5] and the bibliography therein.

We also remark that the previous results and related concepts have been applied to estimate the dimension of certain strange attractors (see [49] and the references therein) and the entropy of some Gibbsian systems [25].

We end this section with a result that connects “wandering rates” of sets in infinite measure preserving systems with their “recurrence rates”. The next theorem follows easily from a result about lower bounds on ergodic averages for measure preserving systems due to Leibman [57]; a weaker form for conservative, ergodic systems can be found in Aaronson [1].

Theorem 10 *Let (X, \mathcal{B}, μ, T) be an infinite measure preserving system, and $A \in \mathcal{B}$ with $\mu(A) < \infty$. Then for all $N \in \mathbb{N}$,*

$$\left(\frac{\mu \left(\bigcup_{n=0}^{N-1} T^{-n} A \right)}{N} \cdot \sum_{n=0}^{N-1} \mu(A \cap T^{-n} A) \right) \geq \frac{1}{2} \cdot (\mu(A))^2.$$

Subsequence Recurrence

In this section we discuss what restrictions one can impose on the set of return times in the various versions of the Poincaré recurrence theorem. We start with:

Definition 11 Let $R \subset \mathbb{Z}$. Then R is a set of:

- (a) *Recurrence* if for any invertible measure preserving system (X, \mathcal{B}, μ, T) , and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in R$ such that $\mu(A \cap T^{-n} A) > 0$.
- (b) *Topological recurrence* if for every compact metric space (X, d) , continuous transformation $T: X \rightarrow X$

and every $\varepsilon > 0$, there are $x \in X$ and nonzero $n \in R$ such that $d(x, T^n x) < \varepsilon$.

It is easy to check that the existence of a single $n \in R$ satisfying the previous recurrence conditions actually guarantees the existence of infinitely many $n \in R$ satisfying the same conditions. Moreover, if R is a set of recurrence then one can see from existence of some T -invariant measure μ that R is also a set of topological recurrence. A (complicated) example showing that the converse is not true was given by Kriz [56].

Before giving a list of examples of sets of (topological) recurrence, we discuss some necessary conditions: A set of topological recurrence must contain infinitely many multiples of every positive integer, as one can see by considering rotations on \mathbb{Z}_d , $d \in \mathbb{N}$. Hence, the sets $\{2n + 1, n \in \mathbb{N}\}$, $\{n^2 + 1, n \in \mathbb{N}\}$, $\{p + 2, p \text{ prime}\}$ are not good for (topological) recurrence. If $(s_n)_{n \in \mathbb{N}}$ is a lacunary sequence (meaning $\liminf_{n \rightarrow \infty} (s_{n+1}/s_n) = \rho > 1$), then one can construct an irrational number α such that $\{s_n \alpha\} \in [\delta, 1 - \delta]$ for all large $n \in \mathbb{N}$, where $\delta > 0$ depends on ρ (see [54], for example). As a consequence, the sequence $(s_n)_{n \in \mathbb{N}}$ is not good for (topological) recurrence.

Lastly, we mention that by considering product systems, one can immediately show that any set of (topological) recurrence R is partition regular, meaning that if R is partitioned into finitely many pieces then at least one of these pieces must still be a set of (topological) recurrence. Using this observation, one concludes for example that any union of finitely many lacunary sequences is not a set of recurrence.

We present now some examples of sets of recurrence:

Theorem 12 *The following are sets of recurrence:*

- (i) Any set of the form $\bigcup_{n \in \mathbb{N}} \{a_n, 2a_n, \dots, na_n\}$ where $a_n \in \mathbb{N}$.
- (ii) Any IP-set, meaning a set that consists of all finite sums of the members of some infinite set.
- (iii) Any difference set $S - S$, meaning a set that consists of all possible differences of the members of some infinite set S .
- (iv) The set $\{p(n), n \in \mathbb{N}\}$ where p is any nonconstant integer polynomial with $p(0) = 0$ [35,66] (In fact we only have to assume that the range of the polynomial contains multiples of an arbitrary positive integer [53]).
- (v) The set $\{p(n), n \in S\}$, where p is an integer polynomial with $p(0) = 0$ and S is any IP-set [12].
- (vi) The set of values of an admissible generalized polynomial (this class contains in particular the

smallest function algebra G containing all integer polynomials having zero constant term and such that if $g_1, \dots, g_k \in G$ and $c_1, \dots, c_k \in \mathbb{R}$ then $\|\sum_{i=1}^k c_i g_i\| \in G$, where $\|x\| = [x + \frac{1}{2}]$ denotes the integer nearest to x) [13].

- (vii) The set of shifted primes $\{p-1, p \text{ prime}\}$, and the set $\{p+1, p \text{ prime}\}$ [66].
- (viii) The set of values of a random non-lacunary sequence. (Pick $n \in \mathbb{N}$ independently with probability b_n where $0 \leq b_n \leq 1$ and $\lim_{n \rightarrow \infty} nb_n = \infty$. The resulting set is almost surely a set of recurrence. If $\limsup_{n \rightarrow \infty} nb_n < \infty$ then the resulting set is almost surely a finite union of sets, each of which is the range of some lacunary sequence, hence is not a set of recurrence). Follows from [22].

Showing that the first three sets are good for recurrence is a straightforward modification of the argument used to prove Theorem 1. Examples (iv) – (viii) require more work.

A criterion of Kamae and Mendés-France [53] provides a powerful tool that may be used in many instances to establish that a set R is a set of recurrence. We mention a variation of their result:

Theorem 13 (Kamae and Mendés-France [53]) Suppose that $R = \{a_1 < a_2 < \dots\}$ is a subset of \mathbb{N} such that:

- (i) The sequence $\{a_n \alpha\}_{n \in \mathbb{N}}$ is uniformly distributed in \mathbb{T} for every irrational α .
- (ii) The set $R_m = \{n \in \mathbb{N} : m|a_n\}$ has positive density for every $m \in \mathbb{N}$.

Then R is a set of recurrence.

We sketch a proof for this result. First, recall Herglotz's theorem: if $(a_n)_{n \in \mathbb{Z}}$ is a positive definite sequence, then there is a unique measure σ on the torus \mathbb{T} such that $a_n = \int_{\mathbb{T}} e^{2\pi i n t} d\sigma(t)$. The case of interest to us is $a_n = \int_{\mathbb{T}} f(x) \cdot f(T^n x) d\mu$, where T is measure preserving and $f \in L^\infty(\mu)$; (a_n) is positive definite, and we call $\sigma = \sigma_f$ the spectral measure of f .

Let now (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Putting $f = \mathbf{1}_A$, one has

$$\begin{aligned} \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int f(x) \cdot f(T^{a_n} x) d\mu \\ = \int_{\mathbb{T}} \lim_{N \rightarrow \infty} \left(\frac{1}{N} \sum_{n=1}^N e^{2\pi i a_n t} \right) d\sigma_f(t). \quad (1) \end{aligned}$$

For t irrational the limit inside the integral is zero (by condition (i)), so the last integral can be taken over the ratio-

nal points in \mathbb{T} . Since the spectral measure of a function orthogonal to the subspace

$$\mathcal{H} = \overline{\{f \in L^2(\mu) : \text{there exists } k \in \mathbb{N} \text{ with } T^k f = f\}} \quad (2)$$

has no rational point masses, we can easily deduce that when computing the first limit in (1), we can replace the function f by its orthogonal projection g onto the subspace \mathcal{H} (g is again nonnegative and $g \neq 0$). To complete the argument, we approximate g by a function g' such that $T^m g' = g'$ for some appropriately chosen m , and use condition (ii) to deduce that the limit of the average (1) is positive.

In order to apply Theorem 13, one uses the standard machinery of uniform distribution. Recall Weyl's criterion: a real-valued sequence $(x_n)_{n \in \mathbb{N}}$ is uniformly distributed mod 1 if for every non-zero $k \in \mathbb{Z}$,

$$\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N e^{2\pi i k x_n} = 0.$$

This criterion becomes especially useful when paired with van der Corput's so-called third principal property: if, for every $h \in \mathbb{N}$, $(x_{n+h} - x_n)_{n \in \mathbb{N}}$ is uniformly distributed mod 1, then $(x_n)_{n \in \mathbb{N}}$ is uniformly distributed mod 1. Using the foregoing criteria and some standard (albeit non-trivial) exponential sum estimates, one can verify for example that the sets (iv) and (vii) in Theorem 12 are good for recurrence.

In light of the connection elucidated above between uniform distribution mod 1 and recurrence, it is not surprising that van der Corput's method has been adapted by modern ergodic theorists for use in establishing recurrence properties directly.

Theorem 14 (Bergelson [7]) Let $(x_n)_{n \in \mathbb{N}}$ be a bounded sequence in a Hilbert space. If

$$D\text{-}\lim_{m \rightarrow \infty} \left(\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \langle x_{n+m}, x_n \rangle \right) = 0,$$

then

$$\lim_{N \rightarrow \infty} \left\| \frac{1}{N} \sum_{n=1}^N x_n \right\| = 0.$$

Let us illustrate how one uses this “van der Corput trick” by showing that $S = \{n^2 : n \in \mathbb{N}\}$ is a set of recurrence. We will actually establish the following stronger

fact: If (X, \mathcal{B}, μ, T) is a measure preserving system and $f \in L^\infty(\mu)$ is nonnegative and $f \neq 0$ then

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int f(x) \cdot f(T^{n^2} x) \, d\mu > 0. \quad (3)$$

Then our result follows by setting $f = \mathbf{1}_A$ for some $A \in \mathcal{B}$ with $\mu(A) > 0$.

The main idea is one that occurs frequently in ergodic theory; split the function f into two components, one of which contributes zero to the limit appearing in (3), and the other one being much easier to handle than f . To do this consider the T -invariant subspace of $L^2(X)$ defined by

$$\mathcal{H} = \overline{\{f \in L^2(\mu) : \text{there exists } k \in \mathbb{N} \text{ with } T^k f = f\}}. \quad (4)$$

Write $f = g + h$ where $g \in \mathcal{H}$ and $h \perp \mathcal{H}$, and expand the average in (3) into a sum of four averages involving the functions g and h . Two of these averages vanish because iterates of g are orthogonal to iterates of h . So in order to show that the only contribution comes from the average that involves the function g alone, it suffices to establish that

$$\lim_{N \rightarrow \infty} \left\| \frac{1}{N} \sum_{n=1}^N T^{n^2} h \right\|_{L^2(\mu)} = 0. \quad (5)$$

To show this we will apply the Hilbert space van der Corput lemma. For given $h \in \mathcal{H}$, we let $x_n = T^{n^2} h$ and compute

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \langle x_{n+m}, x_n \rangle \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int T^{n^2+2nm+m^2} h \cdot T^{n^2} h \, d\mu \\ &= \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int T^{2nm} (T^{m^2} h) \cdot h \, d\mu. \end{aligned}$$

Applying the ergodic theorem to the transformation T^{2m} and using the fact that $h \perp \mathcal{H}$, we get that the last limit is 0. This implies (5).

Thus far we have shown that in order to compute the limit in (3) we can assume that $f = g \in \mathcal{H}$ (g is also nonnegative and $g \neq 0$). By the definition of \mathcal{H} , given any $\varepsilon > 0$, there exists a function $f' \in \mathcal{H}$ such that $T^k f' = f'$ for some $k \in \mathbb{N}$ and $\|f - f'\|_{L^2(\mu)} \leq \varepsilon$. Then the limit in (3) is at least $1/k$ times the limit

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int f(x) \cdot f(T^{(kn)^2} x) \, d\mu.$$

Applying the triangle inequality twice we get that this is greater or equal than

$$\begin{aligned} & \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int f'(x) \cdot f'(T^{(kn)^2} x) \, d\mu - c \cdot \varepsilon \\ &= \int (f'(x))^2 \, d\mu - 2\varepsilon \\ &\geq \left(\int f'(x) \, d\mu \right)^2 - c \cdot \varepsilon, \end{aligned}$$

for some constant c that does not depend on ε (we used that $T^k f' = f'$ and the Cauchy-Schwartz inequality). Choosing ε small enough we conclude that the last quantity is positive, completing the proof.

Multiple Recurrence

Simultaneous multiple returns of positive measure sets to themselves were first considered by H. Furstenberg [34], who gave a new proof of Szemerédi's theorem [69] on arithmetic progressions by deriving it from the following theorem:

Theorem 15 (Furstenberg [34]) *Let (X, \mathcal{B}, μ, T) be a measure preserving system and $A \in \mathcal{B}$ with $\mu(A) > 0$. Then for every $k \in \mathbb{N}$, there is some $n \in \mathbb{N}$ such that*

$$\mu(A \cap T^{-n} A \cap \cdots \cap T^{-kn} A) > 0. \quad (6)$$

Furstenberg's proof came by means of a new structure theorem allowing one to decompose an arbitrary measure preserving system into component elements exhibiting one of two extreme types of behavior: *compactness*, characterized by regular, "almost periodic" trajectories, and *weak mixing*, characterized by irregular, "quasi-random" trajectories. On \mathbb{T} , these types of behavior are exemplified by rotations and by the doubling map, respectively. To see the point, imagine trying to predict the initial digit of the dyadic expansion of $T^n x$ given knowledge of the initial digits of $T^i x$, $1 \leq i < n$. We use the case $k = 2$ to illustrate the basic idea.

It suffices to show that if $f \in L^\infty(\mu)$ is nonnegative and $f \neq 0$, one has

$$\liminf_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \int f(x) \cdot f(T^n x) \cdot f(T^{2n} x) \, d\mu > 0. \quad (7)$$

An ergodic decomposition argument enables us to assume that our system is ergodic. As in the earlier case of the squares, we split f into "almost periodic" and "quasi-random" components. Let \mathcal{K} be the closure in L^2 of the subspace spanned by the eigenfunctions of T , i.e. the functions $f \in L^2(\mu)$ that satisfy $f(Tx) = e^{2\pi i \alpha} f(x)$ for some

$\alpha \in \mathbb{R}$. We write $f = g + h$, where $g \in \mathcal{K}$ and $h \perp \mathcal{K}$. It can be shown that $g, h \in L^\infty(\mu)$ and g is again nonnegative with $g \neq 0$. We expand the average in (7) into a sum of eight averages involving the functions g and h . In order to show that the only non-zero contribution to the limit comes from the term involving g alone, it suffices to establish that

$$\lim_{N \rightarrow \infty} \left\| \frac{1}{N} \sum_{n=1}^N T^n g \cdot T^{2n} h \right\|_{L^2(\mu)} = 0, \quad (8)$$

(and similarly with h and g interchanged, and with $g = h$, which is similar). To establish (8), we use the Hilbert space van der Corput lemma on $x_n = T^n g \cdot T^{2n} h$. Some routine computations and a use of the ergodic theorem reduce the task to showing that

$$\text{D-lim}_{m \rightarrow \infty} \left(\int h(x) \cdot h(T^{2m} x) d\mu \right) = 0.$$

But this is well known for $h \perp \mathcal{K}$ (in virtue of the fact that for $h \perp \mathcal{K}$ the spectral measure σ_h is continuous, for example).

We are left with the average (7) when $f = g \in \mathcal{K}$. In this case f can be approximated arbitrarily well by a linear combination of eigenfunctions, which easily implies that given $\varepsilon > 0$ one has $\|T^n f - f\|_{L^2(\mu)} \leq \varepsilon$ for a set of $n \in \mathbb{N}$ with bounded gaps. Using this fact and the triangle inequality, one finds that for a set of $n \in \mathbb{N}$ with bounded gaps,

$$\int f(x) \cdot f(T^n x) \cdot f(T^{2n} x) d\mu \geq \left(\int f d\mu \right)^3 - c \cdot \varepsilon$$

for a constant c that is independent of ε . Choosing ε small enough, we get (7).

The new techniques developed for the proof of Theorem 15 have led to a number of extensions, many of which have to date only ergodic proofs. To expedite discussion of some of these developments, we introduce a definition:

Definition 16 Let $R \subset \mathbb{Z}$ and $k \in \mathbb{N}$. Then R is a set of k -recurrence if for every invertible measure preserving system (X, \mathcal{B}, μ, T) and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in R$ such that

$$\mu(A \cap T^{-n} A \cap \dots \cap T^{-kn} A) > 0.$$

The notions of k -recurrence are distinct for different values of k . An example of a difference set that is a set of 1-recurrence but not a set of 2-recurrence was given in [34]; sets of k -recurrence that are not sets of $(k+1)$ -recurrence for general k were given in [31] ($R_k = \{n \in \mathbb{N} : \{n^{k+1} \sqrt{2}\} \in [1/4, 3/4]\}$ is such).

Aside from difference sets, the sets of (1-)recurrence given in Theorem 12 may well be sets of k -recurrence for every $k \in \mathbb{N}$, though this has not been verified in all cases. Let us summarize the current state of knowledge. The following are sets of k -recurrence for every k : Sets of the form $\bigcup_{n \in \mathbb{N}} \{a_n, 2a_n, \dots, na_n\}$ where $a_n \in \mathbb{N}$ (this follows from a uniform version of Theorem 15 that can be found in [15]). Every IP-set [37]. The set $\{p(n), n \in \mathbb{N}\}$ where p is any nonconstant integer polynomial with $p(0) = 0$ [16], and more generally, when the range of the polynomial contains multiples of an arbitrary integer [33]. The set $\{p(n), n \in S\}$ where p is an integer polynomial with $p(0) = 0$ and S is any IP-set [17]. The set of values of an admissible generalized polynomial [60]. Moreover, the set of shifted primes $\{p-1, p \text{ prime}\}$, and the set $\{p+1, p \text{ prime}\}$ are sets of 2-recurrence [32].

More generally, one would like to know for which sequences of integers $a_1(n), \dots, a_k(n)$ it is the case that for every invertible measure preserving system (X, \mathcal{B}, μ, T) and $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some nonzero $n \in \mathbb{N}$ such that

$$\mu(A \cap T^{-a_1(n)} A \cap \dots \cap T^{-a_k(n)} A) > 0. \quad (9)$$

Unfortunately, a criterion analogous to the one given in Theorem 13 for 1-recurrence is not yet available for k -recurrence when $k > 1$. Nevertheless, there have been some notable positive results, such as the following:

Theorem 17 (Bergelson and Leibman [16]) Let (X, \mathcal{B}, μ, T) be an invertible measure preserving system and $p_1(n), \dots, p_k(n)$ be integer polynomials with zero constant term. Then for every $A \in \mathcal{B}$ with $\mu(A) > 0$, there is some $n \in \mathbb{N}$ such that

$$\mu(A \cap T^{-p_1(n)} A \cap \dots \cap T^{-p_k(n)} A) > 0. \quad (10)$$

Furthermore, it has been shown that the n in (10) can be chosen from any IP set [17], and the polynomials p_1, \dots, p_k can be chosen to belong to the more general class of admissible generalized polynomials [60].

Very recently, a new boost in the area of multiple recurrence was given by a breakthrough of Host and Kra [50]. Building on work of Conze and Lesigne [26,27] and Furstenberg and Weiss [41] (see also the excellent survey [52], exploring close parallels with [45] and the seminal paper of Gowers [43]), they isolated the structured component (or factor) of a measure preserving system that one needs to analyze in order to prove various multiple recurrence and convergence results. This allowed them, in particular, to prove existence of L^2 limits for the so-called “Furstenberg ergodic averages” $\frac{1}{N} \sum_{n=1}^N \prod_{i=0}^k f(T^{i n} x)$,

which had been a major open problem since the original ergodic proof of Szemerédi's theorem. Subsequently Ziegler in [75] gave a new proof of the aforementioned limit theorem and established minimality of the factor in question. It turns out that this minimal component admits of a purely algebraic characterization; it is a *nilsystem*, i. e. a rotation on a homogeneous space of a nilpotent Lie group. This fact, coupled with some recent results about nilsystems (see [58,59] for example), makes the analysis of some otherwise intractable multiple recurrence problems much more manageable. For example, these developments have made it possible to estimate the size of the multiple intersection in (6) for $k = 2, 3$ (the case $k = 1$ is Theorem 4):

Theorem 18 (Bergelson, Host and Kra [14]) *Let (X, \mathcal{B}, μ, T) be an ergodic measure preserving system and $A \in \mathcal{B}$. Then for $k = 2, 3$ and for every $\varepsilon > 0$,*

$$\mu(A \cap T^{-n}A \cap \dots \cap T^{-kn}A) > \mu^{k+1}(A) - \varepsilon \quad (11)$$

for a set of $n \in \mathbb{N}$ with bounded gaps.

Based on work of Ruzsa that appears as an appendix to the paper, it is also shown in [14] that a similar estimate fails for ergodic systems (with any power of $\mu(A)$ on the right hand side) when $k \geq 4$. Moreover, when the system is nonergodic it also fails for $k = 2, 3$, as can be seen with the help of an example in [6]. Again considering the doubling map $Tx = 2x$ and the set $A = [0, 1/2]$, one sees that the positive results for $k \leq 3$ are sharp. When the polynomials $n, 2n, \dots, kn$ are replaced by linearly independent polynomials p_1, p_2, \dots, p_k with zero constant term, similar lower bounds hold for every $k \in \mathbb{N}$ without assuming ergodicity [30]. The case where the polynomials $n, 2n, 3n$ are replaced with general polynomials p_1, p_2, p_3 with zero constant term is treated in [33].

Connections with Combinatorics and Number Theory

The combinatorial ramifications of ergodic-theoretic recurrence were first observed by Furstenberg, who perceived a correspondence between recurrence properties of measure preserving systems and the existence of structures in sets of integers having positive upper density. This gave rise to the field of ergodic Ramsey theory, in which problems in combinatorial number theory are treated using techniques from ergodic theory. The following formulation is from [8].

Theorem 19 *Let A be a subset of the integers. There exists an invertible measure preserving system (X, \mathcal{B}, μ, T) and*

a set $A \in \mathcal{B}$ with $\mu(A) = \bar{d}(A)$ such that

$$\begin{aligned} & \bar{d}(A \cap (A - n_1) \cap \dots \cap (A - n_k)) \\ & \geq \mu(A \cap T^{-n_1}A \cap \dots \cap T^{-n_k}A), \end{aligned} \quad (12)$$

for all $k \in \mathbb{N}$ and $n_1, \dots, n_k \in \mathbb{Z}$.

Proof The space X will be taken to be the sequence space $\{0, 1\}^{\mathbb{Z}}$, \mathcal{B} is the Borel σ -algebra, while T is the shift map defined by $(Tx)(n) = x(n+1)$ for $x \in \{0, 1\}^{\mathbb{Z}}$, and A is the set of sequences x with $x(0) = 1$. So the only thing that depends on A is the measure μ which we now define. For $m \in \mathbb{N}$ set $\Lambda^0 = \mathbb{Z} \setminus \Lambda$ and $\Lambda^1 = \Lambda$. Using a diagonal argument we can find an increasing sequence of integers $(N_m)_{m \in \mathbb{N}}$ such that $\lim_{m \rightarrow \infty} |\Lambda \cap [1, N_m]|/N_m = \bar{d}(\Lambda)$ and such that

$$\lim_{m \rightarrow \infty} \frac{|\{(\Lambda^{i_1} - n_1) \cap (\Lambda^{i_2} - n_2) \cap \dots \cap (\Lambda^{i_r} - n_r) \cap [1, N_m]\}|}{N_m} \quad (13)$$

exists for every $n_1, \dots, n_r \in \mathbb{Z}$, and $i_1, \dots, i_r \in \{0, 1\}$. For $n_1, n_2, \dots, n_r \in \mathbb{Z}$, and $i_1, i_2, \dots, i_r \in \{0, 1\}$, we define the measure μ of the cylinder set $\{x(n_1) = i_1, x(n_2) = i_2, \dots, x(n_r) = i_r\}$ to be the limit (13). Thus defined, μ extends to a premeasure on the algebra of sets generated by cylinder sets and hence by Carathéodory's extension theorem [24] to a probability measure on \mathcal{B} . It is easy to check that $\mu(A) = \bar{d}(\Lambda)$, the shift transformation T preserves the measure μ and (12) holds. \square

Using this principle for $k = 1$, one may check that any set of recurrence is *intersective*, that is intersects $E - E$ for every set E of positive density. Using it for $n_1 = n, n_2 = 2n, \dots, n_k = kn$, together with Theorem 15, one gets an ergodic proof of Szemerédi's theorem [69], stating that every subset of the integers with positive upper density contains arbitrarily long arithmetic progressions (conversely, one can easily deduce Theorem 15 from Szemerédi's theorem, and that intersective sets are sets of recurrence). Making the choice $n_1 = n^2$ and using part (iv) of Theorem 13, we get an ergodic proof of the surprising result of Sárközy [66] stating that every subset of the integers with positive upper density contains two elements whose difference is a perfect square. More generally, using Theorem 19, one can translate all of the recurrence results of the previous two sections to results in combinatorics. (This is not straightforward for Theorem 18 because of the ergodicity assumption made there. We refer the reader to [14] for the combinatorial consequence of this result). We mention explicitly only the combinatorial consequence of Theorem 17:

Theorem 20 (Bergelson and Leibman [16]) *Let $\Lambda \subset \mathbb{Z}$ with $\bar{d}(\Lambda) > 0$, and p_1, \dots, p_k be integer polynomials with zero constant term. Then Λ contains infinitely many configurations of the form $\{x, x + p_1(n), \dots, x + p_k(n)\}$.*

The ergodic proof is the only one known for this result, even for patterns of the form $\{x, x + n^2, x + 2n^2\}$ or $\{x, x + n, x + n^2\}$.

Ergodic-theoretic contributions to the field of geometric Ramsey theory were made by Furstenberg, Katznelson, and Weiss [40], who showed that if E is a positive upper density subset of \mathbb{R}^2 then: (i) E contains points with any large enough distance (see also [21] and [29]), (ii) Every δ -neighborhood of E contains three points forming a triangle congruent to any given large enough dilation of a given triangle (in [21] it is shown that if the three points lie on a straight line one cannot always find three points with this property in E itself). Recently, a generalization of property (ii) to arbitrary finite configurations of \mathbb{R}^m was obtained by Ziegler [76].

It is also worth mentioning some recent exciting connections of multiple recurrence with some structural properties of the set of prime numbers. The first one is in the work of Green and Tao [45], where the existence of arbitrarily long arithmetic progressions of primes was demonstrated, the authors, in addition to using Szemerédi's theorem outright, use several ideas from its ergodic-theoretic proofs, as appearing in [34] and [39]. The second one is in the recent work of Tao and Ziegler [70], where a quantitative version of Theorem 17 was used to prove that the primes contain arbitrarily long polynomial progressions. Furthermore, several recent results in ergodic theory, related to the structure of the minimal characteristic factors of certain multiple ergodic averages, play an important role in the ongoing attempts of Green and Tao to get asymptotic formulas for the number of k -term arithmetic progressions of primes up to x (see for example [46] and [47]). This project has been completed for $k = 3$, thus verifying an interesting special case of the Hardy–Littlewood k -tuple conjecture predicting the asymptotic growth rate of $N_{a_1, \dots, a_k}(x) =$ the number of configurations of primes having the form $\{p, p + a_1, \dots, p + a_k\}$ with $p \leq x$.

Finally, we remark that in this article we have restricted attention to multiple recurrence and Furstenberg correspondence for \mathbb{Z} actions, while in fact there is a wealth of literature on extensions of these results to general commutative, amenable and even non-amenable groups. For an excellent exposition of these and other recent developments the reader is referred to the surveys [9] and [11]. Here, we give just one notable combinatorial corollary to

some work of this kind, a density version of the classical Hales–Jewett coloring theorem [48].

Theorem 21 (Furstenberg and Katznelson [38]) *Let $W_n(A)$ denote the set of words of length n with letters in the alphabet $A = \{a_1, \dots, a_k\}$. For every $\varepsilon > 0$ there exists $N_0 = N_0(\varepsilon, k)$ such that if $n \geq N_0$ then any subset S of $W_n(A)$ with $|S| \geq \varepsilon k^n$ contains a combinatorial line, i.e., a set consisting of k n -letter words, having fixed letters in l positions, for some $0 \leq l < n$, the remaining $n - l$ positions being occupied by a variable letter x , for $x = a_1, \dots, a_k$. (For example, in $W_4(A)$ the sets $\{(a_1, x, a_2, x) : x \in A\}$ and $\{(x, x, x, x) : x \in A\}$ are combinatorial lines).*

At first glance, the uninitiated reader may not appreciate the importance of this “master” density result, so it is instructive to derive at least one of its immediate consequences. Let $A = \{0, 1, \dots, k - 1\}$ and interpret $W_n(A)$ as integers in base k having at most n digits. Then a combinatorial line in $W_n(A)$ is an arithmetic progression of length k – for example, the line $\{(a_1, x, a_2, x) : x \in A\}$ corresponds to the progression $\{m, m + n, m + 2n, m + 3n\}$, where $m = a_1 + a_2 d^2$ and $n = d + d^3$. This allows one to deduce Szemerédi's theorem. Similarly, one can deduce from Theorem 21 multidimensional and IP extensions of Szemerédi's theorem [36, 37], and some related results about vector spaces over finite fields [37]. Again, the only known proof for the density version of the Hales–Jewett theorem relies heavily on ergodic theory.

Future Directions

In this section we formulate a few open problems relating to the material in the previous three sections. It should be noted that this selection reflects the authors' interests, and does not strive for completeness.

We start with an intriguing question of Katznelson [54] about sets of topological recurrence. A set $S \subset \mathbb{N}$ is a set of Bohr recurrence if for every $\alpha_1, \dots, \alpha_k \in \mathbb{R}$ and $\varepsilon > 0$ there exists $s \in S$ such that $\{s\alpha_i\} \in [0, \varepsilon] \cup [1 - \varepsilon, 1)$ for $i = 1, \dots, k$.

Problem 1 *Is every set of Bohr recurrence a set of topological recurrence?*

Background for this problem and evidence for a positive answer can be found in [54, 72]. As we mentioned in Sect. “Subsequence Recurrence”, there exists a set of topological recurrence (and hence Bohr recurrence) that is not a set of recurrence.

Problem 2 *Is the set $S = \{l!2^m 3^n : l, m, n \in \mathbb{N}\}$ a set of recurrence? Is it a set of k -recurrence for every $k \in \mathbb{N}$?*

It can be shown that S is a set of Bohr recurrence. Theorem 13 cannot be applied since the uniform distribution condition fails for some irrational numbers α . A relevant question was asked by Bergelson in [9]: “Is the set $S = \{2^m 3^n : m, n \in \mathbb{N}\}$ good for single recurrence for weakly mixing systems?”

As we mentioned in Sect. “Multiple Recurrence”, the set of primes shifted by 1 (or -1) is a set of 2-recurrence [32].

Problem 3 Show that the sets $\mathbb{P} - 1$ and $\mathbb{P} + 1$, where \mathbb{P} is the set of primes, are sets of k -recurrence for every $k \in \mathbb{N}$.

As remarked in [32], a positive answer to this question will follow if some uniformity conjectures of Green and Tao [47] are verified.

We mentioned in Sect. “Subsequence Recurrence” that random non-lacunary sequences (see definition there) are almost surely sets of recurrence.

Problem 4 Show that random non-lacunary sequences are almost surely sets of k -recurrence for every $k \in \mathbb{N}$.

The answer is not known even for $k = 2$, though, in unpublished work, Wierdl and Lesigne have shown that the answer is positive for random sequences with at most quadratic growth. We refer the reader to the survey [65] for a nice exposition of the argument used by Bourgain [22] to handle the case $k = 1$.

It was shown in [31] that if S is a set of 2-recurrence then the set of its squares is a set of recurrence for circle rotations. The same method shows that it is actually a set of Bohr recurrence.

Problem 5 If $S \subset \mathbb{Z}$ is a set of 2-recurrence, is it true that $S^2 = \{s^2 : s \in S\}$ is a set of recurrence?

A similar question was asked in [23]: “If S is a set of k -recurrence for every k , is the same true of S^2 ?”

One would like to find a criterion that would allow one to deduce that a sequence is good for double (or higher order) recurrence from some uniform distribution properties of this sequence.

Problem 6 Find necessary conditions for double recurrence similar to the one given in Theorem 13.

It is now well understood that such a criterion should involve uniform distribution properties of some generalized polynomials or 2-step nilsequences.

We mentioned in Sect. “Connections with Combinatorics and Number Theory” that every positive density subset of \mathbb{R}^2 contains points with any large enough distance. Bourgain [21] constructed a positive density subset E of \mathbb{R}^2 , a triangle T , and numbers $t_n \rightarrow \infty$, such that

E does not contain congruent copies of all t_n -dilations of T . But the triangle T used in this construction is degenerate, which leaves the following question open:

Problem 7 Is it true that every positive density subset of \mathbb{R}^2 contains a triangle congruent to any large enough dilation of a given non-degenerate triangle?

For further discussion on this question the reader can consult the survey [44].

The following question of Aaronson and Nakada [1] is related to a classical question of Erdős concerning whether every $K \subset \mathbb{N}$ such that $\sum_{n \in K} 1/n = \infty$ contains arbitrarily long arithmetic progressions:

Problem 8 Suppose that (X, \mathcal{B}, μ, T) is a $\{1/n\}$ -conservative ergodic measure preserving system. Is it true that for every $A \in \mathcal{B}$ with $\mu(A) > 0$ and $k \in \mathbb{N}$ we have $\mu(A \cap T^{-n}A \cap \dots \cap T^{-kn}A) > 0$ for some $n \in \mathbb{N}$?

The answer is positive for the class of Markov shifts, and it is remarked in [1] that if the Erdős conjecture is true then the answer will be positive in general. The converse is not known to be true. For a related result showing that multiple recurrence is preserved by extensions of infinite measure preserving systems see [61].

Our next problem is motivated by the question whether Theorem 21 has a polynomial version (for a precise formulation of the general conjecture see [9]). Not even this most special consequence of it is known to hold.

Problem 9 Let $\varepsilon > 0$. Does there exist $N = N(\varepsilon)$ having the property that every family \mathcal{P} of subsets of $\{1, \dots, N\}^2$ satisfying $|\mathcal{P}| \geq \varepsilon 2^{N^2}$ contains a configuration $\{A, A \cup (\gamma \times \gamma)\}$, where $A \subset \{1, \dots, N\}^2$ and $\gamma \subset \{1, \dots, N\}$ with $A \cap (\gamma \times \gamma) = \emptyset$?

A measure preserving action of a general countably infinite group G is a function $g \rightarrow T_g$ from G into the space of measure preserving transformations of a probability space X such that $T_{gh} = T_g T_h$. It is easy to show that a version of Khintchine’s recurrence theorem holds for such actions: if $\mu(A) > 0$ and $\varepsilon > 0$ then $\{g : \mu(A \cap T_g A) > (\mu(A))^2 - \varepsilon\}$ is syndetic. However it is unknown whether the following ergodic version of Roth’s theorem holds.

Problem 10 Let (T_g) and (S_g) be measure preserving G -actions of a probability space X that commute in the sense that $T_g S_h = S_h T_g$ for all $g, h \in G$. Is it true that for all positive measure sets A , the set of g such that $\mu(A \cap T_g A \cap S_g A) > 0$ is syndetic?

We remark that for general (possibly amenable) groups G not containing arbitrarily large finite subgroups nor elements of infinite order, it is not known whether one can

find a *single* such $g \neq e$. On the other hand, the answer is known to be positive for general G in case $(T_g^{-1}S_g)$ is a G -action [18]; even under such strictures, however, it is unknown whether a triple recurrence theorem holds.

Bibliography

- Aaronson J (1981) The asymptotic distribution behavior of transformations preserving infinite measures. *J Analyse Math* 39:203–234
- Aaronson J (1997) An introduction to infinite ergodic theory. *Mathematical Surveys and Monographs* 50. American Mathematical Society, Providence
- Aaronson J, Nakada H (2000) Multiple recurrence of Markov shifts and other infinite measure preserving transformations. *Israel J Math* 117:285–310
- Barreira L (2001) Hausdorff dimension of measures via Poincaré recurrence. *Comm Math Phys* 219:443–463
- Barreira L (2005) Poincaré recurrence: old and new. XIVth International Congress on Mathematical Physics, World Sci Publ Hackensack, NJ, pp 415–422
- Behrend F (1946) On sets of integers which contain no three in arithmetic progression. *Proc Nat Acad Sci* 23:331–332
- Bergelson V (1987) Weakly mixing PET. *Ergod Theory Dynam Syst* 7:337–349
- Bergelson V (1987) Ergodic Ramsey Theory. In: Simpson S (ed) *Logic and Combinatorics*. Contemporary Mathematics 65. American Math Soc, Providence, pp 63–87
- Bergelson V (1996) Ergodic Ramsey Theory – an update. In: Pollicot M, Schmidt K (eds) *Ergodic theory of \mathbb{Z}^d -actions*. Lecture Note Series 228. London Math Soc, London, pp 1–61
- Bergelson V (2000) The multifarious Poincaré recurrence theorem. In: Foreman M, Kechris A, Louveau A, Weiss B (eds) *Descriptive set theory and dynamical systems*. Lecture Note Series 277. London Math Soc, London, pp 31–57
- Bergelson V (2005) Combinatorial and diophantine applications of ergodic theory. In: Hasselblatt B, Katok A (eds) *Handbook of dynamical systems*, vol 1B. Elsevier, pp 745–841
- Bergelson V, Furstenberg H, McCutcheon R (1996) IP-sets and polynomial recurrence. *Ergod Theory Dynam Syst* 16:963–974
- Bergelson V, Håland-Knutson I, McCutcheon R (2006) IP Systems, generalized polynomials and recurrence. *Ergod Theory Dynam Syst* 26:999–1019
- Bergelson V, Host B, Kra B (2005) Multiple recurrence and nilsequences. *Inventiones Math* 160(2):261–303
- Bergelson V, Host B, McCutcheon R, Parreau F (2000) Aspects of uniformity in recurrence. *Colloq Math* 84/85(2):549–576
- Bergelson V, Leibman A (1996) Polynomial extensions of van der Waerden's and Szemerédi's theorems. *J Amer Math Soc* 9:725–753
- Bergelson V, McCutcheon R (2000) An ergodic IP polynomial Szemerédi theorem. *Mem Amer Math Soc* 146:viii–106
- Bergelson V, McCutcheon R (2007) Central sets and a noncommutative Roth theorem. *Amer J Math* 129:1251–1275
- Birkhoff G (1931) A proof of the ergodic theorem. *Proc Nat Acad Sci* 17:656–660
- Boshernitzan M (1993) Quantitative recurrence results. *Invent Math* 113:617–631
- Bourgain J (1986) A Szemerédi type theorem for sets of positive density in \mathbb{R}^k . *Israel J Math* 54(3):307–316
- Bourgain J (1988) On the maximal ergodic theorem for certain subsets of the positive integers. *Israel J Math* 61:39–72
- Brown T, Graham R, Landman B (1999) On the set of common differences in van der Waerden's theorem on arithmetic progressions. *Canad Math Bull* 42:25–36
- Carathéodory C (1968) *Vorlesungen über reelle Funktionen*, 3rd edn. Chelsea Publishing Co, New York
- Chazottes J, Ugalde E (2005) Entropy estimation and fluctuations of hitting and recurrence times for Gibbsian sources. *Discrete Contin Dyn Syst Ser B* 5(3):565–586
- Conze J, Lesigne E (1984) Théorèmes ergodiques pour des mesures diagonales. *Bull Soc Math France* 112(2):143–175
- Conze J, Lesigne E (1988) Sur un théorème ergodique pour des mesures diagonales. *Probabilités, Publ Inst Rech Math Rennes* 1987-1, Univ Rennes I, Rennes, pp 1–31
- Evans D, Searles D (2002) The fluctuation theorem. *Adv Phys* 51:1529–1585
- Falconer K, Marstrand J (1986) Plane sets with positive density at infinity contain all large distances. *Bull Lond Math Soc* 18:471–474
- Frantzikinakis N, Kra B (2006) Ergodic averages for independent polynomials and applications. *J Lond Math Soc* 74(1):131–142
- Frantzikinakis N, Lesigne E, Wierdl M (2006) Sets of k -recurrence but not $(k+1)$ -recurrence. *Annales de l'Institut Fourier* 56(4):839–849
- Frantzikinakis N, Host B, Kra B (2007) Multiple recurrence and convergence for sets related to the primes. *J Reine Angew Math* 611:131–144. Available at <http://arxiv.org/abs/math/0607637>
- Frantzikinakis N (2008) Multiple ergodic averages for three polynomials and applications. *Trans Am Math Soc* 360(10):5435–5475. Available at <http://arxiv.org/abs/math/0606567>
- Furstenberg H (1977) Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions. *J Anal Math* 71:204–256
- Furstenberg H (1981) *Recurrence in ergodic theory and combinatorial number theory*. Princeton University Press, Princeton
- Furstenberg H, Katznelson Y (1979) An ergodic Szemerédi theorem for commuting transformations. *J Analyse Math* 34:275–291
- Furstenberg H, Katznelson Y (1985) An ergodic Szemerédi theorem for IP-systems and combinatorial theory. *J Analyse Math* 45:117–168
- Furstenberg H, Katznelson Y (1991) A density version of the Hales–Jewett theorem. *J Analyse Math* 57:64–119
- Furstenberg H, Katznelson Y, Ornstein D (1982) The ergodic theoretical proof of Szemerédi's theorem. *Bull Amer Math Soc (NS)* 7(3):527–552
- Furstenberg H, Katznelson Y, Weiss B (1990) Ergodic theory and configurations in sets of positive density. *Mathematics of Ramsey theory*. Algorithms Combin 5. Springer, Berlin, pp 184–198
- Furstenberg H, Weiss B (1996) A mean ergodic theorem for $(1/N) \sum_{n=1}^N f(T^n x)g(T^{n^2} x)$. *Convergence in ergodic theory and probability* (Columbus, OH, 1993), Ohio State Univ Math Res Inst Publ 5, de Gruyter, Berlin, pp 193–227
- Galatolo S, Kim DH, Park KK (2006) The recurrence time for ergodic systems with infinite invariant measures. *Nonlinearity* 19:2567–2580

43. Gowers W (2001) A new proof of Szemerédi's theorem. *Geom Funct Anal* 11:465–588
44. Graham RL (1994) Recent trends in Euclidean Ramsey theory. *Trends in discrete mathematics*. *Discret Math* 136(1–3):119–127
45. Green B, Tao T (2008) The primes contain arbitrarily long arithmetic progressions. *Ann Math* 167:481–547. Available at <http://arxiv.org/abs/math/0404188>
46. Green B, Tao T (to appear) Quadratic uniformity of the Möbius function. *Annales de l'Institut Fourier*. Available at <http://arxiv.org/abs/math.NT/0606087>
47. Green B, Tao T () Linear equations in primes. *Ann Math* (to appear). Available at <http://arxiv.org/abs/math.NT/0606088>
48. Hales A, Jewett R (1963) Regularity and positional games. *Trans Amer Math Soc* 106:222–229
49. Hasley T, Jensen M (2004) Hurricanes and butterflies. *Nature* 428:127–128
50. Host B, Kra B (2005) Nonconventional ergodic averages and nilmanifolds. *Ann Math* 161:397–488
51. Kac M (1947) On the notion of recurrence in discrete stochastic processes. *Bull Amer Math Soc* 53:1002–10010
52. Kra B (2006) The Green-Tao theorem on arithmetic progressions in the primes: an ergodic point of view. *Bull Amer Math Soc (NS)* 43:3–23
53. Kamae T, Mendés-France M (1978) Van der Corput's difference theorem. *Israel J Math* 31:335–342
54. Katznelson Y (2001) Chromatic numbers of Cayley graphs on \mathbb{Z} and recurrence. *Paul Erdős and his mathematics* (Budapest, 1999). *Combinatorica* 21(2):211–219
55. Khintchine A (1934) Eine Verschärfung des Poincaréschen "Wiederkehrrsatzes". *Comp Math* 1:177–179
56. Kriz I (1987) Large independent sets in shift invariant graphs. Solution of Bergelson's problem. *Graphs Combinatorics* 3:145–158
57. Leibman A (2002) Lower bounds for ergodic averages. *Ergod Theory Dynam. Syst* 22:863–872
58. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial sequences of rotations of a nilmanifold. *Ergod Theory Dynam Syst* 25:201–213
59. Leibman A (2005) Pointwise convergence of ergodic averages for polynomial actions of \mathbb{Z}^d by translations on a nilmanifold. *Ergod Theory Dynam Syst* 25:215–225
60. McCutcheon R (2005) FVIP systems and multiple recurrence. *Israel J Math* 146:157–188
61. Meyerovitch T (2007) Extensions and multiple recurrence of infinite measure preserving systems. Preprint. Available at <http://arxiv.org/abs/math/0703914>
62. Ornstein D, Weiss B (1993) Entropy and data compression schemes. *IEEE Trans Inform Theory* 39:78–83
63. Petersen K (1989) *Ergodic theory*. Cambridge Studies in Advanced Mathematics 2. Cambridge University Press, Cambridge
64. Poincaré H (1890) Sur le problème des trois corps et les équations de la dynamique. *Acta Math* 13:1–270
65. Rosenblatt J, Wierdl M (1995) Pointwise ergodic theorems via harmonic analysis. *Ergodic theory and its connections with harmonic analysis* (Alexandria, 1993). London Math Soc, Lecture Note Series 205, Cambridge University Press, Cambridge, pp 3–151
66. Sárközy A (1978) On difference sets of integers III. *Acta Math Acad Sci Hungar* 31:125–149
67. Shkredov I (2002) Recurrence in the mean. *Mat Zametki* 72(4):625–632; translation in *Math Notes* 72(3–4):576–582
68. Sklar L (2004) Philosophy of Statistical Mechanics. In: Zalta EN (ed) *The Stanford Encyclopedia of Philosophy* (Summer 2004 edn), <http://plato.stanford.edu/archives/sum2004/entries/statphys-statmech/>
69. Szemerédi E (1975) On sets of integers containing no k elements in arithmetic progression. *Acta Arith* 27:299–345
70. Tao T, Ziegler T () The primes contain arbitrarily long polynomial progressions. *Acta Math* (to appear). Available at <http://www.arxiv.org/abs/math.DS/0610050>
71. Walters P (1982) *An introduction to ergodic theory*. Graduate Texts in Mathematics, vol 79. Springer, Berlin
72. Weiss B (2000) *Single orbit dynamics*. CBMS Regional Conference Series in Mathematics 95. American Mathematical Society, Providence
73. Wyner A, Ziv J (1989) Some asymptotic properties of the entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans Inform Theory* 35:1250–1258
74. Zermelo E (1896) Über einen Satz der Dynamik und die mechanische Wärmetheorie. *Annalen der Physik* 57:485–94; English translation, On a theorem of dynamics and the mechanical theory of heat. In: Brush SG (ed) *Kinetic Theory*. Oxford, 1966, II, pp 208–17
75. Ziegler T (2007) Universal characteristic factors and Furstenberg averages. *J Amer Math Soc* 20:53–97
76. Ziegler T (2006) Nilfactors of \mathbb{R}^m -actions and configurations in sets of positive upper density in \mathbb{R}^m . *J Anal Math* 99:249–266

Ergodic Theory: Rigidity

VIOREL NIȚICĂ^{1,2}

¹ West Chester University, West Chester, USA

² Institute of Mathematics, Bucharest, Romania

Article Outline

Glossary

Definition of the Subject

Introduction

Basic Definitions and Examples

Differentiable Rigidity

Local Rigidity

Global Rigidity

Measure Rigidity

Future Directions

Acknowledgment

Bibliography

Glossary

Differentiable rigidity Differentiable rigidity refers to finding invariants to the differentiable conjugacy of dynamical systems, and, more general, group actions.

Local rigidity Local rigidity refers to the study of perturbations of homomorphisms from discrete or continuous groups into diffeomorphism groups.

Global rigidity Global rigidity refers to the classification of all group actions on manifolds satisfying certain conditions.

Measure rigidity Measure rigidity refers to the study of invariant measures for actions of abelian groups and semigroups.

Lattice A lattice in a Lie group is a discrete subgroup of finite covolume.

Conjugacy Two elements g_1, g_2 in a group G are said to be conjugated if there exists an element $h \in G$ such that $g_1 = h^{-1}g_2h$. The element h is called conjugacy.

C^k Conjugacy Two diffeomorphisms ϕ_1, ϕ_2 acting on the same manifold M are said to be C^k -conjugated if there exists a C^k diffeomorphism h of M such that $\phi_1 = h^{-1} \circ \phi_2 \circ h$. The diffeomorphism h is called C^k conjugacy.

Definition of the Subject

As one can see from this volume, chaotic behavior of complex dynamical systems is prevalent in nature and in large classes of transformations. Rigidity theory can be viewed as the counterpart to the generic theory of dynamical systems which often investigates chaotic dynamics for a typical transformation belonging to a large class. In rigidity one is interested in finding obstructions to chaotic, or generic, behavior. This often leads to rather unexpected classification results. As such, rigidity in dynamics and ergodic theory is difficult to define precisely and the best approach to this subject is to study various results and themes that developed so far. A classification is offered below in local, global, differentiable and measurable rigidity. One should note that all branches are strongly intertwined and, at this stage of the development of the subject, it is difficult to separate them.

Rigidity is a well developed and prominent topic in modern mathematics. Historically, rigidity has two main origins, one coming from the study of lattices in semi-simple Lie groups, and one coming from the theory of hyperbolic dynamical systems. From the start, ergodic theory was an important tool used to prove rigidity results, and a strong interdependence developed between these fields. Many times a result in rigidity is obtained by combining techniques from the theory of lattices in Lie groups with techniques from hyperbolic dynamical systems and ergodic theory. Among other mathematical disciplines using results and contributing to this field one can mention representation theory, smooth, continuous and mea-

surable dynamics, harmonic and spectral analysis, partial differential equations, differential geometry, and number theory. Additional details about the appearance of rigidity in ergodic theory as well as definitions for some terminology used in the sequel can be found in the articles ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#) by Kleinbock, ► [Ergodic Theory: Recurrence](#) by Frantzikinakis, McCutcheon, and ► [Ergodic Theory: Interactions with Combinatorics and Number Theory](#) by Ward. The theory of hyperbolic dynamics is presented in the article ► [Hyperbolic Dynamical Systems](#) by Viana and in the article ► [Smooth Ergodic Theory](#) by Wilkinson.

Introduction

The first results about classification of lattices in semi-simple Lie groups were local, aimed at trying to understand the space of small perturbations of a given linear representation. A major contributor was Weil [110,111,112], who proved local rigidity of linear representations for large classes of groups, in particular lattices. Another breakthrough was the contribution of Kazhdan [66], who introduced property (T), allowing to show that large classes of lattices are finitely generated. Rigidity theory matured due to the remarkable global rigidity results obtained by Mostow [90] and Margulis [82], leading to a complete classification of lattices in large classes of semi-simple Lie groups.

Briefly, a hyperbolic (Anosov) dynamical system is one that exhibits strong expansion and contraction along complementary directions. An early contribution introducing this class of objects is the paper of Smale [106], in which basic examples and techniques are introduced. A breakthrough came with the results of Anosov [1], who proved structural stability of the Anosov systems and ergodicity of the geodesic flow on a manifold of negative curvature. Motivated by questions arising in mathematical physics, chaos theory and other areas, hyperbolic dynamics emerged as one of the major fields of contemporary mathematics. From the beginning, a major unsolved problem in the field was the classification of Anosov diffeomorphisms and flows.

In the 80s a change in philosophy occurred, partially motivated by a program introduced by Zimmer [114]. The goal of the program was to classify the smooth actions of higher rank semi-simple Lie groups and of their (irreducible) lattices on compact manifolds. It was expected that any such lattice action that preserves a smooth volume form and is ergodic can be reduced to one of the following standard models: isometric actions, linear actions on infranilmanifolds, and left translations on compact ho-

ogeneous spaces. This original conjecture was disproved by Katok, Lewis (see [56]): by blowing up a linear nilmanifold-action at some fixed points they exhibit real-analytic, volume preserving, ergodic lattice actions on manifolds with complicated topology.

Nevertheless, imposing extra assumptions on a higher rank action, for example the existence of some hyperbolicity, allows local and global classification results. The concept of *Anosov action*, that is, an action that contains at least one Anosov diffeomorphism, was introduced for general groups by Pugh, Shub [99]. The significant differences between the classical \mathbb{Z} and \mathbb{R} cases and those of higher rank lattices, or at a more basic level, of higher rank abelian groups, went unnoticed for a while. The surge of activity in the 80s allowed these differences to surface: for lattices in the work of Hurder, Katok, Lewis, Zimmer (see [43,55,56,63]); and for higher rank abelian groups in the work of Katok, Lewis [55] and Katok, Spatzier [59]. As observed in these papers, local and global rigidity are typical for such Anosov actions. This generated additional research which is summarized in Sects. “[Local Rigidity](#)” and “[Global Rigidity](#)”.

Differentiable rigidity is covered in Sect. “[Differentiable Rigidity](#)”. An interesting problem is to find moduli for the C^k conjugacy, $k \geq 1$, of Anosov diffeomorphisms and flows. This was tackled so far only for low dimensional cases ($n = 2$ for diffeomorphisms and $n = 3$ for flows). Another direction that can be included here refers to finding obstructions for higher transverse regularity of the stable/unstable foliation of a hyperbolic system. A spin-off of the research done so far, which is of high interest by itself, and has applications to local and global rigidity, consists of results lifting the regularity of solutions of cohomological equations over hyperbolic systems. In turn, these results motivated a more careful study of analytic questions about lifting the regularity of real valued continuous functions that enjoy higher regularity along webs of foliations. We also include in this section rigidity results for cocycles over higher rank abelian actions. These are crucial to the proof of local rigidity of higher rank abelian group actions. A more detailed presentation of the material relevant to differentiable rigidity can be found in the forthcoming monograph [58].

Measure rigidity refers to the study of invariant measures under actions of abelian groups and semigroups. If the actions are hyperbolic, higher-rank, and satisfy natural algebraic and irreducibility assumptions, one expects the invariant measures to be rare. This direction was started by a question of Furstenberg, asking if any nonatomic probability measure on the circle, invariant and ergodic under multiplications by 2 and 3, is the Lebesgue measure. An

early contribution is that of Rudolph [103], who answered positively if the action has an element of strictly positive entropy. Katok, Spatzier [61] extended the question to more general higher rank abelian actions, such as actions by linear automorphisms of tori and Weyl chamber flows. A related direction is the study of the invariant sets and measures under the action of horocycle flows, where important progress was made by Ratner [100,101] and earlier by Margulis [16,81,83]. An application of these results present in the last papers is the proof of the long standing Oppenheim’s conjecture, about the density of the values of the quadratic forms at integer points. Recent developments due to Einsiedler, Katok, Lindenstrauss [20] give a partial answer to another outstanding conjecture in number theory, Littlewood’s conjecture, and emphasize measure rigidity as one of a more promising directions in rigidity. More details are shown in Sect. “[Measure Rigidity](#)”.

Four other recent surveys of rigidity theory, each one with a fair amount of overlap but also complementary in part to the present one, that discuss various aspects of the field and its significance are written by Fisher [23], Lindenstrauss [68], Nițică, Török [95], and Spatzier [107]. Among these, [23] concentrates mostly on local and global rigidity, [95] on differentiable rigidity, [68] on measure rigidity, and [107] gives a general overview.

Here is a word of caution for the reader. Many times, instead of the most general results, we present an example that contains the essence of what is available. Also, several important facts that should have been included, are left out. This is because stating complete results would require more space than allocated to this material. The limited knowledge of the author also plays a role here. He apologizes for any obvious omissions and hopes that the bibliography will help fill the gaps.

Basic Definitions and Examples

A detailed introduction to the theory of Anosov systems and hyperbolic dynamics is given in the monograph [51]. The proofs of the basic results for diffeomorphisms stated below can be found there. The proofs for flows are similar. Surveys about hyperbolic dynamics in this volume are the article ▶ [Hyperbolic Dynamical Systems](#) by Viana and the article ▶ [Smooth Ergodic Theory](#) by Wilkinson.

Consider a compact differentiable manifold M and $f: M \rightarrow M$ a C^1 diffeomorphism. Let TM be the tangent bundle of M , and $Df: TM \rightarrow TM$ be the derivative of f . The map f is said to be an *Anosov diffeomorphism* if there is a smooth Riemannian metric $\|\cdot\|$ on M , which induces a metric d_M called *adapted*, a number $\lambda \in (0, 1)$, and

a continuous Df invariant splitting $TM = E^s \oplus E^u$ such that

$$\|Df v\| \leq \lambda \|v\|, v \in E^s, \quad \|Df^{-1} v\| \leq \lambda \|v\|, v \in E^u.$$

For each $x \in M$ there is a pair of embedded C^1 discs $W_{\text{loc}}^s(x)$, $W_{\text{loc}}^u(x)$, called the local stable manifold and the local unstable manifold at x , respectively, such that:

1. $T_x W_{\text{loc}}^s(x) = E^s(x)$, $T_x W_{\text{loc}}^u(x) = E^u(x)$;
2. $f(W_{\text{loc}}^s(x)) \subset W_{\text{loc}}^s(fx)$, $f^{-1}(W_{\text{loc}}^u(x)) \subset W_{\text{loc}}^u(f^{-1}x)$;
3. For any $\mu \in (\lambda, 1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned} d_M(f^n x, f^n y) &\leq C \mu^n d_M(x, y), \quad \text{for } y \in W_{\text{loc}}^s(x), \\ d_M(f^{-n} x, f^{-n} y) &\leq C \mu^n d_M(x, y), \quad \text{for } y \in W_{\text{loc}}^u(x). \end{aligned}$$

The local stable (unstable) manifolds can be extended to global stable (unstable) manifolds $W^s(x)$ and $W^u(x)$ which are well defined and smoothly injectively immersed. These global manifolds are the leaves of global foliations W^s and W^u of M . In general, these foliations are only continuous, but their leaves are differentiable.

Let $\phi: \mathbb{R} \times M \rightarrow M$ be a C^1 flow. The flow ϕ is said to be an *Anosov flow* if there is a Riemannian metric $\|\cdot\|$ on M , a constant $0 < \lambda < 1$, and a continuous Df invariant splitting $TM = E^s \oplus E^0 \oplus E^u$ such that for all $x \in M$ and $t > 0$:

1. $\frac{d}{dt}|_{t=0} \phi^t \in E_x^c \setminus \{0\}$, $\dim E_x^c = 1$,
2. $\|D\phi^t v\| \leq \lambda^t \|v\|$, $v \in E^s$,
3. $\|D\phi^{-t} v\| \leq \lambda^t \|v\|$, $v \in E^u$.

For each $x \in M$ there is a pair of embedded C^1 discs $W_{\text{loc}}^s(x)$, $W_{\text{loc}}^u(x)$, called the local (strong) stable manifold and the local (strong) unstable manifold at x , respectively, such that:

1. $T_x W_{\text{loc}}^s(x) = E^s(x)$, $T_x W_{\text{loc}}^u(x) = E^u(x)$;
2. $\phi^t(W_{\text{loc}}^s(x)) \subset W_{\text{loc}}^s(\phi^t x)$,
 $\phi^{-t}(W_{\text{loc}}^u(x)) \subset W_{\text{loc}}^u(\phi^{-t} x)$ for $t > 0$;
3. For any $\mu \in (\lambda, 1)$, there exists a constant $C > 0$ such that for all $n \in \mathbb{N}$,

$$\begin{aligned} d_M(\phi^t x, \phi^t y) &\leq C \mu^t d_M(x, y), \\ &\quad \text{for } y \in W_{\text{loc}}^s(x), t > 0, \\ d_M(\phi^{-t} x, \phi^{-t} y) &\leq C \mu^t d_M(x, y), \\ &\quad \text{for } y \in W_{\text{loc}}^u(x), t > 0. \end{aligned}$$

The local stable (unstable) manifolds can be extended to global stable (unstable) manifolds $W^s(x)$ and $W^u(x)$. These global manifolds are the leaves of global foliations W^s and W^u of M . One can also define weak stable and

weak unstable foliations with leaves given by $W^{cs}(x) = \bigcup_{t \in \mathbb{R}} (W^s(x))$ and $W^{cu}(x) = \bigcup_{t \in \mathbb{R}} (W^u(x))$, which have as tangent distributions $E^{cs} = E^c \oplus E^s$ and $E^{cu} = E^c \oplus E^u$. In general, all these foliations are only continuous, but their leaves are differentiable.

Any Anosov diffeomorphism is *structurally stable*, that is, any C^1 diffeomorphism that is C^1 close to an Anosov diffeomorphism is topologically conjugate to the unperturbed one via a Hölder homeomorphism. An Anosov flow is structurally stable in the orbit equivalence sense: any C^1 small perturbation of an Anosov flow has the orbit foliation topologically conjugate via a Hölder homeomorphism to the orbit foliation of the unperturbed flow.

Let $SL(n, \mathbb{R})$ be the group of all n -dimensional square matrices with real valued entries of determinant 1. Let $SL(n, \mathbb{Z}) \subset SL(n, \mathbb{R})$ be the subgroup with integer entries. Basic examples of Anosov diffeomorphisms are automorphisms of the n -torus $\mathbb{T}^n = \mathbb{R}^n / \mathbb{Z}^n$ induced by hyperbolic matrices in $SL(n, \mathbb{Z})$. A hyperbolic matrix is one that has only nonzero eigenvalues, all away in absolute value from 1. A specific example of such matrix in $SL(2, \mathbb{Z})$ is

$$\begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

Basic examples of Anosov flows are given by the geodesic flows of surfaces of constant negative curvature. The unitary bundle of such a surface can be realized as $M = \Gamma \backslash PSL(2, \mathbb{R})$, where $PSL(2, \mathbb{R}) = SL(2, \mathbb{R}) / \{\pm 1\}$ and Γ is a cocompact lattice in $PSL(2, \mathbb{R})$. The action of the geodesic flow on M is induced by right multiplication with elements in the diagonal one-parameter subgroup

$$\left\{ \begin{pmatrix} e^{t/2} & 0 \\ 0 & e^{-t/2} \end{pmatrix}, t \in \mathbb{R} \right\}.$$

A related transformation, which is not hyperbolic, but will be of interest in this presentation, is the *horocycle flow* induced by right multiplication on M by elements in the one parameter subgroup

$$\left\{ \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, t \in \mathbb{R} \right\}.$$

Of interest in this survey are also actions of more general groups than \mathbb{Z} and \mathbb{R} . Typical examples of higher rank \mathbb{Z}^k Anosov actions are constructed on tori using groups of units in number fields. See [65] for more details about this construction. A particular example of Anosov \mathbb{Z}^2 -action on \mathbb{T}^3 is induced by the hyperbolic matrices:

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 8 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \\ 1 & 8 & 4 \end{pmatrix}.$$

One can check, by looking at the eigenvalues, that A and B are not multiples of the same matrix. Moreover, A and B commute.

Typical examples of higher rank Anosov \mathbb{R}^k -actions are given by Weyl chamber flows, which we now describe using some notions from the theory of Lie groups. A good reference for the background in Lie group theory necessary here is the book of Helgason [40]. Note that for a hyperbolic element of such an action the center distribution is k -dimensional and coincides with the tangent distribution to the orbit foliation of \mathbb{R}^k .

Let G be a semi-simple connected real Lie group of the noncompact type, with Lie algebra \mathfrak{g} . Let $K \subset G$ be a maximal compact subgroup that gives a Cartan decomposition $\mathfrak{g} = \mathfrak{k} + \mathfrak{p}$, where \mathfrak{k} is the Lie algebra of K and \mathfrak{p} is the orthogonal complement of \mathfrak{k} with respect to the Killing form of \mathfrak{g} . Let $\alpha \subset \mathfrak{p}$ be a maximal abelian subalgebra and $A = \exp \alpha$ be the corresponding subgroup. The simultaneous diagonalization of $\text{ad}_{\mathfrak{g}}(\alpha)$ gives the decomposition

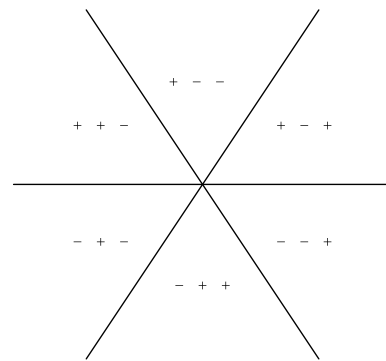
$$\mathfrak{g} = \mathfrak{g} + \sum_{\lambda \in \Lambda} \mathfrak{g}_{\lambda}, \quad \mathfrak{g}_0 = \alpha + \mathfrak{m},$$

where Λ is the set of restricted roots. The spaces \mathfrak{g}_{λ} are called *root spaces*. A point $H \in \alpha$ is called *regular* if $\lambda(H) \neq 0$ for all $\lambda \in \Lambda$. Otherwise it is called *singular*. The set of regular elements consists of the complement of a union of finitely many hyperplanes. Its components are cones in α called *Weyl chambers*. The faces of the Weyl chambers are called *Weyl chamber walls*.

Let M be the centralizer of A in K . Suppose Γ is an irreducible torsion-free cocompact lattice in G . Since A commutes with M , the action of A by right translations on $\Gamma \backslash G$ descends to an A -action on $N := \Gamma \backslash G/M$. This action is called a *Weyl chamber flow*. Any Weyl chamber flow is an Anosov action, that is, has an element that acts hyperbolically transversally to the orbit foliation of A . Note that all maximal connected \mathbb{R} diagonalizable subgroups of G are conjugate and their common dimension is called the \mathbb{R} -rank of G . If the \mathbb{R} -rank k of G is higher than 2, then the Weyl chamber flow is a higher rank hyperbolic \mathbb{R}^k -action.

An example of semi-simple Lie group is $SL(n, \mathbb{R})$. Let A be the diagonal subgroup of matrices with positive entries in $SL(n, \mathbb{R})$. An example of Weyl chamber flow that will be discussed in the sequel is the action of A by right translations on $\Gamma \backslash SL(n, \mathbb{R})$, where Γ is a cocompact lattice. In this case the centralizer M is trivial. The rank of this action is $n - 1$. The picture of the Weyl chambers for $n = 3$ is shown in Fig. 1. The signs that appear in each chamber are the signs of half of the Lyapunov exponents of a regular element from the chamber with respect

ERGODIC THEORY: RIGIDITY



Ergodic Theory: Rigidity, Figure 1
Weyl chambers for $SL(3, \mathbb{R})$

to a certain fixed basis. For this action, the Lyapunov exponents appear in pairs of opposite signs.

An example of higher rank lattice Anosov action that will be discussed in the sequel is the standard action of $SL(n, \mathbb{Z})$ on the torus \mathbb{T}^n , $(A, x) \mapsto Ax, A \in SL(n, \mathbb{Z}), x \in \mathbb{T}^n$. $SL(n, \mathbb{Z})$ is a (noncocompact!) lattice in $SL(n, \mathbb{R})$. As shown in [55], this action is generated by Anosov diffeomorphisms.

We describe now a class of dynamical systems more general than the hyperbolic one. A C^1 diffeomorphism f of a compact differentiable manifold M is called *partially hyperbolic* if there exists a continuous invariant splitting of the tangent bundle $TM = E^s \oplus E^0 \oplus E^u$ such that the derivative of f expands E^u much more than E^0 , and contracts E^s much more than E^0 . See [9,41] and [10] for the theory of partially hyperbolic diffeomorphisms. E^s and E^u are called stable, respectively unstable distributions, and are integrable. E^0 is called center distribution and, in general, it is not integrable. A structural stability result proved by Hirsch, Pugh, Shub [41], that is a frequently used tool in rigidity, shows that, if E^0 is integrable to a smooth foliation, then any perturbation \tilde{f} of f is partially hyperbolic and has an integrable center foliation. Moreover, the center foliations of \tilde{f} of f are mapped one into the other by a homeomorphism that conjugates the maps induced on the factor spaces of the center foliations by \tilde{f} of f respectively.

We review now basic facts about cocycles. These basic definitions refer to several regularity classes: measurable, continuous, or differentiable. Let G be a group acting on a set M , and denote the action $G \times M \rightarrow M$ by $(g, x) \mapsto gx$; thus $(g_1 g_2)x = g_1(g_2 x)$. Let Γ be a group with unit 1_Γ . M is usually endowed with a measurable, continuous, or differentiable structure. A *cocycle* β over

the action is a function $\beta: G \times M \rightarrow \Gamma$ such that

$$\beta(g_1 g_2, x) = \beta(g_1, g_2 x) \beta(g_2, x), \quad (1)$$

for all $g_1, g_2 \in G, x \in M$. Note that any group representation $\pi: G \rightarrow \Gamma$ defines a cocycle called *constant cocycle*. The trivial representation defines the *trivial cocycle*.

A natural equivalence relation for the set of cocycles is given by cohomology. Two cocycles $\beta_1, \beta_2: G \times M \rightarrow \Gamma$ are cohomologous if there exists a map $P: M \rightarrow \Gamma$, called *transfer map*, such that

$$\beta_1(g, x) = P(gx) \beta_2(g, x) P(x)^{-1}, \quad (2)$$

for all $g \in G, x \in M$.

Differentiable Rigidity

We start by reviewing cohomological results. Several basic notions are already defined in Sect. “[Basic Definitions and Examples](#)”. In this section we assume that the cocycles are at least continuous. A cocycle $\beta: G \times M \rightarrow \Gamma$ over an action $(g, x) \mapsto gx, g \in G, x \in M$, is said to satisfy *closing conditions* if for any $g \in G$ and $x \in M$ such that $gx = x$, one has $\beta(g, x) = 1_\Gamma$. Note that closing conditions are necessary in order for a cocycle to be cohomologous to the trivial one. Since a \mathbb{Z} -cocycle is determined by a function $\beta: M \rightarrow \Gamma, \beta(x) := \beta(1, x)$, the closing conditions become

$$f^n x = x \text{ implies } \beta(f^{n-1}x) \dots \beta(x) = 1_\Gamma, \quad (3)$$

where $f: M \rightarrow M$ is the function that implements the \mathbb{Z} -action.

The first cohomological results for hyperbolic systems were obtained by Livshits [70,71]. Let M be a compact Riemannian manifold with a \mathbb{Z} -action implemented by a topologically transitive Anosov C^1 diffeomorphism f . Then an α -Hölder function $\beta: M \rightarrow \mathbb{R}$ determines a cocycle cohomologous to a trivial cocycle if and only if β satisfies the closing conditions (3). The transfer map is α -Hölder. Moreover, for each Hölder class α and each finite dimensional Lie group Γ , there is a neighborhood U of the identity in Γ such that an α -Hölder function $\beta: M \rightarrow \Gamma$ determines a cocycle cohomologous to the trivial cocycle if and only if β satisfies the closing conditions (3). The transfer map is again α -Hölder. Similar results are true for Anosov flows.

Using Fourier analysis, Veech [108] extended Livshits's result to real valued cocycles over \mathbb{Z} -actions induced by ergodic endomorphisms of an n -dimensional torus, not necessarily hyperbolic. For cocycles with values in abelian groups, the question of two arbitrary cocycles being cohomologous reduces to the question of an

arbitrary cocycle being cohomologous to a trivial one. This is not the case for cocycles with values in non-abelian groups. Parry [98] extended Livshits's criteria to one for cohomology of two arbitrary cocycles with values in compact Lie groups. Parry's result was generalized by Schmidt [104] to cocycles with values in Lie groups that, in addition, satisfy a center bunching condition. Nițică, Török [92] extended Livshits's result to cocycles with values in the group $\text{Diff}^k(M)$ of C^k diffeomorphism of a compact manifold M with stably trivial bundle, $k \geq 3$. Examples of such manifolds are the tori and the spheres. In this case, the transfer map takes values in $\text{Diff}^{k-3}(M)$, and it is Hölder with respect to a natural metric on $\text{Diff}^{k-3}(M)$. In [92] one can also find a generalization of Livshits's result to generic Anosov actions, that is, actions generated by families of Anosov diffeomorphisms that do not interchange the stable and unstable directions of elements in the family. An example of such an action is the standard action of $SL(n, \mathbb{Z})$ on the n -dimensional torus.

A question of interest is the following: if two C^k cocycles, $1 \leq k \leq \omega$, over a hyperbolic action, are cohomologous through a continuous/measurable transfer map P , what can be said about the higher regularity of P ? For real valued cocycles the question can be reduced to one about cohomologically trivial cocycles. Livshits showed that for a real valued C^1 cocycle cohomologous to a constant the transfer map is C^1 . He also obtained C^∞ regularity results if the action is given by hyperbolic automorphisms of a torus. After preliminary results by Guillemin and Kazhdan for geodesic flows on surfaces of negative curvature, for general hyperbolic systems the question was answered positively by de la Llave, Marco, Moriyon [76] in the C^∞ case and by de la Llave [74] in the real analytic case. Nițică, Török [93] considered the lift of regularity for a transfer map between two cohomologous cocycles with values in a Lie group or a diffeomorphism group. In contrast to the case of cocycles cohomologous to trivial ones, here one needs to require for the transfer map a certain amount of Hölder regularity that depends on the ratio between the expansion/contraction that appears in the base and the expansion/contraction introduced by the cocycle in the fiber. This assumption is essential, as follows from a counterexample found by de la Llave's [73].

Useful tools in this development have been results from analysis that lift the regularity of a continuous real valued function which is assumed to have higher regularity along pairs of transverse Hölder foliations. Many times the foliations are the stable and unstable ones associated to a hyperbolic system. Journé [46] proved the $C^{n,\alpha}$ regularity of a continuous function that is $C^{n,\alpha}$ along two transverse continuous foliations with $C^{n,\alpha}$ leaves. If one

is interested only in C^∞ regularity, a convenient alternative is a result of Hurder, Katok [44]. This has a simpler proof and can be applied to the more general situation in which the function is regular along a web of transverse foliations. A real analytic regularity result along these lines belongs to de la Llave [74]. In certain problems, for example when working with Weyl chamber flows, it is difficult to control the regularity in enough directions to span the whole tangent space. Nevertheless, the tangent space can be generated if one consider higher brackets of good directions. A C^∞ regularity result for this case belongs to Katok, Spatzier [60]. In order to apply this result, the foliations need to be C^∞ not only along the leaves, but also transversally.

An application of the above regularity results is to questions about transverse regularity of the stable and unstable foliations of the geodesic flow on certain C^∞ surfaces of nonpositive curvature. For compact negatively curved C^∞ surfaces, E. Hopf showed that these foliations are C^1 , and it follows from the work of Anosov that individual leaves are C^∞ . Hurder, Katok [44] showed that once the weak-stable and weak-unstable foliations of a volume-preserving Anosov flow on a compact 3-manifold are C^2 , they are C^∞ .

Another application of the regularity results is to the study of invariants for C^k conjugacy of hyperbolic systems. By structural stability, a small C^1 perturbation of a hyperbolic system is C^0 conjugate to the unperturbed one. The conjugacy, in general, is only Hölder. If the conjugacy is C^1 then it preserves the eigenvalues of the derivative at the periodic orbit. The following two results describe the invariants of smooth and real analytic conjugacy of low dimensional hyperbolic systems. They are proved in a series of papers written in various combinations by de la Llave, Marco, Moryion [72,74,75,79,80].

Let X, Y be two $C^\infty(C^\omega)$ transitive Anosov vector fields on a compact three-dimensional manifold. If they are C^0 conjugate and the eigenvalues of the derivative at the corresponding periodic orbits are the same, then the conjugating homeomorphism is $C^\infty(C^\omega)$. In particular, any C^1 conjugacy is $C^\infty(C^\omega)$.

Assume now that f, g are two $C^\infty(C^\omega)$ Anosov diffeomorphisms on a compact two dimensional manifold. If they are C^0 conjugate and the eigenvalues of the derivative at the corresponding periodic orbits are the same, then the conjugating diffeomorphism is $C^\infty(C^\omega)$. In particular, any C^1 conjugacy is $C^\infty(C^\omega)$.

An important direction was initiated by Katok, Spatzier [59] who studied cohomological results over hyperbolic \mathbb{Z}^k or \mathbb{R}^k -actions, $k \geq 2$. They show that real valued smooth/Hölder cocycles over typical classes of hyper-

bolic \mathbb{Z}^k or \mathbb{R}^k , $k \geq 2$, actions are smoothly/Hölder cohomologous to constants. These results cover, in particular, actions by hyperbolic automorphisms of a torus, and Weyl chamber flows. The proofs rely on harmonic analysis techniques, such as Fourier transform and group representations for semi-simple Lie groups.

A geometric method for cocycle rigidity was developed in [64]. One constructs a differentiable form using invariant structures along stable/unstable foliations, and the commutativity of the action. The form is exact if and only if the cocycle is cohomologous to a constant one. The method covers actions on nilmanifolds satisfying a condition called TNS (Totally Non-Symplectic). This condition means that the action is higher rank abelian hyperbolic, and that the tangent space is a direct sum of invariant distributions, with each pair of these included in the stable distribution of a hyperbolic element of the action. The method was also applied to small (i.e. close to identity on a set of generators) Lie group valued cocycles. A related paper is [96] which contains rigidity results for cocycles over (TNS) actions with values in compact Lie groups. In this situation the number of cohomology classes is finite. An example of (TNS) action is given by the action of a maximal diagonalizable subgroup of $SL(n, \mathbb{Z})$ on \mathbb{T}^n .

Recently Damjanović, Katok [14] developed a new method that was applied to the action of the matrix diagonal group on $\Gamma \backslash SL(n, \mathbb{R})$. They use techniques from [54], where one finds cohomology invariants for cocycles over partially hyperbolic actions that satisfy accessibility property. Accessibility means that one can connect any two points from the manifold supporting the partially hyperbolic dynamical system by transverse piecewise smooth paths included in stable/unstable leaves. This notion was introduced by Brin, Pesin [9] and it is playing a crucial role in the recent surge of activity in the field of partially hyperbolic diffeomorphisms. See [10] for a recent survey of the subject. The cohomology invariants described in [54] are heights of the cocycle over cycles constructed in the base out of pieces inside stable/unstable leaves. They provide a complete set of obstructions for solving the cohomology equation. A new tool introduced in [14] is algebraic K -theory [88]. The method can be extended to cocycles with non-abelian range. In [57] one finds related results for small cocycles with values in a Lie group or the diffeomorphism group of a compact manifold.

The equivalent of the Livshits theorem in the higher-rank setting appears to be a description of the highest cohomology rather than the first cohomology. Indeed, for higher rank partially hyperbolic actions of the torus, the intermediate cohomologies are trivial, while for the high-

est one the closing conditions characterize the cohomology classes. This behavior provides a generalization of Veech cohomological result and of Katok, Spatzier cohomological result for toral automorphisms, and was discovered by A. Katok, S. Katok [52,53].

Flaminio, Forni [27] studied the cohomological equation over the horocycle flow. It is shown that there are infinitely many obstructions to the existence of a smooth solution. Moreover, if these obstructions vanish, then one can solve the cohomological equation. In [28] it is shown a similar result for cocycles over area preserving flows on compact higher-genus surfaces under certain assumptions that hold generically. Mieczkowski [87] extended these techniques and studied the cohomology of parabolic higher rank abelian actions. All these results rely on non-commutative Fourier analysis, more specifically representation theory of $SL(2, \mathbb{R})$ and $SL(2, \mathbb{C})$.

Local Rigidity

Let Γ be a finitely (or compactly) generated group, G a topological group, and $\pi: \Gamma \rightarrow G$ a homomorphism. The target of local rigidity theory is to understand the space of perturbations of various homomorphisms π . Trivial perturbations of a homomorphism arise from conjugation by an arbitrary element of G . In order to rule them out, one says that π is *locally rigid* if any nearby homomorphism π' , (that is, π' close to π on a finite or compact set of generators of Γ), is conjugate to π by an element $g \in G$, that is, $\pi(\gamma) = g\pi'(\gamma)g^{-1}$ for all $\gamma \in \Gamma$. If G is path-wise connected, one can also consider *deformation rigidity*, meaning that any nearby continuous path of homomorphisms is conjugated to the initial one via a continuous path of elements in G that has an end in the identity.

Initial results on local rigidity are about embeddings of lattices into semi-simple Lie groups. The main results belong to Weil [110,111,112]. He showed that if G is a semi-simple Lie group that is not locally isomorphic to $SL(2, \mathbb{R})$ and if $\Gamma \subset G$ is an irreducible cocompact lattice, then the natural embedding of Γ into G is locally rigid. Earlier results were obtained by Selberg [105], Calabi, Vesentini [12], and Calabi [11]. Selberg proved the local rigidity of the natural embedding of cocompact lattices into $SL(n, \mathbb{R})$. His proof used the dynamics of iterates of matrices, in particular the existence of singular directions, or walls of Weyl chambers, in the maximal diagonalizable subgroups of $SL(n, \mathbb{R})$. Selberg's approach inspired Mostow [90] to use the boundaries at infinity in his proof of strong rigidity of lattices, which in turn was crucial to the development of superrigidity due to Margulis [82]. See Sect. "Global Rigidity" for more details.

Recall that hyperbolic dynamical systems are structurally stable. Thus they are, in a certain sense, locally rigid. We introduce now a precise definition of local rigidity in the infinite-dimensional setup. The fact that for general group actions one needs to consider different regularities for the actions, perturbations and conjugacies is apparent from the description of structural stability for Anosov systems.

A C^k action α of a finitely generated discrete group Γ on a manifold M , that is, a homomorphism $\alpha: \Gamma \rightarrow \text{Diff}^k(M)$, is said to be $C^{k,l,p,r}$ *locally rigid* if any C^l perturbation $\tilde{\alpha}$ which is C^p close to α on a family of generators, is C^r conjugate to α , i.e. there exists a C^r diffeomorphism $H: M \rightarrow M$ which conjugates $\tilde{\alpha}$ to α , that is, $H \circ \alpha(g) = \tilde{\alpha}(g) \circ H$ for all $g \in \Gamma$. Note that for Anosov \mathbb{Z} -actions, $C^{1,1,1,0}$ rigidity is known as *structural stability*. One can also introduce the notion of *deformation rigidity* if the initial action and the perturbation are conjugate by a continuous path of diffeomorphisms that has an end coinciding to the identity.

A weaker notion of local rigidity can be defined in the presence of invariant foliations for the initial group action and for the perturbation. The map H is now required to preserve the leaves of the foliations and to conjugate only after factorization by the invariant foliations. The importance of this notion is apparent from the leafwise conjugacy structural stability theorem of Hirsch, Pugh, Shub [41]. See Sect. "Basic Definitions and Examples". Moreover, for Anosov flows this is the natural notion of structural stability, and appears by taking the invariant foliation to be the one-dimensional orbit foliation. For more general actions, of lattices or higher rank abelian groups, this property is often used in combination to cocycle rigidity in order to show local rigidity. We discuss more about this when we review local rigidity results for partially hyperbolic actions.

We summarize now several developments in local rigidity that emerged in the 80s. Initial results [67,115] were about *infinitesimal rigidity*, that is, a weaker version of local rigidity suitable for discrete groups representations in infinite-dimensional spaces of smooth vector fields. Then Hurder [43] proved $C^{\infty,\infty,1,\infty}$ deformation rigidity and Katok, Lewis, Zimmer [55,56,63] proved $C^{\infty,\infty,1,\infty}$ local rigidity of the standard action of $SL(n, \mathbb{Z})$, $n \geq 3$, on the n -dimensional torus. In these results crucial use was made of the presence of an Anosov element in the action. Due to the uniqueness of the conjugacy coming from structural stability, one has a continuous candidate for the conjugacy between the actions. Margulis, Qian [85] used the existence of a spanning family of directions that are hyperbolic for certain elements of the action to show lo-

cal rigidity of partially hyperbolic actions that are not hyperbolic. Another important tool present in many proofs is Margulis and Zimmer superrigidity. These results allow one to produce a measurable conjugacy for the perturbation. Then one shows that the conjugacy has higher regularity using the presence of hyperbolicity. Having enough directions to span the whole tangent space is essential to lift the regularity. A cocycle to which superrigidity can be applied is the derivative cocycle.

The study of local rigidity of partially hyperbolic actions that contain a compact factor was initiated by Nițică, Török [92,94]. Let $n \geq 3$ and $d \geq 1$. Let ρ be the action of $SL(n, \mathbb{Z})$ on $\mathbb{T}^{n+d} = \mathbb{T}^n \times \mathbb{T}^d$ given by $\rho(A)(x, y) = (Ax, y)$, $x \in \mathbb{T}^n$, $y \in \mathbb{T}^d$, $A \in SL(n, \mathbb{Z})$. Then, for $K \geq 1$, [92] shows that ρ is $C^{\infty, \infty, 5, K-1}$ deformation rigid. The proof is based on three results in hyperbolic dynamics: the generalization of Livshits's cohomological results to cocycles with values in diffeomorphism groups, the extension of Livshits's result to general Anosov actions, and a version of the Hirsch, Pugh, Shub structural stability theorem improving the regularity of the conjugacy.

Assume now $n \geq 3$ and $K \geq 1$. If ρ is the action of $SL(n, \mathbb{Z})$ on $\mathbb{T}^{n+1} = \mathbb{T}^n \times \mathbb{T}$ given by $\rho(A)(x, y) = (Ax, y)$, $x \in \mathbb{T}^n$, $y \in \mathbb{T}$, $A \in SL(n, \mathbb{Z})$, [94] shows that the action ρ is $C^{\infty, \infty, 2, K-1}$ locally rigid. Ingredients in the proof are two rigidity results, one about TNS actions, and one about actions of property (T) groups. A locally compact group has property (T) if the trivial representation is isolated in the Fell topology. This means that if G acts on a Hilbert space unitarily and it has almost invariant vectors, then it has invariant vectors. Hirsch–Pugh–Shub theorem implies that perturbations of abelian partially hyperbolic actions of product type are conjugated to skew-products of abelian Anosov actions via cocycles with values in diffeomorphism groups. In addition, the TNS property implies that the sum of the stable and unstable distributions of any regular element of the perturbation is integrable. The leaves of the integral foliation are closed, covering the base simply. Thus one obtains a conjugacy between the perturbation and a product action. Property (T) is used to show that the conjugacy reduces the perturbed action to a family of perturbations of hyperbolic actions. But the last ones are already known to be conjugate to the hyperbolic action in the base.

Recent important progress in the question of local rigidity of lattice actions was made by Fisher, Margulis [24,25,26]. Their proofs are modeled along the proof of Weil's local rigidity result [112] and use an analog of Hamilton's [39] hard implicit function theorem. Let G be a connected semi-simple Lie group with all simple factors of rank at least two, and $\Gamma \subset G$ a lattice. The main result

shows that a volume preserving affine action ρ of G or Γ on a compact smooth manifold X is $C^{\infty, \infty, \infty, \infty}$ locally rigid. Lower regularity results are also available. A component of the proof shows that if Γ is a group with property (T), X a compact smooth manifold, and ρ a smooth action of Γ on X by Riemannian isometries, then ρ is $C^{\infty, \infty, \infty, \infty}$ locally rigid. An earlier local rigidity result for this type of actions by cocompact lattices was obtained by Benveniste [5].

Many lattices act naturally on “boundaries” of type G/P , where G is a semi-simple algebraic Lie group and P is a parabolic subgroup. An example is given by $G = SL(2, \mathbb{R})$ and P the subgroup in G consisting of upper triangular matrices. Local rigidity results for this type of actions were found by Ghys [34], Kanai [50] and Katok, Spatzier [62].

Starting with the work of Katok and Lewis, a related direction was the study of local rigidity for higher rank abelian actions. They prove in [55] the $C^{\infty, \infty, 1, \infty}$ local rigidity of the action of a \mathbb{Z}^n maximal diagonalizable (over \mathbb{R}) subgroup of $SL(n+1, \mathbb{Z})$, $n \geq 2$, acting on the torus \mathbb{T}^{n+1} . These type of results were later pushed forward by Katok, Spatzier [62]. Using the theory of nonstationary normal forms developed in [38] and [37] by Katok, Guysinsky, they proved several local rigidity results. The first one assumes that G is a semisimple Lie group with all simple factors of rank at least two, Γ a lattice in G , N a nilpotent Lie group and Λ a lattice in N . Then any Anosov affine action of Γ on N/Λ is $C^{\infty, \infty, 1, \infty}$ locally rigid. Second, let \mathbb{Z}^d be a group of affine transformations of N/Λ for which the derivatives are simultaneously diagonalizable over \mathbb{R} with no eigenvalues on the unit circle. Then the \mathbb{Z}^d -action on N/Λ is $C^{\infty, \infty, 1, \infty}$ locally rigid. A related result for continuous groups is the $C^{\infty, \infty, 1, \infty}$ local rigidity (after factorization by the orbit foliation) of the action of a maximal abelian \mathbb{R} -split subgroup in an \mathbb{R} -split semi-simple Lie group of real rank at least two on G/Λ , where Λ is a cocompact lattice in G .

One can also study rigidity of higher rank abelian partially hyperbolic actions that are not hyperbolic. Natural examples appear as automorphisms of tori and as variants of Weyl chamber flows. For the case of ergodic actions by automorphisms of a torus, this was investigated using a version of the KAM (Kolmogorov, Arnold, Moser) method by Damianović, Katok [15]. As usual in the KAM method, one starts with a linearization of the conjugacy equation. At each step of the iterative KAM scheme, some twisted cohomological equations are solved. The existence of the solutions is forced by the ergodicity of the action and the higher rank assumptions. Diophantine conditions present in this case allow to control the fixed loss of regu-

larity which is necessary for the convergence of these solutions to a conjugacy.

Global Rigidity

The first remarkable result in global rigidity belongs to Mostow [90]. For G a connected non-compact semi-simple Lie group not locally isomorphic to $SL(2, \mathbb{R})$, and two irreducible cocompact lattices $\Gamma_1, \Gamma_2 \subset G$, Mostow showed that any isomorphism θ from Γ_1 into Γ_2 extends to an isomorphism of G into itself. G has an involution σ whose fixed set is a maximal compact subgroup K . One constructs the symmetric Riemannian space $X = G/K$. To each chamber of X corresponds a parabolic group and these parabolic groups are endowed with a Tits geometry similar to the projective geometry of lines, planes etc. formed in the classical case when $G = PGL(n, \mathbb{R})$. The proof of Mostow's result starts by building a θ -equivariant pseudo-isometric map $\phi: G/K_1 \rightarrow G/K_2$. The map ϕ induces an incidence preserving θ -equivariant isomorphism ϕ_0 of the Tits geometries. By Tits' generalized fundamental theorem of projective geometry, ϕ_0 is induced by an isomorphism of G . Finally, $\theta(\gamma) = \phi_0 \cdot \gamma \cdot \phi_0^{-1}$ gives the desired conclusion.

The next remarkable result in global rigidity is Margulis' superrigidity theorem. An account of this development can be found in the monograph [82]. For large classes of irreducible lattices in semi-simple Lie groups, this result classifies all finite dimensional representations. Let G be a semi-simple simply connected Lie group of rank higher than two and $\Gamma < G$ an irreducible lattice. Then any linear representation π of Γ is almost the restriction of a linear representation of G . That is, there exists a linear representation π_1 of G and a bounded image representation π_2 of Γ such that $\pi = \pi_1 \pi_2$. The possible representations π_2 are also classified by Margulis up to some facts concerning finite image representations. As in the case of Mostow's result, the proof involved the analysis of a map defined on the boundary at infinity. In this case the map is studied using deep results from dynamics like the multiplicativity ergodic theorem of Osledec [97] or the theory of random walks on groups developed by Furstenberg [31]. An important consequence of Margulis superrigidity result is the arithmeticity of irreducible lattices in connected semi-simple Lie groups of rank higher than two. A basic example of arithmetic lattice can be obtained by taking the integer points in a semi-simple Lie group that is a matrix group, like taking $SL(n, \mathbb{Z})$ inside $SL(n, \mathbb{R})$. Special cases of superrigidity theorems were proved by Corlette [13] and Gromov, Schoen [36] for the rank one groups $Sp(1, n)$ and respec-

tively F_4 using the theory of harmonic maps. A consequence is the arithmeticity of lattices in these groups. Some of these results are put into differential geometric setting in [89].

Margulis superrigidity result was extended to cocycles by Zimmer. A detailed exposition, including a self contained presentation of several rigidity results of Margulis, can be found in the monograph [113]. We mention here a version of this result that can be found in [24]. Let M be a compact manifold, H a matrix group, $P = M \times H$, and Γ a lattice in a simply connected, semi-simple Lie group with all factors of rank higher than two. Assume that Γ acts on M and H in a way that makes the projection from P to M equivariant. Moreover, the action of Γ on P is measure preserving and ergodic. Then there exists a measurable map $s: M \rightarrow H$, a representation $\pi: G \rightarrow H$, a compact subgroup $K < H$ which commute with $\pi(G)$ and a measurable map $k: \Gamma \times M \rightarrow K$ such that $\gamma \cdot s(m) = k(\gamma, m) \cdot \pi(\gamma) \cdot s(\gamma \cdot m)$. One can easily check from the last equation that k is a cocycle. So, up to a measurable change of coordinates given by the map s , the action of Γ on P is a compact extension via a cocycle of a linear representation of G .

Developing further the method of Mostow for studying the Tits building associated to a symmetric space of non-positive curvature led Ballman, Brin, Eberlein, Spatzier [2,3] to a number of characterizations of symmetric spaces. In particular, they showed that if M is a complete Riemannian manifold of non-positive curvature, finite volume, with simply connected cover, irreducible and of rank at least two, then M is isometric to a symmetric space with the connected component of $\text{Isom}(M)$ having no compact factors.

A topological rigidity theorem has been proved by Farrell, Jones [21]. They showed that if N is a complete connected Riemannian manifold whose sectional curvature lies in a closed interval included in $(-\infty, 0]$, and M is a topological manifold of dimension greater than 5, then any proper homotopy equivalence $f: M \rightarrow N$ is properly homotopic to a homeomorphism. In particular, if M and N are both compact connected negatively curved Riemannian manifolds with isomorphic fundamental groups, then M and N are homeomorphic.

Likewise to the case of local rigidity, a source of inspiration for results in global rigidity was the theory of hyperbolic systems, in particular their classification. The only known examples of Anosov diffeomorphisms are hyperbolic automorphisms of infranilmanifolds. Moreover, any Anosov diffeomorphism on an infranilmanifold is topologically conjugate to a hyperbolic automorphism [29,78]. It is conjectured that any Anosov diffeomorphism is topo-

logically conjugate to a hyperbolic automorphism of an infranilmanifold. Partial results are obtained in [91], where the conjecture is proved for Anosov diffeomorphisms with codimension one stable/unstable foliation. The proof of the general conjecture eluded efforts done so far. It is not even known if any Anosov diffeomorphism is topologically transitive, that is, if it has a dense orbit. A few positive results are available. Let M be a C^∞ compact manifold endowed with a C^∞ affine connection. Let f be a topologically transitive Anosov diffeomorphism preserving the connection and such that the stable and unstable distributions are C^∞ . Then Benoist, Labourie [4] proved that f is C^∞ conjugate to a hyperbolic automorphism of an infranilmanifold.

The situation for Anosov flows is somehow different. As shown in [30], there exist Anosov flows that are not topologically transitive, so a general analog of the conjecture is false. Nevertheless, for the case of codimension one stable or unstable foliation, it is conjectured in [109] that any Anosov flow on a manifold of dimension greater than three admits a global cross-section. This would imply that the flow is topologically conjugate to the suspension of a linear automorphism of a torus.

For actions of groups larger than \mathbb{Z} , or \mathbb{R} , global classification results are more abundant. A useful strategy in these type of results, which are quite technical, is to start by obtaining a measurable description of the action, most of the time using Margulis–Zimmer superrigidity results, and then use extra assumptions on the action, such as the presence of a hyperbolic element, or the presence of an invariant geometric structure, or both, in order to show that the measurable model is actually continuous or even differentiable. For actions of higher rank Lie groups and their lattices some representative papers are by Katok, Lewis, Zimmer [63] and Goetze, Spatzier [35]. For actions of higher rank abelian groups see Kalinin, Spatzier [49].

Measure Rigidity

Measure rigidity is the study of invariants measures for actions of one parameter and multiparameter abelian groups and semigroups acting on manifolds. Typical situations when interesting rigidity phenomena appear are for one parameter unipotent actions and higher rank hyperbolic actions, discrete or continuous.

A unipotent matrix is one all of whose eigenvalues are one. An important case where the action of a unipotent flow appears is that of the horocycle flow. The invariant measures for it were studied by Furstenberg [33], who showed that the horocycle flow on a compact surface is uniquely ergodic, that is, the ergodic measure is unique.

Dani and Smillie [17] extended this result to the case of non-compact surfaces, with the only other ergodic measures appearing being those supported on compact horocycles. An important breakthrough is the work of Margulis [81], who solved a long standing question in number theory, Oppenheim’s conjecture. The conjecture is about the density properties of the values of an indefinite quadratic form in three or more variables, provided the form is not proportional to a rational form. The proof of the conjecture is based on the study of the orbits for unipotent flows acting by translations on the homogenous space $SL(n, \mathbb{Z}) \backslash SL(n, \mathbb{R})$. All these results were special cases of the Raghunathan conjecture about the structure of the orbits of the actions of unipotent flows on homogeneous spaces. Raghunathan’s conjecture was proved in full generality by Ratner [100,101]. Borel, Prasad [8] raised the question of an analog of Raghunathan’s conjecture for S -algebraic groups. S -algebraic groups are products of real and p -adic algebraic groups. This was answered independently by Margulis, Tomanov [86] and Ratner [102].

A basic example of higher rank abelian hyperbolic action is given by the action of $S_{m,n}$, the multiplicative semigroup of endomorphisms generated by the multiplication by m and n , two nontrivial integers, on the one dimensional torus \mathbb{T}^1 . In a pioneering paper [32] Furstenberg showed that for m, n that are not powers of the same integer the action of $S_{m,n}$ has a unique closed, infinite invariant set, namely \mathbb{T}^1 itself. Since there are many closed, infinite invariant sets for multiplication by m , and by n , this result shows a remarkable rigidity property of the joint action. Furstenberg’s result was generalized later by Berend for other group actions on higher dimensional tori and on other compact abelian groups in [6] and [7].

Furstenberg further opened the field by raising the following question:

Conjecture 1 *Let μ be a $S_{m,n}$ -invariant and ergodic probability measure on \mathbb{T}^1 . Then μ is either an atomic measure supported on a finite union of (rational) periodic orbits or μ is the Lebesgue measure.*

While the statement appears to be simple, proving it has been elusive. The first partial result was given by Lyons [77] under the strong additional assumption that the measure makes one of the endomorphisms generating the action exact. Later Rudolph [103] and Johnson [45] weaken the exactness assumption and proved that μ must be the Lebesgue measure provided that multiplication by m (or multiplication by n) has positive entropy with respect to μ . Their results have been proved again using slightly different methods by Feldman [22]. A further extension was given by Host [42].

Katok proposed another example for which measure rigidity can be fruitfully tested, the \mathbb{Z}^2 -action induced by two commuting hyperbolic automorphisms on the torus \mathbb{T}^3 . An example of such action is shown in Sect. “Basic Definitions and Examples”. One can also consider the action induced by a \mathbb{Z}^{n-1} maximal abelian group of hyperbolic automorphisms acting on the torus \mathbb{T}^n . Katok, Spatzier [61] developed a more geometric technique allowing to prove measure rigidity for these actions if they have an element acting with positive entropy. Their technique can be applied in higher generality if the action is irreducible in a strong sense and, in addition, it has individual ergodic elements or it is TNS. See also [47]. This method is based on the study of conditional measures induced by the invariant measure on certain invariant foliations that appear naturally in the presence of a hyperbolic action. Besides stable and unstable foliations, one can also consider various intersections of them. Einsiedler, Lindenstrauss [19] were able to eliminate the ergodicity and TNS assumptions.

Yet another interesting example of higher rank abelian action is given by Weyl chamber flows. These do not satisfy the TNS condition. Einsiedler, Katok [18] proved that if G is $SL(n, \mathbb{R})$, $\Gamma \subset G$ is a lattice, H is the subgroup of positive diagonal matrices in G , and μ a H -invariant and ergodic measure on G/Γ such that the entropy of μ with respect to all one parameter subgroups of H is positive, then μ is the G invariant measure on G/Γ .

These results are useful in the investigation of several deep questions in number theory. Define $X = SL(3, \mathbb{Z}) \backslash SL(3, \mathbb{R})$ the diagonal subgroup of matrices with positive entries in $SL(3, \mathbb{R})$. This space is not compact but is endowed with a unique translation invariant probability measure. The diagonal subgroup

$$H = \left\{ \begin{pmatrix} e^s & 0 & 0 \\ 0 & e^t & 0 \\ 0 & 0 & e^{-s-t} \end{pmatrix} : s, t \in \mathbb{R} \right\}$$

acts naturally on X by right translations. It was conjectured by Margulis [83] that any compact H -invariant subset of X is a union of compact H -orbits. A positive solution to this conjecture implies a long standing conjecture of Littlewood:

Conjecture 2 Let $\|x\|$ denote the distance between the real number x and the closest integer. Then

$$\liminf_{n \rightarrow \infty} n \|n\alpha\| \|n\beta\| = 0 \quad (4)$$

for any real numbers α and β .

A partial result was obtained by Einsiedler, Katok, Lindenstrauss [20] who proved that the set of pairs $(\alpha, \beta) \in \mathbb{R}^2$

for which (4) is not satisfied has Hausdorff dimension zero. Applications of these techniques to questions in quantum ergodicity were found by Lindenstrauss [69].

A current direction in measure rigidity is attempting to classify the invariant measures under rather weak assumptions about the higher rank abelian action, like the homotopical data for the action. Kalinin and Katok [48] proved that any \mathbb{Z}^k , $k \geq 2$, smooth action α on a $k+1$ -dimensional torus whose elements are homotopic to the corresponding elements of an action by hyperbolic automorphisms preserves an absolutely continuous measure.

Future Directions

An important open problem in differential rigidity is to find invariants for the C^k conjugacy of the perturbations of higher dimensional hyperbolic systems. For Anosov diffeomorphisms, de la Llave counterexample [73] shows that this extension is not possible for a four dimensional example that appears as a direct product of two dimensional Anosov diffeomorphisms. Indeed, there are C^∞ perturbations of the product that are only C^k conjugate to the unperturbed system for any $k \geq 1$. In the positive direction, Katok conjectured that generalizations are possible for the diffeomorphism induced by an irreducible hyperbolic automorphism of a torus. One can also investigate this question for Anosov flows.

Examples of higher rank Lie groups can be obtained by taking products of rank one Lie groups. Many actions of irreducible lattices in these type of groups are believed to be locally rigid, but the techniques available so far cannot be applied. A related problem is to study local rigidity in low regularity classes, for example the local rigidity of homomorphisms from higher rank lattices into homeomorphism groups. More problems related to local rigidity can be found in [23].

An important problem in global rigidity, emphasized by Katok and Spatzier, is to show that, up to differentiable conjugacy, any higher rank Anosov or partially hyperbolic action is algebraic under the assumption that it is sufficiently irreducible. The irreducibility assumption is needed in order to exclude actions obtained by successive application of products, extensions, restrictions and time changes from basic ingredients which include some rank one actions.

Another problem of current research in measure rigidity is to develop a counterpart of Ratner's theory for the case of actions by hyperbolic higher rank abelian groups on homogenous spaces. It was conjectured by Katok, Spatzier [61] that the invariant measures for such actions given by toral automorphisms or Weyl chamber flows are

essentially algebraic, that is, supported on closed orbits of connected subgroups. Margulis in [84] extended this conjecture to a rather general setup addressing both the topological and measurable aspects of the problem. More details about actions on homogenous spaces, as well as connections to diophantine approximation, can be found in the article ► [Ergodic Theory on Homogeneous Spaces and Metric Number Theory](#) by Kleinbock.

Acknowledgment

This research was supported in part by NSF Grant DMS-0500832.

Bibliography

Primary Literature

- Anosov DV (1967) Geodesic flows on closed Riemannian manifolds with negative curvature. *Proc Stek Inst* 90:1–235
- Ballman W, Brin M, Eberlein P (1985) Structure of manifolds of non-negative curvature. I. *Ann Math* 122:171–203
- Ballman W, Brin M, Spatzier R (1985) Structure of manifolds of non-negative curvature. II. *Ann Math* 122:205–235
- Benoist Y, Labourie F (1993) Flots d'Anosov à distributions stable et instable différentiables. *Invent Math* 111:285–308
- Benveniste EJ (2000) Rigidity of isometric lattice actions on compact Riemannian manifolds. *Geom Func Anal* 10:516–542
- Berend D (1983) Multi-invariant sets on tori. *Trans AMS* 280:509–532
- Berend D (1984) Multi-invariant sets on compact abelian groups. *Trans AMS* 286:505–535
- Borel A, Prasad G (1992) Values of isotropic quadratic forms at S -integral points. *Compositio Math* 83:347–372
- Brin MI, Pesin YA (1974) Partially hyperbolic dynamical systems. *Izvestia* 38:170–212
- Burns K, Pugh C, Shub M, Wilkinson A (2001) Recent results about stable ergodicity. In: Katok A, Pesin Y, de la Llave R, Weiss H (eds) *Smooth ergodic theory and its applications*, Seattle, 1999. *Proc Symp Pure Math*, vol 69. AMS, Providence, pp 327–366
- Calabi E (1961) On compact Riemannian manifolds with constant curvature. I. *Proc Symp Pure Math*, vol 3. AMS, Providence, pp 155–180
- Calabi E, Vesentini E (1960) On compact, locally symmetric Kähler manifolds. *Ann Math* 17:472–507
- Corlette K (1992) Archimedean superrigidity and hyperbolic geometry. *Ann Math* 135:165–182
- Damianović D, Katok A (2005) Periodic cycle functionals and cocycle rigidity for certain partially hyperbolic actions. *Disc Cont Dynam Syst* 13:985–1005
- Damianović D, Katok A (2007) Local rigidity of partially hyperbolic actions KAM, I -method and \mathbb{Z}^k -actions on the torus. Preprint available at http://www.math.psu.edu/katok_a
- Dani SG, Margulis GA (1990) Values of quadratic forms at integer points: an elementary approach. *Enseign Math* 36:143–174
- Dani SG, Smillie J (1984) Uniform distributions of horocycle orbits for Fuchsian groups. *Duke Math J* 51:185–194
- Einsiedler M, Katok A (2003) Invariant measures on G/Γ for split simple Lie groups. *Comm Pure Appl Math* 56:1184–1221
- Einsiedler M, Lindenstrauss E (2003) Rigidity properties of \mathbb{Z}^d -actions on tori and solenoids. *Elec Res Ann AMS* 9:99–110
- Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood conjecture. *Ann Math* 164:513–560
- Farrell FT, Jones LE (1989) A topological analog of Mostow's rigidity theorem. *J AMS* 2:237–370
- Feldman J (1993) A generalization of a result of R Lyons about measures on $[0,1]$. *Isr J Math* 81:281–287
- Fisher D (2006) Local rigidity of group actions: past, present, future. In: *Dynamics, Ergodic Theory and Geometry* (2007). Cambridge University Press
- Fisher D, Margulis GA (2003) Local rigidity for cocycles. In: *Surv Diff Geom VIII*. International Press, Cambridge, pp 191–234
- Fisher D, Margulis GA (2004) Local rigidity of affine actions of higher rank Lie groups and their lattices. 2003
- Fisher D, Margulis GA (2005) Almost isometric actions, property T, and local rigidity. *Invent Math* 162:19–80
- Flaminio L, Forni G (2003) Invariant distributions and time averages for horocycle flows. *Duke Math J* 119:465–526
- Forni G (1997) Solutions of the cohomological equation for area-preserving flows on compact surfaces of higher genus. *Ann Math* 146:295–344
- Franks J (1970) Anosov diffeomorphisms. In: Chern SS, Smale S (eds) *Global Analysis (Proc Symp Pure Math, XIV, Berkeley 1968)*. AMS, Providence, pp 61–93
- Franks J, Williams R (1980) Anomalous Anosov flows. In: *Global theory of dynamical systems*, Proc Inter Conf Evanston, 1979. *Lecture Notes in Mathematics*, vol 819. Springer, Berlin, pp 158–174
- Furstenberg H (1963) A Poisson formula for semi-simple Lie groups. *Ann Math* 77:335–386
- Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math Syst Theor* 1:1–49
- Furstenberg H (1973) The unique ergodicity of the horocycle flow. In: *Recent advances in topological dynamics*, Proc Conf Yale Univ, New Haven 1972. *Lecture Notes in Mathematics*, vol 318. Springer, Berlin, pp 95–115
- Ghys E (1985) Actions localement libres du groupe affine. *Invent Math* 82:479–526
- Goetze E, Spatzier R (1999) Smooth classification of Cartan actions of higher rank semi-simple Lie groups and their lattices. *Ann Math* 150:743–773
- Gromov M, Schoen R (1992) Harmonic maps into singular spaces and p -adic superrigidity for lattices in groups of rank one. *Publ Math IHES* 76:165–246
- Guysinsky M (2002) The theory of nonstationary normal forms. *Erg Th Dyn Syst* 22:845–862
- Guysinsky M, Katok A (1998) Normal forms and invariant geometric structures for dynamical systems with invariant contracting foliations. *Math Res Lett* 5:149–163
- Hamilton R (1982) The inverse function theorem of Nash and Moser. *Bull AMS* 7:65–222
- Helgason S (1978) *Differential Geometry, Lie Groups and Symmetric Spaces*. Academic Press, New York
- Hirsch M, Pugh C, Shub M (1977) *Invariant Manifolds*. *Lecture Notes in Mathematics*, vol 583. Springer, Berlin

42. Host B (1995) Nombres normaux, entropie, translations. *Isr J Math* 91:419–428
43. Hurder S (1992) Rigidity of Anosov actions of higher rank lattices. *Ann Math* 135:361–410
44. Hurder S, Katok A (1990) Differentiability, rigidity and Godbillon-Vey classes for Anosov flows. *Publ Math IHES* 72:5–61
45. Johnson AS (1992) Measures on the circle invariant under multiplication by a nonlacunary subsemigroup of integers. *Isr J Math* 77:211–240
46. Journé JL (1988) A regularity lemma for functions of several variables. *Rev Mat Iberoam* 4:187–193
47. Kalinin B, Katok A (2002) Measurable rigidity and disjointness for \mathbb{Z}^k -actions by toral automorphisms. *Erg Theor Dyn Syst* 22:507–523
48. Kalinin B, Katok A (2007) Measure rigidity beyond uniform hyperbolicity: invariant measures for Cartan actions on tori. *J Modern Dyn* 1:123–146
49. Kalinin B, Spatzier R (2007) On the classification of Cartan actions. *Geom Func Anal* 17:468–490
50. Kanai M (1996) A new approach to the rigidity of discrete group actions. *Geom Func Anal* 6:943–1056
51. Katok A, Hasselblatt B (1995) Introduction to the modern theory of dynamical systems. *Encyclopedia of Mathematics and its Applications* 54. Cambridge University Press, Cambridge
52. Katok A, Katok S (1995) Higher cohomology for abelian groups of toral automorphisms. *Erg Theor Dyn Syst* 15:569–592
53. Katok A, Katok S (2005) Higher cohomology for abelian groups of toral automorphisms. II. The partially hyperbolic case, and corrigendum. *Erg Theor Dyn Syst* 25:1909–1917
54. Katok A, Kononenko A (1996) Cocycles' stability for partially hyperbolic systems. *Math Res Lett* 3:191–210
55. Katok A, Lewis J (1991) Local rigidity for certain groups of toral automorphisms. *Isr J Math* 75:203–241
56. Katok A, Lewis J (1996) Global rigidity results for lattice actions on tori and new examples of vol preserving actions. *Isr J Math* 93:253–280
57. Katok A, Nițică V (2007) Rigidity of higher rank abelian cocycles with values in diffeomorphism groups. *Geometriae Dedicata* 124:109–131
58. Katok A, Nițică V () Differentiable rigidity of abelian group actions. Cambridge University Press (to appear)
59. Katok A, Spatzier R (1994) First cohomology of Anosov actions of higher rank abelian groups and applications to rigidity. *Publ Math IHES* 79:131–156
60. Katok A, Spatzier R (1994) Subelliptic estimates of polynomial differential operators and applications to rigidity of abelian actions. *Math Res Lett* 1:193–202
61. Katok A, Spatzier R (1996) Invariant measures for higher-rank abelian actions. *Erg Theor Dyn Syst* 16:751–778; Katok A, Spatzier R (1998) Corrections to: Invariant measures for higher-rank abelian actions.; (1996) *Erg Theor Dyn Syst* 16:751–778; *Erg Theor Dyn Syst* 18:503–507
62. Katok A, Spatzier R (1997) Differential rigidity of Anosov actions of higher rank abelian groups and algebraic lattice actions. *Trudy Mat Inst Stek* 216:292–319
63. Katok A, Lewis J, Zimmer R (1996) Cocycle superrigidity and rigidity for lattice actions on tori. *Topology* 35:27–38
64. Katok A, Nițică V, Török A (2000) Non-abelian cohomology of abelian Anosov actions. *Erg Theor Dyn Syst* 2:259–288
65. Katok A, Katok S, Schmidt K (2002) Rigidity of measurable structure for \mathbb{Z}^d -actions by automorphisms of a torus. *Comm Math Helv* 77:718–745
66. Kazhdan DA (1967) On the connection of a dual space of a group with the structure of its closed subgroups. *Funkc Anal Prilozn* 1:71–74
67. Lewis J (1991) Infinitesimal rigidity for the action of $SL_n(\mathbb{Z})$ on \mathbb{T}^n . *Trans AMS* 324:421–445
68. Lindenstrauss E (2005) Rigidity of multiparameter actions. *Isr Math J* 149:199–226
69. Lindenstrauss E (2006) Invariant measures and arithmetic quantum unique ergodicity. *Ann Math* 163:165–219
70. Livshits A (1971) Homology properties of Y -systems. *Math Zametki* 10:758–763
71. Livshits A (1972) Cohomology of dynamical systems. *Izvestia* 6:1278–1301 he Livšic cohomology equation. *Ann Math* 123:537–611
72. de la Llave R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. I. *Comm Math Phys* 109:369–378
73. de la Llave R (1992) Smooth conjugacy and S-R-B measures for uniformly and non-uniformly hyperbolic dynamical systems. *Comm Math Phys* 150:289–320
74. de la Llave R (1997) Analytic regularity of solutions of Livshits's cohomology equation and some applications to analytic conjugacy of hyperbolic dynamical systems. *Erg Theor Dyn Syst* 17:649–662
75. de la Llave R, Moriyo R (1988) Invariants for smooth conjugacy of hyperbolic dynamical systems. IV. *Comm Math Phys* 116:185–192
76. de la Llave R, Marco JM, Moriyo R (1986) Canonical perturbation theory of Anosov systems and regularity results for the Livšic cohomology equation. *Ann Math* 123:537–611
77. Lyons R (1988) On measures simultaneously 2- and 3-invariant. *Isr J Math* 61:219–224
78. Manning A (1974) There are no new Anosov diffeomorphisms on tori. *Amer J Math* 96:422–429
79. Marco JM, Moriyo R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. I. *Comm Math Phys* 109:681–689
80. Marco JM, Moriyo R (1987) Invariants for smooth conjugacy of hyperbolic dynamical systems. III. *Comm Math Phys* 112:317–333
81. Margulis GA (1989) Discrete subgroups and ergodic theory. In: *Number theory, trace formulas and discrete groups*, Oslo, 1987. Academic Press, Boston, pp 277–298
82. Margulis GA (1991) Discrete subgroups of semi-simple Lie groups. Springer, Berlin
83. Margulis GA (1997) Oppenheim conjecture. In: *Fields Medalists Lectures*, vol 5. World Sci Ser 20th Century Math. World Sci Publ, River Edge, pp 272–327
84. Margulis GA (2000) Problems and conjectures in rigidity theory. In: *Mathematics: frontiers and perspectives*. AMS, Providence, pp 161–174
85. Margulis GA, Qian N (2001) Local rigidity of weakly hyperbolic actions of higher rank real Lie groups and their lattices. *Erg Theor Dyn Syst* 21:121–164
86. Margulis GA, Tomanov G (1994) Invariant measures for actions of unipotent groups over local fields of homogenous spaces. *Invent Math* 116:347–392

87. Mieczkowski D (2007) The first cohomology of parabolic actions for some higher-rank abelian groups and representation theory. *J Modern Dyn* 1:61–92
88. Milnor J (1971) Introduction to algebraic K-theory. Princeton University Press, Princeton
89. Mok N, Siu YT, Yeung SK (1993) Geometric superrigidity. *Invent Math* 113:57–83
90. Mostow GD (1973) Strong rigidity of locally symmetric spaces. *Ann Math Studies* 78. Princeton University Press, Princeton
91. Newhouse SE (1970) On codimension one Anosov diffeomorphisms. *Amer J Math* 92:761–770
92. Niţică V, Török A (1995) Cohomology of dynamical systems and rigidity of partially hyperbolic actions of higher rank lattices. *Duke Math J* 79:751–810
93. Niţică V, Török A (1998) Regularity of the transfer map for cohomologous cocycles. *Erg Theor Dyn Syst* 18:1187–1209
94. Niţică V, Török A (2001) Local rigidity of certain partially hyperbolic actions of product type. *Erg Theor Dyn Syst* 21:1213–1237
95. Niţică V, Török A (2002) On the cohomology of Anosov actions. In: *Rigidity in dynamics and geometry*, Cambridge, 2000. Springer, Berlin, pp 345–361
96. Niţică V, Török A (2003) Cocycles over abelian TNS actions. *Geom Ded* 102:65–90
97. Oseledec VI (1968) A multiplicative ergodic theorem. Characteristic Lyapunov exponents of dynamical systems. *Trudy Mosk Mat Obsc* 19:179–210
98. Parry W (1999) The Livšic periodic point theorem for non-abelian cocycles. *Erg Theor Dyn Syst* 19:687–701
99. Pugh C, Shub M (1972) Ergodicity of Anosov actions. *Invent Math* 15:1–23
100. Ratner M (1991) On Ragunathan's measure conjecture. *Ann Math* 134:545–607
101. Ratner M (1991) Ragunathan's topological conjecture and distributions of unipotent flows. *Duke Math J* 63:235–280
102. Ratner M (1995) Ragunathan's conjecture for Cartesians products of real and p-adic Lie groups. *Duke Math J* 77:275–382
103. Rudolph D (1990) $\times 2$ and $\times 3$ invariant measures and entropy. *Erg Theor Dyn Syst* 10:395–406
104. Schmidt K (1999) Remarks on Livšic' theory for nonabelian cocycles. *Erg Theor Dyn Syst* 19:703–721
105. Selberg A (1960) On discontinuous groups in higher-dimensional symmetric spaces. In: *Contributions to function theory. Inter Colloq Function Theory*, Bombay. Tata Institute of Fundamental Research, pp 147–164
106. Smale S (1967) Differentiable dynamical systems. *Bull AMS* 73:747–817
107. Spatzier R (1995) Harmonic analysis in rigidity theory. In: *Ergodic theory and its connections with harmonic analysis*. Alexandria, 1993. London Math Soc Lect Notes Ser, vol 205. Cambridge University Press, Cambridge, pp 153–205
108. Veech WA (1986) Periodic points and invariant pseudomeasures for toral endomorphisms. *Erg Theor Dyn Syst* 6:449–473
109. Verjovsky A (1974) Codimension one Anosov flows. *Bul Soc Math Mex* 19:49–77
110. Weil A (1960) On discrete subgroups of Lie groups. I. *Ann Math* 72:369–384
111. Weil A (1962) On discrete subgroups of Lie groups. II. *Ann Math* 75:578–602
112. Weil A (1964) Remarks on the cohomology of groups. *Ann Math* 80:149–157
113. Zimmer R (1984) *Ergodic theory and semi-simple groups*. Birkhäuser, Boston
114. Zimmer R (1987) Actions of semi-simple groups and discrete subgroups. *Proc Inter Congress of Math* (1986). AMS, Providence, pp 1247–1258
115. Zimmer R (1990) Infinitesimal rigidity of smooth actions of discrete subgroups of Lie groups. *J Diff Geom* 31:301–322

Books and Reviews

- de la Harpe P, Valette A (1989) La propriété (T) de Kazhdan pour les groupes localement compacts. *Astérisque* 175
- Feres R (1998) *Dynamical systems and semi-simple groups: An introduction*. Cambridge Tracts in Mathematics, vol 126. Cambridge University Press, Cambridge
- Feres R, Katok A (2002) Ergodic theory and dynamics of G-spaces. In: *Handbook in Dynamical Systems*, 1A. Elsevier, Amsterdam, pp 665–763
- Gromov M (1988) Rigid transformation groups. In: Bernard D, Choquet-Bruhat Y (eds) *Géométrie Différentielle* (Paris, 1986). Hermann, Paris, pp 65–139; *Travaux en Cours*. 33
- Kleinbock D, Shah N, Starkov A (2002) Dynamics of subgroup actions on homogeneous spaces of Lie groups and applications to number theory. In: *Handbook in Dynamical Systems*, 1A. Elsevier, Amsterdam, pp 813–930
- Knapp A (2002) *Lie groups beyond an introduction*, 2nd edn. Progress in Mathematics, 140. Birkhäuser, Boston
- Ragunathan MS (1972) *Discrete subgroups of Lie groups*. Springer, Berlin
- Witte MD (2005) *Ratner's theorems on unipotent flows*. Chicago Lectures in Mathematics. University of Chicago Press, Chicago

Evacuation as a Communication and Social Phenomenon

DOUGLAS GOUDIE

Australian Centre for Disaster Studies, School of Earth and Environmental Science, James Cook University, Townsville, Australia

Article Outline

Glossary

Definition of the Subject

Introduction

Concepts, Language and Mathematically Modeling the Propensity to Evacuate

Mathematical Modeling

Effective Risk Communication

Integrating Theory and Implementation

Institutional Barriers to Greater Community Self-Help

Experiences and Lessons – Some Case Studies

Discussion

Acknowledgment

Bibliography

Glossary

Community A group of neighbors or people with a commonality of association and generally defined by location, shared experience, or function [59].

Community empowerment Internally and externally nurtures a community to accept that residents live in a hazard zone, and they choose to do things as a group to maximize their safety.

Community safety group Existing community groups (such as Neighborhood Watch) and individuals, working with formal response organizations form a coherent affiliation in and near a hazard zone, to help maximize safety and care for all community members.

Disaster The interface between an extreme physical event and a vulnerable human population [81].

Disaster lead time The time taken from first detection of a natural disaster threat to the likely time of impact on humans or human structures.

Disaster threat A natural extreme event which may impact on a community.

Effective risk communication . That which motivates people to maximize their own safety.

Emergency An actual or imminent event which endangers or threatens to endanger life, property or the environment, and which requires a significant and coordinated response [55].

Evacuation People relocating to safely escape hazardous disaster impacts. To move from a high danger zone to relative safety.

Hazard A source of potential harm or a situation with a potential to cause loss. A situation or condition with potential for loss or harm to the community or environment [55]. Hazard is synonymous with 'source of risk' [25].

Hazard zone Defined geographic areas which may be subject to a natural disaster impact of flood, bush-fire, storm surge, destructive winds, earthquake, landslide or damaging hail. Hazard zones include major accident sites, including industrial, transport or mining precincts; or biological or terrorist threat or impact, or from-source predicted area(s) of pandemic spread.

Mitigation Any efforts taken which may reduce the impact of a threat.

Prevention Measures to eliminate or reduce the incidence or severity of emergencies [55].

Ramp-up preparations The final set of preparations and precautionary evacuations taken ahead of a forecast disaster impact. This includes earlier final actions than precipitated by formal organizations.

Risk treatment options Measures that modify the characteristics of hazards, communities and environments to reduce risk, e. g. prevention, preparedness, response and recovery [55].

Vulnerability comprises 'resilience' and 'susceptibility'. 'Resilience' is related to 'existing controls' and the capacity to reduce or sustain harm. 'Susceptibility' is related to 'exposure' [25].

Definition of the Subject

This article intends to show how system and complexity science can contribute to an understanding and improvement of evacuation processes, especially considering the roles of engaged communities at risk, the concepts of community self-help, and clear communication about local threats and remedies.

This article shows researchers in Complexity and Systems Science (CSS) a social sciences approach to maximize effective and precautionary evacuation, maximize safety, minimize loss and speed full recovery. The computational and analytical modeling tools of CSS may be considered to apply to a complex interaction of community awareness, inclination to accept the reality of a natural disaster threat, along with achieving background and final preparations to maximize safety and recovery from a natural disaster impact. This article may stimulate CSS researchers to develop detailed models of the complex systems and complexity of melding information from Weather Bureaus and Disaster Managers, via contacts and intervening media to communities at risk, with the shared social goal of maximizing safety. This social sciences task requires cross-disciplinary approaches of respect and response.

The old disaster management model lacked the predictive and rapid communication systems now available and developing in disaster predictive models (such as a flood maps). An approach to modeling the great complexity of human behavior responding to threat is provided. Such a model must include people's prior knowledge of a threat type, and consider such fine detail as the overarching language used in a country with threat zones, and the dominant languages of all under threat. It is hoped this article stimulates CSS models to further engage in this social good of helping people get safe and stay safe through natural disasters by providing predictive tool to Authorities to better inform and encourage those at risk to action, including the possible need for precautionary evacuations ahead of a predicted impact.

Disaster management in Australia, and increasingly, globally, is focused on mitigation as part of a 'threat continuum', from acceptance that some locations are vulner-

able to a hazard impact, through to recovery [13]. Emergency warnings and a possible need to evacuate are embedded as ‘spikes’ on that continuum. Thus, this article stresses the importance of developing ways; incentives, to mobilize aware at-risk community members to precautionary self-evacuation. For this to happen, people need to know and internalize the reality that they are in a hazard zone.

Thus, in the cost-effective philosophy of engendering self-help, the process of understanding the complexity of achieving the shared social goal in maximizing safety and minimizing loss is to engender creation of empowered communities with a high motivation for safety-oriented and precautionary action. This is likely to lead to minimized loss and disruption, and maximized recovery. This article details many elements of that process, and invites detailed development of the Sustainability Implementation Research to achieve that goal through CSS.

To model the path to collective safety, the complexity of the dynamics at play need to be clarified: impact preparedness, including possible evacuation, is a communication and social issue.

This article demonstrates that acknowledgment of hazard zones, developing community acceptance of threat and needed action needs to be at the individual, household and community levels. Evacuation modeling is needed only for those whose homes may be at real threat of a disaster impact. For those living in a hazard zone, a fully informed community, who have internalized the reality of the threat and have worked for maximum background preparation, and have mechanisms to receive alerts and warnings of a looming threat; a community predisposed to precautionary evacuations will result.

Capturing this complexity is the challenge for modelers. Evacuation is about hazard zone residents actively monitoring a looming threat via refined communication channels detailed in this article, within a developed social predisposition to act. Some examples are given. For consideration by scientists and students internationally, this article introduces the Communication Safety Triangle and the Seven Steps to Community Safety on the Preparedness Continuum, within the new research frame of Sustainability Implementation Research (SIR).

Introduction

The purpose of this article is to share with modelers and complexity and system scientists the social and communication issues of modeling effective safety strategies to a natural disaster threat. It is hoped, through the approaches and processes described in this article, that mod-

elers will more clearly link physical threats with warnings and community engagement.

This article first looks at the definitions and language used in risk communication and effective warnings, leading to informal and formal evacuations, then considers some Australian policies relating to emergency management. Theories related to risk communication are presented, with examples of evacuation issues provided from Indigenous communities, and from non English speaking households. The needed conceptual shift to self-help is placed within the larger theoretical frame of paradigms and paradigms shifts.

An example of including residents to internalize threats is given, followed by a more general example of transport evacuation.

A discussion of international evacuation issues precedes a broader view of some of the institutional barriers which may restrict the uptake of the seven step approach to an aware, informed community, relying on accurate information and choosing to self evacuate as a precaution. These issues are discussed, considering effective ways of allowing people to know that they are at risk so they are inclined to evacuate themselves, as a practice. These approaches can be used or tested by other scholars. Recommendations and a summary of the key issues to maximize voluntary and safe evacuations finish this article.

Some Key Evacuation Issues

In an era of increasing social self-help [7,55] and community empowerment [13,25,26], a part of global efforts to embrace Ecologically Sustainable Development [8,62] is to increasingly see evacuation as a social phenomenon.

This article provides a conceptual frame, the Communication Safety Triangle (CST) which includes responsible media telling people at risk what they need to know and *seven steps to community safety* (7SCS). The 7SCS help guide emergency managers and modelers to treat the possible need for evacuation as a decision-making process where the community should be aware of the potential threats and receive clear, detailed and reliable information on the possible need to evacuate, so most residents in a high impact zone self-evacuate in a precautionary way, as a practice. Examples provided in this article illustrate this new, *sustainability implementation research* (SIR) approach to disaster management.

Warnings precede a perceived need to evacuate. In the USA, the need for an integrative approach to warnings is identified: “There is a major need for better coordination among the warning providers, more effective delivery mechanisms, better education of those at risk, and new

ways for building partnerships among the many public and private groups involved” [63].

Sustainability implementation is the way to achieve a sustainable future. Disaster management, effective risk communication and community self-help provide a stark and comprehensible example of what sustainability implementation means, how it will benefit societies, and will help channel us into a safer, more sustainable future. Within this approach, scientists and students become major agents for sustainability implementation, as models can illustrate the needed paths to achievement.

Evacuation Overview

There are three types of disasters which may require evacuation: human induced and natural disasters with and without lead times. This article is focused on precautionary evacuations ahead of natural disasters with lead times (Table 1). The research and conceptual frames draw on and are applicable to all communities in hazard zones. Defining, modeling and effectively communicating to at-risk residents and travelers about specific geographic hazards and safety-oriented behaviors are core elements of successful precautionary evacuations.

Within Table 1, evacuation may be a response to an emerging hazard of indicated intensity, direction and speed. This article advocates development of informed and directed communities which will actively respond to a communicated threat, with the vulnerable moving them-

selves early to places of safety, or being helped by other community members to move.

Concepts, Language and Mathematically Modeling the Propensity to Evacuate

Precautionary self evacuation pivots on knowing who is at risk. With the help of researchers, modelers; planning and community involvement can prepare people and their valuables to minimize loss. Very public hazard maps will help people internalize that they are in a hazard zone, and what they should do. The concept is not new or alarming: every public building has emergency exit routes marked, with all that implies. Air flight comes with the mandatory emergency preparation presentation. Minute or recurrent risks to where we live or travel to should be no less public, nor more alarming than a fire drill in a public building.

Hazard Types – Little or Considerable Warning Times

Hazards may or may not have lead times (Table 1). Sudden onset threats include: tsunami, earthquake, major eruption, major toxic spill or discharge, mine disaster, terrorist threat or attack. Signaled threats include: – cyclone, flood, fire and destructive winds. This article focuses on disasters with sufficient warning periods to be able to evacuate the vulnerable away from the predicted worst impact areas. Table 1 that with sudden onset impacts, sheltering to sur-

Evacuation as a Communication and Social Phenomenon, Table 1
Evacuation decision matrix – short to long warning times

Evacuation decision matrix, Evacuation around <i>sudden onset impact</i>			
Hazard:	Landslide	Earthquake	Tsunami
Possible safety-oriented response	Stay in strong structure. Move across slope as soon as possible.	Get into the open or shelter under strong structure.	Immediately flee to higher ground.

Precautionary Evacuation (PE) decisions with lead time – signaled threats			
Considerations for evacuation decision	Hazard		
	Destructive wind/cyclone	Fire	Flood
1. Vulnerability of present environment	If likely to be in a storm surge, must PE; If shelter weak, must PE	House material, surrounds, water available. If poor, PE	May be inundated= PE; may be cut off, consider PE
2. If 1 OK: vulnerability of individuals (e. g. weak)	PE first	Asthmatics and less able: PE	Judgments of flood height. If in any doubt, PE
3a. Distance to safe shelter	The further, PE earlier	The further, PE earlier	The further, PE earlier
3b. Safety along exit route	Know in preparation	Know in preparation	Know in preparation
3c. Means of travel	Reliability and suitability	Reliability and suitability	Reliability and suitability
4. Community cohesion	Help available	Help available	Help available

vive the first few minutes is critical, then moving to open; stable ground is important to avoid further aftershocks or landslide. Always take direction from local authorities. For natural disasters with lead times, Table 1 shows that, if you will clearly be safe where you are, stay put, and prepare as best you can. If, in the worst case, you may not be safe, move early (Table 1). The vulnerability of individuals needs to be considered. In a bush fire, for instance, the general advice is: if the property is well prepared [7], stay and defend. If, however, you may not be able to cope with the psychological terror of staying through the fire front, fully prepare your property, and then leave early (Table 1). If the exit route may be blocked by flood waters, or by dense smoke or fire, evacuation needs to precede that obstruction. In the ‘background’ phase of community disaster preparation, all such possible obstacles to a clear escape route (including gridlock congestion) need to be factored in to the timing of precautionary evacuation (Table 1).

Finally, as seen in the Woodgate Beach example of Sect. “Experiences and Lessons – Some Case Studies”, the level of community support in ensuring all residents are aware of and prepared for the ‘ramp-up’ phase of a possible disaster impact, and receive the warning as soon as possible, the able-bodied will help the less able to get out of harm’s way early in the threat period, as a safety-oriented practice, minimizing the demands of the formal response groups as the likelihood of impact increases.

All effective communication involves sending signals and having them received and processed, then incorporated into the receiver’s world view [27,40,41,76,81,82]. Effective risk communication [4,13], Sect. “Effective Risk Communication”, motivates people to act to maximize their own safety. This may not happen if people have not internalized that the threat is real. They are in denial. Alternatively, people may be ignorant of the threat, or why it should be taken seriously. Salter [77] categorizes ignorance from pure ignorance to acts of ignoring.

Disaster Definitions

A disaster may be seen as a negative impact of a hazard on a community as measure of vulnerability. The language of disaster mitigation evolved and is increasingly practiced since the late 1990s i. e. [13,96]. Risk is seen as a function of probability and consequence, related to exposure and the level of force embedded in the threatening hazard.

Boughton [6] points out that natural disasters are usually extremely rare for the individuals concerned but they can cause massive impacts. Because Australia is so vast, overall there are reasonably frequent natural disas-

ters. However, in most locations they are rare indeed. Boughton [6] argues that a “natural disaster” is a natural event in which the community life is seriously and traumatically disrupted. Embracing the way forward with ‘structural mitigation’, “... a key step in preventing natural disasters is to prevent building damage.” [6]. Like “disasters”, “community” is difficult to define [84]. As developed in Sect. “Integrating Theory and Implementation”, “community” is the collection of people in a close geographic area, particularly focused on near neighbors and supportive friends of those in or near a known hazard zone.

What to Communicate

Having the right words or approaches in place *as policy* does not automatically guarantee community safety-oriented responses to disruptive warnings. Since 1989 the approach to cope with disasters has been *prevention, preparedness, response and recovery* training courses. This helps focus all concerned on the temporal sequence. The language could perhaps be refined to talk about acceptance that a threat exists, background, then ‘ramp-up’ (final) preparations to safe shelter; impact, then orderly return and recovery. Classically, ‘response’ was seen as the final, near impact flurry of ‘lock down’ activity [57].

Deeper than language is our conceptual frame (Sect. “Integrating Theory and Implementation”). If researchers encouraged disaster managers to move ‘response’ behavior back a day; a few hours to ‘precautionary, early response’, many of Lewis’s [57] legitimate concerns would be addressed. Some disaster managers may see themselves as dramatic figures in the early impact phase of a disaster, rather than calm, precautionary minimizers of risk. This culture has changed greatly over the last decade (Sect. “Integrating Theory and Implementation”). For many reasons, there is a clear divide between the US Federal response to Cyclone Katrina (2005) and to the California fires (October, 2007). This shift from passive to active can be encouraged by modelers. As part of effective risk communication to encourage precautionary impact preparedness, research with remote Indigenous communities and recently arrived, non-English speaking refugees (Sect. “Integrating Theory and Implementation”) showed the importance of accurate, plain English, and the use of images.

Yates [93] argues mitigation efforts need to be refined to make sure they are focused on issues from the relevant local communities. Much of the problem of non response seems that ‘the message’ to take care does not effectively get through to the target [68]. It is to do with communica-

tion, with signals sent, signals received, and their interpretation.

The types of received and interpreted messages are explored specifically in Sect. “[Integrating Theory and Implementation](#)”, which develops the intellectual foundation to more fully understand the semiotics of risk communication. The following outlines the concepts and relevance of semiotics.

Semiotics is the study of signs including words, sounds and such things as ‘body language’. By the ‘message conveying’ principles of semiotics we can understand how various authors on disasters and evacuations approach the topic, and which cultural signs and symbols they manipulate. For example, a sign showing a person running up-slope ahead of an exaggerated tsunami wave contains all the preferred imagery to tell people what to do in a tsunami warning. Images, as seen for locating oneself in airports or finding evacuation stairs, do not use words.

The next section lays the foundations for a mathematical modeling of propensity for householder self-evacuation, centered on knowledge and refined communications acting on residents in hazard zones who are predisposed to act in their own safety.

Mathematical Modeling

Foundations of the Household Safety Preparedness and Action Index: the HSPA Index

This section does not consider refined algorithms to define how successfully people may respond to *Authority’s Perceived Need to Evacuate (APNE)*. This section considers the form and ingredients of how this could be done; the elements and their possible interactions to predict the success of APNE. Unfortunately, in the previously cited Canberra 2003 bush fires, the APNE did not appear universally, nor was it properly conveyed to the residents in dire risk.

Rohrmann [75] has clearly mapped a process from the warning signal to the hoped-for response. The information and processing flow to internalized inferences is called intuitive heuristics [74]. Renn & Rohrmann [74] basically argue that people process probabilistic information on the likelihood of them needing to move, aligned and convergent with researchers like Thompson [87]. Hence the need for the seven step process (Fig. 8) to trigger a paradigm shift for people in danger zones.

As seen in formula 1, there are finite evacuation triggers (ET), ranging from no lead time LT_0 , such as an earthquake, to a LT of days (LT_{xd}), such as a cyclone. So a key component of a successful response to an APNE is the time

in which to disseminate the warning (LT_{0-xd} ; Table 1) Also, Impact Severity (IS) is central to safety and loss.

As detailed in Sect. “[Integrating Theory and Implementation](#)” and Fig. 8; *The seven steps to community safety*, there is a convincing body of research which indicated that People’s Propensity to Respond (PPR) to an emergency warning (whether that response be to finalize preparation to stay and defend the property or evacuate as a precaution) is strongly linked to their Acceptance they are In a Hazard Zone (AIHZ).

Community Resilience (linkage and inter-support and communication – CR) is also important; along with a potential evacuee’s Knowledge Base of the hazard and safety-maximizing behavior (KB). Thinking Through to Recovery (TR) is also important. As per points 6 and 7 in the 7SCS; Fig. 8, the medium of warning delivery, and the quality of the warning information (Medium & Message: MM) is a factor in the response of those under threat. Overall, Sorenson and Mileti [81] believe that increased credibility of the warning source means the specific warning will be more effective. They also provide evidence that the electronic mass media produced the most believable public warnings. This underlines why development of formal links between all types and levels of media information about hazards and preparation from the weather bureau and emergency managers is so important. Having formal links with the media, coupled with web-delivered ‘real-terrain’ simulations of the hazard will help produce a powerful and effective ‘active warning’ regime.

Issues of Exit Route(s) (ER), Possible Travel Mode (PTM) may be crucial. A normally free-flowing exit route may be blocked by the disaster itself, or others trying to flee. This may involve accidents. People’s psychological and physical states of Well-Being (WB) will also affect potential effective evacuation decisions and accomplishment. Authors like Rohrmann [75] have attempted to stylistically model some of this complexity.

Many of the earlier recognized impediments to full preparedness, like unrealistic optimism (Heller et al., [46]) are incorporated into one overarching factor of formula (1): People’s Propensity to Respond (PPR), including the centrally important feature of individual households fully accepting they are in a hazard zone, and that they need to engage in background preparation, ‘listen up’ around a hazard threat, and undertake their own, precautionary final preparations.

Cutter et al. [17] developed 11 factors indicating vulnerability caused by a major disaster, using principal components analysis within factor analysis. The resultant Social Vulnerability Index (SoVI), necessarily weights the 11 components according to the percentage of variation

of analysis of counties on the USA as to resident's vulnerability. Personal wealth, age and housing density head up the contributions of the 11 variables used to determine the SoVI of USA counties. Cutter's SoVI contributes to the likelihood that proper preparations; including for a timely evacuation, would occur, either at a broad regional or more individual level. This SoIV is included in the HPSAI below. Its weighting remains to be tested. A final issue overarching many others is Institutional Barriers to change (IB), detailed in Sect. "Institutional Barriers to Greater Community Self-Help". The support or otherwise of the media can play a critical role in effective crisis communication – most messages are likely to be through the media, so the media is part of the systems complexity blend: Media Support (MS).

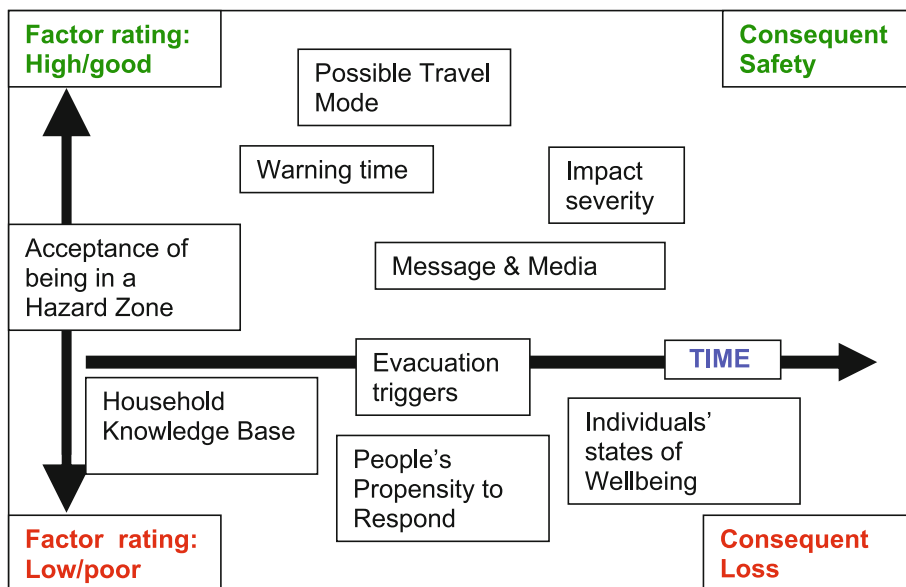
From the weight of SIR and the conceptual framework described in this article, the following generic formula expressing the above factors likely to impact on householder propensity for precautionary evacuations may entice other, more theoretical researchers, to develop the needed algorithms. What follows is purely a design base for others to develop predictive modeling of *People's Propensity and Capacity to Successfully Evacuate (PPCSE – or safely and actively stay)* PPCSE is better named the *Household Preparedness and Safety Action Index: the HPSA Index*. The following is a synthesis of Bayesian Logic (Hoeting et al. [45]) and Eigenvalues.

People's Propensity and Capacity to Successfully Evacuate (PPCSE) or Household Preparedness and Safety Actions Index (HPSAI):

$$\begin{aligned} \text{HPSAI} \propto & f(\text{PPR})(\text{AIHZ})(\text{LT}_{0-xd}) \\ & (\text{APNE})(\text{ET})(\text{IS})(\text{AIHZ})(\text{CR})(\text{TR})(\text{MM})(\text{ER}) \\ & (\text{PTM})(\text{WB})(\text{SoIV})(\text{IB})(\text{MS}) \end{aligned} \quad (1)$$

Formula (1) anticipates constants to 'weigh' each factor according to its contributing importance on the resultant HPSAI (see [45] p 384) for detail of the Bayesian modeling approach.

Resultant safety (*safety* being a combination of successfully evacuating to a safe place, or sheltering within the impact zone in a safe place) of those under threat will be greatest (Fig. 1) where the factors of formula (1) tend to intersect in the positive quadrant of an Eigen plane (Fig. 1) [94]. Equally, outcomes of loss will occur with, for example, short lead times, insufficient warnings of evacuation triggers or lack of belief that residents or travelers are in a hazard zone. A journal search showed few links between disasters and Eigenvalues. An exception is Fowler et al. (2007, [32]). With global warming and increased populations increasingly encroaching into obvious hazard zones, modeling to influence planning and to maximize HPSAI must be a major growth industry.



Evacuation as a Communication and Social Phenomenon, Figure 1
Factors in the Eigen plane determining Household Preparedness and Safety Action Index

Calibrating the Model: the Greek Fires, August '07

[98] or California in late October '07 [99] show that if various of the above variables or likely variable clusters or Eigenvalues from factor analysis have low values, people will not be placed in a strong, precautionary position of safety. This was true of fire-threatened residents in Canberra in 2003 (<http://www.abc.net.au/canberra/bushfires/>), all ultimately leading back to the mathematical weight which needs to be given to each point on in the 7SCS: especially accepting that the threat is real (and possibly imminent).

Once the above-described preparedness paradigm is internalized, it is difficult to experience any disasters news coverage without feeling some despair: cars submerged or floating in flood waters; people fleeing for their lives ahead of imminent immolation; any reports of lost lives through natural disasters with lead times of hours or longer. There are few excuses for this to remain so. In most parts of the developed world, there are prevailing technologies to detect and rapidly convey threat. Researchers are well placed to conceive, trial and link the threat information directly to prepared populations; to close the gap between knowing and acting to maximize safety, as a social exercise of community engagement and communication refinement.

Cyclone Larry [54] showed that, with general community acceptance of the reality of the threat, good warnings and a lead time of about 20 h, combined with well developed disaster management and media cooperation in getting the needed messages to people well primed to the 7SCS, there was no loss of life, although about 5000 people were directly subjected to destructive winds of a category 4/5 cyclone. Residents of the impact zone generally followed the 7SCS, supporting the general approach of formula (1): the community was prepared and acted in a precautionary way to maximize their own safety.

The above model design offers elements/factors to maximize *People's Propensity and Capacity to Successfully Evacuate* (PPCSE – or *safely and actively stay*); the *Household Safety Preparedness and Action Index* (HSPAI), so strengthening a paradigm among Agencies and hazard-zone residents of 'self-help' and 'community engagement' approaches; less passive than earlier approaches used in considering people's vulnerability. This generic approach is supported by model testing by Schadschneider et al. (► [Evacuation Dynamics: Empirical Results, Modeling and Applications](#)).

The Australian examples detailed in this article, including the national shift to mitigation (COAG 2004 [13]), best illustrated in the Woodgate Beach example (Sect. "Experiences and Lessons – Some Case Studies"), show that

the conceptual frame and thus variables chosen to mathematically model evacuation propensity and capacity are more important than the results of any model using 'passive' variables or variable clusters. The advantages and means of achieving this shift from community passivity to partnerships; place- and people-based community empowerment and self-help, form the basis of this article. The above mathematical approach may help validate and guide this necessary conceptual shift.

The next section considers policy, laws, concepts and possible safety-oriented actions that words, images and stories convey to vulnerable residents in hazard zones. In Australia, policy and legislative requirements are indicated, which aim to responsibly maximize community safety [13,24,25,26,27,28,44,70]. Some new concepts, epitomized by new language phrases (with their embedded meanings) are discussed in Sect. "Integrating Theory and Implementation", especially *Community safety groups*. New phrases include 'Social Burnoffs' and 'Practice evacuations'. Like "The Communication Safety Triangle" and "Seven steps to community safety", it is hoped these expressions enter local, state and national lexicons and modeling in disaster management.

Effective Risk Communication

Policy and Laws on Hazard Preparedness and Evacuation

All risk communication operates within government policies and legislation. Forced evacuation may be normal in some jurisdictions for all threats; outlawed in others. Policy may encourage passivity, or be energetically pro-active in achieving the 7SCS, both in urban planning and building materials, and community engagement. In Australia, for instance, there are many federal and regional guides to urban growth, e.g. [2]. A new urban paradigm is, perhaps, best expressed as: "Think globally, act locally, respond personally" ([2] pnp, [3] p 1). This paradigm shift can easily embrace and be informed by computer modeling of individual's behavior, demonstrated in recent Springer's Journal of Systems Science and Complexity articles, such as Kikuchi T & Nakamori Y [97]. Agent model analysis to explore effects of interaction and environment on individual performance [97]

From all decisions of urban development being in the hands of 'experts', government policy increasingly requires a dialog with local residents, through public participation. The first barrier to ecologically sustainable urban development is "... belief systems – doubting a problem exists, or supporting the status quo. [Solutions include] ... consciousness raising campaigns, public partici-

pation in decision-making, demonstration projects and incentives and disincentives” ([76] p 46). This is the KB of formula (1).

This article stresses that response models, Authorities and all residents are constrained (and empowered) by national and regional laws and policies. Research can only work within a prevailing frame, and will be welcomed by or influence prevailing policy. It is thus useful to have clear goals to achieve social or environmental ‘good’ [13,44,70]. This requires greater community involvement in mitigation, and responsible media helping to mitigate disaster impacts (MS of formula 1). In Australia, the national radio broadcaster has a binding agreement with the weather bureau and disaster managers to read issued warnings verbatim; report and engage community members in safety oriented behavior, by providing relevant information. Local community media are often willing to develop the same responsible functions.

Risk Communication’s Core Goals

Risk communicators advise “individuals and communities to respond appropriately to a threat in order to reduce the risk of death, injury, property loss and damage.” [42]. Risk communicators and disaster managers need to formally work closely with the media to maximize social benefit.

Although it may be intended that information flows and is received accurately; that desired behavior will result, and that only communication techniques are important [49], humans tend to be irrational and optimistic, and only hear what they want to hear. It is not what our message *is*, but what, if anything, the listener *does* with our message. To have any chance of ‘success’, information needs to have meaning which is shared between those who construct and send the warning, and those for whom the warning is meant to inform and motivate to action.

‘Action statements’ (what the at-risk person, family or community should actually do to minimize damage) are seen as central to the whole purpose of risk communication [78]. Kasperson & Stallen [49], along with Salter [77,78] and others detail risk communication messages in terms of content, clarity, understandability, consistency, relevance, accuracy, certainty, frequency, channel, credibility, public participation, ethnicity, age, gender, roles, responsibility, elements, sequencing, synopsis, prognosis, location, action, warning timing, and action statements. It is not a case of saying: “a category three cyclone will pass over Smithfield”. It is more a case of making sure the members of Smithfield hear that message, and feel moved to and competent to take well-understood personal risk-minimizing actions.

Knowledge, Self, World Views and Messages

For precautionary evacuations to be successful, the seven steps to safety (Sect. “[Integrating Theory and Implementation](#)”) have occurred; or people have been coerced by authorities, or they have seen others depart, and felt insecure, so they leave as a follower of the ‘innovation’ of the ‘norm’: precautionary evacuation. This occurred in the desktop exercise in an isolated settlement in March 2007, detailed in Sect. “[Experiences and Lessons – Some Case Studies](#)”.

It is now difficult to understand why people needed prompting to take evasive action against the forecast Brisbane floods in 1974, but many ignored the prompts. Authorities believed houses and roads would be flooded, so people should finalize travel or evacuation early [10]. For those in the flood zone, 88% of a later survey sample reported evacuating their home. Some took this step very early, but 67% of respondents only made preparations immediately before leaving home. Twelve percent only left on foot or in boats *after waters entered the main living areas of their homes*. Almost 22% of respondents said they made no preparations, mainly because the threat was not recognized in time [10]. This well documented ‘poor’ crisis communication can help calibrate formula 1.

A key safety message surrounding floods is not to enter flood waters. Figure 2 is one example of using images to help convey the reality of a threat, and the pitfalls of belated action.

Effectively Conveying Meaning

Because language is pivotal to acceptance of risk and conveyed warnings of need for evacuations, it is important that all participants reasonably agree on word meanings. This section starts in the comfort zone of simple definitions, then considers semiotics, considers core issues of world views (paradigms), and finishes with the uneasy realities of our knowledge base, our epistemological orientation. Sect. “[Integrating Theory and Implementation](#)” shows that some cultures do not carry the background cultural experience to absorb the meaning of cyclone or bushfire – there is no shared experience or ‘stories’ of the power of these extreme forces of nature. Some of the underlying knowledge foundations of context, intent and behavioral motivation need to be considered – how humans construct, transfer, acquire and use knowledge.

Imparting meteorological knowledge, then warnings, to target audiences to engender safety-oriented responses is a complex exercise in social empowerment, explored in Sect. “[Experiences and Lessons – Some Case Studies](#)”. As information promulgators, information and warning



Evacuation as a Communication and Social Phenomenon, Figure 2

What flood-threatened residents need to consider (Photos courtesy Townsville City Council)

sources should understand a little of *Perception* (the raw data from the outside world entering an organism via any of the five senses), *Cognition* (internal processing, analyzing, information storage and processing), *Attitudes* (how we think and feel about particular issues, implying a predisposition to specific action), *Language use* and links to *Behavior*.

An intellectual framework to risk communication is provided by Rohrmann [75]. It considers ‘the message features’: the recipient features, social influences and context which influences individual risk assessment and management, including preventative action. Rohrmann and Handmer’s publications [41,42,43,75] inform the Communication Safety Triangle and *the seven steps to community safety* provided in this article.

Philosophy for Policy Review: Crying Wolf or Worse – Applying the Precautionary Principle

From the 1990s a strong issue of debate in risk communication has been “the right to know” [4]. Some disaster managers wish to avoid false alarms, which may cause ‘concern fatigue’. This can be seen as an institutional barrier to change (Sect. “[Institutional Barriers to Greater Community Self-Help](#)”). ‘Avoiding undue alarm’ is in conflict with the right to know, and the precautionary principle of ESD.

The ‘precautionary’ approach is supported by the Economic Commission for Asia and the Pacific (ECAP), the World Meteorological Organization and the Red Cross Societies. The alerting of the community and its responsible authorities must begin, at least provisionally, as soon as the existence of a tropical cyclone over the seas bordering the country is known ([24] p 16). According to ECAP et al. [24], the warning challenge is less clear for predicted localized downpours and flash flooding – how much effort should be taken to warn – what is the message, how do you

keep it to the affected area, and what do you want people to do? These questions resonated in Australia after billion dollar hail damage in Sydney in April 1999, or damaging flash floods in Melbourne in December 2003.

Precautionary Evacuations

Handmer [43] reports an evacuation of 250,000 Dutch ahead of a flood threat in 1995. Eighty-eight percent of people surveyed in broad post-emergency surveys “believe that evacuation was appropriate.” [43] p 24. In part this may be because of floods experienced two years prior. Good skills in dealing with the mass media appear to have helped in the effective precautionary evacuation. The Dutch experience showed a willingness to evacuate again in future, even though the threatening flood waters did not inundate to the level feared. This compares favorably with the 2005 boat owners’ responses to Cyclone Ingrid in Port Douglas, North Queensland (Sect. “[Institutional Barriers to Greater Community Self-Help](#)”). Both Handmer and Goudie’s research [33,36,38,41,42,43,52,53,54] show clearly that people would rather practice (make a precautionary evacuation) than incur loss. It was treated as a learning experience. The ‘boy who cried wolf’ argument is not acceptable. This is an important message researchers can test and convey to partnering Disaster Managers.

Have Clear, Consistent Messages

Part of the *seven steps to community safety* is clear, reliable, explicit languages and images. Salter et al. [78] point out that the use of meteorological category systems such as ‘minor’, ‘moderate’ or ‘major’ carry unambiguous information about the level of disruption likely from a particular flood. Language used should not be for the convenience of the warning agencies. Its function is to convey clear unambiguous messages to the threatened public.

Use Past Events to ‘Make the Threat Real’

The purpose of risk communication is to make people perceive the threat as real, and to successfully motivate safety-oriented action. Boughton (1992, p 6) argues that awareness of hazards and disasters can be fostered by “drawing attention to media coverage of hazards in other places”. Images of large scale floods and evacuation in Holland in 1995 [42] may help prompt a future flood evacuation.

The Changing Politics of Risk Communication

Risk communication is often laden with values and political implications (Sect. “[Institutional Barriers to Greater Community Self-Help](#)”). For instance, in the 1990s in Cairns, north Queensland, it was argued that the reason for not having detailed local cyclone surge inundation maps made available at the corner store level was that such information may have a negative impact on local land prices. This appeared more a political decision than an attempt at effective risk communication. In 2007, such maps are freely available on the local Council web site [9].

Sirens or Not

No or short lead time disasters are a powerful argument for warning sirens to alert people, perhaps at 2 am, to “listen to local media now”. Lives will be saved in future tsunamis and bush fires with the reintroduction of sirens, as insistent triggers to “find out more now”. However, “large numbers of sirens are needed to cover populated areas and to be loud enough to be heard indoors by most people. Sirens are expensive to install and maintain and can only provide limited information” [63] p 33. Fixed public address systems or those on vehicles may be used. People who hear such sirens will be encouraged to tune to local media, and to phone others in the threat zone, to warn them of the alert. Sorenson and Mileti [81] believe sirens are most effective if used on populations without other ways of receiving the warning. Wider use would appear prudent, especially with short lead-time threats.

Media Roles

As community media moves from ‘sensationalism’ [13, 71], 2007 research by Goudie confirms that community media organizations now want to become responsible conveyors of safety-oriented information to people at risk. After the Canberra fires of 2003, where 300 homes and four lives were lost in the nation’s ‘bush capital’, all local media signed binding agreements with emergency authorities to faithfully convey provided safety information.

The Communication Safety Triangle envisages local media and householders drawing detailed local threat information from the internet, with media conveying that directly to readers, watchers and listeners. This forms the basis of future ‘world best practice’ risk communication. In the preparation for threat impact, the reliable information will help people make informed decisions, rather than remaining trapped and inactive in uncertainty.

Risk communication is complex, involving many values, predispositions and distorting lenses. Rohrmann’s [75] fine risk communication explanations show that we may tell the people at risk, but they may not interpret as intended. Clear, consistent warnings in plain English, with clear images of the threat, showing safety-oriented behavior are needed from reliable sources. Warnings, seen on a continuum of risk and preparation actions, should be able to be discussed and reinforced with information from other sources. This is most likely to produce safety-oriented behavior, with the constrained and clear assistance of the media, as responsible agents for community safety.

Given the previously fraught nature of risk communication, Sect. “[Institutional Barriers to Greater Community Self-Help](#)” provides a conceptual framework, with Australian examples of current community safety theory and implementation, preceding some detailed examples.

Integrating Theory and Implementation

This section introduces a triangular model to best ensure community safety. With disaster managers central, the three elements are the community, local media and the internet. Also, there is a continuous spectrum of seven steps, from accepting there is a risk, through early preparation, final (ramp up) preparation, which may include evacuation, surviving the hazard impact and achieving smooth recovery. An exploration of motivation leads to a discussion of ‘world view’, including some insights into remote Indigenous communities [36], and recently arrived, non-English speaking immigrants. The conceptual frame and steps are intended to help guide risk communicators and potential modelers through the issues of acceptance, preparations, evacuation and functional recovery. Within CSS, all these factors have some bearing on consequent behavior.

Changing Values and Roles

In an attempt to understand why we support or ignore certain messages relevant to our safety, and thus facilitate modeling, this section considers the Dominant Social Paradigm and the New Environmental Paradigm.

A paradigm is a coherent world view, “a mental image of social reality that guides expectations in a society” ([23] p 10, [29]). Shared paradigms change. The underlying philosophy shift toward disaster mitigation generates a distinct policy move toward safe, self-helping communities. Thus, the goal of disaster managers, effective risk communicators and computer modelers becomes one of championing or demonstrating the efficacy of self-help techniques and information to and within communities.

World Views Which Accept Responsibilities of Living in a Hazard Zone

‘Why do we do what we do?’ has long been a central exploration of psychology. Precautionary safety behavior, including self-evacuations, may depend on a person’s world view. Stern et al. [83] developed a simple and elegant model to help explain human behavior, starting with a person’s position in a social structure, with constraints and incentives which generate values. Values determine general beliefs, leading to a consistent world view, specific beliefs and attitudes, which predisposes intent and helps explain behavior ([83] Fig. 3). Loosely translated, behavior results from a linked sequence starting with: the where, when and to whom of an individual’s birth, the surrounding circumstances and ‘cultural sheath’ of their early childhood experiences, leading to acceptance or rebellion against prevailing social norms, determining an evolving world view. Once we have our coherent world view, beliefs and values give us our ‘predisposition to act’, our ‘intent’ which precedes action/behavior.

Many things now mold and modify world views, including the media, and normative world views are muta-

ble, changeable. Aligned starkly with our shared and collective biological survival urge, and much concerted efforts by scientists and conservationists for decades; publicizing by people like Al Gore, and the authority of England’s uptake of the 2006 Stern Report, there is now a global paradigm shift to the urgent need for behavioral and technological change to minimize the gross impacts of global warming, pertinent to increased disasters and needed evacuation readiness. Modeling specific hazard zones’ strengths and weaknesses and broadcasting prediction simulations to the internet and TV will help mobilize those at risk.

Why We Do What We Do

A behavioral model of causality ([91] Fig. 4) shows relationships between reported attitudes and actual behavior. However, a complex model proposed by Kitchin [51], with the strength of including social and environmental interactions shows why we do what we do: Kitchin’s model includes a person’s ‘working and long term’ memory. Internal information is processed within ‘real world’ context, such as cost [82], which may be processed within the ‘it can’t happen to me’ cognitive frame [61] of subjective reality. These complex but quantifiable attributes will contribute to modeling maximum likelihood to accept, prepare and act to maximize safety.

Learning from 18 Remote Indigenous Communities

Much of the rest of this section underlines why modelers will need to deal in great detail with ‘cultural’ aspects of massed responses to safety threats.

The Stern model (Fig. 3) helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94] helps explain why there is such a strong sense of self-help in remote Indigenous communities [27,29,80,94]. Elders decide responses to threats, there are historic and immediate constraints, generating a value system where community members need to look out for each other.

In many traditional Australian Aboriginal stories, most people drown in flood disasters, often as punishment when people do not take care of each other [36,39,64,87, 88]. If general beliefs embrace self help, including the felt need to ensure safety, and a resultant intent to achieve community safety, this should lead to safety-centered behavior.

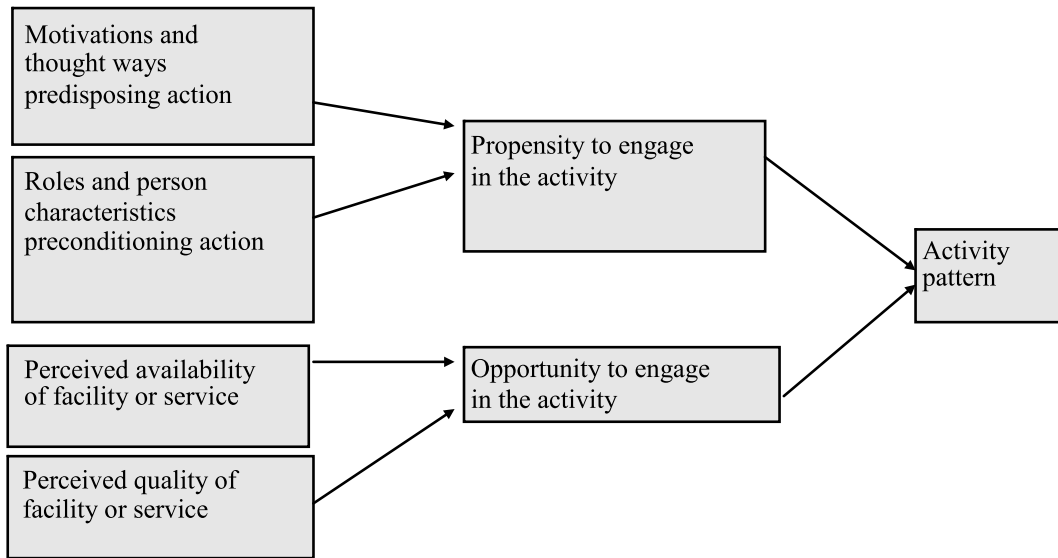
Each community ([36] Fig.5) was not greatly concerned by weather extremes (values), but each relied on and respected their traditional reading of threats, and information from the Bureau of Meteorology (the Bureau).

Behavior is explained by:

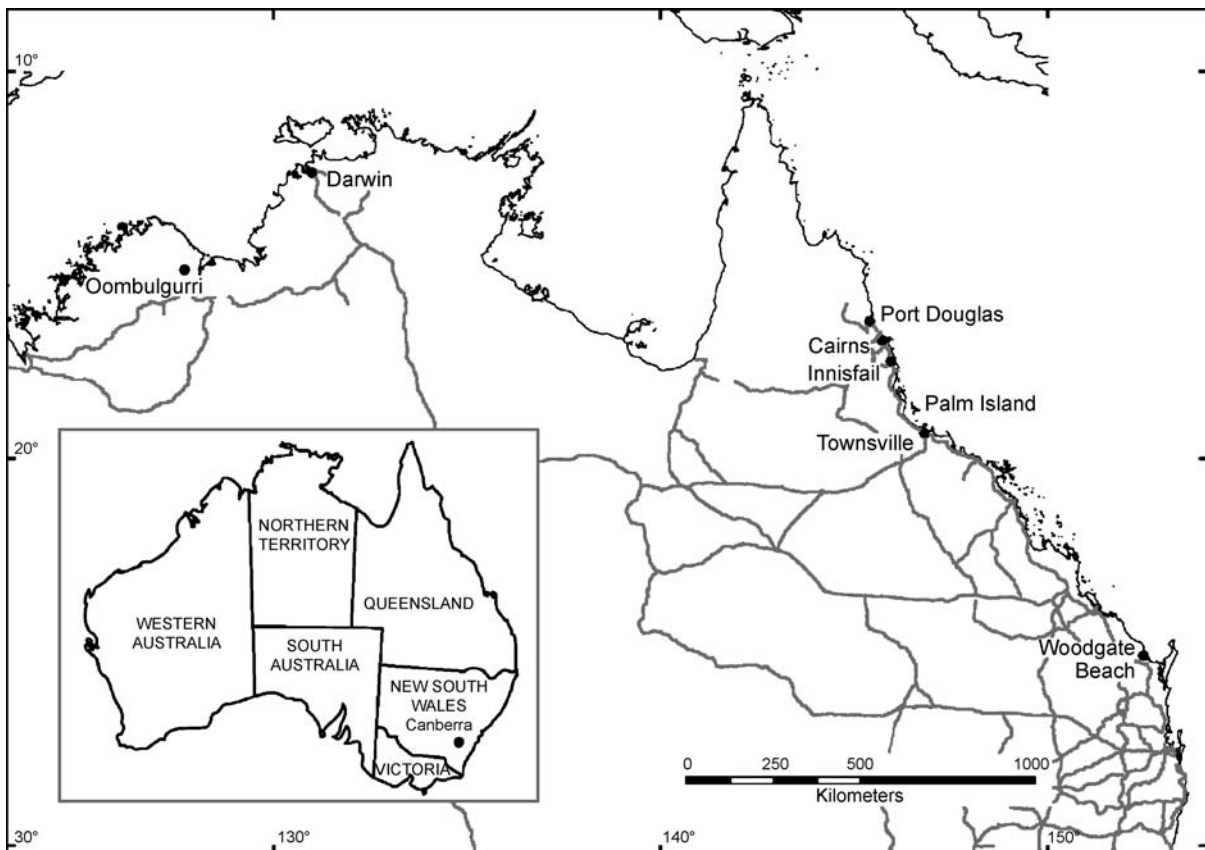
1. a person’s *position in a social structure*,
2. with *constraints and incentives* as generators of *values*, which lead to
3. *general beliefs*,
4. *world view*,
5. *specific beliefs and attitudes*, generating
6. *intent*, which helps explain
7. *Behavior*.

Evacuation as a Communication and Social Phenomenon, Figure 3

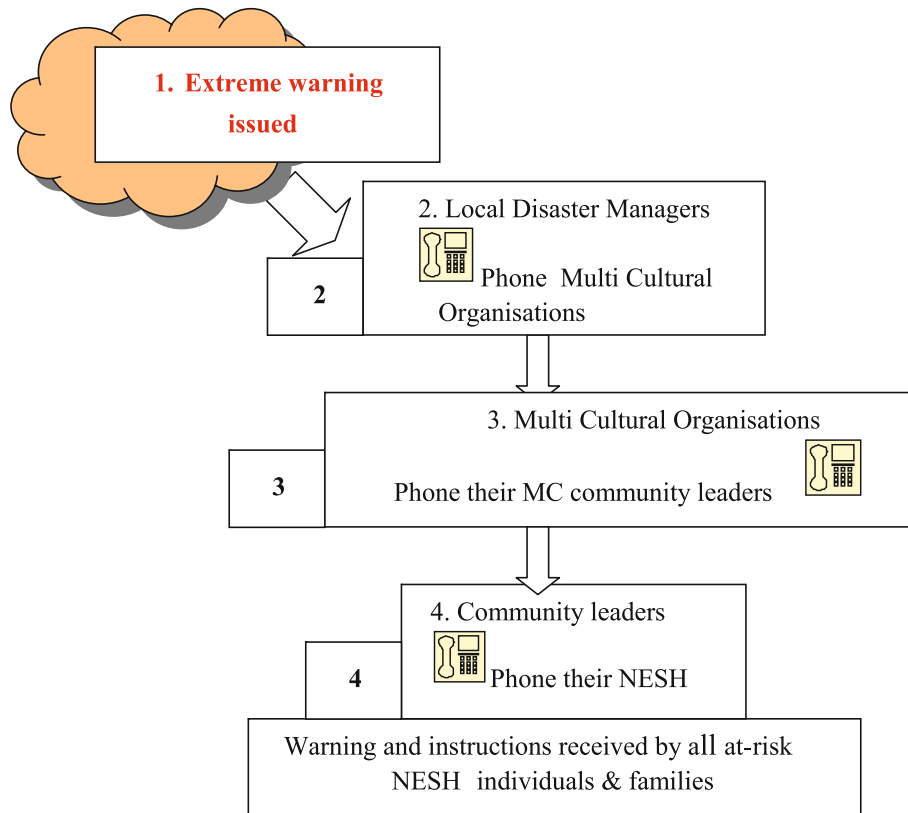
Stern’s 1995 behavioral explanation model



Evacuation as a Communication and Social Phenomenon, Figure 4
Possible determinants of activity patterns (from [91])



Evacuation as a Communication and Social Phenomenon, Figure 5
Australian study sites



Evacuation as a Communication and Social Phenomenon, Figure 6
Multicultural phone tree disaster warnings model

The world view is that flooding or worse may happen, and that the Bureau will provide adequate warning. Responses may be inhibited by the Social Vulnerability Index (SoVI) of formula 1.

Community Links, Phone Trees and the Web in Risk Communication for Non-English Speaking Households (NESH)

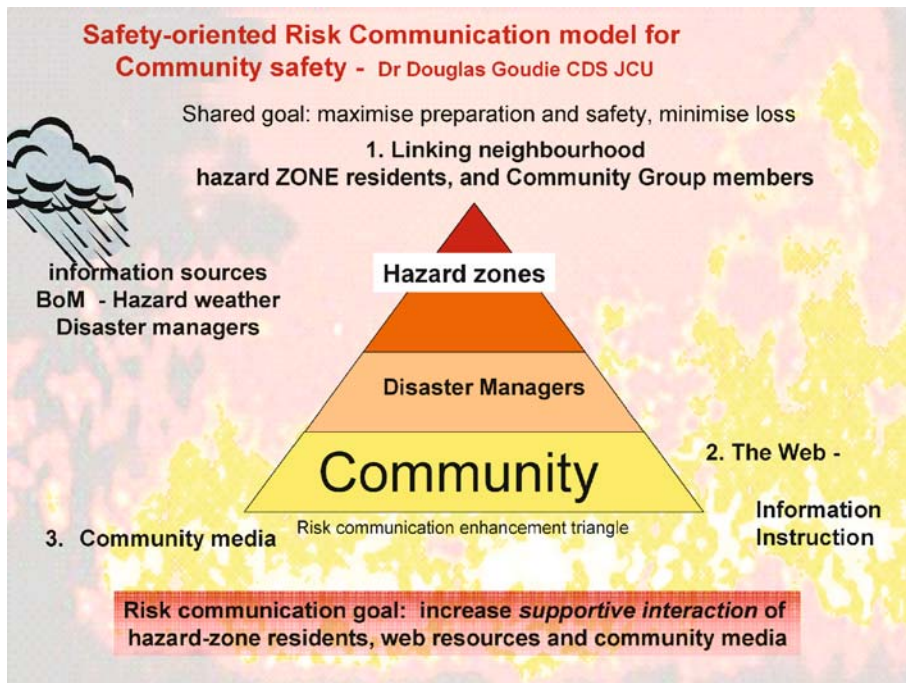
Many multicultural organizations and focus group meetings in 2005 helped develop the warnings phone tree model for NESH in Fig. 6. NESH rarely listen to the English language media. With about 30,000 NES people arriving in Australia each year [18], there is Federal government recognition of special emergency management needs [28]. This way of getting evacuation warnings through was developed once interviews revealed that practically all NESH have a mobile phone, and are closely connected with their nearest government funded Multicultural Organization and their 'community leader'; hence the phone tree. Modeling is of and for the real world, so re-

search like this reported NESH study is needed to see what 'complex systems' may be possible. Further recommendations to develop a multilingual warning web site with a simple guide to the seven steps, in relevant languages, to be accessed and used by MCO and NESH training sessions is being considered by authorities. This provides another example of the internet's latent role in effective risk communication.

The Australian Northern Territory has cyclone preparedness kits with action guides in 8 languages; the NSW Rural Fire Service has information in 27 languages. This shows the importance of modeling all possible communication avenues, encapsulated as Medium & Message: MM in the general formula for Household Preparedness and Safety Actions Index (HPSAI).

An Holistic Approach to Community Risk Management

The seven steps for effective warnings involves community networking, 'responsible' media, and, increased web



Evacuation as a Communication and Social Phenomenon, Figure 7
The Communication Safety Triangle

use. A ‘continuum’ approach to hazards which may lead to a need to evacuate is important to modelers because it is important to all people in hazard zones: they will be able to access more accurate, detailed and timely information of any looming threat, assisting Emergency Managers because a self-help public will decrease demands on formal evacuation, rescue and property protection [93]. Recovery will be less arduous [67] if hazard impact is minimized. Reducing disaster impacts will reduce costs, to national benefit. This gives strength to supporting the very demanding goal of developing the model of formula 1.

Risk Managers

Researchers can work with risk managers to achieve “... better, timely warnings and advice on safe action during fire events” [3]. With risk managers central, Fig. 7 suggests that enhancing community links [7]; web information for residents and media outlets, and cooperation of community media with fire managers [89] will more empower householders to embrace self-help in fire safety. Figure 7, the Community Safety Triangle, with the “Seven steps to community safety” (Fig. 8) provides the conceptual frame to enable maximum safety modeling.

Core of Risk Communication for Community Safety Through Natural Disasters

Building Community Links and Refining Media Delivery will Change the Household Preparedness and Safety Actions Index

The disaster safety benefits of enhanced community links fit positively with broader social policy [47]. Stronger community links will help ensure that threat information is easily accessed and shared; and the need to actively self-protect is internalized at the neighborhood level. Residents will benefit by enhanced community links (with such innovations as social burn offs and Community Safety Groups) in improving their general quality of local social interaction. Changing Community Resilience (CR of the formula) will then change overall preparedness, so pilot tests of change to CR will show quantifiable changes to the Household Preparedness and Safety Actions Index. Disaster web site managers will be providing a product which is rationalized to reduce national duplication of core information. This will positively change Medium & Message, MM, of the formula. Residents with few English language skills will benefit by having disaster information delivered,

Goal: Maximise safety and recovery, and minimise loss in hazard zones.

1. Encourage those in hazard zones to accept that the risks are real.
2. Help create an aware, informed community, predisposed to safety-oriented action, as a precaution; as a practice.
3. Encourage information-sharing and support among friends, neighbours, family.
4. Provide ‘what to do’ (action) information, via reliable sources, including web and community media delivered for background and preparation.
5. Encourage people to think right through to impact and recovery.
6. When a threat is closing in, warning messages will clearly convey: *this is real, this is coming at me. I need to make safe where I am, or move early to somewhere much safer. I will not travel during the impact period.*
7. Provide timely, effective threat warnings and fine location and forecast weather detail, and recommended local responses.

Evacuation as a Communication and Social Phenomenon, Figure 8
Seven steps to community safety

via the web, in language and concepts that can be internalized and acted upon.

Handmer [43] suggests that a flood, for instance is actually ‘owned’ by the communities at risk. Individuals and organizations within these communities actively seek out information and mobilize their personal networks for action. In this way of looking at the warning process, the warning specialists act as mediators between the threat and the threatened. Local knowledge is used and the whole response process remains focused on safety and loss-minimization. For this plausible vision to become fact, local capacity building will need to proceed apace with these more formal efforts to inform and maximize community safety.

The importance of ‘informal’ information sources and community links are shown in Table 2, from about 200 people interviewed immediately after Cyclone Larry in 2006 [53], showing that about 90% are deeply dependent on social and family support, if only for reassurance.

Finally for community enhancement within the CST model, there is current and lucid national disaster mitigation policy [13] <http://www.dotars.gov.au/localgovt/ndr/>

Evacuation as a Communication and Social Phenomenon, Table 2

Cyclone zone people have much personal contact

Contact with other relatives	% of total response (rounded)
Yes	25
Lots	20
Mobile contact	15
Landline Phone	30
No	10
Frequency of neighbor contact	
Often or lots	50
A bit	20
Helped/contacted during eye	15
None	15

[nat_disaster_report/naturaldis.pdf](#): in the statement of the paradigm shift [13], p 13, “Principle 7 – Reform Commitment 7: develop jointly improved national practices in community awareness, education, and warnings which can be tailored to suit State, Territory and local circumstances.”

Community Media

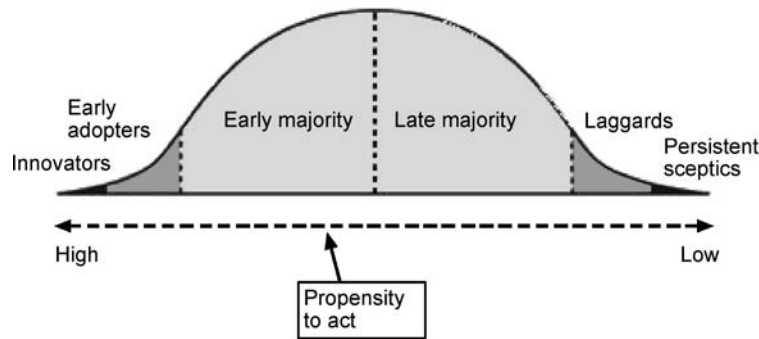
“The Media: Media organizations, particularly public and private radio and television organizations, have responsibilities in ensuring that timely and appropriate warnings and advice on disasters is broadcast to communities at the request of relevant authorities. They also have a role to play in educating the community about natural disaster issues” ([13] p. 18).

The Web

Rationalized web information delivery: One national generic provider and up keeper of core information, known to all, is recommended; to which state and local government will add their unique detail. A rationalized disaster information web delivery can properly incorporate broad contents, fine (and zoom-in findable) and multi lingual information.

Simulations of predicted near-term fire or other threat movement based on the above, akin to micro scale modeling of cyclone impact forecasts will help revolutionize the way people react to fire and flood threat. With about one in three households having the internet, prior to likely power loss residents can draw on enhanced fine detail of the current fire situation, to assist in their actual decision making. Before Cyclone Larry struck, residents with web access to the Bureau copied cyclone forecast maps and distributed

The stages of change from the Diffusion of Innovations theory. The vortex shape has no mathematical basis. Its purpose is to illustrate the point.



Evacuation as a Communication and Social Phenomenon, Figure 9
O'Neil's 2004 innovation uptake model

them to neighbors [53]. All the preceding is encapsulated in Figs. 8 and 9.

Images Tell a Story

Images such as cyclone forecast track maps in risk communication convey more information than words alone [36,53]. With this feedback from real residents, it becomes clear that linking fire zone residents, for instance, to educative web fire images becomes an important goal (e.g. [100]).

Community Story Telling Bonds Neighbors

The multi-pronged approach [33,52,54] fosters community 'story-telling' links [50,59,90] relating to community self-help and safety. Strengthening community bonds, working with enhanced web resources and community media [14,72] will increase community internalization of risk [31], enhancing the likelihood of safety-oriented responses, leading to and possibly including evacuation. People need to accept the reality of the threat, indeed, feel some anxiety about the threat to help drive the intent to seek more information, or the intent to prepare [68].

With easier access to relevant web information; with greater detail of current fire behavior and *nowcasts*; fire zone residents and community radio announcers can describe the looming threat, helping timely preparations and the monumental 'stay or go' decision. Community radio, like the National public radio, the ABC, will deliver authoritative and timely risk communication directly from the refined web information.

Theory and emerging practice converge on using refined web-delivered material to households and their neighbors, and to local media, to inform and motivate res-

idents through the whole continuum of the seven steps to community safety. The next section considers institutional barriers to change, followed by extensive research findings and lessons from Australia in disaster management to develop effective warnings and self evacuations.

Institutional Barriers to Greater Community Self-Help

Institutional barriers to change take many forms, from 'unconsciously' avoiding consideration of the extreme event as 'too hard', to a misuse of the power relationships within bureaucracies because of fear of change, or malice. Clear examples of the 'too hard syndrome' mingled with entrenched vested interests has been denial of links between smoking and cancer, or delayed uptake of global warming mitigation.

There were Institutional Barriers for disaster managers to consider land-based flooding from torrential rains preceding a cyclone (hurricane/typhoon) [34] in many cyclone-prone areas. The Queensland State Planning policy [70] now explicitly refers to Probable Maximum Flood. Including maximum flooding is now part of emergency manager's internal planning base. This means the Exit Routes (ER, formula 1) are now considered in planned evacuations.

Since the early 90s researchers like Boughton [6] have suggested having drills for schools and other institutions in readiness for possible earthquakes, cyclones or other hazards. This author supports precautionary evacuation practices. There are, of course, liability barriers to easily undertaking practice evacuations. There is also uncertainty of threat, which may restrain some risk managers [86].

Institutions have cultures which may passively express antipathy to a paradigm shift toward sustainability, whilst being required to usher in sustainability [19,37]. Emergency Management Australia, the peak national body, has the slogan: *Safe, sustainable communities*. However, there is a multitude of conservative forces representing the Dominant Social Paradigm restricting innovation, despite sustainability policy. People engaging in sustainability implementation may come into conflict with institutional representatives of the old paradigm.

With most development taking place in urban settings, concepts of urban sustainability attempt to merge two different fields of human endeavor – how we modify our landscape, our built environment, and how we behave in that environment. Disaster mitigation is obliged to fuse these two seemingly disparate fields. There may be resistance to that. That resistance needs to be included in any mass disaster movement modeling. The following subsections indicate some reasons for IBs which modelers may need to consider in the IB factor.

Political Insecurity, Real Estate Values, or Undue Alarm

The tension between people's right to know, government duty of care and politicians' perceptions of probability of risk on 'their shift' form a complex interplay. If maximizing safety is the shared goal, all risk communication theory suggests people will behave appropriately with the realistic threat information, motivation and 'how to' instruction on safety-oriented behavior leading into, through and recovering from a natural disaster impact.

Paradigms of Politics: the Real Estate Industry and Vested Interests

Broughton [6] argues that if people know of the threats, they are likely to support politicians who make sound decisions for community survival. Some plans are made but they are hidden away for various financial and political reasons. This is a case in which attitudinal changes on the part of those communities may change the priorities of the decision makers and promote the interests of the community.

Those most empowered to assist in implementing projects that require independence, initiative, local links and knowledge may be the ones who prove most obstructive to innovation [1,48,56,69].

Action Research [16] sets out to research, develop and document socially and environmentally desired outcomes within sustainability principles. In this inclusive and 'engagement-focused' research, gaining support from the top

is crucial to the success of any organizations' efforts at 'social change' [66]. If innovation uptake is viewed as a normative uptake process [12,65] and Fig. 9, then the gradual, long term paradigm shift to sustainability [30] is plausible and has a conceptual frame.

Uptake of the Communication Safety Triangle and the seven steps to community safety (Figs. 7 and 8) will make them the disaster managers' 'norm' over time, simply because that is the direction of social evolution in disaster management. It fits social policy and advanced disaster management approaches, and the technology and willingness of regional media and residents is there. All that is needed is the testing and roll-out by more innovative disaster managers. The same applies for Community Safety Groups and variants of *over-managed, locally controlled near-house "social burn-offs"* in the case of fire risk management.

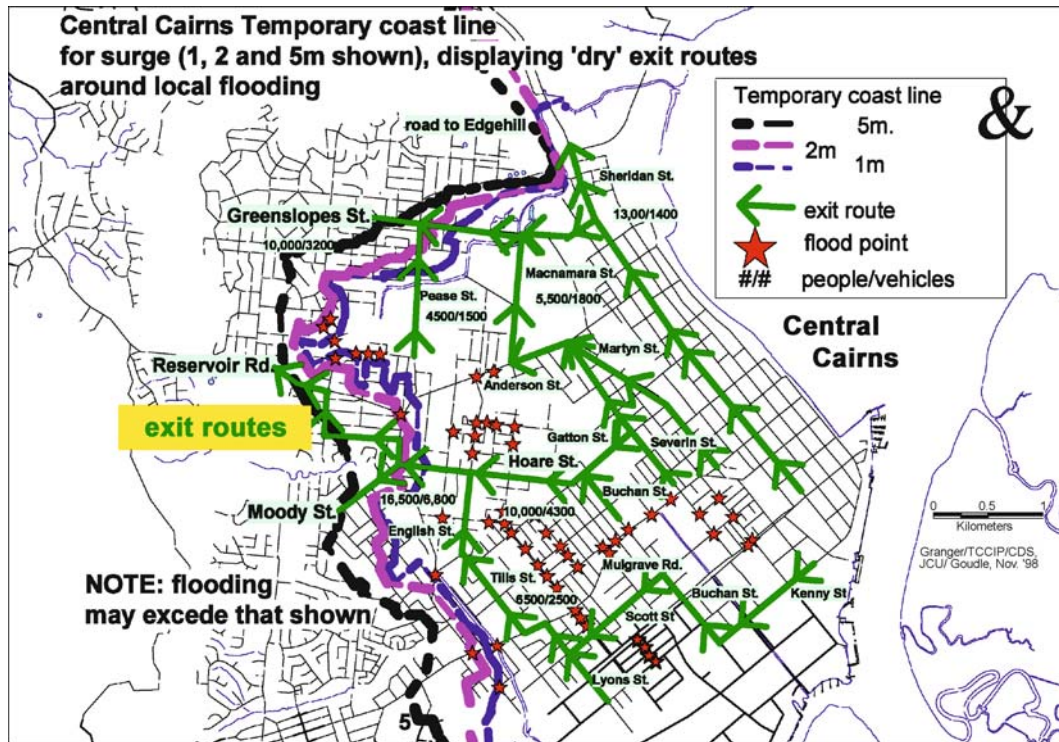
Bureaucratic processes need to support the stated goals of their own work units, but individuals may be ambiguous, contradictory, belated, bullying, ill informed and, due perhaps to time pressure or arrogance, quite destructive. Such individuals may be 'corporate psychopaths' (<http://www.abc.net.au/catalyst/stories/s1360571.htm>). Instead, innovators may be seen as threatening; troublesome. Any innovative pilot project must develop strategies that will maximize change within their unique situation. Ideally, that happens within a supportive parent organization. The next section discusses the emergent issues of effective risk communication and precautionary self-evacuations.

Experiences and Lessons – Some Case Studies

This section considers disaster impacts and impacts from north Queensland and from fire zones in SE Australia to illustrate the generally positive points of approaches already outlined: living flesh for modelers to understand how subtle, complex but do-able successful community preparedness and willingness to act for maximum safety can be. Researching these events, gaining community and disaster managers feedback on risk communication; working closely with the Australian Bureau of Meteorology over 15 years, through the Federal disaster 'paradigm shift' to mitigation ([13] p 13) all inform the emergent CST and seven steps to community safety.

Cairns and Storm Surge Considering Exit Routes (ER, Formula 1)

Cairns City, North Queensland, is centered on land less than 2 m above high tide, and subject to cyclone (storm) surge of up to 5 m. A storm surge tracks just behind the eye of a cyclone, a low mound of sea water, perhaps 50 km



Evacuation as a Communication and Social Phenomenon, Figure 10

Cairns flood prone evacuation routes and schemata of wave reach and temporary coast line from 1998 information

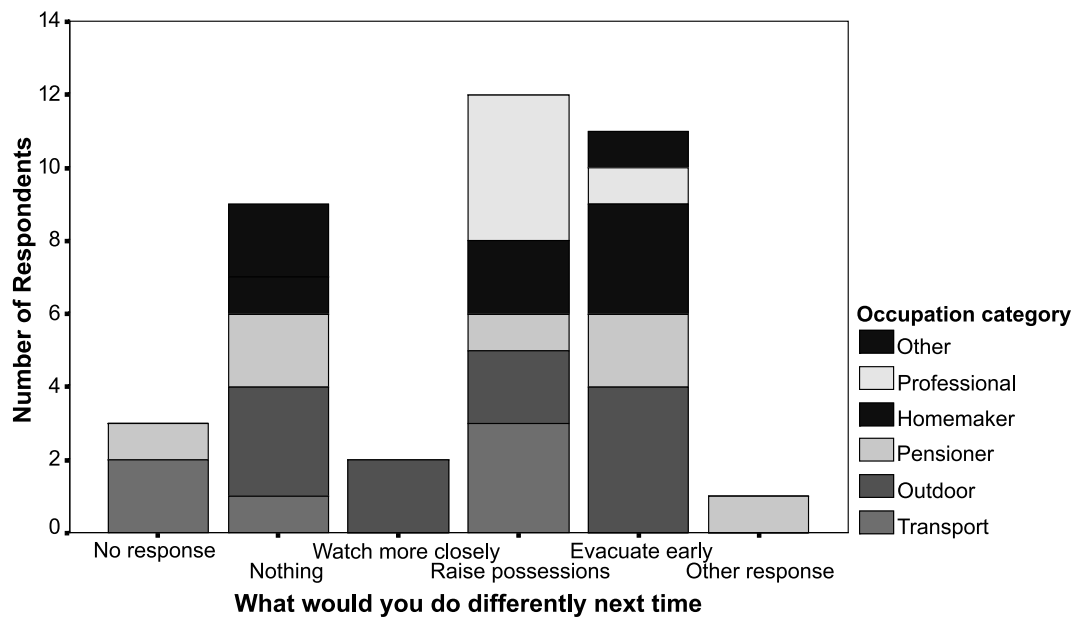
wide and up to 3 m above normal sea level. It may flow overland for perhaps 4 h, as destructive winds to 280 km/h tear at structures, and churning seas; both laden with pounding debris, behave as battering rams and missiles. Because cyclones have such long lead times from modern electronic detection to landfall, results of this deadly combination [60] will be widespread destruction and loss of infrastructure, but not life, with precautionary evacuations [38,72]. At the macro-observation scale, people and vehicles in motion within confining environments tend to behave like fluids, capable of modeling. Constrictions of flow paths cause congestion, as described by Helbing et al., (► *Pedestrian, Crowd and Evacuation Dynamics*), so early and precautionary evacuation will help minimize the likelihood of grid-lock.

Land-based flooding may be a core issue to effective evacuation [34,38,95], with up to 40,000 people in the vulnerable central city area and northern beach suburbs needing early, precautionary evacuation [79], Fig. 10. Cairns City Council now has a storm surge map on its public web site ([101], posted in 2006), like the public and informative flood map for the Redlands Shire, SE of Brisbane, Queensland ([102]).

Within the philosophical and moral frame of people's "right to know", these local governments are using the internet to inform people they are in a hazard zone, the first step in making the threat real to those residents. They have made the paradigm shift to providing fine-detailed background information which says to residents: "you are in a hazard zone, you may need to do things. Listen for warnings and be prepared to act in a precautionary way."

Figure 11 shows that after a devastating flood in 1997, the flood-affected residents of Cloncurry, NW Queensland town would do things differently, with better warnings, when faced with rising flood waters. Remote automatic flood monitoring devices were requested, but the local downpour over a fully flooded, vast and flat landscape appeared to be the cause of the flood rising 2 m higher than any flood in the prior hundred years. Precautionary evacuation of the low-lying homes would have prevented much heartache and loss of valued possessions [52].

In 2005, north Queensland was threatened by cyclone Ingrid. Newspaper-reading residents were left in no doubt about the threat (Fig. 12); a good example of the clear warning role played by the media, and non-language image used to convey meaning (Sect. "Discussion").



Evacuation as a Communication and Social Phenomenon, Figure 11
Cloncurry reflective of likely responses to future floods

A portion of a Weather Bureau warning media bulletin for Cyclone Ingrid follows, showing, verbatim, what the nation radio broadcaster, the Indigenous Radio Network and most responsible media outlet relayed on. This is clear information, including the possible impacts (M& M of formula 1).

Media: For immediate broadcast. Transmitters in the area Cape Grenville to Cooktown are requested to use the Standard Emergency Warning Signal.

TOP PRIORITY

TROPICAL CYCLONE ADVICE NUMBER 14

Issued by the Bureau of Meteorology, Brisbane

Issued at 10:56 am on Wednesday the 9th of March 2005

A Cyclone WARNING is current for communities between Cape Grenville and

Cooktown. The warning extends inland across central Cape York Peninsula.

A Cyclone WATCH is current for coastal and island communities on the eastern Gulf of Carpentaria between Weipa and Kowanyama.

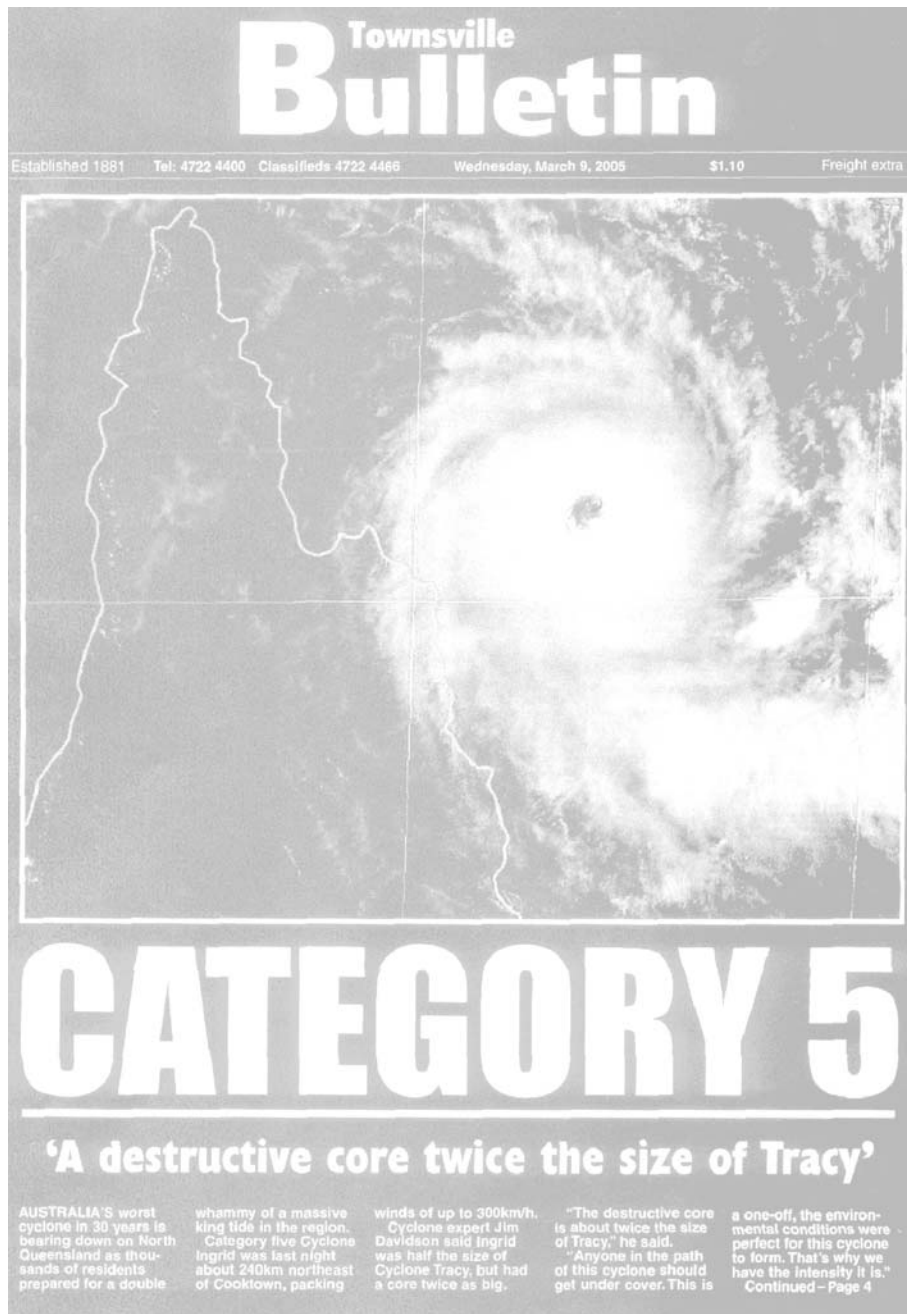
The watch south to the Gilbert River Mouth has been canceled.

At 10:00 am EST SEVERE TROPICAL CYCLONE Ingrid, Category 4, with central pressure 935 hPa, was located near latitude 13.5 south longitude 145.5 east, which is about 140 km northeast of Cape Melville and 260 km east of Coen. The cyclone was moving westward at 11 km/h.

Severe Tropical Cyclone Ingrid poses a serious threat to the far north.

Queensland coast with very destructive wind gusts to 280 km/hr near the center. Gales are expected to develop between Cape Grenville and Cooktown during the afternoon. Destructive winds are expected between Coen and Cape Flattery overnight. The very destructive core of the cyclone is expected near the coast between Coen and Cape Melville on Thursday morning.

Coastal residents between Coen and Cape Flattery are specifically warned of the dangerous storm tide as the cyclone crosses the coast early Thursday. The sea is likely to rise steadily to a level significantly above the highest tides of the year with damaging waves, strong currents and flooding of low-lying areas extending some way inland. People living in areas likely to be affected by this flooding should be prepared to evacuate if advised to do so.

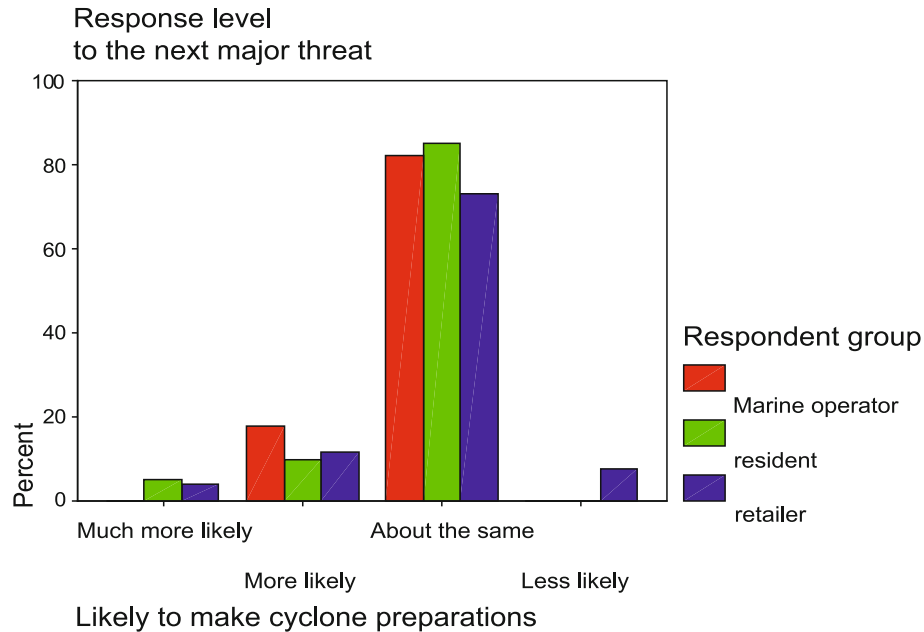


Evacuation as a Communication and Social Phenomenon, Figure 12
Media coverage of threatening Cyclone Ingrid, 2005

Very heavy rain can be expected to develop on the coast and ranges north of Cooktown.

Research in Port Douglas, NQ was conducted immediately following Cyclone Ingrid, to help clarify the 'Boy

who cried wolf' hypothesis about 'concern fatigue' over precautionary evacuations (Sect. "[Institutional Barriers to Greater Community Self-Help](#)"). Feedback from marine tourist operators and tourist businesses that removed about 60 large vessels to safe, up-creek moorings, or fully



Evacuation as a Communication and Social Phenomenon, Figure 13
People of Port Douglas appreciated a precautionary evacuation

shuttered their premises and raised all stock. These preparations were arduous, and costly. The cyclone veered away from Port Douglas, with virtually no impact. Despite that, people were glad of the precautionary evacuation, as a practice (Fig. 13), reinforcing the findings of the Netherlands flood study [43]. People appreciate practice, as part of the new paradigm of a precautionary approach to hazards.

From Those Impacted by Larry, 2006

After Cyclone Larry, a powerful Category 4/5 severe cyclone which damaged the Innisfail region, Goudie led a social science research team into the impact area, interviewing households for 4 days (King et al. [53,54]), from the 150 householders survey, focused on risk communication, we learned: as with fire zone residents and those who experienced Cyclone Larry, that most people, once they have experienced a major disaster, maintain a healthy respect and inclination to act ahead of any future such threat. Hence the advocated merit of encouraging people with experience to 'tell their story'; and community (Media and Message and Community Resilience of formula 1) to help make the threat real to others. The next section is, perhaps, at the leading edge of where disaster management will go: living community self-help. Such communities can provide pilot locations to 'calibrate' formula 1.

Woodgate Beach and Community Safety Groups (CSGs)

This section reports a collaborative process between the author and many others, mainly residents of Woodgate Beach, Queensland (Fig. 5), to develop a *Preparation and Evacuation Plan* for residents, formal response groups, the Local Government Disaster Management Group (LDMG) and Shire Council.

Through the 18 month consultation and development process, the plan needed to be realistic and achievable, relying mainly on locals working together under the LDMG and local SES group, aiming to optimize each element of formula 1.

Community networking and self-responsibility for community safety is profoundly developed in the isolated, 700 strong, Central-coast Queensland settlement. The volunteer community safety groups were made up of dedicated people. Goudie led development of a community-based evacuation plan, aiming to: ensure maximum preparation and minimum impact on property and residents; optimizing formula 1.

There were 5 meetings, of up to 80 residents and representatives of all formal response groups, weather and earthquake experts. Meetings included researcher-led two-day 'table-top' evacuation exercise. Woodgate Beach did not develop an evacuation plan, but a preparations and evacuation plan:

Preparations and Evacuation Self-Help Approach

To nurture aware, informed residents preparing to be safe through, and recover from natural disaster impacts.

Overarching Evacuation Approach

- Identify vulnerable areas or houses
- Evacuate caravan park, visitors, people at risk unable to easily move themselves

The Planning Process:

1. DEFINE THE THREATS – Fire, flood, wind, cyclone surge, earth movement, tsunami.
2. MOVE FROM WHERE TO WHERE? Fires can be fought; floods, storm surge and cyclonic winds, tsunami or earthquake cannot be. In all threats, make sure your property is as secure from impact as possible.
3. IDENTIFY THE VULNERABLE – Which buildings, infrastructure and people may be in the threat zone.

Threats and Treatments An all-of-community approach has developed an annual round of public education projects, press releases and pamphlets to inform residents and tourists of the annual cycle of dangers to be wary of, from cyclones to fire care and management. The Council newsletter will be a consistent source of information on threats and how to minimize risks (M & M), from the flying debris of a wind storm to being patient at a flood-swollen creek crossing.

A sequenced approach, where the aged and vulnerable are moved first from the highest risk areas will be taken, as a practice, as a precaution. To highlight the stages of disaster and possible evacuation planning, the background preparations phase is included in this article:

Background Preparations

1. Woodgate Beach residents recognize and act on the need for background preparations to minimize the impacts of all hazard impacts. This includes property maintenance and upkeep. This maximizes PPR of formula 1.
2. Provide newcomers with an information pack, including a copy of extracts of this Community preparations and evacuation plan (CPEP). All community members, including tourists, the elderly, infirm, and needy are incorporated into this CPEP. This enhances the KB of formula 1.
3. Provide dot points on evacuation for local residents in the *Disaster Preparedness Information Kit*, delivered to each household.

4. Expose tourists to the essence of local threats and what will be expected of them: leave early, unless they or their vehicles can actively help, under direction.
5. Define safe shelters – preferably with friends or compatible households. Organizing possible billets for any major impact on portions or all of Woodgate Beach can form a key function of the Community Safety Group. This is the CR of formula 1.
6. Go through whole plan, and address matters like the caravan park needing auxiliary power for fuel pumping before the cyclone season.

The Community Safety Group Approach to Disaster Management This approach is detailed to provide a guide for researchers to use as a framework for other communities:

The Community Safety Group

Purpose encourage:

1. Early Warning Alert The CSG is an affiliation of existing community groups and neighborhood-level residents who make first contact with ‘walking-distance’ neighbors as soon as anyone hears of a warning that a natural disaster may be approaching their area.

2. Final Preparations (Ramp-Up) Activation The neighbor-level CSG will provide early local motivation for final safety preparations.

Recovery – Thinking Through to Recovery (TR) of formula 1.

Recovery is now seen as part of the preparations package, rather than just looking to minimizing impact. The developing approach is to see the whole threat event as one continuous process: from awareness and structural preparedness, through initial communication of threat, to final precautionary preparations and impact and rapid recovery to a fully functional community.

International Snaps

The Center for Disaster Development within The Northumbria University, Newcastle on Tyne, coastal north east England specializes in recovery and response with an emphasis on development long term recovery and resilience-building. Embedded in this approach is to undertake mitigation. Interviewed by Goudie in May 2005, the Director reported “We use the approach that local knowledge is always drawn on, and that people involved should be in control”, KB and CR of formula (1) and agrees “all disaster management is under the umbrella of sustainable development”. Other interviewed staff support precautionary

evacuation as a practice, and, independent of the media, an alert signal for people to tune in to the media.

The Shetland Islands

Like Australia's emerging approach to self-help communities, The Shetland Council's Emergency Management Planner (2005 interview) gave strong practical support to the approach of an informed, aware, self-activating community ready to act on reliable, clear, how-to emergency warnings, ranging from making safe if intrinsically out of the main impact zones, to early 'self-evacuation'. This convergence from differently evolved and slightly different disaster management systems is grounds for optimism that the evacuation approach and formula expounded in this article has widely applicable merit.

Remote communities like the Shetlands have lessons for the mainstream in taking responsibility and perforce being self-reliant and oriented to robust self-help. Such isolated communities represent matured examples of the informed, aware communities, predisposed to precautionary action that mainstream populations now aspire to. Community-building is an international aspiration, achievable in urban settings.

Hurricane Katrina

Hurricane Katrina, USA late August, 2005 (Fig. 14) has deliberately not been included in this evacuation analy-

sis. With days of clear warning, general formal approaches to precautionary action which underpin the CST and the seven steps to safety and recovery were underplayed.

Figure 14 is included to remind all readers that hazards are real threats to real people in real hazard zones. If all the parameters to maximizing formula 1, such as Institutional Barriers are not optimal to safety, great distress and loss can ensue.

This section has provided a cross-section of disaster threats, evolving reasons to empower communities, and some factors to include in modeling greater community safety. After discussion in the following section, some recommendations and future directions are provided.

Modeling not only the natural hazard but the various social and communication parameters will provide a good basis for not only simulating impact; say, of the extent of flood waters; but will also highlight which impact areas are the most and least likely to properly act in their own best safety.

Discussion

Risk Communication Theory and Residents in Hazard Zones

People need to know that an impact is possible before they will willingly evacuate. The concept of risk characterization [42,46,68,75,77,83] makes clear that people need a practical understanding (*the possibility of impact is real,*



Evacuation as a Communication and Social Phenomenon, Figure 14

What no-one should have to stay through – Katrina '05 http://news.bbc.co.uk/1/hi/in_pictures/4194032.stm

I fully accept that) to then illuminate practical choices. ‘Internalizing’ that *a threat is real and what you need to do to maximize your safety* may be the core goal of risk communication. This internalization and safety-oriented action may lead to a choice of precautionary evacuations. People need to make informed decisions, thus leading to decision-driven activity.

Delivering Real-Time Warnings

The idea of subjective uncertainty [74] is well displayed by fire-zone residents who, despite all authority efforts to have them commit to an early decision to stay and defend or to leave their properties early, recurrently reported that they would decide on the day whether they would stay or go. There are tragic and recent international examples of resident’s decisions made on too little understanding or information.

The bushfire ‘stay or go’ decision point explored in 2007 research by the author with Australian fire-zone residents clarifies that decision impact on those under threat – packing up their valuables, children and pets and fleeing their house, almost certainly putting it at risk of burning down; as opposed to staying in their house and experiencing the terror of the developing bushfire, was a decision many preferred to delay. Further, many acknowledged that to leave early with all of the disruption, only to learn that the fire was not an actual threat to their property helped induces people to delay that decision point.

Subjective uncertainty is a psychological problem. The outcome of current research is to encourage the threat information gatherers and providers; the weather bureau and disaster managers, to refine information detail and use all currently available modes of dissemination to provide the threatened with the maximum amount of fine details to make that evacuation decision in an informed and timely manner. This is a clear example of where theory and the practical views of hazard zone residents converge.

The 2007 southeast Australian fire zone research shows that people who are more obviously at risk from bushfire are far more prepared than people who are at risk from an occasional but as potentially destructive bushfire. The literature [13,24,31] shows that the less frequent the event, the less prepared people are likely to be. Aligned to the goals of this publication, the formula explained in this article on the realities and complex elements of combining physical with human geography to the ‘social good’ of maximizing safety around disaster impacts is a great potential application of complexity and systems science. This complexity includes the Media (Media Support MS of formula 1).

The Media

The media plays an ongoing role in socialization and the development of normative values. Dominick [20] argues that the media plays a key role in the cognitive development of individuals. Cognition is the act of coming to know something. Mass media can play a defining role in people’s awareness and responses to disaster threats. The CST embraces local media as an active agent in providing needed local information. Local media outlets can draw on and enhance Internet information to inform readers and listeners of ‘how to’ actions to maximize their safety. In this way the mass media can help clarify facts to help make contentious decisions (which may result in life or major property loss) by providing the fine details to reduce the uncertainty surrounding decisions that householders under threat need to make. The seven steps to community safety (Sect. “[Integrating Theory and Implementation](#)”) underline this core need for individuals in hazard zones to accept they are at risk and for the Internet and local media to act as providers of relevant, local, timely information for informed decision-making.

The media can play a powerful role in mobilizing communities [58], but they need to have accurate and timely information from disaster managers. The information which fire zone residents in southeastern Australia in 2007 asked for from the Internet is in full keeping with theory, and a recognition that greater access to current facts will reduce uncertainty in decision-making.

Triggering Action: Modeling and Simulating This Geographic, Communications and Psychological Complexity

The reason step one in the seven step process is so important is to counteract a psychological defense against accepting threat. We need to have a sense of personal invulnerability to get us through each day [75]. If we were scared of all possible problems we would cease to operate. The risk literature from Handmer, Goudie, Rohrmann, Salter [34,42,75,77] and many others underlines the importance of the clarity in the description of the threat and safety oriented action. Thus Acceptance they are In a Hazard Zone (AIHZ) of formula 1 is central to possible consequent safety-oriented behavior.

One reason given for risk communication failing is that it is not clear what should be done. The seven steps include a requirement that ‘what to do’ information be an integral part of any warning as a prelude to evacuation. The clear message is that the process of people excepting that a risk is real needs to well precede any actual impact. This is underlined by Kasperson & Stallen [49], who also

stress the importance of time perception and time horizons. Hence the importance of Knowledge Base of the hazard and safety-maximizing behavior (KB) of formula 1.

Timely Preparation, Final Actions and Evacuations

Timeliness of warnings is paramount to the effectiveness of warnings [85]. In the early 1990s there was much psychology research on perceptions of risk probabilities, individual psychometric studies on perceptions of future time, time orientation, planning horizons, and prediction of future events. In 2007 Australian fire zone residents say they want fine, detailed information as a threat approaches, so the uncertainty of what they may face is minimized. Their planning horizons are generally well oriented to maximum safety, but the crucial 'stay or go' decision will be based on information as close to impact as allows them adequate time to act safely.

Svenson [85] and others make clear that it is not only what you may chose to do, but also when you do it. In the case of remote area warnings of flooding, for instance, when flood waters may take a week to block down-catchment roads, traveling earlier than planned may be the best way to avoid the fate of the truck shown in Fig. 15. This is an example of an 'Active warning image', along with the image of the person standing on the car roof in the flooded road crossing (Fig. 16). Images like these are of use to remind people to travel before or after expected flooding – not during the flood.

In 2007, in a two day threat and evacuation exercise in Woodgate Beach operated on the underlined imperative, from the outset, that evacuation ahead of a storm surge would need to be completed not as cyclonic winds struck, but six hours before landfall. Once the winds gusted above 100 km per hour, all peoples, including disaster managers, must be in safe shelter above cyclone surge height.

Disasters and evacuations are issues of people, information, time and space. People in threat zones are entitled to accurate and timely localized information, detailing the threat and likely time-linked movement. People also want to know, via the internet or battery operated radios, what authorities are doing. That information may influence their own actions. What they ask for is the enhancement of their 'risk decision landscape' [83].

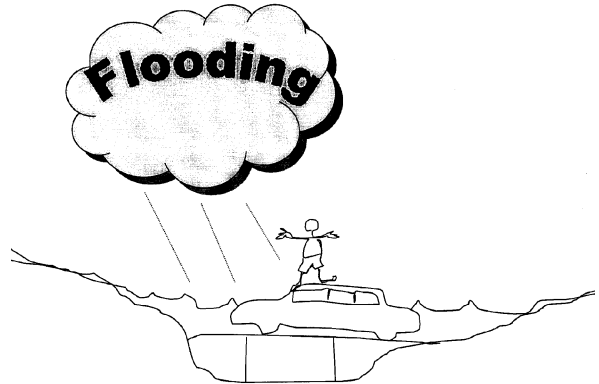
In 1974 communication management of and responses to disasters in Darwin and Brisbane shocked Australians. Darwin's Cyclone Tracy killed 65 people, and the Brisbane floods killed 16 [10,11]. These catastrophes taught Australian emergency planners much about the importance of effective warnings, sharing honest, complete and open information in a timely way to emergency managers, emer-



Evacuation as a Communication and Social Phenomenon, Figure 15

Nearly crossing a flooded road – Tully NQ 2004, From: Townsville Bulletin, 28/4/4.

Conveying weather threat information - flood



Evacuation as a Communication and Social Phenomenon, Figure 16

An image dissuading people from driving in to flood waters

gency workers, and those at risk to ensure sound preparation and responses. The Australian natural disasters of 1974, with major flooding elsewhere in Australia that year, also reinforced the importance of community and family ties to get people through the often profound emotional trauma allied with major natural disasters. The detailed documentation of Cyclone Tracey can help guide development of predictive models of disaster preparedness, communications and responses. With so few well document cases, a Bayesian Logic approach could be used, some well documented disasters, communications and response cases to develop the model, others to test and refine it.

The more recent and current disaster impacts all point to the importance of communities being properly informed. Indeed, the literature on knowledge, risk and inclusion of residents in the bushfire planning process is well described by Goldammer [32]. Issues of increased encroachment on to bush edges along with climate change are causing increased international concern over fire threats, often to places without much collective wisdom over the reality of those fire threats, or the nature of the strategies needed to maximize community safety.

South Eastern Australia is seen as one of the most bushfire-prone environments in the world [92]. Boura [7], an Australian fire manager, describes the development of community fireguard groups in Victoria. These groups are managed and instructed by the Country Fire Authority, but they are residents within very localized neighborhoods. They have their counterparts in the ACT and some other Australian States. They draw on detailed, often technically advanced information from the weather bureau and fire agencies to increase local awareness and action. This is a real playing out of the theories of social amplification developed by psychologists and risk communication theorists [73]; advanced planning techniques [16] and aligned with the way many feel about the direction of emergency management in Australia [15].

Linking the People with the Message

The foci on culture, community and social frameworks, either formal or informal neighborly links, is considered by Douglas [21]. Douglas argues that there is a need for people to understand that risk and danger exist where they live, despite a low probability that an impact may occur in any given hazard season. There is a large body of psychology which considers why people ignore clear messages

which may maximize their safety [12]. Douglas speaks of artificially distorted world views. Douglas posits such bias is rooted in over-simple views of heroic and bourgeois fiction. Changing such a normative world view has been discussed in this article in terms of a paradigm shift, well displayed by the Australian Government [13]. A model constructed from formula 1 needs to include all the subtle complex issues of social profiling.

Woodgate Beach (Sect. “Experiences and Lessons – Some Case Studies”), perhaps because of physical isolation, is fully embracing self-help, also displaying a deep culture of volunteerism that can be nurtured and emulated in ‘urban villages’. Sustainable urban planning concepts of nodes or activity centers are now well evolved as the hub of ‘urban villages’ [35]. What is happening in Woodgate Beach can be used as a model for any formula of developed social collective will to self-help, ultimately whether urban and surrounded by other neighborhoods, or more physically isolated.

Like the Ferny Creek (Victoria) community who agitated for a fixed bushfire siren after three of their neighbors were burned to death in the flash fire of 1997, community action needs a few individuals of vision and drive, residents of ‘neighborhoods’ can initiate and embrace safety-oriented behaviors and structures. The best thing the insti-



Evacuation as a Communication and Social Phenomenon, Figure 17
Initiatives to empower residents to be prepared for fire

tutional systems can do is draw out, nurture and encourage these individuals to mobilize a focus on gaining threat knowledge and acting to increase community responses. Part of the community engagement in the fact they are in a threat zone, as step 1 to action, is shown in Fig. 17; a great road-side banners initiative by one section of the Victorian Country Fire Authority to make the fire threat and implications real for fire-prone residents. Figure 17 shows, in few words, that if you intend to evacuate, do it early. Figure 17 says that the support and creativity to maximize communication effectiveness (the opposite of Institutional Barriers) helps contribute to a positive Medium and Message (MM) of formula 1.

Traditional Weather Warnings

Traditional weather warning signs include ant movement for looming flood, and general bird movement for strong winds, including cyclones [64,87,88]. On Palm Island, when the birds and animals go quiet, it signals that a major storm may be on the way [36,39]. The buildup to the summer monsoon rains is the universal experience of still, humid and hot conditions, continuing in intensity and cloud until the rains come. All cultures in all hazard zones have their traditional knowledge. Cultural Knowledge Bases (KB of formula 1) needs to be modeled into any mathematical prediction of likely propensity to respond to a natural disaster threat.

Words and Images

Word and image use are critical to effective communication. A newspaper graphic ahead of the terrorist-threatened 2004 Olympic Games (Fig. 18) conveys much about the world we now occupy, and uses no words.

The need for clear messages, most likely to provoke a precautionary response, is supported by the risk communication literature of Sect. “Effective Risk Communication” and the communication and cognitive theory of

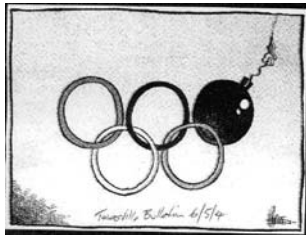
Sect. “Integrating Theory and Implementation”. If this knowledge is melded to requests of remote Indigenous peoples and NESH, the safety goal of clear, plain words and images will become the risk communication norm. There are many examples of images conveying meaning, such as the internationally used figure running upstairs, with an arrow; signifying an evacuation route. The symbol for tsunami; a stylized figure racing up a steep slope with a very large wave following, threatening to engulf them is also without words, but conveys all we need to know about evacuating ahead of a tsunami: get up slope immediately.

Images and Warning Maps

Risk communication is usually about attempts to prompt considered action by a person or community. Effective communication should make the future threat real in present thinking. Alternative responses should be outlined, along with likely consequences. To further prompt a considered and active response to the ‘action warning’, the consequences of inaction or a range of defensive actions should be made lucid. ‘Preferred’ behavior should seem reasonably attractive to ‘target individuals’ [85]. The theoretical overview of risk communication in Sect. “Effective Risk Communication” explains why the message must be clear to the target audience. It needs to have some cognitive content to get people thinking about how it may impact on them, and what the alternative outcomes for them may be if the predicted impact strikes where they are. Stimulus and local risk simulations, as embedded in the Redlands shire flood map [102], should be saturation broadcast into all hazard zones. The CST should be attempted with at-risk groups; people in threat zones, encouraging them to properly think about the real threat and to indicate to people what inaction may bring. A range of safety-oriented actions should also be presented.

Modeling Risk Communication into and Through Communities

Efforts have been made since the 1960s to see how well people understood the hazard. These studies continue [5]. The model of hearing, understanding, believing and feeling that the information is personalized so that those at risk will act has been well understood since the early 90s [81]. The general issues of credibility [18,69,70] still apply. Studies reported by Sorenson and Mileti [81] show increased knowledge as a result of risk communication efforts. People do become more aware of hazards and their personal place within the hazard threats. Unfortunately the link between knowledge and behavior remains tenuous [35].



Evacuation as a Communication and Social Phenomenon, Figure 18

An image of terror in 2004 – Source: Townsville Bulletin

Knowing that the normal first warning of major disruptive weather comes via the evening TV news in remote settlements, simulations will carry a high embedded likelihood of safety-oriented community response.

Sorenson and Mileti [81] showed that the believability of warnings increases as people get more warnings from officials with high credibility, and that women tend to believe emergency warnings more than men. They also showed that people higher up the socioeconomic ladder tend to believe warnings more than their counterparts. Minority groups have lower than average belief in warnings while people with a high knowledge of the hazard tend to find warnings more credible.

New Cultures to Hazard Zones

The issues discussed in this article from remote Indigenous communities and from non-English-speaking households make clear the importance of using plain English. Work with representatives of both groups made clear that plain English and images were necessary to convey the concept of danger to people in hazard zones, concurring with the expositions of Douglas [21]. The simpler the language the better. Douglas argues that even the word 'risk' could be dropped and 'danger' used instead. Goudie's work with recent Somali arrivals in Australia showed that, like Japanese tourists and many other people from non-English-speaking backgrounds, they understood the word and concept of 'danger', but did not, for instance, understand the word 'severe'. The importance of language clarity *to the intended audiences* in the MM expression of formula 1 needs emphasis.

For risk communication and evacuation advice to penetrate to all people in hazard zones, the language and images must be clear, simple, and compelling. If risk communicators construct campaigns of engagement which work for such 'marginalized' groups, the 'mainstream' can be easily included in that campaign.

Playing the Odds

Renn and Rohrmann [74] make clear that the assumption that risk judgments and evaluations are universal processes independent of social status or national heritage are unreasonable. Goudie's Australian research with multicultural organizations shows most clearly that there are many populations in hazard zones and that each of those groups must be addressed in ways that make the hazard real to them. Findings from the Cyclone Larry research [53] also show that there are many groups other than cultural, such as 'social isolates' – the socially disenfranchised, who need

special attention around disaster threats to ensure their safety.

Douglas [21] argues that risk perception and thus response is also an issue of moral and political issues. Westerners tend to consider the probability of an impact in terms of gambling whereas other cultures may need other triggers to internalize the threat and motivate them to action. The concept of a low probability event needing to be taken seriously is discussed by Douglas and others. Residents of Florida are well used to a near-annual evacuation ahead of frequent cyclones. Residents of the coastal communities of Queensland may be less likely to take the warning seriously, because they believe that there is no real chance that they will be personally affected by a cyclone. The whole issue of low probability and high impact events needs to be stressed to relevant impact zone residents.

The first and hardest step in an effective community evacuation remains convincing people in the projected impact zone that the looming threat is real. Renn and Rohrmann [74] suggest that to help make that threat real, project the expected number of fatalities or the catastrophic potential or where the threat may come from.

Spreading the Warning

Handmer [43] recommends that the professional warning agencies should attempt to harness the "personal informational networks of individuals within formal communication systems, and by assuring that formal warning advice is consistent with local norms and behavior" ([43] p 27). Shifting the normative values of recalcitrant disaster managers, and residents in ignorance or denial in urban risk zones becomes the key task of the *seven steps to community safety*, along with the amalgamating approach of the *community safety triangle*.

Part of the strength of the CST is that it taps into and informs an existing social predisposition for people to talk to each other, particularly in times of common threat (King and Goudie 2006). Using emergent technologies to provide real-time information to community media and households helps realize aware informed communities, predisposed to action.

With literature and research results discussed throughout this article, this discussion underlines the importance of modeling local norms, and interactive, safety-focused behavior. Emerging concepts which may be tested in further research are Community Safety groups (Sect. "Experiences and Lessons – Some Case Studies") and Social Burnoffs or flood evacuation practices. The main thrust of this work is to encourage interested modelers to accurately simulate how communities understand and act on

the need for safety-oriented action, where people are inclusive at the neighborhood level, acting as a self-preserving group, no matter what the ethnicity or state of surrounding populations. Neighborhood cohesion and empowerment is current social policy consistent with ESD philosophies and principles, enshrined in current planning laws such as IPA 1997 [46].

To illustrate the power of the internet, consider that any online reader globally can click on [103] and see the social aspects of Queensland's sustainability planning law. Indeed, click on and draw from any of the web sites given. If we were not suffering from information overload and time paucity, the paradigm shift to sustainability implementation would be rapid, the modeling acted on by disaster managers and financial cost-benefit analysts. Until that time, models will help convince and guide people in testable and demonstrable ways to increase community safety. Our very busyness is perhaps the main barrier to reducing global warming and natural disaster impacts. An alternative way into a more sustainable future is centered on local needs-meeting, including nurturing more local community cohesion and sharing of reliable, safety-oriented information and safety-oriented actions.

Future Directions in Modeling Disaster Preparedness and Evacuations as a Social Issue

As global warming and climate change intensifies, the need to model for preparedness and evacuations will increase with more frequent and extreme weather events. The total and fine detail of all needed background preparations and ramp-up preparations are too numerous and arduous for formal disaster management organizations to implement alone, hence the increased need to promote and strongly support the role of community engagement, of community empowerment and nurturing self-help in maximizing effective natural disaster preparation, including evacuations.

Australia has matured disaster policy, law and evolving practice, all embedded in concepts of ecologically sustainable development. Researchers can access, model and test local applicability of some of the Australian experiences and culture of community self-help. Sustainability Implementation Research will become universal as the era of last-minute organizational flurries whilst goading an ignorant and passive population at threat is superseded by prepared and bonded communities who are primed to receive well-delivered warnings, often sourced from the web, and who move themselves to safety in plenty of time. Modeling this will help sell the approach to conservative disaster managers.

The Sustainability Implementation Research, including data-based simulations introduced in this article, is the logical next step to 'action research' of the social sciences and will become the norm in approaching all issues of sustainability where the policies are mature, but agencies are unsure how to implement the paradigm shift; the behavioral and technological shift, to agreed long-term goals.

Meaningful consultation is a defining requirement of sustainability planning and good modeling, so all future social research of merit will take the Human Geographer's approach and "ask the local residents, involve the local residents." Community empowerment, and its modeling means local residents are entitled to help mold their own future. With hazard management, the shared SIR goal is to empower hazard-zone communities to be aware of the threat, be basically prepared for any warning, and act, as a community, in a precautionary way to maximize safety, minimize loss, and speed recovery.

As we move forward, plain language(s), clear, widely displayed hazard maps and images of like hazard impacts will help energize hazard-zone residents to the threats and their own needed safety-oriented behavior.

Researchers and Complexity and Systems Scientists have a moral if onerous obligation to challenge any layer of government clearly exercising or imposing barriers to helping people internalize any threats, then be supported in getting safe and staying safe, recognizing that change may threaten some individuals or sections within bureaucracies. As disaster management moves from a militaristic model of command and control to community empowerment and self help, the developing potential of the web as a key information source into hazard zones – ultimately web-to-air – will bear the fruit of greater community safety and minimized loss. Web-to-TV in disaster warning will be a fine conduit for showing residents-at-risk where the threat may be in relation to their home in a few hours. That immediacy of moving images is a compelling motivator for safety-oriented action.

In future, researchers, modelers and others involved in maximizing community safety will embrace some variations of the *communications safety triangle* and the *seven steps to community safety*, simply because they make sense; are highly cost-effective and easily web-refined from international to local conditions, populations and threats.

Acknowledgment

I most thank my research guardian and mentor over 15 years, Prof. David King, Director of both the Australian Center for Disaster Studies and of the Center for Tropi-

cal Urban and Regional Planning. David allowed me freedom to develop as an 'evidence-based' scientist, to conceive core approaches to sustainability implementation research; productive in helping render positive change in both disaster management and sustainability planning.

Thanks to Australian Bureau of Meteorology staff for their interactive support to improve risk communications, listening to what real people in real hazard zones experience, how they hear warnings, and how the medium and the messages can be and are optimized. The Bureau embraces the core goal to motivate safety-oriented action by people in hazard zones. The Bureau listens and improves the message and the delivery.

The bushfire research of '06 & 07 was funded by the Australian Bushfire Cooperative Research Center, supported by the University of Tasmania.

The 14 years of research reported in this article was not possible without the contributions of Authorities and more than 1000 Australians, old and new, who opened their organizations or doors to myself or research team members and shared their hazard experiences and specific warning needs. Thanks all.

Bibliography

Primary Literature

- ACF (2004) Institutions for Sustainability. Australian Conservation Foundation ACF 1–37. http://www.acfonline.org.au/uploads/res/res_tp007.pdf. Accessed 2008
- AMCORD (1995) AMCORD95, Australian Model Code of Residential Development. Department of Housing and Regional Development. Aust Govt Printing Service, Canberra
- Australian Bushfire CRC (2005) Bushfire CRC/Research/Community Self Sufficiency for Fire Safety/Effective Risk Communication. <http://www.bushfirecrc.com/research/c41/c41.html>. Accessed 2008
- Baram M (1991) Rights and duties concerning the availability of environmental risk information to the public. In: Kasperson RE, Stallen PJM (eds) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
- Berry L, King D (1998) Tropical cyclone awareness and education issues for far north Queensland school students – Storm Watchers. Aust J Emerg Manag 13:6
- Boughton GN (1992) Education on Natural Hazards, The Macedon Digest. Aust J Disaster Manag 7(2):4–7
- Boura J (1998) Community Fireguard: Creating partnerships with the community to minimise the impact of bushfire. Aust J Emerg Manag 13:59–64
- Brundtland GH (1988) Our Common Future. The World Commission for the Environment and Development. Alianza Publications. http://tilastokeskus.fi/abo2004/foredrag/hoglund_pp.pdf, <http://www.erf.es/eng/empresa/brundtland.html>, <http://web.uvic.ca/~stucraw/Lethbridge/MyArticles/Brundtland.htm>. Accessed 2008
- Cairns City Council (2007) Storm surge maps. <http://www.cairns.qld.gov.au/cairns/files/StormTideMaps/index.pdf>. Accessed 2008
- Chamberlain ER, Hartshorn AE, Mugglestone H, Short P, Svensson H, Western JS (1981a) Queensland flood report Australia Day (1974). Australian Government Publishing Service, Canberra, p 38
- Chamberlain ER, Doube L, Milne G, Rolls M, Western JS (1981b) The Experience of cyclone Tracy. Australian Government Publishing Service, Canberra, p 150
- Cialdini R, Reno R, Kallgren C (1990) A focus theory of normative conduct: recycling the concept of norms to reduce littering in public places. *J Pers Soc Psychol* 58(6):1015–1026
- COAG (2004) Natural Disasters in Australia, reforming mitigation, relief and recovery arrangements. [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(756EDFD270AD704EF00C15CF396D6111\)~COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf/\\$file/COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(756EDFD270AD704EF00C15CF396D6111)~COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf/$file/COAG+Report+on+Natural+Disasters+in+Australia+--+August+2002.pdf). Accessed 2008, Council of Australian Governments Commonwealth of Australia, 201
- Cohen E, White P, Hughes P (2006) Bushfire and the Media Reports 1–3. Latrobe University and BCRC, La Trobe University
- Cronan K (1998) Foundations of emergency management. *The Aust J Emerg Manag* 1(13):20–23
- Cuthill M (2004) Community well-being – the ultimate goal of democratic governance. *Qld Plan* 44(2):8–11
- Cutter SL, Boruff BJ, Shirley WL (2003) Social Vulnerability to Environmental Hazards. *Soc Sci Q* 84(2):242–261
- DIMIA (2004) Department of Immigration, Multicultural and Indigenous Affairs. http://www.humanrights.gov.au/racial_discrimination/face_facts/mig.htm#q. 2005
- Dolphin RR, Richard R, Ying F (2000) Is Corporate Communications a Strategic Function. *Manag Decis* 38(2):99–106
- Dominick JR (1993) The dynamics of mass communication. McGraw-Hill Inc, Columbus, p 616
- Douglas M (1992) Risk and Blame. *Essays in cultural theory*. Routledge, London/New York
- Drabek TE (1994) Disaster evacuation and the tourist industry. Program on environment and behaviour monograph 57. University of Colorado, Colorado
- Dunlap RE, Van Liere KD (1978) The 'new environmental paradigm': a proposed measuring instrument and preliminary results. *J Envtl Ed* 9(4):10–19
- ECAP (1997) Guidelines for disaster prevention and preparedness in tropical cyclone areas. Economic Commission for Asia and the Pacific, the World Meteorological Organisation and the Red Cross Societies, Geneva/Bangkok
- EMA (2000) Emergency Risk Management Applications Guide. The Australian Emergency Manuals Series. Emergency Management Australia. Dickson ACT:Emergency Management Australia, p 4
- EMA (2002) Research agenda for Emergency Management. March. EMA Research and development Strategy for "Safer Sustainable Communities". Dickson ACT:Emergency Management Australia
- EMA (2002) Indigenous Communities and Emergency Management. Emergency Management Australia, Canberra, p 22
- EMA (2002) Guidelines for Emergency Managers working with Culturally and Linguistically Diverse Communities. <http://www.ema.gov.au/ema/rwpattach>.

- [nsf/viewasattachmentpersonal/AFD7467016783EA8CA256CB30036EF42/\\$file/caldsept2002.pdf](#). Accessed 2008
29. EMAI (1998) Emergency Management Australia Information Service. Report of the strategic planning conference on the development of enhanced awareness education programs and materials for remote Aboriginal and Torres Strait Islander communities. Darwin, May 1997. Conference Proceedings, EMA, Mt Macedon Victoria
 30. Fien J (1993) Education for the environment – critical curriculum theorising and environmental education. Deakin University, Geelong, Victoria Australia
 31. Finnis K, Johnston D, Paton D (2004) Volcanic Hazard Risk Perceptions in New Zealand. Tephra, Earth and atmospheric sciences University of Alberta Edmonton, Alberta, Canada, pp 60–64, [http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/\\$file/volcanichazardrisk.pdf](http://www.civildefence.govt.nz/memwebsite.nsf/Files/Tephra%20v21%20chapters/$file/volcanichazardrisk.pdf), 2008
 32. Fowler KL, Kling ND, Larson MD (2007) Eigenvalues. *Bus Soc* 46:1. March. 88–103. Sage Publications
 33. Goldammer J (2005) Wildland fire – rising threats and vulnerabilities at the interface with society. In: Know risk. United Nations Tudor Rose, T. Jeggle, United Nations, ed; UN International Strategy for Disaster Reduction (UN-ISDR), Geneva, 376 p, pp 322–3
 34. Goudie D, King D (1999) Cyclone surge and community preparedness. *Aust J Emerg Manag* 13:1:454–60. [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/viewasattachmentpersonal/\(85FE07930A2BB4482E194CD03685A8EB\)~Cyclone_surge_and_community_preparedness.pdf/\\$file/Cyclone_surge_and_community_preparedness.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/viewasattachmentpersonal/(85FE07930A2BB4482E194CD03685A8EB)~Cyclone_surge_and_community_preparedness.pdf/$file/Cyclone_surge_and_community_preparedness.pdf), 2008
 35. Goudie D (2001) Toward Sustainable Urban Travel. Ph D thesis. James Cook University. <http://eprints.jcu.edu.au/967/>, 2008
 36. Goudie D (2004) Disruptive weather warnings and weather knowledge in remote Australian Indigenous communities. Web-based report http://www.tesag.jcu.edu.au/CDS/Pages/reports/Gou_iwwrpt/index.shtml?id=23, 2008
 37. Goudie D (2005) Sustainability planning: pushing against institutional barriers. *Ecosystems and Sustainable Development V*. WIT Press. WIT Transactions on Ecology and the Environment 81(5):215–224, www.witpress.com, 2008
 38. Goudie D (2007) Transport and Evacuation Planning. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 48–62
 39. Goudie D (2007) Oral histories about weather hazards in northern Australia. In: King D, Cottrell A (eds) Communities Living With Hazards. Centre for Disaster Studies, James Cook University with Queensland Department of Emergency Services, 293, James Cook University, North Queensland Australia, pp 102–125
 40. Gurmankin AD, Baron J, Armstrong K (2004) Intended message versus received message in hypothetical physician risk communication: exploring the gap. *Risk Analysis* 24(5):1337–1347
 41. Handmer J (1992) Can we have too much warning time? A study of Rockhampton, Australia. The Macedon Digest. *Aust J Disaster Manag* 7:2 p 8–10
 42. Handmer J (2000) Are Flood Warnings Futile? *Risk Commun emergencies* 2(e):1–14. <http://www.massey.ac.nz/~trauma/issues/2000-2/handmer.htm>, 2008
 43. Handmer J (2001) Improving flood warnings in Europe: a research and policy agenda. *Environ Hazards* 3(2001):19–28
 44. Heller K, Alexander DB, Gatz M, Knight BG, Rose T (2005) Social And Personal Factors As Predictors Of Earthquake Preparation: The Role Of Support Provision, Network Discussion, Negative Affect, Age, and Education. *J Appl Soc Psychol* 35(2):399–422
 45. Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian Model Averaging: A Tutorial. *Stat Sci* 14(4):382–417
 46. IPA (1997) The Integrated Planning Act. Queensland Government, <http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf>, 2008
 47. ISR (2007) Australian Policyonline. Institute for Social Research, Swinburne University of Technology, http://www.apo.org.au/linkboard/results.shtml?filename_num=117732, 2008
 48. Jarach M (1989) Overview of the literature on barriers to the diffusion of renewable energy sources in agriculture. *Appl En* 32(2):117–131
 49. Kasperson RE, Stallen PJM (1991) Communicating Risks to the Public International Perspectives. Kluwer, Dordrecht/Boston/London
 50. Kim YC, Ball-Rokeach SJ (2006) Civic engagement from a communication infrastructure perspective. *Commun Theory* 16:173–197
 51. Kitchin RM (1996) Increasing the integrity of cognitive mapping research: appraising conceptual schemata of environment-behaviour interaction. *Progress Hum Geogr* 20(1):56–84
 52. King D, Goudie D (1998) Breaking through the disbelief – the March 1997 floods at Cloncurry. Even the duck swam away. *Aust J Emerg Manag* 4:12 29–33
 53. King D, Goudie D (2006) Cyclone Larry, March 2006 Post Disaster Residents Survey. Centre for Disaster Studies, James Cook University, with the Australian Bureau of Meteorology P77 http://www.tesag.jcu.edu.au/CDS/Pages/reports/Larry_mainreport.pdf, 2008
 54. King D, Goudie D, Dominey-Howes D (2006) Cyclone knowledge and household preparation – some insights from Cyclone Larry Report on how well Innisfail prepared for Cyclone Larry. The *Aust J Emerg Manag* 21:3 52–59 [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(A80860EC13A61F5BA8C1121176F6CC3C\)~AJEM_EMA_Larry_Aug2006.pdf/\\$file/AJEM_EMA_Larry_Aug2006.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(A80860EC13A61F5BA8C1121176F6CC3C)~AJEM_EMA_Larry_Aug2006.pdf/$file/AJEM_EMA_Larry_Aug2006.pdf). Accessed August 2006
 55. Kobb P (2000) Emergency Risk Management Applications Guide. *Emerg Manag Aust*, Dickson, A.C.T.: Emergency Management Australia. Australian Emergency Manuals Series; 05 [http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/\(383B7EDC29CDE21FBA276BBCE12CDC0\)~Manual+05a.pdf/\\$file/Manual+05a.pdf](http://www.ema.gov.au/agd/EMA/rwpattach.nsf/VAP/(383B7EDC29CDE21FBA276BBCE12CDC0)~Manual+05a.pdf/$file/Manual+05a.pdf), 2008
 56. Leibovitz J (2003) Institutional barriers to associative city-region governance: the politics of institution-building and economic governance in 'Canada's Technology Triangle.' *Urban Studies* 40(13):2613–2642
 57. Lewis C (2006) Risk Management and Prevention Strategies. *Aust J Emerg Manag* 21(3):47–51

58. Lichtenberg J, Maclean D (1991) The role of the media in risk communication. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
59. Lidstone J (2006) Blazer to the Rescue! The role of puppetry in enhancing fire prevention and preparedness for young children. *Aust J Emerg Manag* 21(2):17–28
60. Loudness RS (1977) Tropical Cyclones in the Australian Region July 1909 to June 1975. AGPS, Canberra
61. McKenna F (1993) It won't happen to me: Unrealistic optimism or illusion of control. *Brit J Psych* 84:39–50
62. Munro DA (1995) Ecologically sustainable development – is it possible? How will we recognise it? In: Sivakumar M, Messer J (eds) *Protecting the future – ESD in action*. Futureworld, Wollongong
63. NDSI Working Group (2000) *Effective Disaster Warnings*. Working Group on Natural Disaster Information Systems. Subcommittee on Natural Disaster Reduction National Science and Technology Council Committee on Environment and Natural Resources. Executive Office of the President of the United States of America, 56, http://www.incident.com/cap/docs/NDIS_rev_Oct27.pdf, 2008
64. Napurrurlarlur NO, Jakamarrarlur NP (1988) Ngawarra-Kurlu. Yuendumu B.R.D.U., Darwin, p 19
65. O'Neill P (2004) Why don't they listen – Developing a risk communication model to promote community safety behaviour. The International Emergency Management Society, 11th Annual Conference Proceedings, Melbourne, Victoria, Australia, May 18–21 2004, pp 160–169
66. Young J, O'Neill P (1999) A social marketing framework for the development of effective public awareness programs. http://www.ses.nsw.gov.au/multiattachments/2740/documentname/A_Social_Marketing_Framework_for_the_Development_of_Effective_Public_Awareness_Programs.pdf, 2008. NSW Australia
67. Paton D (2003) Stress in Disaster Response: A risk management approach. *Disaster Prev Manag* 12(3):203–209
68. Paton D, Smith L, Johnston D (2005) When good intentions turn bad: promoting natural Hazard preparedness. *Aust J Emerg Manag* 20:25–30
69. Phillips R (1994) Long Range Planning. *Lond* 27(4):143–145
70. QG & QES (2003) State Planning Policy. Mitigating the adverse impacts of flood, bushfire and landslide. State planning policy 1/03. Dept Local Government and planning, & Dept of Emergency Services, <http://www.emergency.qld.gov.au/publications/spp/pdf/spp.pdf>, 2008
71. Quarantelli EL (2002) The role of the Mass Communication system in natural and technological disasters and possible extrapolation to terrorism situations. *Risk Manag Int J* 4(4):7–21
72. Raggatt P, Butterworth E, Morrissey S (1993) Issues in Natural Disaster Management: Community Response to the Threat of Tropical Cyclones in Australia. *Disaster Prev Manag* 2(3):12–21
73. Renn O, Levine D (1991) Credibility and trust in risk communication. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
74. Renn O, Rohrmann B (2000) Cross-cultural risk perception, A survey of empirical studies. Kluwer, Dordrecht, p 240
75. Rohrmann B (2000) A socio-psychological model for analysing risk communication processes. *Australas J Disaster Trauma Stud* 2000(2) <http://www.massey.ac.nz/~trauma/issues/2000-2/rohrmann.htm>, 2008
76. Rounsefell V (1992) Unified human settlement ecology. In: Birkeland J (ed) *Design for sustainability: A sourcebook of integrated eco-logical solutions*. Earthscan Publications, London, S4.2:78–83
77. Salter J (1992) The Nature of the disaster – more than just the meanings of words: Some reflections on definitions, doctrine and concepts. *The Macedon Digest. Aust J Disaster Manag* 7(2):1–3
78. Salter J, Bally J, Elliott J, Packham D (1993) Natural disasters: protecting vulnerable communities. In: Merriman PA, Browitt CW (eds) *Conference Proceedings*. London, 13–15 October 1993, Royal Society (Great Britain)
79. Sheppard E (1986) Modelling and predicting aggregate flows. In: Hanson S (ed) *The geography of urban transportation*. Guilford Press, New York/London, pp 91–110
80. Skertchly A, Skertchly K (2000) *Message Sticks – Hazard Mitigation Visual Language*. EMA, ACT, Dickson, Australian Capital Territory
81. Sorenson J, Mileti D (1991) Risk communication in emergencies. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
82. Stern P (1992) Psychological dimensions of global environmental change. *Ann Rev Psychol* 43:269–302
83. Stern PC, Fineberg HV (1996) *Understanding Risk, Informing decisions in a democratic society*. National Academy Press, Washington DC, pp 249
84. Sullivan M (2003) Communities and their experience of emergencies. *Aust J Emerg Manag* 18(1):19–26
85. Svenson O (1991) The time dimension in perception and communication of risk. In: Kasperson RE, Stallen PJM (ed) *Communicating Risks to the Public – International Perspectives*. Kluwer, Dordrecht/Boston/London, p 481
86. Thompson KM (2002) Variability and uncertainty meet risk management and risk communication. *Risk Analysis* 22(3):647–654
87. Utemorrhah D, Clendon M (2000) Dumbi the owl. In: Kimberley Language Resource Centre (ed) *Worrorra Lalai, Worrorra Dreamtime Stories*. KLRC, Halls Creek WA, pp 113
88. Utemorrhah D (1980) How The People Were All Drown(ed). In: Mowanjum (ed) *Visions of Mowanjum: Aboriginal writings from the Kimberley*. Rigby, Adelaide
89. Wakefield SE, Elliot SJ (2003) Constructing the news: the role of local newspapers in environmental risk communication. *Prof Geogr* 55(2):216–266
90. Wall M (2006) The case study method and management learning: making the most of a strong story telling tradition in emergency services management education. *Aust J Emerg Manag* 21(2):11–16
91. Walmsley DJ (1988) *Urban living, the individual in the city*. Longman scientific and technical, Longman, London, p 104
92. Woods F, Gabriel P (2005) Individual responsibility and state-wide strategies: bushfire in Victoria, Australia. In: Know risk. United Nations Tudor Rose, UK, Jeggel T (ed) Leicester, LE1 5RA, UK 376. pp 326–8
93. Yates J (1992) Assisting the community to plan: A pilot program in Western Australia. *The Macedon Digest. Aust J Disaster Manag* 7(2):12

94. Yates J (1997) Federalism and disaster mitigation in remote Aboriginal communities in Western Australia. Spring, AJEM, 25–32, Emergency Management Australia, Mt Macedon Victoria Australia. Publisher: Grey Worldwide Canberra Australia
95. Yeo S (2002) Natural Hazards. Flooding in Australia: A review of events in 1998. 25:177–191. Department of Physical Geography, Macquarie University, NSW, <http://www.springerlink.com/content/n160g0121800n742/> Natural Hazards 25(2):177–191, 2002. Kluwer Academic Publishers. Printed in the Netherlands. <http://www.springerlink.com/content/n160g0121800n742/fulltext.pdf>
96. Zamecka A, Buchanan G (2000) Disaster risk management. Queensland Department of Emergency Services, Brisbane, p 115
97. Kikuchi T, Nakamori Y (2007) Agent model analysis to explore effects of interaction and environment on individual performance. J Syst Sci Complexity 2007 20:1–17. Springer Science + Business Media, Inc
98. <http://news.bbc.co.uk/1/hi/world/europe/6963373.stm> BBC (2005) Access to 2008, In pictures: Hurricane onslaught
99. <http://news.bbc.co.uk/2/hi/americas/7055721.stm> BBC (2007) Californians flee as fires rage
100. Mornington Peninsula Shire Council (2006) Fire wise fire management. <http://www.mornpen.vic.gov.au/Files/FireWiseFireManagementBooklet.pdf>
101. Cairns City Council (2006) Storm tide maps. <http://www.hazardsaustralia.info/Mapping.html>
102. <http://maps.redland.qld.gov.au/website/redemapexternal%5Fv2%5F03/Default.aspx>, Redlands Shire Accessed 2008 Flood Maps
103. Queensland Government (1997) Integrated planning act. <http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/I/integplana97.pdf>
- Geoscience Australia (2005) Sentinel. Commonwealth of Australia. Canberra, ACT, Australia, <http://sentinel.ga.gov.au/acres/sentinel/index.shtml>
- Rural Fire Service (2008) Fire Safety Information. New South Wales Rural Fire Service. NSW Government, Sydney, http://www.rfs.nsw.gov.au/dsp_content.cfm?CAT_ID=515 NSW Rural Fire Service, with information in 27 languages
- CFA (2008) Country Fire Authority. Melbourne, Victoria, Australia, <http://www.cfa.vic.gov.au>
- ESA (2008) Community Education. Emergency Services Agency, Australian Capital Territory, http://www.esa.act.gov.au/esawebsite/content_esa/community_education/community_education.html

Books and Reviews

- Bushnell S, Cottrell A, Spillman M, Lowe D (2006) Thuringowa bushfire case study. Understanding Communities. Project BCRC Program C Community self-sufficiency for fire safety. Bushfire CRC, JCU CDS, Townsville, Australia
- Granger KJ, Smith DI (1995) Storm tide impact and consequence modelling: some preliminary observations. Math Comput Model 21:9:15–21
- Roth W (1897) Ethnological Studies among the North West Central Queensland Aborigines. Queensland government, Brisbane/London
- FEMA (2008) Prepare for a Wildfire. Federal Emergency Management Agency. U.S. Department of Homeland Security Washington, DC http://www.fema.gov/hazard/wildfire/wf_prepare.shtm
- Australian Bureau of Meteorology (2008) Protecting yourself and your home. Bushfire weather. BoM. Melbourne Australia http://www.bom.gov.au/inside/services_policy/fire_ag/bushfire/protect.htm
- Emergency Management Australia (2003). Community safety Bushfire action guide. EMA. Dickson, Australian Capital Territory <http://www.ema.gov.au/agd/ema/emainternet.nsf/Page/RWP07C6046B98D07DB8CA256C5A00230553>
- ABC (2006) Bushfire summer. Australian Broadcasting Commission. Melbourne, www.abc.net.au/bushfire

Evacuation Dynamics: Empirical Results, Modeling and Applications

ANDREAS SCHADSCHNEIDER^{1,2}, WOLFRAM KLINGSCH³,
HUBERT KLÜPFEL⁴, TOBIAS KRETZ⁵,
CHRISTIAN ROGGSCH³, ARMIN SEYFRIED⁶

¹ Institut für Theoretische Physik, Universität zu Köln, Köln, Germany

² Interdisziplinäres Zentrum für Komplexe Systeme, Bonn, Germany

³ Institute for Building Material Technology and Fire Safety Science, University of Wuppertal, Wuppertal, Germany

⁴ TraffGo HT GmbH, Duisburg, Germany

⁵ PTV Planung Transport Verkehr AG, Karlsruhe, Germany

⁶ Jülich Supercomputing Centre, Research Centre Jülich, Jülich, Germany

Article Outline

Glossary
Definition of the Subject
Introduction
Empirical Results
Modeling
Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Pedestrian A person traveling on foot. In this article, other characterizations are used depending on the context, e. g., agent or particle.

Crowd A large group of pedestrians moving in the same area, but not necessarily in the same direction.

Evacuation The movement of persons from a dangerous place due to the threat or occurrence of a disastrous event. In normal situations this is called “egress” instead.

Flow The flow or current J is defined as the number of persons passing a specified cross-section per unit time. The common unit of flow is “persons per second”. Specific flow is the flow per unit cross-section. The maximal flow supported by a facility (or a part of it) is called “capacity”.

Fundamental diagram In traffic engineering (and physics): density-dependence of the flow: $J(\rho)$. Due to the hydrodynamic relation $J = \rho v b$ equivalent representations used frequently are $v = v(\rho)$ or $v = v(J)$. The fundamental diagram is probably the most important quantitative characterization of traffic systems.

Lane formation In bidirectional flows, lanes are often dynamically formed in which all pedestrians move in the same direction.

Bottleneck A limited resource for pedestrian flows, for example a door, a narrowing in a corridor, or stairs, i. e., a location of reduced capacity. At bottlenecks jamming occurs if the inflow is larger than the capacity. Other phenomena that can be observed are the formation of lanes and the zipper-effect.

Microscopic models Models which represent each pedestrian separately with individual properties like walking velocity or route choice behavior and the interactions between them. Typical models that belong to this class are cellular automata and the social-force model.

Macroscopic models Models which do not distinguish individuals. The description is based on aggregate quantities, e. g., appropriate densities. Typical models belonging to this class are fluid-dynamic approaches. Hand calculation methods which are based on related ideas and are often used in the field of (fire-safety) engineering belong to this class as well.

Crowd disaster An accident in which the specific behavior of the crowd is a relevant factor, e. g., through competitive and non-adaptive behavior. In the media, it is often called “panic” which is a controversial concept in crowd dynamics and should thus be avoided.

the participants and for the organizers who must be prepared for any case of emergency or critical situation. Usually in such cases the participants must be guided away from the dangerous area as quickly as possible. Therefore the understanding of the dynamics of large groups of people is very important.

In general, evacuation is egress from an area, a building or a vessel due to a potential or actual threat. In the cases described above, the dynamics of the evacuation processes are quite complex due to the large number of people and their interaction, external factors such as fire, complex building geometries, etc. Evacuation dynamics must be described and understood on different levels: physical, physiological, psychological, and social. Accordingly, the scientific investigation of evacuation dynamics involves many research areas and disciplines. The system “evacuation process” (i. e., the population and the environment) can be modeled on many different levels of detail, ranging from hydro-dynamic models to artificial intelligence and multi-agent systems. There are at least three aspects of evacuation dynamics that motivate its scientific investigation:

- 1) As in most many-particle systems several interesting collective phenomena can be observed that need to be explained;
- 2) Models need to be developed that are able to reproduce pedestrian dynamics in a realistic way, and
- 3) Pedestrian dynamics must be applied to facility design and to emergency preparation and management.

The investigation of evacuation dynamics is a difficult problem that requires close collaboration between different fields. The origin of the apparent complexity lies in the fact that one is concerned with a many-‘particle’ system with complex interactions that are not fully understood. Typically the systems are far from equilibrium and so are, e. g., sensitive to boundary conditions. Motion and behavior are influenced by several external factors and often crowds can be rather inhomogeneous.

In this article we want to deal with these problems from different perspectives and will not only review the theoretical background, but will also discuss some concrete applications.

Definition of the Subject

Today, there are many occasions on which a large number of people gathers in a rather small area. Office buildings and apartment houses grow larger and more complex. Very large events related to sports, entertainment or cultural and religious events are held all over the world on a regular basis. This brings about serious safety issues for

Introduction

The awareness that emergency exits are one of the most important factors to ensure the safety of persons in buildings can be traced back more than 100 years. Disasters due to the fires in the Ring theater in Vienna and the urban theater in Nizza in 1881 resulted in several hundred fatal-

ities and led to a rethinking of the safety in buildings [24]. First, attempts were made to improve safety by using non-flammable building materials. However, the disaster at the Troquois Theater in Chicago with more than 500 fatalities, where only the decorations burned, demonstrated the need for more effective measures. This was a starting point for studying the influences of emergency exits and thus the dynamics of pedestrian streams [24,32].

In recent years there have been two major evacuation incidents which gained immense global attention. First, there was the capsizing of the Baltic Sea ferry MV Estonia (September 28, 1994, 852 casualties) [100] and, of course, the terrorist attacks of 9/11 (2,749 casualties). Other prominent examples of the possible tragic outcomes of the dynamics of pedestrian crowds are the Hillsborough stadium disaster in Sheffield (April 15, 1989, 96 casualties) [182], the accident at Bergisel (December 4, 1999, 5 casualties) [189], the stampede in Baghdad (August 30, 2005, 1,011 casualties), the tragedy at the concert of “The Who” (December 3, 1979, 11 casualties) [73] and – very early – the events at the crowning ceremony of Tsar Nicholas II in St. Petersburg in May 1896 with 1,300 to 3,000 fatalities (sources vary considerably) [168]. In the past, tragic accidents have happened frequently in Mecca during the Hajj (1990: 1,426, 1994: 270, 1997: 343, 1998: 107, 2001: 35, 2003: 14, 2004: 244, and 2006: 364 casualties). What stands out is that the initiating events are very diverse and range from external human aggression (terrorism) to external physical dangers (fire) and rumors to various shades of greedy behavior in absence of any external danger.

Many authors have pointed out that the results of experts’ investigations and the way the media typically reports about an accident very often differ strongly [17,77,109,155,156,178]. Public discussion has a much greater tendency to identify “panic” as the cause of a disaster, while expert commissions often conclude that there either was no panic at all, or panic was merely a result of some other preceding phenomenon.

This article first discusses the empirical basis of pedestrian dynamics in Sect. “[Empirical Results](#)”. Here we introduce the basic observables and describe the main qualitative and quantitative results, focusing on collective phenomena and the fundamental diagram. It is emphasized that even for the most basic quantities, no consensus about basic behavior has been reached.

In Sect. “[Modeling](#)” various model approaches that have been applied to the description of pedestrian dynamics are reviewed.

Section “[Applications](#)” discusses more practical issues and gives a few examples for applications to safety analysis.

In this regard, prediction of evacuation times is an important problem as legal regulations must often be fulfilled. Here, commercial software tools are available. A comparison shows that the results must be interpreted with care.

Empirical Results

Overview

Pedestrians are three-dimensional objects and a complete description of their highly developed and complicated motion sequence is rather difficult. Therefore, in pedestrian and evacuation dynamics, pedestrian motion is usually treated as two-dimensional by considering the vertical projection of the body.

In the following sections we review the present knowledge of empirical results. These are relevant not only as a basis for the development of models, but also for applications such as safety studies and legal regulations.

We start with the phenomenological description of collective effects. Some of these are known from everyday experience and will serve as benchmark tests for any kind of modeling approach. Any model that does not reproduce these effects is missing some essential part of the dynamics. Next, the foundations of a quantitative description are laid by introducing the fundamental observables of pedestrian dynamics. Difficulties arise from different conventions and definitions. Then pedestrian dynamics in several simple scenarios (corridor, stairs etc.) are discussed. Surprisingly, even for these simple cases no consensus about the basic quantitative properties exists. Finally, more complex scenarios are discussed which are combinations of the simpler elements. Investigations of scenarios such as evacuations of large buildings or ships suffer even more from lack of reliable quantitative and sometimes even qualitative results.

Collective Effects

One of the reasons why the investigation of pedestrian dynamics is attractive for physicists is the large variety of interesting collective effects and self-organization phenomena that can be observed. These macroscopic effects reflect the individuals’ microscopic interactions and thus give important information for any modeling approach.

Jamming Jamming and clogging typically occur for high densities at locations where the inflow exceeds capacity. Locations with reduced capacity are called *bottle-necks*. Typical examples are exits (Fig. 1) or narrowings. This kind of jamming phenomenon does not depend strongly on the microscopic dynamics of the



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 1

Clogging near a bottleneck. The shape of the clog is discussed in more detail in Subsect. “Theoretical Results”

particles. Rather it is a consequence of an exclusion principle: space occupied by one particle is not available for others.

This clogging effect is typical for a bottleneck situation. It is important for practical applications, especially evacuation simulations.

Other types of jamming occur in the case of counterflow where two groups of pedestrians mutually block each other. This happens typically at high densities and when it is not possible to turn around and move back, e. g., when the flow of people is large.

Density waves Density waves in pedestrian crowds can be generally characterized as quasi-periodic density variations in space and time. A typical example is the movement in a densely crowded corridor (e. g., in subway-stations close to the density that causes a complete halt of motion) where phenomena similar to stop-and-go vehicular traffic can be observed, e. g., density fluctuations in a longitudinal direction that move backwards (opposite to the movement direction of the crowd) through the corridor. More specifically, for the situation on the Jamarat Bridge in Makkah (during the Hajj pilgrimage 2006), stop-and-go waves have been

reported. At densities of 7 persons per m^2 upstream, moving stop-and-go waves of period 45 s have been observed that lasted for 20 minutes [59]. Fruin reports, that “at occupancies of about 7 persons per square meter the crowd becomes almost a fluid mass. Shock waves can be propagated through the mass sufficient to lift people off their feet and propel them distances of 3 m (10 ft) or more.” [36].

Lane formation In counterflow, i. e., two groups of people moving in opposite directions, (dynamically varying) lanes are formed where people move in just one direction [135,139,197]. In this way, strong interactions with oncoming pedestrians are reduced which is more comfortable and allows higher walking speeds. The occurrence of lane formation does not require a preference of moving on one side. It also occurs in situations without left- or right-preference. However, cultural differences for the preferred side have been observed. Although this preference is not essential for the phenomenon itself, it has an influence on the kind of lanes formed and their order.

Several quantities for the quantitative characterization of lane formation have been proposed. Yamori [197] has introduced a band index which is basically the ratio of pedestrians in lanes to their total number. In [13] a characterization of lane formation through the (transversal) velocity profiles at fixed positions has been proposed. Lane formation has also been predicted to occur in colloidal mixtures driven by an external field [15,28,158]. Here, an order parameter $\phi = \frac{1}{N} \langle \sum_{j=1}^N \phi_j \rangle$ has been introduced where $\phi_j = 1$ if the lateral distance to all other particles of the other type is larger than a typical density-dependent length scale, and $\phi_j = 0$ otherwise.

The number of lanes can vary considerably with the total width of the flow. Figure 2 shows a street in the city center of Cologne during World Youth Day in Cologne (August 2005) where two comparatively large lanes have been formed.

The number of lanes usually is not constant and might change in time, even if there are relatively small changes in density. The number of lanes in opposite directions is not always identical. This can be interpreted as a sort of spontaneous symmetry breaking.

Quantitative empirical studies of lane formation are rare. Experimental results have been reported in [94] where two groups with varying relative sizes had to pass each other in a corridor with a width of 2 m. On one hand, similar to [197] a variety of different lane patterns were observed, ranging from 2 to 4 lanes. On the other hand, in spite of this complexity, surprisingly



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 2

The “Hohe Straße” in Cologne during World Youth Day 2005. The yellow line is the border of the two walking directions

large flows could be measured: the sum of (specific) flow and counterflow was between 1.8 and 2.8 persons per meter per second and exceeded the specific flow for one-directional motion (≈ 1.4 P/ms).

Oscillations In counterflow at bottlenecks, e.g., doors, one can sometimes observe oscillatory changes of the direction of motion. Once a pedestrian is able to pass the bottleneck it becomes easier for others to follow in the same direction until somebody is able to pass the bottleneck (e.g., through a fluctuation) in the opposite direction.

Patterns at intersections At intersections, various collective patterns of motion can be formed. A typical example is short-lived roundabouts which make motion more efficient. Even if these are connected with small detours, the formation of these patterns can be favorable since they allow for “smoother” motion.

Emergency situations, “panic” In emergency situations various collective phenomena have been reported that have sometimes misleadingly been attributed to *panic behavior*. However, there is strong evidence that this is not the case. Although a precise accepted definition of *panic* is missing, usually certain aspects are associated with this concept [77]. Typically “panic” is assumed to occur in situations where people compete for scarce or dwindling resources (e.g., safe space or access to an exit) which leads to selfish, asocial or even completely irrational behavior and contagion that affects large groups. A closer investigation of many crowd disasters has revealed that most of the above characteristics have played almost no role and most of the time have not been observed at all (see e.g. [73]). Often

the reason for these accidents is much simpler, e.g., in several cases the capacity of the facilities was too small for the actual pedestrian traffic, e.g., Luschniki Stadium Moskau (October 20, 1982), Bergisel (December 4, 1999), pedestrian bridge Kobe (Akashi) (July 21, 2001) [186]. Therefore the term “panic” should be avoided, *crowd disaster* being a more appropriate characterization. Also it should be kept in mind that in dangerous situations it is *not* irrational to fight for resources (or for your own life), if everybody else does this [18,113]. Only from the outside is this behavior perceived as irrational since it might lead to a catastrophe [178]. The latter aspect is therefore better described as *non-adaptive behavior*. We will discuss these issues in more detail in Subsect. “Evacuations: Empirical Results”.

Observables

Before we review experimental studies in this section, the commonly used observables are introduced.

The flow J of a pedestrian stream gives the number of pedestrians crossing a fixed location of a facility per unit of time. Usually it is taken as a scalar quantity since only the flow normal to some cross-section is considered. There are various methods to measure flow. The most natural approach is to determine the times t_i at which pedestrians pass a fixed measurement location. The time gaps $\Delta t_i = t_{i+1} - t_i$ between two consecutive pedestrians i and $i + 1$ are directly related to the flow

$$J = \frac{1}{\langle \Delta t_i \rangle} \quad \text{with} \quad \langle \Delta t_i \rangle = \frac{1}{N} \sum_{i=1}^N (t_{i+1} - t_i) = \frac{t_{N+1} - t_1}{N}. \quad (1)$$

Another possibility for measuring the flow of a pedestrian stream is borrowed from fluid dynamics. The flow through a facility of width b is determined by the average density ρ and the average speed v of a pedestrian stream as

$$J = \rho v b = J_s b. \quad (2)$$

where the *specific flow*¹

$$J_s = \rho v \quad (3)$$

gives the flow per unit-width. This relation is also known as *hydrodynamic relation*.

There are several problems concerning the way in which velocities, densities or time gaps are measured and

¹In strictly one-dimensional motion often a line density (dimension: 1/length) is used. Then the *flow* is given by $J = \rho v$.

the conformance of the two definitions of flow. The flow according to Eq. (1) is usually measured as a mean value over time at a certain location, while the measurement of the density in Eq. (2) is connected with an instantaneous mean value over space. This can lead to a bias caused by underestimation of fast moving pedestrians at the average over space compared to the mean value of the flow over time at a single measurement line (see the discussion for vehicular traffic e. g., in [51,81,102]). Furthermore, most experimental studies measuring the flow according to Eq. (2) combine for technical reasons an *average* velocity of a single pedestrian over time with an *instantaneous* density. To ensure a correspondence of the mean values the average velocity of all pedestrians contributing to the density at a certain instant has to be considered. However this procedure is very time consuming and not realized in practice up to now. Moreover, the fact that the dimension of the test section has usually the same order of magnitude as the extent of the pedestrians can influence the averages over space. These all are possible factors why different measurements can differ in a large way, see discussion in Subsect. “Fundamental Diagram”.

Another way to quantify the pedestrian load of facilities has been proposed by Fruin [35]. The “pedestrian area module” is given by the reciprocal of the density. Thompson and Marchant [184] introduced the so-called “inter-person distance” d , which is measured between center coordinates of the assessing and obstructing persons. According to the “pedestrian area module” Thompson and Marchant call $\sqrt{1/\rho}$ the “average inter-person distance” for a pedestrian stream of evenly spaced persons [184]. An alternative definition is introduced in [58] where the local density is obtained by averaging over a circular region of radius R ,

$$\rho(\mathbf{r}, t) = \sum_j f(\mathbf{r}_j(t) - \mathbf{r}), \quad (4)$$

where $\mathbf{r}_j(t)$ are the positions of the pedestrians j encompassed by \mathbf{r} and $f(\dots)$ is a Gaussian, distance-dependent weight function.

In contrast to the density definitions above, Predtechenskii and Milinskii [151] consider the ratio of the sum of the projection area f_j of the bodies and the total area of the pedestrian stream A , defining the (dimensionless) density $\bar{\rho}$ as

$$\bar{\rho} = \frac{\sum_j f_j}{A}, \quad (5)$$

a quantity known as *occupancy* in the context of vehicular traffic. Since the projection area f_j depends strongly on the

type of person (e. g., it is much smaller for a child than for an adult), the densities for different pedestrian streams consisting of the same number of persons and the same stream area can be quite different.

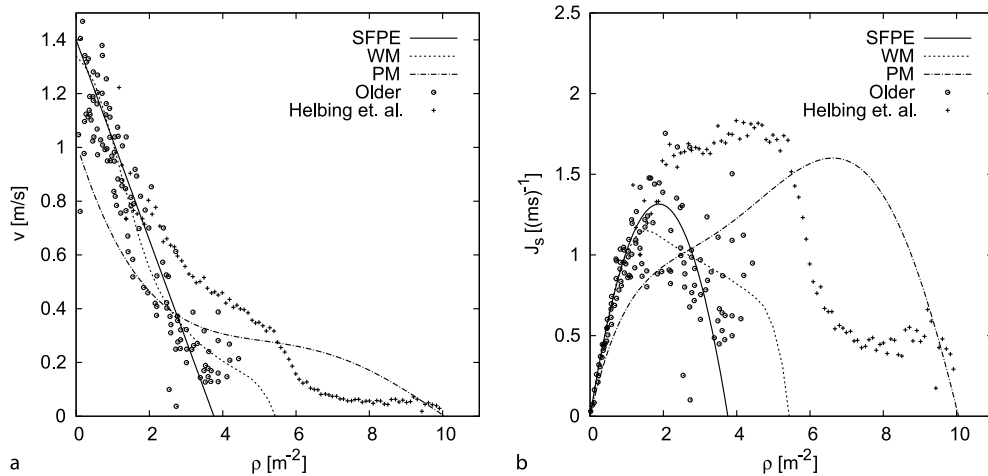
Beside technical problems due to camera distortions and camera perspective there are several conceptual problems, such as the association of averaged with instantaneous quantities, the need to choose an observation area in the same order of magnitude as the extent of a pedestrian together with the definition of the density of objects with nonzero extent and much more. A detailed analysis of the ways in which measurement influences the relations is necessary but still lacking.

Fundamental Diagram

The fundamental diagram describes the empirical relation between density ρ and flow J . The name indicates its importance and naturally it has been the subject of many investigations. Due to the hydrodynamic relation (3) there are three equivalent forms: $J_s(\rho)$, $v(\rho)$ and $v(J_s)$. In applications the relation is a basic input for engineering methods developed for the design and dimensioning of pedestrian facilities [35,136,150]. Furthermore, it is a quantitative benchmark for models of pedestrian dynamics [21,86,112,175].

In this section we will concentrate on planar facilities such as sidewalks, corridors and halls. For various facilities such as floors, stairs or ramps, the shape of the diagrams differ, but in general it is assumed that the fundamental diagrams for the same type of facilities but having different widths merge into one diagram for the specific flow J_s . In first order this is confirmed by measurements on different widths [49,135,139,142]. However, Navin and Wheeler observed in narrow sidewalks more orderly movement leading to slightly higher specific flows than for wider sidewalks [135]. A natural lower bound for the independence of the specific flow from the width is given by the body size and the asymmetry in movement possibilities of the human body. Surprisingly, Kretz et al. found an increase of the specific flow for bottlenecks with $b \leq 0.7$ m [93]. This will be discussed in more detail later. For the following discussion we assume facility widths larger than $b = 0.6$ m and use the most common representations $J_s(\rho)$ and $v(\rho)$.

Figure 3 shows various fundamental diagrams used in planning guidelines and measurements of two selected empirical studies representing the overall range of the data. The comparison reveals that specifications and measurements disagree considerably. In particular, the maximum of the function giving the capacity $J_{s,\max}$ ranges from 1.2 (ms)^{-1} to 1.8 (ms)^{-1} , the density value where the max-



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 3

Fundamental diagrams for pedestrian movement in planar facilities. The lines refer to specifications according to planning guidelines (SFPE Handbook [136]), Predtechenskii and Milinskii (PM) [150], Weidmann (WM) [192]. Data points give the range of experimental measurements (Older [142] and Helbing [58])

imum flow is reached (ρ_c) ranges from 1.75 m^{-2} to 7 m^{-2} and, most notably, the density ρ_0 , where the velocity approaches zero due to overcrowding, ranges from 3.8 m^{-2} to 10 m^{-2} .

Several explanations for these deviations have been suggested, including cultural and population differences [58,116], differences between uni- and multidirectional flow [99,135,154], short-ranged fluctuations [154], influence of psychological factors given by the incentive of the movement [150] and, partially related to the latter, the type of traffic (commuters, shoppers) [139].

It seems that the most elaborate fundamental diagram is given by Weidmann who collected 25 data sets. An examination of the data which were included in Weidmann's analysis shows that most measurements with densities larger than $\rho = 1.8 \text{ m}^{-2}$ are performed on multidirectional streams [135,139,140,142,148]. But data gained by measurements on strictly unidirectional streams has also been considered [35,49,188]. Thus Weidmann neglected differences between uni- and multidirectional flow in accordance with Fruin, who states in his often cited book [35] that the fundamental diagrams of multidirectional and unidirectional flow differ only slightly. This disagrees with results of Navin and Wheeler [135] and Lam et al. [99] who found a reduction of the flow in dependence of directional imbalances. Here lane formation in bidirectional flow has to be considered. Bidirectional pedestrian flow includes unordered streams as well as lane-separated and thus quasi-unidirectional streams in opposite directions. A more detailed discussion and data can be found

in [99,135,154]. A surprising finding is that the sum of flow and counterflow in corridors is larger than the unidirectional flow and for equally distributed loads it can be twice the unidirectional flow [94].

Another explanation is given by Helbing et al. [58] who argue that cultural and population differences are responsible for the deviations between Weidmann and their data. In contrast to this interpretation the data of Hanking and Wright [49] gained by measurements in the London subway (UK) are in good agreement with the data of Mori and Tsukaguchi [115] measured in the central business district of Osaka (Japan), both on strictly uni-directional streams. This brief discussion clearly shows that up to now there is no consensus about the origin of the discrepancies between different fundamental diagrams and how one can explain the shape of the function.

However, all diagrams agree in one characteristic: velocity decreases with increasing density. As the discussion above indicates there are many possible reasons and causes for the velocity reduction. For the movement of pedestrians along a line, a linear relation between speed and the inverse of the density was measured in [174]. The speed for walking pedestrians depends also linearly on the step size [192] and the inverse of the density can be regarded as the required length for one pedestrian to move. Thus it seems that smaller step sizes caused by a reduction of the available space with increasing density is, at least for a certain density region, one cause for the decrease of speed. However, this is only a starting point for a more elaborated modeling of the fundamental diagram.

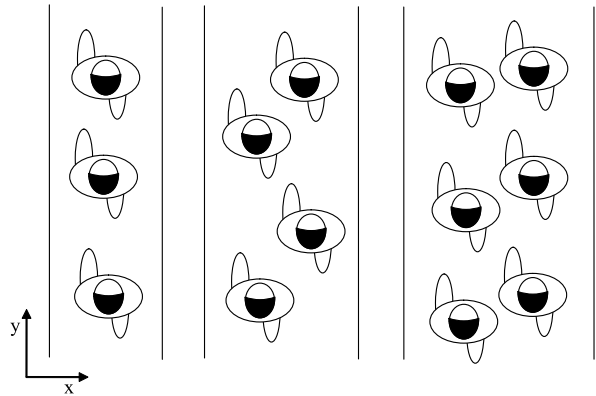
Bottleneck Flow

The flow of pedestrians through bottlenecks shows a rich variety of phenomena, e.g., the formation of lanes at the entrance to the bottleneck [64,66,93,176], clogging and blockages at narrow bottlenecks [24,57,93,121,122,150] or some special features of bidirectional bottleneck flow [57]. Moreover, the estimation of bottleneck capacities by the maxima of fundamental diagrams is an important tool for the design and dimensioning of pedestrian facilities.

Capacity and Bottleneck Width One of the most important practical questions is how the capacity of a bottleneck rises with increasing width. Studies of this dependence can be traced back to the beginning of the last century [24,32] and, up to now, have been discussed controversially. As already mentioned in the context of the fundamental diagram there are multiple possible influences on pedestrian flow and thus on the capacity. In the following, the major findings are outlined, demonstrating the complexity of the system and documenting a controversial discussion over one hundred years.

At first sight, a stepwise increase of capacity with the width appears to be natural if lanes are formed. For independent lanes, where pedestrians in one lane are not influenced by those in others, the capacity increases only if an additional lane can be formed. This is reflected in the stepwise enlargement of exit width, which has been a requirement of several building codes and design recommendations. See e.g., the discussion in [146] for the USA and GB and [130] for Germany. e.g.; the German building code requires an exit width (e.g., for a door) to be at least 90 cm plus 60 cm for every 200 persons. Independently from this simple lane model, Hoogendoorn and Daamen [64,66] measured by a laboratory experiment the trajectories of pedestrians passing a bottleneck. The trajectories show that inside a bottleneck the formation of lanes occurs, resulting from the zipper effect occurring on entry to the bottleneck. Due to the zipper effect, a self-organization phenomenon leading to an optimization of the available space and velocity; the lanes are not independent and thus do not allow passing (Fig. 4). The empirical results of [64,66] indicate a distance between lanes of $d \approx 0.45$ m, independent of the bottleneck width b , implying a stepwise increase of capacity. However, the investigation was restricted to two values ($b = 1.0$ m and $b = 2.0$ m) of the width.

In contrast, the study [176] considered more values of the width and found that the lane distance increases continuously as illustrated in Fig. 4. Moreover it was shown that a continuous increase of the lane distance leads to



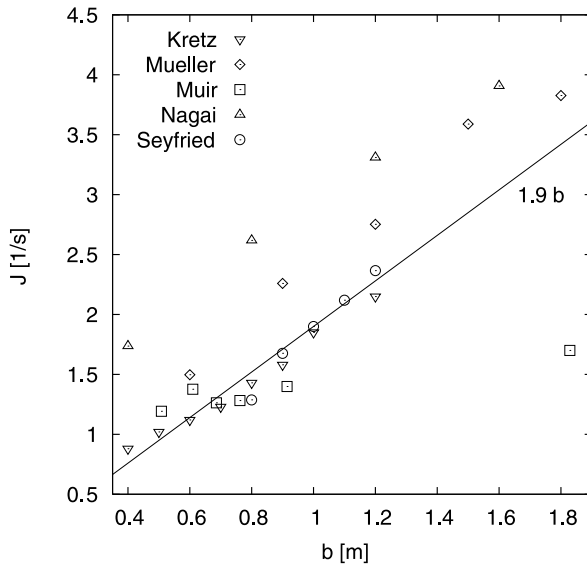
Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 4

A sketch of the zipper effect with continuously increasing lane distances in x : The distance in the walking direction decreases with increasing lateral distance. Density and velocities are the same in all cases, but the flow increases continuously with the width of the section

a very weak dependence on its width of the density and velocity inside the bottleneck. Thus in reference to Eq. (2) the flow does not necessarily depend on the number of lanes. This is consistent with common guidelines and handbooks² which assume that the capacity is a linear function of the width [35,136,150,192]. It is given by the maximum of the fundamental diagram and in reference to the specific flow concept introduced in Subsect. “Observables”, Eqs. (2), (3), the maximum grows linearly with the facility width. To find a conclusive judgment on the question if the capacity grows continuously with the width the results of different laboratory experiments [93,121,122,132,176] are compared in [176].

In the following we discuss the data of flow measurement collected in Fig. 5. The corresponding setups are sketched in Fig. 6. First, note that all presented data are taken under laboratory conditions where the test persons are advised to move normally. The data by Muir et al. [121], who studied the evacuation of airplanes (see Fig. 6b), seem to support the stepwise increase of flow with the width. They show constant flow values for $b > 0.6$ m. But the independence of flow over the large range from $b = 0.6$ m to $b = 1.8$ m indicates that in this special setup the flow is not restricted by the bottleneck width. Moreover, it was shown in [176] by determination of the trajectories that the distance between lanes changes continuously, invalidating the basic assumption leading to a stepwise increasing flow. Thus all collected data for flow measurements in Fig. 5 are compatible with a continuous

²One exception is the German MVStättV [130], see above.



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 5

Influence of the width of a bottleneck on the flow. Experimental data [121,122,132,176] of different types of bottlenecks and initial conditions. All data are taken under laboratory conditions where the test persons are advised to move normally

and almost linear increase with the bottleneck width for $b > 0.6$ m.

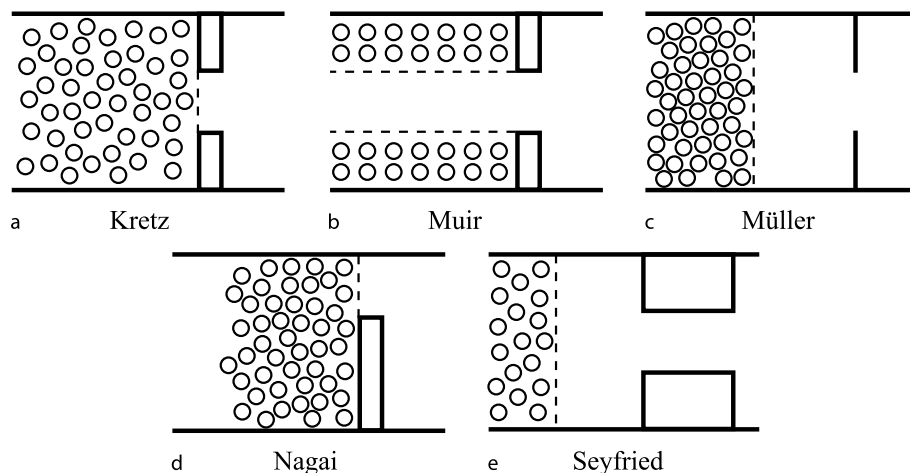
The data in Fig. 5 differ considerably in values of bottleneck capacity. In particular, the flow values of Nagai [132] and Müller [122] are much higher than the maxima of empirical fundamental diagrams (see Sub-

sect. “Fundamental Diagram”). The influence of “panic” or pushing can be excluded since in all experiments the participants were instructed to move normally. The comparison of the different experimental setups (Fig. 6) shows that the exact geometry of the bottleneck is of only minor influence on the flow, while a high initial density in front of the bottleneck can increase the resulting flow values. This is confirmed by the study of Nagai et al., see Figure 6 in [132]. There it is shown that for $b = 1.2$ m the flow grows from $J = 1.04 \text{ s}^{-1}$ to 3.31 s^{-1} when the initial density is increased from 0.4 m^{-2} to 5 m^{-2} .

The linear dependence of the flow on the width has a natural limitation due to the nonzero body-size and the asymmetry given by the sequence of movement in steps. Movement of pedestrians through bottlenecks smaller than shoulder width requires a rotation of the body. Kretz et al. found in their experiment [93] that the specific flow J_s increases if the width decreases from $b = 0.7$ m to $b = 0.4$ m.

Connection Between Bottleneck Flow and Fundamental Diagrams

An interesting question is how the bottleneck flow is connected to the fundamental diagram. General results for driven diffusive systems [149] show that boundary conditions only *select* between the states of the undisturbed system instead of creating completely different ones. Therefore it is surprising that the measured maximal flow at bottlenecks can exceed the maximum of the empirical fundamental diagram. These questions are related to the common jamming criterion. Generally, it is assumed that a jam occurs if the incoming flow exceeds the capacity of the bottleneck. In this case one expects the



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 6

Outlines of the experimental arrangements under which the data shown in Fig. 5 were taken

flow through the bottleneck to continue with the capacity (or lower values).

The data presented in [176] show a more complicated picture. While the density in front of the bottleneck amounts to $\rho \approx 5.0(\pm 1) \text{ m}^{-2}$, the density inside the bottleneck tunes around $\rho \approx 1.8 \text{ m}^{-2}$. The observation that the density inside the bottleneck is lower than in front of the bottleneck is consistent with measurements of Daamen and Hoogendoorn [20] and the description given by Predtechenskii and Milinskii in [150]. The latter assumes that in the case of a jam the flow through the bottleneck is determined by the flow in front of the bottleneck. The density inside the jam will be higher than the density associated with the capacity. Thus the reduced flow in front of the bottleneck causes a flow through the bottleneck smaller than the bottleneck capacity. Correspondingly the associated density is also smaller than that at capacity. But the discussion above cannot explain why the capacities measured at bottlenecks are significantly higher than the maxima of empirical fundamental diagrams and cast doubts on the common jamming criterion. Possible unconsidered influences are stochastic flow fluctuations, non-stationarity of the flow, flow interferences due to the necessity of local organization or changes of the incentive during the access into the bottleneck.

Blockages in Competitive Situations As stated above all data collected in Fig. 5 are gained by runs where the test persons were instructed to move normally. By definition a bottleneck is a limited resource and it is possible that under competitive situations pedestrian flow through bottlenecks is different from the flow in normal situations. One qualitative difference to normal situations is the occurrence of blockages. Regarding the term ‘panic’ one has to bear in mind that for the occurrence of blockages some kind of reward is essential, while the emotional state of the test persons is not. This was a result of a very interesting and often cited study by Mintz [113]. First experiments with real pedestrians have been performed by Dieckmann [24] in 1911 as a reaction to many fatalities in theater fires at the end of the 19th century. In these small scale experiments test persons were instructed to go through great trouble to pass the door as fast as possible. Even in the first run he observed a stable “wedging”. In [150] it is described how these obstruction occurs due to the formation of arches in front of the door under high pressure. This is very similar to the well-known phenomenon of *arching* occurring in the flow of granular materials through narrow openings [194].

Systematic studies including the influence of the shape and width of the bottleneck and comparisons with flow

values under normal situations have been performed by Müller and Muir et al. [121,122]. Müller found that funnel-like geometries support the formation of arches and thus blockages. For further discussion, one must distinguish between temporary blockages and stable blockages leading to a zero flow. For the setup sketched in Fig. 6c Müller found that temporary blockages occur only for $b < 1.8 \text{ m}$. For $b \leq 1.2 \text{ m}$ the flow shows strong pulsing due to unstable blockages. Temporal disruptions of the flow appear for $b \leq 1.0 \text{ m}$. In comparison to normal situations the flow is higher, and in general the occurrence of blockages decreases with width. However a surprising result is that for narrow bottlenecks, increasing the width can be counterproductive since it also increases the probability of blockages. Muir et al. for example note that in their setup (Fig. 6b) the enlargement of the width from $b = 0.5 \text{ m}$ to $b = 0.6 \text{ m}$ leads to an increase of temporary blockages. The authors explain this by differences in the perception of the situation by the test persons. While the smaller width is clearly passable only for one person, the wider width may lead to the perception that the bottleneck is sufficiently wide to allow two persons to pass through. How many people have direct access to the bottleneck is clearly influenced by the width of the corridor in front of the bottleneck. Also, Müller found hints that flow under competitive situations did not increase in general with the bottleneck width. He notes an optimal ratio of 0.75:1 between the bottleneck width and the width of the corridor in front of the bottleneck.

To reduce the occurrence of blockages, and thus evacuation times, Helbing et al. [54,55,83] suggested putting a column (asymmetrically) in front of a bottleneck. It should be emphasized that this theoretical prediction was made under the assumption that the system parameters, i. e., the basic behavior of the pedestrians, does not change in the presence of the column. This is highly questionable in real situations where a column can be perceived as an additional obstacle or can make it difficult to find the exit. In experiments [57] an increase of the flow of about 30% has been observed for a door with $b = 0.82 \text{ m}$. But this experiment was performed only for one width and the discussion above indicates the strong influence of the specific setup used. Independent of this uncertainty this concept is limited, as the occurrence of stable arches, to narrow bottlenecks. In practice narrow bottlenecks are not suitable for a large number of people and an opening in a room has other important functionalities, which would be restricted by a column.

Another finding is the observation that the total flow at bottlenecks with bidirectional movement is higher than it is for unidirectional flows [57].

Stairs

In most evacuation scenarios stairs are important elements that are a major determinant for the evacuation time. Due to their physical dimension, which is often smaller than other parts of a building, or due to a reduced walking speed, stairs generally must be considered as bottlenecks for the flow of evacuees. For the movement on stairs, just as for the movement on flat terrain, the fundamental diagram is of central interest. Compared to the latter there are more degrees of freedom, which influence the fundamental diagram:

- One has to distinguish between upward and downward movement.
- The influence of riser height and tread width (which determine the incline) has to be taken into account.
- For upward motion exhaustion effects lead to a strong time dependence of the free speed.

It is probably a consequence of the existence of a continuum of fundamental diagrams in dependence of the incline that there are no generally accepted fundamental diagrams for movement on stairs. However, there are studies on various details—mostly the free speed—of motion on stairs in dependence of the incline [35,38,39,46], conditions (comfortable, normal, dangerous) [151], age and sex [35], tread width [33], and the length of a stair [95]; and in consideration of various disabilities [11].

In addition there are some compilations or “meta studies”: Graat [46] compiled a list of capacity measurements and Weidmann [192] built an average of 58 single studies and found an average for the horizontal upstairs speed—the speed when the motion is projected to the horizontal level—of 0.610 m/s.

Depending on various parameters, the aforesaid studies report horizontal upward walking speeds varying over a wide range from 0.391 to 1.16 m/s. Interestingly, on one and the same short stairs it could be observed [95] that people on average walked faster up- than downwards.

There is also a model where the upstairs speed is calculated from the stair geometry (riser and tread) [183] and an empirical investigation of the collision avoidance behavior on stairs [37].

On stairs (up- as well as downward) people like to put their hand on the handrail, i. e., they tend to walk close to walls, even if there is no counterflow. This is in contrast to movement on flat terrain, where at least in situations of low density there is a tendency to keep some distance from walls.

The movement on stairs is typically associated with a reduction of the walking speed. For upward motion this

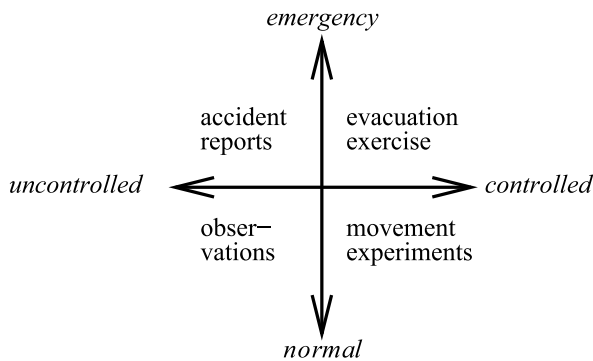
follows from the increased physical effort required. This has two aspects: first, there is the physical potential energy that a pedestrian has to supply if he wants to rise in height; second, the motion process itself is more exertive – the leg has to be lifted higher – than during motion on a level, even if this motion process is executed only on the spot. Concerning the potential energy there is no comparable effect for people going downstairs. But still one can observe jams forming at the upper end of downstairs streams. These are due to the slight hesitation that occurs when pedestrians synchronize their steps with the geometry of the (down-)stairs ahead. Therefore the bottleneck character of downstairs is less a consequence of the speed on the stairs itself and more of the transition from planar to downward movement, at least as long as the steps are not overly steep.

Evacuations: Empirical Results

Up to now this section has focused on empirical results for pedestrian motion in rather simple scenarios. As we have seen there are many open questions where no consensus has been reached, sometimes even about the qualitative aspects. This becomes even more relevant for full-scale descriptions of evacuations from large buildings or cruise ships. These are typically a combination of many of the simpler elements, so a lack of reliable information is not surprising. In the following we will discuss several complex scenarios in more detail.

Evacuation Experiments In the case of an emergency, the movement of a crowd usually is more straightforward than in the general case. Commuters in a railway station, for example, or visitors of a building might have complex itineraries which are usually represented by origin-destination matrices. In the case of an evacuation, however, the aims and routes are known and usually the same, i. e., the exits and the egress routes. This is the reason why an evacuation process is rather strictly limited in space and time, i. e., its beginning and end are well-defined: the sound of the alarm, initial position of all persons, safe areas (final position of all persons), and the time at which the last person reaches the safe area. When all people have left a building or vessel and reached a safe area (or the lifeboats or life-rafts), then the evacuation is finished. Therefore, it is also possible to perform evacuation trials and measure overall evacuation times. Before we go into details, we will clarify three different aspects of data on evacuation processes:

- (1) The definition and parts of evacuation time,
- (2) The different sources of data, and
- (3) The application of these data.



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 7

Empirical data can be roughly classified according to controlled/uncontrolled and emergency/normal situations

Concerning the evacuation time, five different phases can be distinguished [48,118,153]:

- (1) Detection time,
- (2) Awareness time,
- (3) Decision time,
- (4) Reaction time, and
- (5) Movement time.

In IMO's regulations [118,119], the first four are grouped together into *response time*. Usually, this time is called *pre-movement time*.

One possible scheme for the classification of data on evacuation processes is shown in the following Fig. 7.

Please note that not only data obtained from uncontrolled or emergency situations can be used in the context of evacuation assessment. Knowledge about bottleneck capacities (i. e., flows through doors and on stairs) is especially important when assessing the layout of a building with respect to evacuation. The purpose of empirical data in the context of evacuation processes (and modeling in general) is threefold [43,71]:

- (1) Identify parameters (factors that influence the evacuation process, e. g., bottleneck widths and capacities),
- (2) Quantify (calibrate) those parameters, e. g., flow through a bottleneck in persons per meter per second, and
- (3) Validate simulation results, e. g., compare the overall evacuation time measured in an evacuation with simulation or calculation results.

The validation is usually based on data from the evacuation of complete buildings, aircraft, trains or ships. These are available from two different sources:

- (1) Full scale evacuation trials and
- (2) Real evacuations.

Evacuation trials are usually observed and videotaped. Reports of real evacuation processes are obtained from eyewitness records and a posteriori incident investigations. Since the setting of a complete evacuation is not experimental, it is hardly possible to measure microscopic features of the crowd motion. Therefore, calibration of parameters is usually not the main purpose in evacuation trials; rather, they are carried out to gain knowledge about the overall evacuation process, the behavior of the persons, to identify the governing influences/parameters and to validate simulation results.

One major concern in evacuation exercises is the well-being of the participants. Due to practical, financial, and ethical constraints, an evacuation trial cannot be, by nature, realistic. Therefore, an evacuation exercise does not convey the increased stress of a real evacuation. To draw conclusions on the evacuation process, the walking speed observed in an exercise should not be assumed to be higher in a real evacuation [145]. Along the same lines of argument, a simplified evacuation analysis based on, e. g., a hydro-dynamic model can predict an evacuation exercise, and the same constraints apply for its results concerning the prediction of evacuation times and the evacuation process. If population parameters (such as gender, age, walking speed, etc.) are explicitly stated in the model, increased stress can be simulated by adapting these parameters.

In summary, evacuation exercises are just too expensive, time consuming, and dangerous to be a standard measure for evacuation analysis. An evacuation exercise organized by the UK Marine Coastguard Agency on the Ro-Ro ferry "Stena Invicta" held in Dover Harbor in 1996 cost more than 10,000 GBP [117]. This is one major argument for the use of evacuation simulations based on hydro-dynamic models and calculations.

Panic, Herding, and Similar Conjectured Collective Phenomena

As already mentioned earlier in Subsect. "Collective Effects", the concept of "panic" and its relevance for crowd disasters is rather controversial. It is usually used to describe irrational and unsocial behavior. In the context of evacuations, empirical evidence shows that this type of behavior is rare [3,17,77,178]. On the other hand there are indications that fear might be "contagious" [22]. Related concepts like "herding" and "stampede" imply a certain similarity between the behavior of human crowds and animal behavior. This terminology is quite often used in the public media. *Herding* has been described in animal experiments [166] and is difficult to measure in human crowds. However, it seems to be natural that herding exists in certain situations, e. g., limited

visibility due to failing lights or strong smoke when exits are hard to find.

Panic As stated earlier, “panic” behavior is usually characterized by selfish and anti-social behavior which through contagion affects large groups and even leads to completely irrational actions. Often it is assumed, especially in the media, to occur in situations where people compete for scarce or dwindling resources, which in the case of emergencies are safe space or access to an exit. However, this point of view does not stand close scrutiny and it has turned out that this behavior has played no role at all in many tragic events [73,77]. For these incidents *crowd disaster* is a much more appropriate characterization.

Furthermore, lack of social behavior seems to be more frequent during so called “acquisitive panics” or “crazes” [179] than during “flight panics”. That is, social behavior seems to be less stable if there is something to gain than if there is some external danger which threatens all members of a group. Examples of crazes (acquisitive panics) include the Victoria Hall Disaster (1883) [150], the crowning ceremony of Tsar Nicholas II (1896) [168], a governmental Christmas celebration in Aracaju (2001), the distribution of free Saris in Uttar Pradesh (2004), and the opening of an IKEA store in Jeddah (2004). Crowd accidents which occur at rock concerts and religious events as well bear more similarities with crazes than with panics.

However, it is not the case that altruism and cooperation increase with danger. The events during the capsizing of the MV Estonia (see Sect. 16.6 of [100]) show some behavioral threshold: faced with immediate life-threatening danger, most people struggle for their own survival or that of close relatives.

Herding Herding in a broad context means “go with the flow” or “follow the crowd”. Like “panic”, the term “herding” is often used in the context of stock market crashes, i. e., causing an avalanche effect. Like “panic” the term is usually not well defined and is used in an allegoric way. Therefore, it is advisable to avoid the term in a scientific context (apart from zoology, of course). Furthermore, “herding”, “stampede”, and “panic” have a strong connotation of “deindividuation”. The conjecture of an automatic deindividuation caused by large crowds [101] has been replaced by a social attachment theory (“the typical response to a variety of threats and disasters is not to flee but to seek the proximity of familiar persons and places”) [109].

Stampede Stampede is – like herding – a term from zoology where herds of large mammals, such as buffalo, collectively run in one direction and might overrun any obsta-

cles. This is dangerous for human observers if they cannot get out of the way. The term “stampede” is sometimes used for crowd accidents [73], too. It is furthermore assumed to be highly correlated with panic. When arguing along those lines, a stampede might be the result of “crowd panic” or vice versa.

Shock or Density Waves Shock waves are reported for rock concerts [180] and religious events [2,58]. They might result in people standing close to each other falling down. Pressures in dense crowds of up to 4, 450 N/m² have been reported.

Although empirical data on crowd disasters exist, e. g., in the form of reports from survivors or even video footage, it is almost impossible to derive quantitative results from them. Models that aim at describing such scenarios make predictions for certain counter-intuitive phenomena that should occur. In the faster-is-slower effect [54] a higher desired velocity leads to a slower movement of a large crowd. In the freezing-by-heating effect [53] increasing the fluctuations can lead to a more ordered state. For a thorough discussion we refer to [54,55] and references therein. However, from a statistical point of view there is insufficient data to decide the relevance of these effects in real emergency situations, not least because it is almost impossible to perform “realistic” experiments.

Sources of Empirical Data on Evacuation Processes

The evacuation of a building can either be an isolated process (due to fire restricted to this building, a bomb threat, etc.) or it can be part of the evacuation of a complete area. We will focus on the single building evacuation, here. For the evacuation of complete areas, e. g., because of flooding or hurricanes, cf. [157] and references therein.

For passenger ships, a distinction between High Speed Craft (HSC), Ro-Ro passenger ferries, and other passenger vessels (cruise ships) is made. High Speed Craft do not have cabins and the seating arrangement is similar to aircraft. Therefore, there is a separate guideline for HSC [119]. A performance-based evacuation analysis at an early stage of design is required for HSC and Ro-Pax. There is currently no such requirement for cruise ships. For an overview of IMO’s requirements and the historical development up to 2001 cf. [27]. In addition to the five components for the overall evacuation time listed above, there are three more specific to ships:

- (6) Preparation time (for the life-saving appliances, i. e., lifeboats, life-rafts, davits, chutes),
- (7) Embarkation time, and
- (8) Launching time.

Therefore, the evacuation procedure on ships is more complex than for buildings. Additionally, SAR (Search And Rescue) is an integral part of ship evacuation.

For High Speed Craft, the time limit is 17 minutes for evacuation [70], for Ro-Ro passenger ships it is 60 minutes [118], and for all other passenger ships (e.g., cruise ships) it is 60 minutes if the number of main vertical zones is less or equal to five and 80 minutes otherwise [118]. For HSC, no distinction is made between assembly and embarkation phases.

For aircraft, the approach can be compared to that of HSC. First, an evacuation test is mandatory and there is a time limit of 90 seconds that has to be complied to in the test [31].

In many countries there is no strict criterion for the maximum evacuation time of buildings. The requirements are based on minimum exit widths and maximum escape path lengths.

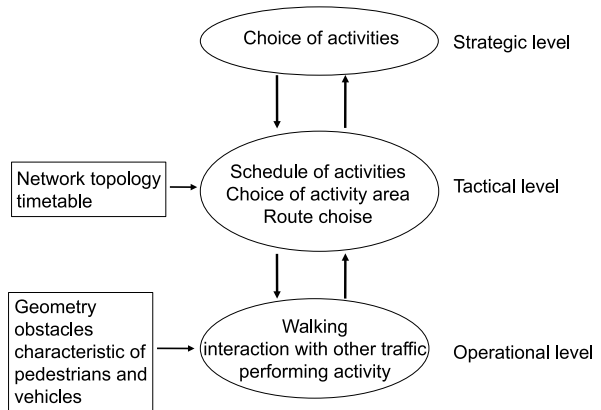
A number of real evacuations has been investigated and reports are publicly available. Among the most recent ones are: Beverly Hills Club [12], MGM Grand Hotel, [12], retail store [4], department store [1], World Trade Center [47] and www.wtc.nist.gov, high-rise buildings [144, 173], theater [191] for buildings, High Speed Craft “Sleipner” [138] for HSC, an overview up to 1998 [143], exit width variation [121], double deck aircraft [74], another overview for aircraft [120], and for trains [43,169].

Modeling

A comprehensive theory of pedestrian dynamics has to take into account three different levels of behavior (Fig. 8). At the *strategic level*, pedestrians decide which activities they like to perform and the order of these activities. With the choices made at the strategic level, the *tactical level* concerns the short-term decisions made by the pedestrians, e.g., choosing the precise route taking into account obstacles, density of pedestrians etc. Finally, the *operational level* describes the actual walking behavior of pedestrians, e.g., their immediate decisions necessary to avoid collisions etc.

Processes at the strategic and tactical level are usually considered to be exogenous to pedestrian simulation. Here information from other disciplines (sociology, psychology etc.) is required. In the following we will mostly be concerned with the operational level, although some of the models that we are going to describe allow us to take into account certain elements of behavior at the tactical level as well.

Modeling on the operational level is usually based on variations of models from physics. Indeed the motion of



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 8

The different levels of modeling pedestrian behavior (after [19, 65])

pedestrian crowds shares certain similarities with fluids and the flow of granular materials. The goal is to find models which are as simple as possible, but at the same time can reproduce “realistic” behavior in the sense that the empirical observations are reproduced. Therefore, based on the experience from physics, pedestrians are often modeled as simple “particles” that interact with each other.

There are several characteristics which can be used to classify the modeling approaches:

Microscopic vs. macroscopic In microscopic models each individual is represented separately. Such an approach allows us to introduce different types of pedestrians with individual properties as well as issues such as route choice. In contrast, in macroscopic models, individuals cannot be distinguished. Instead the state of the system is described by densities, usually a mass density derived from the positions of the persons and a corresponding locally averaged velocity.

Discrete vs. continuous Each of the three basic variables for a description of a system of pedestrians, namely space, time and state variable (e.g., velocities), can be either discrete (i.e., an integer number) or continuous (i.e., a real number). Here all combinations are possible. In a cellular automaton approach all variables are by definition discrete, whereas in hydrodynamic models all are continuous. These are the most common choices, but other combinations are used as well. Sometimes for a cellular automata approach a continuous time variable is also allowed. In computer simulation this is realized through a *random-sequential update* where at each step the particle or site to be up-

dated (moved) is chosen randomly (from *all* particles or sites, respectively). A discrete time is usually realized through a *parallel* or *synchronous update* where all particles or sites are moved at the same time. This introduces a timescale. In so-called coupled map lattices time is discrete, whereas space and state variables are continuous.

Deterministic vs. stochastic The dynamics of pedestrians can either be deterministic or stochastic. In the first case the behavior at a certain time is completely determined by the present state. In stochastic models, behavior is controlled by certain probabilities such that the agents can react differently in the same situation. This is one of the lessons learnt from the theory of complex systems where it has been shown for many examples that through introduction of stochasticity into rather simple systems very complex behavior can be generated. On the other hand, the stochasticity in the models reflects our lack of knowledge of the underlying physical processes that, e. g., determine the decision-making of the pedestrians. Through stochastic behavioral rules it often becomes possible to generate a rather realistic representation of complex systems such as pedestrian crowds.

This “intrinsic” stochasticity should be distinguished from “noise”. Sometimes external noise terms are added to the *macroscopic* observables, such as position or velocity. Often the main effect of these terms is to avoid certain special configurations which are considered to be unrealistic, like completely blocked states. Otherwise the behavior is very similar to the deterministic case. For true stochasticity, on the other hand, the deterministic limit usually has very different properties from the generic case.

Rule-based vs. force-based Interactions between the agents can be implemented in at least two different ways: In a rule-based approach agents make “decisions” based on their current situation, the nature of their neighborhood as well as their goals, etc. It focuses on the *intrinsic properties* of the agents and thus the rules are often justified from psychology. In force-based models, agents “feel” a force exerted by others and the infrastructure. They therefore emphasize *extrinsic properties* and their relevance for the motion of the agents. This is a physical approach based on the observation that the presence of others leads to deviations from a straight motion. In analogy to Newtonian mechanics a force is made responsible for these accelerations.

Cellular automata are typically rule-based models, whereas, e. g., the social-force model belongs to the

force-based approaches. However, sometimes a clear distinction cannot be made; many models combine aspects of both approaches.

High vs. low fidelity *Fidelity* here refers to the apparent realism of the modeling approach. High fidelity models try to capture the complexity of decision making, actions, etc. that constitute pedestrian motion in a realistic way. In contrast, in the simplest models pedestrians are represented by particles without any intelligence. Usually the behavior of these particles is determined by “forces”. This approach can be extended, e. g., by allowing different “internal” states of the particles so that they react differently to the same force depending on the internal state. This can be interpreted as some kind of “intelligence” and leads to more complex approaches, like multi-agent models. Roughly speaking, the number of parameters in a model is a good measure for fidelity in the sense introduced here, but note that higher fidelity does not necessarily mean that empirical observations are reproduced better!

It should be mentioned that a clear classification according to the characteristics outlined here is not always possible. In the following we will describe some model classes in more detail.

Fluid-dynamic and Gas kinetic Models

Pedestrian dynamics has some obvious similarities with fluids. For example, the motion around obstacles appears to follow “streamlines”. Motion at intermediate densities is restricted (short-ranged correlations). Therefore it is not surprising that, very much like for vehicular dynamics, the earliest models of pedestrian dynamics took inspiration from hydrodynamics or gas-kinetic theory [50,61,68,69]. Typically these macroscopic models are deterministic, force-based and of low fidelity.

Henderson [60,61] has tried to establish an analogy of large crowds with a classical gas. From measurements of motion in different crowds in a low density (“gaseous”) phase he found good agreement of the velocity distribution functions with Maxwell–Boltzmann distribution [60].

Motivated by this observation, he later developed a fluid-dynamic theory of pedestrian flow [61]. Describing the interactions between the pedestrians as a collision process where the particles exchange momenta and energy, a homogeneous crowd can be described by the well-known kinetic theory of gases. However, the interpretation of the quantities is not entirely clear, e. g., what the analogues of pressure and temperature are in the context of pedestrian

motion. Temperature could be identified with the velocity variance, which is related to the distribution of desired velocities, whereas the pressure expresses the desire to move against a force in a certain direction.

The applicability of classical hydrodynamic models is based on several conservation laws. The conservation of mass, corresponding to conservation of the total number of pedestrians, is expressed through a continuity equation of the form

$$\frac{\partial \rho(\mathbf{r}, t)}{\partial t} + \nabla \cdot \mathbf{J}(\mathbf{r}, t) = 0, \quad (6)$$

which connects the local density $\rho(\mathbf{r}, t)$ with the current $\mathbf{J}(\mathbf{r}, t)$. This equation can be generalized to include source and sink terms. However, the assumption of conservation of energy and momentum is not true for interactions between pedestrians which in general do not even satisfy Newton's Third Law ("actio = reaction"). In [50] several other differences to normal fluids were pointed out, e. g., the anisotropy of interactions or the fact that pedestrians usually have an individual preferred direction of motion.

In [50] a better founded fluid-dynamical description was derived on the basis of a gas kinetic model which describes the system in terms of a density function $f(\mathbf{r}, \mathbf{v}, t)$. The dynamics of this function are determined by Boltzmann's transport equation that describes its change for a given state as difference of inflow and outflow due to binary collisions.

An important new aspect in pedestrian dynamics is the existence of desired directions of motion which allows us to distinguish different groups μ of particles. The corresponding densities f_μ change in time due to four different effects:

1. A relaxation term with characteristic time τ describes tendency of pedestrians to approach their intended velocities.
2. The interaction between pedestrians is modeled by a Stosszahlansatz as in the Boltzmann equation. Here, pair interactions between types μ and ν occur with a total rate that is proportional to the densities f_μ and f_ν .
3. Pedestrians are allowed to change from type μ to ν which, e. g., accounts for turning left or right at a crossing.
4. Additional gain and loss terms allow us to model entrances and exits where pedestrian can enter or leave the system.

The resulting fluid-dynamic equations derived from this gas kinetic approach are similar to that of ordinary fluids.

However, due to the different types of pedestrians, corresponding to individuals who have approximately the same desired velocity, one actually obtains a set of coupled equations describing several interacting fluids. These equations contain additional characteristic terms describing the approach to the intended velocity and the change of fluid-type due to interactions in avoidance maneuvers.

Equilibrium is approached through the tendency to walk with the intended velocity, not through interactions as in ordinary fluids. Momentum and energy are not conserved in pedestrian motion, but the relaxation towards the intended velocity describes a tendency to restore these quantities.

Unsurprisingly for a macroscopic approach, the gas-kinetic models have problems at low densities. For a discussion, see e. g. [50].

Hand Calculation method For practical applications effective engineering tools have been developed from the hydrodynamical description. In engineering these are often called *hand calculation methods*. One could also classify some of them as queuing models since the central idea is to describe pedestrian dynamics as flow on a network with links of limited capacities. These methods allow us to calculate evacuation times in a relatively simple way that does not require any simulations. Parameters entering in the calculations can be adapted to the situation that is studied. Often they are based on empirical results, e. g., evacuation trials. Details about this kind of model can be found in Subsect. "Calculation of Evacuation Times".

Social-Force Models

The social-force model [52] is a deterministic continuum model in which interactions between pedestrians are implemented by using the concept of a *social force* or *social field* [103]. It is based on the idea that changes in behavior can be understood in terms of fields or forces. Applied to pedestrian dynamics, the social force $\mathbf{F}_j^{(\text{soc})}$ represents the influence of the environment (other pedestrians, infrastructure) and changes the velocity \mathbf{v}_j of pedestrian j . Thus it is responsible for acceleration which justifies the interpretation as a force. The basic equation of motion for a pedestrian of mass m_j is then of the general form

$$\frac{d\mathbf{v}_j}{dt} = \mathbf{f}_j^{(\text{pers})} + \mathbf{f}_j^{(\text{soc})} + \mathbf{f}_j^{(\text{phys})} \quad (7)$$

where $\mathbf{f}_j^{(\text{soc})} = \frac{1}{m_j} \mathbf{F}_j^{(\text{soc})} = \sum_{l \neq j} \mathbf{f}_{jl}^{(\text{soc})}$ is the total (specific) force due to other pedestrians. $\mathbf{f}_j^{(\text{pers})}$ denotes a "personal" force which makes a pedestrian attempt to move with his or her own preferred velocity $\mathbf{v}_j^{(0)}$ and thus acts as a driving

term. It is given by

$$\mathbf{f}_j^{(\text{pers})} = \frac{\mathbf{v}_j^{(0)} - \mathbf{v}_j}{\tau_j} \quad (8)$$

where τ_j is reaction or acceleration time. In high density situations, physical forces $\mathbf{f}_{jl}^{(\text{phys})}$ also become important, e.g., friction and compression when pedestrians make contact.

The most important contribution to the social force $\mathbf{f}_j^{(\text{soc})}$ comes from the territorial effect, i.e., the private sphere. Pedestrians feel uncomfortable if they get too close to others, which effectively leads to a repulsive force between them. Similar effects are observed for the environment, e.g., people prefer not to walk too close to walls.

Since social forces are difficult to determine empirically, some assumptions must be made. Usually an exponential form is assumed. Describing the pedestrians as disks of radius R_j and position (of the center of mass) \mathbf{r}_j , the typical structure of the force between the pedestrians is described by [54]

$$\mathbf{f}_{jl}^{(\text{soc})} = A_j \exp \left[\frac{R_{jl} - \Delta r_{jl}}{\xi_j} \right] \mathbf{n}_{jl} \quad (9)$$

with $R_{jl} = R_j + R_l$, the sum of the disk radii, $\Delta r_{jl} = |\mathbf{r}_j - \mathbf{r}_l|$, the distance between the centers of mass, $\mathbf{n}_{jl} = \mathbf{r}_j - \mathbf{r}_l / \Delta r_{jl}$, the normalized vector pointing from pedestrian l to j . A_j can be interpreted as strength, ξ_j as the range of the interactions.

The appeal of the social-force model is given mainly by analogy to Newtonian dynamics. For the solution of the equations of motion of Newtonian many-particle systems, well-founded molecular dynamics techniques exist. However, in most studies so far the distinctions between pedestrian and Newtonian dynamics are not discussed in detail. A straightforward implementation of the equations of motion neglecting these distinctions can lead to unrealistic movement of single pedestrians. For example, negative velocities in the main moving direction cannot be excluded in general even if asymmetric interactions (violating Newton's Third Law) between the pedestrians are chosen. Another effect is the occurrence of velocities higher than the preferred velocity $v_j^{(0)}$ due to the forces on pedestrians in the moving direction. To prevent this effect, additional restrictions for the degrees of freedom must be introduced, see for example [52], or the superposition of forces has to be discarded [175]. A general discussion of the limited analogy between Newtonian dynamics and the social-force model as well as the consequences for model implementation is still missing.

Apart from the ad hoc introduction of interactions, the structure of the social-force model can also be derived from an extremal principle [62,63]. It follows under the assumption that pedestrian behavior is determined by the desire to minimize a certain cost function which takes into account not only kinematic aspects and walking comfort, but also deviations from a planned route.

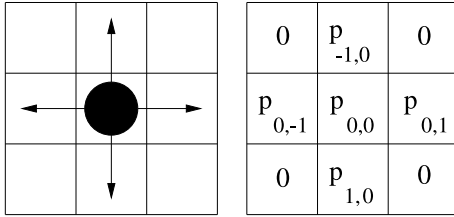
Cellular Automata

Cellular automata (CA) are rule-based dynamical models that are discrete in space, time and state variable which in the case of traffic usually corresponds to velocity. Discreteness in time means that the positions of the agents are updated in well defined steps. In computer simulations this is realized through a *parallel* or *synchronous* update where all pedestrians move at the same time. The time step corresponds to a natural timescale Δt which could be identified, e.g., with some reaction time. This can be used for the calibration of the model which is essential for making quantitative predictions. A natural space discretization can be derived from the maximal densities observed in dense crowds which gives the minimal space requirement of one person. Usually each cell in the CA can be occupied by only one particle (exclusion principle) so that this space requirement can be identified with the cell size. In this way, a maximal density of 6.25 P/m^2 [192] leads to a cell size of $40 \times 40 \text{ cm}^2$. Sometimes finer discretizations are more appropriate (see Subsect. "Theoretical Results"). In this case pedestrians correspond to extended particles that occupy more than one cell (e.g., four cells). The exclusion principle and the modeling of humans as non-compressible particles mimics short-range repulsive interactions, i.e., the "private-sphere".

The dynamics are usually defined by rules which specify transition probabilities for motion to one of the neighboring cells (Fig. 9). The models differ in the specification of these probabilities as well in that of the "neighborhood". For deterministic models, all but one are of probability zero.

The first cellular automata (CA) models [7,41,89,129] for pedestrian dynamics can be considered two-dimensional variants of the asymmetric simple exclusion process (ASEP) (for reviews, see [9,23,172]) or models for city or highway traffic [6,16,133] based on it. Most of these models represent pedestrians by particles without any internal degrees of freedom. They can move to one of the neighboring cells based on certain transition probabilities which are determined by three factors:

- (1) The desired direction of motion, e.g., to find the shortest connection,



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 9

A particle, its possible directions of motion and the corresponding transition probabilities p_{ij} for the case of a von Neumann neighborhood

- (2) Interactions with other pedestrians, and
- (3) Interactions with the infrastructure (walls, doors, etc.).

Fukui–Ishibashi Model One of the first CA models for pedestrian dynamics was proposed by Fukui and Ishibashi [40,41] and is based on a two-dimensional variant of the ASEP. They studied bidirectional motion in a long corridor where particles moving in opposite directions were updated alternately. Particles move deterministically in their desired direction; only if the desired cell is occupied by an oppositely moving particle do they make a random sidestep.

Various extensions and variations of the model have been proposed, e.g., an asymmetric variant [129] where walkers prefer lane changes to the right, different update types [193], simultaneous (exchange) motion of pedestrians standing “face-to-face” [72], or the possibility of backstepping [107]. The influence of the shape of the particles has been investigated in [131]. Also other geometries [128, 181] and extensions to full 2-dimensional motion have been studied in various modifications [106,107,127]

Blue–Adler Model The model of Blue and Adler [7,8] is based on a variant of the Nagel–Schreckenberg model [133] of highway traffic. Pedestrian motion is considered in analogy to a multi-lane highway. The structure of the rules is similar to the basic two-lane rules suggested in [159]. The update is performed in four steps which are applied to all pedestrians in parallel. In the first step each pedestrian chooses a preferred lane. In the second step the lane changes are performed. In the third step the velocities are determined based on the available gap in the new lanes. Finally, in the fourth step the pedestrians move forward according to the velocities determined in the previous step.

In counterflow situations head-on-conflicts occur. These are resolved stochastically and with some probability opposing pedestrians are allowed to exchange positions within one time step. Note that the motion of a sin-

gle pedestrian (not interacting with others) is deterministic otherwise.

Unlike the Fukui–Ishibashi model, motion is not restricted to nearest-neighbor sites. Instead, pedestrians can have different velocities v_{\max} which correspond to the maximal number of cells they are allowed to move forward. In contrast to vehicular traffic, acceleration to v_{\max} can be assumed to be instantaneous in pedestrian motion.

In order to study the effects of inhomogeneities, the pedestrians are assigned different maximal velocities v_{\max} . Fast walkers have $v_{\max} = 4$, standard walkers $v_{\max} = 3$ and slow walkers $v_{\max} = 2$. The cell size is assumed to be 50 cm \times 50 cm. The best agreement with empirical observations has been achieved with 5% slow and 5% fast walkers [8]. Furthermore the fundamental diagram in more complex situations, such as bi- or four-directional flows, has been investigated.

Gipps–Marksjös Model A more sophisticated discrete model was suggested by Gipps and Marksjös [45] in 1985. One motivation for developing a discrete model was the limited computer power at that time. Therefore a discrete model, which reproduces the properties of pedestrian motion realistically, was in many respects a real improvement over the existing continuum approaches.

Interactions between pedestrians are assumed to be repulsive, anticipating the idea of social forces (see Subsect. “Social-Force Models”). The pedestrians move on a grid of rectangular cells of size 0.5 \times 0.5 m. To each cell a score is assigned based on its proximity to other pedestrians. This score represents the repulsive interactions and actual motion is then determined by the competition between these repulsions and the gain of approaching the destination. Applying this procedure to all pedestrians, a potential value is assigned to each cell which is the sum of the individual contributions. The pedestrian then selects the cell of its nine neighbors (Moore neighborhood) which leads to the maximum benefit. This benefit is defined as the difference between the gain of moving closer to the destination and the cost of moving closer to other pedestrians as represented by the potential. This requires a suitable chosen gain function P .

The updating is done sequentially to avoid conflicts of several pedestrians trying to move to the same position. In order to model different velocities, faster pedestrians are updated more frequently. Note that the model dynamics are deterministic.

Floor Field CA Floor field CA [13,14,83,167] can also be considered as an extension of the ASEP. However, the transition probabilities to neighboring cells are no longer

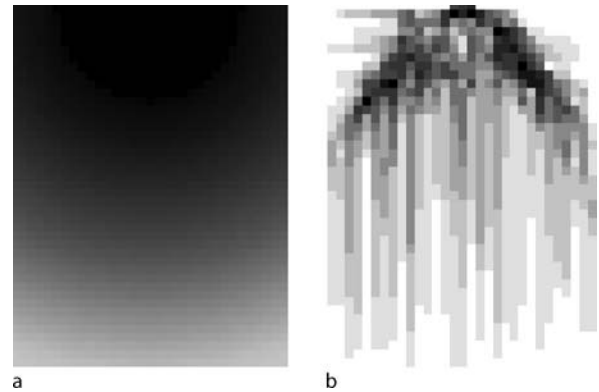
fixed but vary dynamically. This is motivated by the process of chemotaxis (see [5] for a review) used by some insects (e. g., ants) for communication. They create a chemical trace to guide other individuals to food sources. In this way a complex trail system is formed that has many similarities with human transport networks.

In the approach of [13] the pedestrians also create a trace. In contrast to chemotaxis, however, this trace is only virtual, although one could assume that it corresponds to some abstract representation of the path in the mind of the pedestrians. Although this is mainly a technical trick which reduces interactions to local ones that allow efficient simulations in arbitrary geometries, one could also think of the trail as representation of the paths in the mind of a pedestrian. The locality becomes important in complex geometries as no algorithm is required to check whether the interaction between particles is screened by walls, etc. The number of interaction terms always grows linearly with the number of particles.

The translation into local interactions is achieved by the introduction of so-called *floor fields*. The transition probabilities for all pedestrians depend on the strength of the floor fields in their neighborhood in such a way that transitions in the direction of larger fields are preferred. The *dynamic floor field* D_{ij} corresponds to a virtual trace which is created by the motion of the pedestrians and in turn influences the motion of other individuals. Furthermore it has its own dynamics, namely through diffusion and decay, which leads to a dilution and finally the vanishing of the trace after some time. The *static floor field* S_{ij} does not change with time since it only takes into account the effects of the surroundings. Therefore it exists even without any pedestrians present. It allows us to model, e. g., preferred areas, walls and other obstacles. Figure 10 shows the static floor field used for the simulation of evacuations from a room with a single door. Its strength decreases with increasing distance from the door. Since the pedestrians prefer motion into the direction of larger fields, this is already sufficient to find the door.

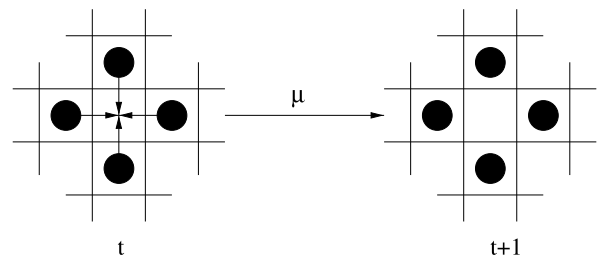
Coupling constants control the relative influence of both fields. For a strong coupling to the static field pedestrians will choose the shortest path to the exit. This corresponds to a ‘normal’ situation. A strong coupling to the dynamic field implies a strong herding behavior where pedestrians try to follow the lead of others. This often happens in emergency situations.

The model uses a fully parallel update. Therefore conflicts can occur where different particles choose the same destination cell. This is relevant for high density situations and happens in all models with parallel update if motion in different directions is allowed. Conflicts have been con-



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 10

Left: Static floor field for the simulation of an evacuation from a large room with a single door. The door is located in the middle of the upper boundary and the field strength increases with increasing intensity. **Right:** Snapshot of the dynamical floor field created by people leaving the room



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 11

Refused movement due to the friction parameter μ (for $m = 4$)

sidered a technical problem for a long time and usually the dynamics have been modified in order to avoid them. The simplest method is to update pedestrians sequentially instead of using fully parallel dynamics. However, this leads to other problems, e. g., the identification of the relevant timescale. Therefore it has been suggested in [84,85] to take these conflicts seriously as an important part of the dynamics.

For the floor field model it has been shown in [85] that the behavior becomes more realistic if not all conflicts are resolved by allowing one pedestrian to move while the others stay at their positions. Instead with probability $\mu \in [0, 1]$, which is called the friction parameter, the movement of *all* involved pedestrians is denied [85] (see Fig. 11).

This allows one to describe clogging effects between the pedestrians in a much more detailed way [85]. μ works as some kind of local pressure between the pedestrians.

If μ is high, the pedestrians handicap each other trying to reach their desired target sites. This local effect can have enormous influence on macroscopic quantities like flow and evacuation time [85]. Note that the kind of friction introduced here only influences interacting particles, not the average velocity of a freely moving pedestrian.

Surprisingly, the qualitative behaviors of the floor field model and the social-force models are very similar despite the fact that the interactions are very different. In the floor field model interactions are attractive, whereas in the social-force model they are repulsive. However, in the latter interactions are between particle densities. In contrast, in the floor field model the particle density interacts with the velocity density.

Other Approaches

Lattice-gas models In 1986, Frisch, Hasslacher, and Pomeau [34] showed that one does not have to take into account detailed molecular motion within fluids in order to obtain a realistic picture of (2d) fluid dynamics. They proposed a lattice gas model [164,165] on a triangular lattice with hexagonal symmetry, which is similar in spirit to CA models, but the exclusion principle is relaxed: particles with different velocities are allowed to occupy the same site. Note that the allowed velocities differ only in the direction, not absolute value. The dynamics are based on a succession of collision and propagation that can be chosen in such a way that the coarse-grained averages of this microscopic dynamic is asymptotically equivalent to the Navier–Stokes equations of incompressible fluids.

In [108] a kind of mesoscopic approach inspired by these lattice gas models has been suggested as a model for pedestrian dynamics. In analogy with the description of transport phenomena in fluids (e. g., the Boltzmann equation) the dynamics are based on a succession of collision and propagation.

Pedestrians are modeled as particles, moving on a triangular lattice, which have a preferred direction of motion \mathbf{c}_F . However, the particles do not strictly follow this direction but also have a tendency to move with the flow. Furthermore, at high densities the crowd motion is influenced by a kind of friction which slows down the pedestrians. This is achieved by reducing the number of individuals allowed to move to neighboring sites.

As in a lattice gas model [165], the dynamics now consists of two steps. In the *propagation step* each pedestrian moves to the neighbor site in the direction of its velocity vector. In the *collision step* the particles interact and new velocities (directions) are determined. In contrast to physical systems, momentum, etc., does not need to be con-

served during the collision step. These considerations lead to a collision step that takes into account the favorite direction \mathbf{c}_F , the local density (the number of pedestrians at the collision site), and a quantity called mobility at all neighbor sites which is a normalized measure of the local flow after the collision.

Optimal-Velocity Model The optimal velocity (OV) model originally introduced for the description of highway traffic can be generalized to higher dimensions [134] which allows its application to pedestrian dynamics.

In the two-dimensional extension of the OV model the equation of motion for particle i is given by

$$\frac{d^2}{dt^2}\mathbf{x}_i(t) = a \left\{ \mathbf{V}_0 + \sum_j \mathbf{V}(\mathbf{x}_j(t) - \mathbf{x}_i(t)) - \frac{d}{dt}\mathbf{x}_i(t) \right\}, \quad (10)$$

where $\mathbf{x}_i = (x_i, y_i)$ is the position of particle i . It can be considered as a special case of the general social-force model (7) without physical forces. The optimal-velocity function

$$\mathbf{V}(\mathbf{x}_j - \mathbf{x}_i) = f(r_{ij})(1 + \cos \varphi)\mathbf{n}_{ij}, \quad (11)$$

$$f(r_{ij}) = \alpha \{ \tanh \beta(r_{ij} - b) + c \}, \quad (12)$$

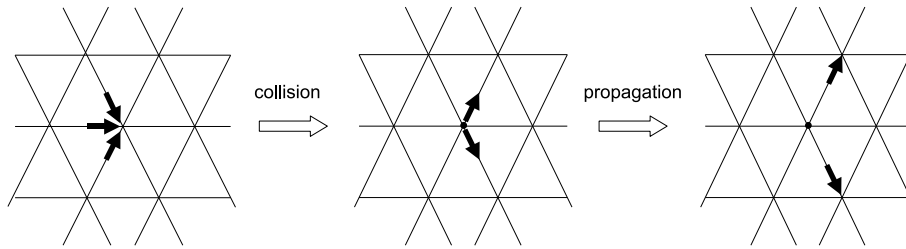
where $r_{ij} = |\mathbf{x}_j - \mathbf{x}_i|$, $\cos \varphi = (\mathbf{x}_j - \mathbf{x}_i)/r_{ij}$ and $\mathbf{n}_{ij} = (\mathbf{x}_j - \mathbf{x}_i)/r_{ij}$ is determined by interactions with other pedestrians. \mathbf{V}_0 is a constant vector that represents a ‘desired velocity’ at which an isolated pedestrian would move. The strength of the interaction depends on the distance r_{ij} between the i th and j th particles, and on the angle φ between the directions of $\mathbf{x}_j - \mathbf{x}_i$ and the current velocity $d\mathbf{x}_i/dt$. Due to the term $(1 + \cos \varphi)$, a particle reacts more sensitively to particles in front than to those behind.

Now two cases can be distinguished: repulsive and attractive interactions. The former is relevant for pedestrian dynamics whereas the latter is more suitable for biological motion. Therefore, for pedestrian motion one chooses $c = 1$ which implies $f < 0$.

A detailed analysis [134] shows that the model exhibits a rich phase diagram including the formation of various patterns.

Other Models We briefly mention a few other model approaches that have been suggested. In [10] a discretized version of the social-force model has been introduced and shown to reproduce qualitatively the observed collective phenomena.

In [141] a magnetic force model has been proposed where pedestrians and their goals are treated as magnetic poles of opposite sign.



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 12

The dynamics of lattice gas models proceed in two steps. Pedestrians coming from neighboring sites interact in the collision step where velocities are redistributed. In the propagation step the pedestrians move to neighbor sites in directions determined by the collision step

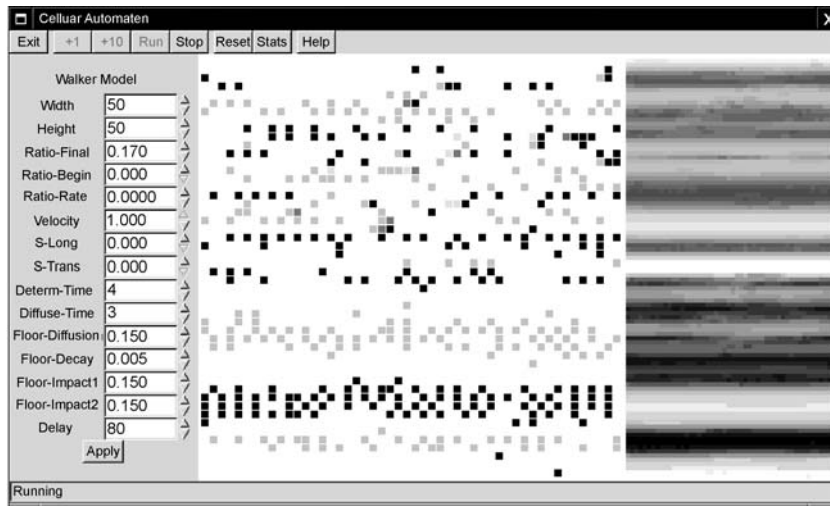
Another class of models is based on ideas from queuing theory. In principle, some hand calculation methods can be considered as macroscopic queuing models. Typically, rooms are represented as nodes in the queuing network and links correspond to doors. In microscopic approaches, in the movement process each agent chooses a new node, e. g., according to some probability [105].

Theoretical Results

As emphasized in Subsect. “Collective Effects”, the collective effects observed in the motion of pedestrian crowds are a direct consequence of microscopic dynamics. These effects are reproduced quite well by some models, e. g., the social-force and floor-field model, at least on a qualitative level. As mentioned before, the qualitative behavior of the two models is rather similar despite the very different im-

plementation of the interactions. This indicates a certain robustness of the collective phenomena observed.

As an example we discuss the formation of lanes in counterflow formation. Empirically one observes a strong tendency to follow immediately in the “wake” of another person heading in the same direction. Such lane formation was reproduced in the social-force model [52,53] as well as in the floor-field model [13,76] (see Fig. 13). While the formation of lanes in general is essential to avoid deadlocks and thus keep the chance of reproducing realistic fluxes, the number of direction changes per meter cross section is a parameter which in reality crucially depends on the situation [76]: the longer a counterflow situation is assumed to persist, the fewer lanes per meter cross section can be found. The correct reproduction of counterflow is an issue for an accommodating animation, but more or less unimportant for macroscopic observables. This is probably the



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 13

Lane formation in the floor-field model. The central window is the corridor and the light and dark squares are right- and left-moving pedestrians, respectively. In the bottom part well-separated lanes can be observed whereas in the top part the motion is still disordered. The right part of the figure shows the floor fields for the right-movers (upper half) and left-movers (lower half)

main reason why there seems to have been little effort put into the attempt to reproduce different “kinds” of lane formation in a controlled, situation-dependent manner.

On the quantitative side, the fundamental diagram is the first and most serious test for any model. Since most quantitative results rely on the fundamental diagram, it can be considered the most important characteristic of pedestrian dynamics. It is not only relevant for movement in a corridor or through a bottleneck, but also as an important determinant of evacuation times. However, as emphasized earlier, there is currently no consensus on the empirical form of the fundamental diagram. Therefore, a calibration of the model parameters is currently difficult.

Most cellular automata models are based on the asymmetric simple exclusion process. This strictly one-dimensional stochastic process has a fundamental diagram which is symmetric around density $\rho = 1/2$. Lane changes in two-dimensional extensions lead to only a small shift towards smaller densities. Despite the discrepancies in the empirical results, an almost symmetric fundamental diagram can be excluded.

Based on the experience with modeling of highway traffic [16,133], models with higher speeds have been introduced which naturally lead to an asymmetric fundamental diagram. Typically this is implemented by allowing the agents to move more than one cell per update step [82, 86,87,92,195,196]. These model variants have been shown to be flexible enough to reproduce, e.g., Weidmann’s fundamental diagram for the flow in a corridor [192] with high precision. Usually in the simulations a homogeneous population is assumed. However, in reality, different pedestrians have different properties such as walking speed, motivation, etc. This is easily taken into account in every microscopic model. There are many parameters that could potentially have an influence on the fundamental diagram. However, the current empirical situation does not allow to decide this question.

Another problem occurring in CA models has its origin in the discreteness of space. Through the choice of the lattice discretization, space is no longer isotropic. Motion in directions not parallel to the main axis of the lattice are difficult to realize and can only be approximated by a sequence of steps parallel to the main directions.

Higher velocities also require the extension of the neighborhood of a particle which is no longer identical to the cells adjacent to the current position. A natural definition of “neighborhood” corresponds to those cells that could be reached within one time step. In this way the introduction of higher velocities also reduces the problem of space isotropy as the neighborhoods become more isotropic for larger velocities.

Other solutions to this problem have been proposed. One way is to count the number of diagonal steps and let the agent suspend from moving following certain rules which depend on the number of diagonal steps [171]. A similar idea is to sum up the real distance that an agent has moved during one round: a diagonal step counts $\sqrt{2}$ and a horizontal or vertical step counts 1. An agent has to finish its round as soon as this sum is bigger than its speed [87]. A third possibility – which works for arbitrary speeds – is to assign selection probabilities to each of the four lattice positions adjacent to the exact final position [195,196]. Naturally these probabilities are inversely proportional to the square area between the exact final position and the lattice point, as in this case the probabilities are normalized by construction if one has a square lattice with points on all integer number combinations. However, one also could think of other methods to calculate the probability.

For the social-force model, the specification of the repulsive interaction (with and without hard core, exponential or reciprocal with distance) as well as the parameter sets for the forces changes in different publications [52,53, 54,114]. In [55] the authors state that “most observed self-organization phenomena are quite insensitive to the specification of interaction forces”. However, at least for the fundamental diagram, a relation connected with all phenomena in pedestrian dynamics, this statement is questionable. As remarked in [56] the reproduction of the fundamental diagram “requires a less simple specification of the repulsive interaction forces”. Indeed in [175] it was shown that the choice of hard-core forces or repulsive soft interactions as well as the particular parameter set can strongly influence the resulting fundamental diagram regarding qualitative as well as quantitative effects.

Also a more realistic behavior at higher densities requires a modification of the basic model. Here the use of density-dependent desired velocities leads to a reduction of the otherwise unrealistically large number of collisions [10].

The particular specification of forces and the previously mentioned problem with Newton’s Third law can lead in principle to some unwanted effects, such as momentary velocities larger than the preferred velocity [52] or the penetration of pedestrians into each other or into walls [98]. It is possible that these effects can be suppressed for certain parameter sets by contact or friction forces, but the general appearance is not excluded. Only in the first publication [52] are restrictions for the velocity explicitly formulated to prevent velocities larger than the intended speed; other authors tried to improve the model by introducing more parameters [98]. But additional param-

ter and artificial restrictions of variables diminish the simplicity and thus the attractiveness of the model. A general discussion of how to deal with these problems of the social-force model and a verification that the observed phenomena are not limited to a certain specification of the interaction and a special parameter set is up to now still missing.

While realistic reproduction within the empirical range of these macroscopic observables, especially the fundamental diagram, is absolutely essential to guarantee safety standards in evacuation simulations, and while a user should always be distrustful of models where no fundamental diagram has ever been published, it is by no means sufficient to exclusively check for the realism of macroscopic observables. On the microscopic level there are a large number of phenomena which need to be reproduced realistically, be it just to make a simulation animation look realistic or because microscopic effects can often easily influence macroscopic observables.

If one compares simulations of bottleneck flows with real events, one observes that in simulations the form of the queue in front of bottlenecks is often a half-circle, while in reality it is drop- or wedge-shaped. In most cases this discrepancy probably does not have an influence on the simulated evacuation time, but it is interesting to note where it originates from. Most simulation models implicitly or explicitly use some kind of utility maximization to steer the pedestrians – with the utility being foremost inversely proportional to the distance from the nearest exit. This obviously leads to half-circle-shaped queues in front of bottlenecks. So wherever one observes queues different than half-circles, people have exchanged their normal “utility function based on the distance” with something else. One such alternative utility function could be that people are just curious about what is inside or behind the bottleneck, so they seek a position where they can look into it. A more probable explanation would be that in any case it is the time distance not the spatial distance which is sought to be minimized. As anyone knows what the inescapable loss in time a bottleneck means for the whole waiting group, the precise waiting spot is not that important. However, in societies with a strong feeling for equality, people would strongly wish to equally distribute the waiting time and keep a first-in-first-out principle, which can best be accomplished and controlled when the queue is more or less one-dimensional, respectively just as wide as the bottleneck itself.

Finally it should be mentioned that theoretical investigations based on simulations of models for pedestrian dynamics have led to the prediction of some surprising and counter-intuitive collective phenomena, such as the

reduction of evacuation times through additional columns near exits (see Subsect. “[Bottleneck Flow](#)”) or the faster-is-slower [54] and freezing-by-heating effect [53]. However, so far the empirical evidence for the relevance or even the occurrence of these effects in real situations is rather scarce.

Applications

In the following section we discuss more practical aspects of based on the modeling concepts presented in Sect. “[Modeling](#)”. Tools of different sophistication have been developed that are nowadays routinely used in safety analysis. The latter becomes more and more relevant since many public facilities must fulfill certain legal standards. As an example we mention aircrafts which must be evacuated within 90 seconds. The simulations etc. are already used in the planning stages because changes of the design at a later stage are difficult and expensive.

For this kind of safety analysis tools of different sophistication have been developed. Some of them mainly are able to predict just evacuation times whereas others are based on microscopic simulations which allow also to study various external influences (fire, smoke, ...) in much detail.

Calculation of Evacuation Times

The basic idea of hand calculation methods has already briefly been described at the end of Subsect. “[Fluid-Dynamic and Gas Kinetic Models](#)”. Here we want to discuss its practical aspects in more detail.

The approach has been developed since the middle of the 1950s [185]. The basic idea of these methods is the assumption that people can be modeled to behave like fluids. Knowledge of the flow (see Eq. 1) and the technical data of the facility are then sufficient to evaluate evacuation times, etc.

Hand calculation method can be divided into two major approaches: methods with “dynamic” flow [35,42,78,79,80,136,151,152,163,192] and methods with “fixed” flow [110,123,124,125,126,137,145,173,185]. As methods with “dynamic” flow we cite methods where the pedestrian flow is dependent on the density of the pedestrian stream (see Subsect. “[Observables](#)”) in the selected facility, thus the flow can be obtained from fundamental diagrams (see Subsect. “[Fundamental Diagram](#)”) or it is explicitly prescribed in the chosen method. This flow can change during movement through the building, e. g., by using stairs, thus the pedestrian stream has a “dynamic” flow. Methods with “fixed” flow do not use this concept of relationship between density and flow. In these methods selected

facilities (e. g., stairs or doors) have a fixed flow which is independent from the density, which is usually not used in these methods. The “fixed” flow is usually based upon empirical and measured data of flow, which are specified for a special type of building, such as high-rise buildings or railway stations, for example. Because of much simplification, in these “fixed” flow methods a calculation can always be done very quickly.

Methods with “dynamic” flow allow one to describe the condition of the pedestrian flow in every part of a selected building or environment, because they are mostly based upon the continuity equation, thus it is possible to calculate different kinds of buildings. This allows the user to calculate transitions from wide to narrow, floor to door, floor to stair, etc. The disadvantage is that some these methods are very elaborate and time-intensive. But in general, a method with “dynamic” flow is not complicated to calculate, thus we want to divide hand calculation methods in simple [35,42,110,123,124,125,126,136,137,145,152,163,173,185,192] and complex [78,79,80,151] for evacuation calculation. All of these hand calculation methods are able to predict total evacuation times for a selected building, but differences between different methods still exist. Thus the user has to ensure that he is familiar with assumptions made by each method to ensure that a result is interpreted in a correct way [161].

Simulation of Evacuation Processes

Before we go into the details of evacuation simulation, let us briefly clarify its scope and limitations and contrast it to other methods used in evacuation analysis. When analyzing evacuation processes, three different approaches can be identified:

- (1) Risk assessment,
- (2) Optimization, and
- (3) Simulation.

The aim and result of risk-assessment is a list of events and their consequences (e. g., damage, financial loss, loss of life), i. e., usually an event tree with probabilities and expectation values for financial loss. Optimization aims at, roughly speaking, minimizing the evacuation time and reducing the area and duration of congestion. And finally, simulation describes a system with respect to its function and behavior by investigating a model of the system. This model is usually non-analytic, so does not provide explicit equations for the calculation of, e. g., evacuation time. Of course, simulations are used for “optimization” in a more general sense, too, i. e., they can be part of an optimization. This holds for risk assessment, too, if simulations are used

to determine the outcomes of the different scenarios in the event tree.

In evacuation analysis the system is, generally speaking, a group of persons in an environment. More specifically, four components (sub-systems/sub-models) of the system *evacuation process* can be identified:

- (1) Geometry,
- (2) Environment,
- (3) Population, and
- (4) Hazards [43].

Any evacuation simulation must at least take into account (1) and (3). The behavior of the persons (which can be described on the strategic, tactical, and operational level—see Sect. “Modeling”) is part of the population sub-model. An alternative way of describing behavior is according to its algorithmic representation: no behavior modeling – functional analogy – implicit representation (equation) – rule based – artificial intelligence [43].

In the context of evacuation, hazards are first of all fire and smoke, which then require a toxicity sub-model, e. g., the fractional effective dose model (FED), to assess the physiological effect of toxic gases and temperature [25]. Further hazards to take into account might be earthquakes, flooding, or in the case of ships, list, heel, or roll motion. The sub-model environment comprises all other influences that affect the evacuation process, e. g., exit signs, surface texture, public address system, etc.

In summary, aims of an evacuation analysis and simulation are to provide feedback and hints for improvement at an early stage of design, information for safer and more rigorous regulations, improvement of emergency preparedness, training of staff, and accident investigation [43]. They usually do not provide direct results on the probability of a scenario or a systematic search for optimal geometries.

Calculation of Overall Evacuation Time, Identification of Congestion, and Corrective Actions The scope of this section is to show general results that can be obtained by evacuation simulations. They are general in the sense that they can basically be obtained by any stochastic and microscopic model, i. e., apart from these two requirements, the results are not model specific. In detail, five different results of evacuation simulations can be distinguished:

- (1) Distribution of evacuation times,
- (2) Evacuation curve (number of persons evacuated vs. time),

- (3) Sequence of the evacuation (e.g., snapshots/screenshots at specific times, e.g., every minute), and
- (4) Identification of congestion, usually based on density and time.

The last point (4), in particular, needs some more explanation: congestion is defined based on density. Notwithstanding the difficulties of measuring density, we suggest density as the most suitable criterion for the identification of congestion. In addition to the mere occurrence of densities exceeding a certain threshold (say 3.5 persons per square meter), the time this threshold is exceeded is another necessary condition for a sensible definition of congestion. In the case presented here, 10% of the overall evacuation time is used. Both criteria are in accordance with the IMO regulations [118].

Based on these results, evacuation time and areas of congestion, corrective actions can be taken. The most straightforward measure would be a change of geometry, i.e., shorter or wider escape paths (floors, stairs, doors). This can be directly put into the geometry sub-model, the simulation can be re-run, and the result checked. Secondly, the signage, and therefore the orientation capability, can be improved. This is not as straightforward as geometrical changes. It does depend more heavily on the model characteristics as to how these changes influence the evacuation sequence.

We will not go into these details in the following two sections but rather show two typical examples for evacuation simulations and the results obtained. We will also not discuss the results in detail, since they are of an illustrative nature in the context of this article. The following examples are based on investigations that have been performed using a cellular automaton model which is described along with the simulation program in [90,111].

Simulation Example 1 – Hotel The first example we show is a hotel with 8069 persons. In Fig. 15 only the ground floor is shown. There are nine floors altogether. The upper floors influence the ground floor only via the stair landings and the exits adjacent to them. Most of the 8069 persons are initially located in the ground floor, since the theater and conference area is located there. The upper floors are mainly covering bedrooms and some small conference areas.

The first step in our example (which might well be a useful recipe for evacuation analyses in general and is again in accordance with [118]) is to perform a statistical analysis. To this end, 500 samples are simulated. The evacuation time of a single run is the time it takes for all persons to get out. In this context, no fire or smoke are

taken into account. Since there are stochastic influences in the model used, the significant overall evacuation time is taken to be the 95-percentile (cf. Fig. 14). Finally, the maximum, minimum, mean, and significant values for the evacuation curve (number of persons evacuated vs. time) are also shown in Fig. 14.

The next figure (Fig. 16) shows the cumulated density. The thresholds (red areas) are 3.5 persons per square meter and 10% of the overall evacuation time (in this case 49 seconds). The overall evacuation time is 8:13 minutes (493 seconds). This value is obtained by taking the 95-percentile of the frequency distribution for the overall evacuation times (cf. Fig. 14).

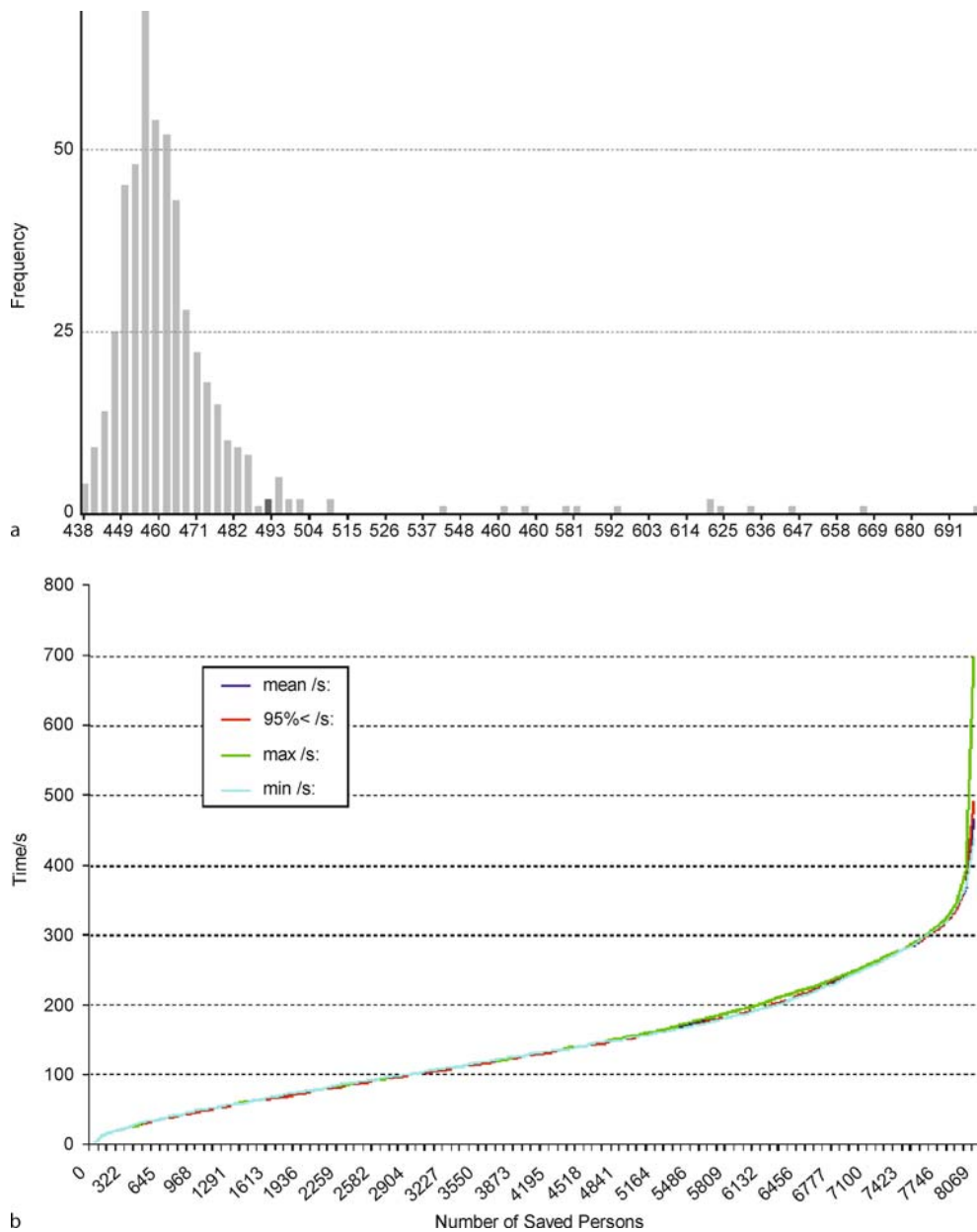
Of course, a distribution of overall evacuation times (for one scenario, i.e., the same initial parameters) can only be obtained by a stochastic model. In a deterministic model only one single value is calculated for the overall evacuation time. The variance of the overall evacuation times is due to two effects in the model used here: the initial position of the persons is determined anew at the beginning of each simulation run since only the statistical properties of the overall population are set and the motion of the persons is governed by partially stochastic rules (e.g., probabilistic parameters).

Simulation Example 2 – Passenger Ship The second example we will show is a ship. The major difference from the previous example is the addition of (1) the assembly phase and (2) embarkation and launching.

$$\begin{aligned}
 T &= A + \frac{2}{3}(E + L) \\
 &= f_{\text{safety}} \cdot (t_{\text{react}} + t_{\text{walk}}) + \frac{2}{3}(E + L) \\
 &\leq 60 \text{ minutes} .
 \end{aligned}$$

Embarkation and launching time ($E + L$) are required to be less than 30 minutes. For the sake of the evacuation analysis at an early design stage, the sum of embarkation and launching time can be assumed to be 30 minutes. Therefore, the requirement for A is 40 minutes. Alternatively, the embarkation and launching time can be determined by an evacuation trial.

Figure 17 shows the layout, initial population distribution (night case), density plot for the day case, and density plot for the night case. The reaction times are different for the day and the night case: 3 to 7 minutes (equally distributed) in the one and 7 to 13 minutes in the other. The longer reaction time in the night case results in less congestion (cf. Fig. 17). Both cases must be done in the analysis according to [118]. Additionally, a secondary night and day case are required (making up four cases altogether). In these secondary cases the main vertical zone (MVZ) lead-



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 14

Frequency distribution for the overall evacuation time (a) and evacuation curve (b)

ing to the longest overall individual assembly time is identified, and then either half of the stairway capacity in this zone is assumed to be not available, or 50% of the persons initially located in this zone must be led via one neighboring zone to the assembly station.

In the same way as shown for the two examples, simulations can be performed for other types of buildings and vessels. This technique has been applied to various passen-

gers ships [112] to football stadiums [88] and the World Youth Day 2005 [88], the Jamarat Bridge in Makkah [88], a movie theater and schools (mainly for calibration and validation) [90] and airports [171]. Of course, many examples of applications based on various models can be found in the literature. For an overview, the proceedings of the PED conference series are an excellent starting point [44, 170,190].



a



b

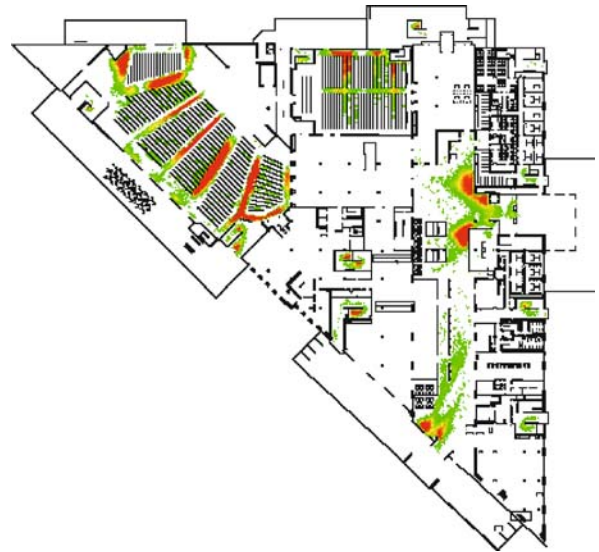
Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 15

Initial population distribution and situation after two minutes

Comparison of Commercial Software Tools

From a practical point of view, application of models for pedestrian dynamics and evacuation processes becomes more and more relevant in safety analysis. This has led to the development of a number of software tools that, with different sophistication, help us study many aspects without risking the health of test persons in evacuation trials.

There are commercial, as well as non-commercial software tools. All tools might be based on different models [97,187]. They have become very popular since the middle of the 1990s. A first comparison of different com-

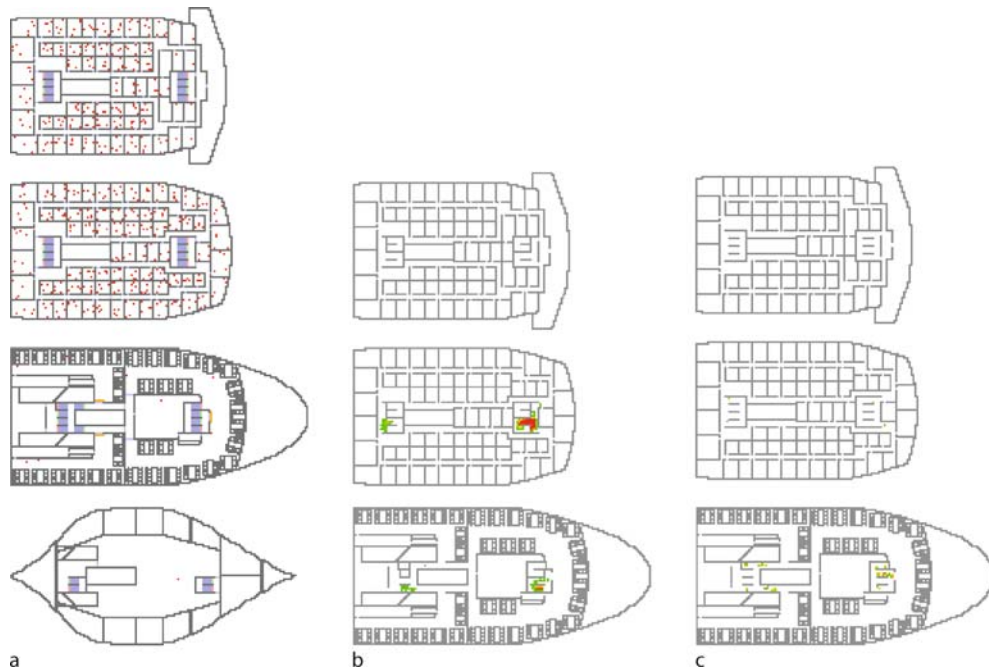


Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 16

Density plot, i.e., cumulated person density exceeding 3.5 persons per square meter and 10% of overall evacuation time

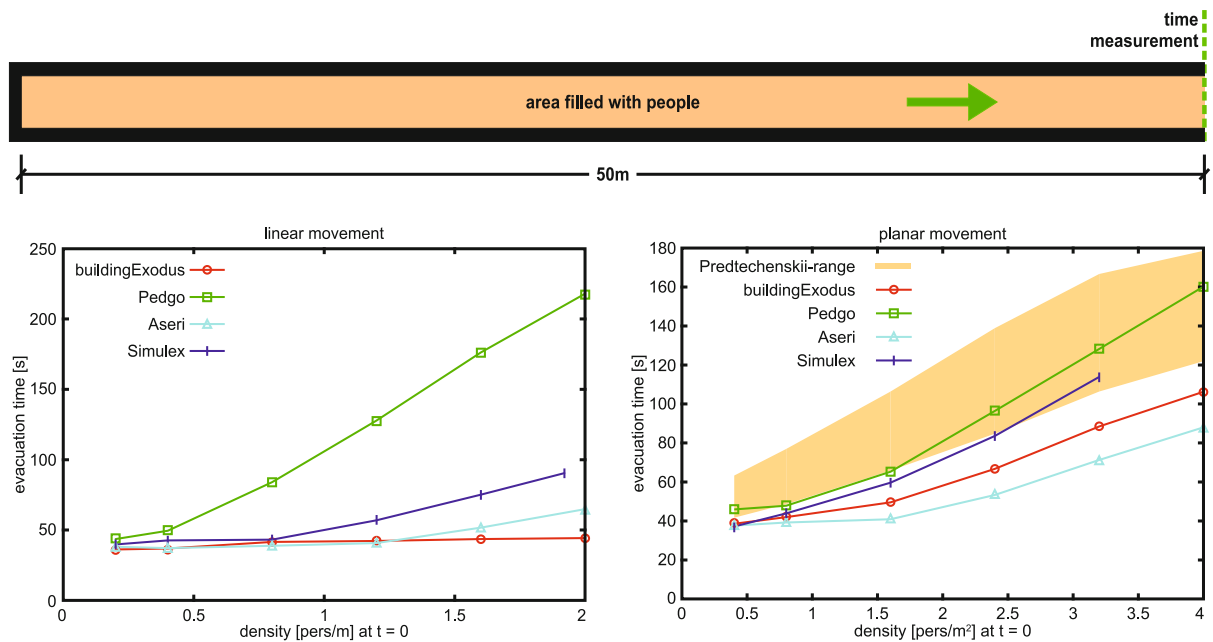
mercial software tools can be found in [191], where they were said to produce “reasonable results”. Further comparisons of real evacuation data with software tools or hand calculation methods can be found in [29,67,91,96,104,160,161,177]. But results predicted by different commercial software tools can differ by up to 40% for the same building [96]. Results may differ, too, when calculating with different assumptions, e.g., different reaction times, use of more or less detailed stair models, or when calculating with a real occupant load in contrast to an uncertainty analysis [96,104]. Contrary to these results, another study [161] shows that calculations with different software tools are able to predict total evacuation times for high-rise buildings and there are no large differences as shown in [96]. In [161] the results of an evacuation trial and simulations with different commercial software tools differed for selected floors of a highrise building. The densities were very low in this instance. In this case human behavior has a very large influence on the evacuation time. By contrast, evacuations at medium or high densities, human behavior has a smaller influence on the evacuation time of selected areas because congestion appears and continues larger than in low density situations – thus people reach the exit while congestion is still a factor [162]. In low density situations congestions are very rare, thus people move narrowly with free walking velocity through the building [162].

But the results presented in [161] also show that commercial software tools sometimes have problems with the



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 17

Initial distribution for the night case, density plot for the day case, and density plot for the night case for the "AENEAS steamliner"

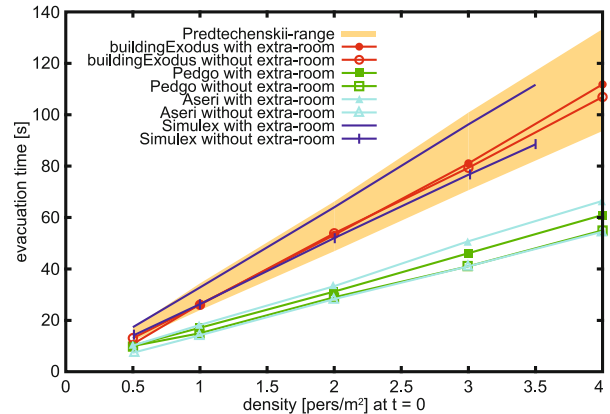
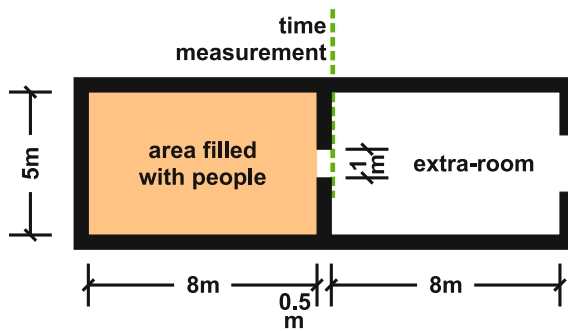


Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 18

Comparison of different software tools by simulating linear (*left*, narrow floor) and planar (*right*, 2 m wide floor) movement [162]

empirical relationship of density and walking speed (see Fig. 18). Furthermore, it is very important how boundary conditions are implemented in these tools (see Fig. 19), and the investigation of a simple scenario of a single room

using different software tools shows results differing by about a factor of two (see Fig. 19) [161]. In this case all software tools predict a congestion at the exit. Furthermore it is possible that the implemented algorithm fails [161].



Evacuation Dynamics: Empirical Results, Modeling and Applications, Figure 19
Comparison of different software tools by simulating a simple room geometry [162]

Thus for the user it is hard to know which algorithms are implemented in closed-source tools so that such a tool must be considered as “black box” [147]. It is also quite difficult to compare results about density and appearing congestions calculated by different software tools [162] and so it is questionable how these results should be interpreted. But, as pointed out earlier, reliable empirical data are often missing so that a validation of software tools or models is quite difficult [162].

Future Directions

The discussion has shown that the problem of crowd dynamics and evacuation processes is far from being well understood. One big problem is experimental basis. As in many human systems, it is difficult to perform controlled experiments on a sufficiently large scale. This would be necessary since data from actual emergency situations is usually not available, at least in sufficient quality. Progress should be possible by using modern video and computer technology which should allow us, in principle, to extract precise data even for the trajectories of individuals.

The full understanding of the complex dynamics of evacuation processes requires collaboration between engineering, physics, computer science, psychology, etc. Engineering in cooperation with computer science will lead to an improved empirical basis. Methods from physics allow us to develop simple but realistic models that capture the main aspects of the dynamics. Psychology is then needed to understand the interactions between individuals in sufficient detail to get a reliable set of ‘interaction’ parameters for the physical models.

In the end, we hope these joint efforts will lead to realistic models for evacuation processes that not only allow us to study these in the planning stages of facilities, but

even allow for dynamical real-time evacuation control in case an emergency occurs.

Acknowledgments

The authors would like to acknowledge the contribution of Tim Meyer-König (the developer of PedGo) and Michael Schreckenberg, Ansgar Kirchner, Bernhard Steffen for many fruitful discussions and valuable hints.

Bibliography

Primary Literature

1. Abe K (1986) *The Science of Human Panic*. Brain, Tokyo (in Japanese)
2. AlGadhi SAH, Mahmassani HS, Herman R (2002) A speed-concentration relation for bi-directional crowd movements with strong interaction. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 3–20
3. American Sociological Association (2002) In disasters, panic is rare, altruism dominates. Technical report, American Sociological Association
4. Ashe B, Shields TJ (1999) Analysis and modelling of the unannounced evacuation of a large retail store. *Fire Mater* 23: 333–336
5. Ben-Jacob E (1997) From snowflake formation to growth of bacterial colonies, Part II. Cooperative formation of complex colonial patterns. *Contemp Phys* 38:205
6. Biham O, Middleton AA, Levine D (1992) Self-organization and a dynamical transition in traffic-flow models. *Phys Rev A* 46:R6124
7. Blue VJ, Adler JL (2000) Cellular automata microsimulation of bi-directional pedestrian flows. *J Trans Res Board* 1678: 135–141
8. Blue VJ, Adler JL (2002) Flow capacities from cellular automata modeling of proportional splits of pedestrians by direction. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 115–121

9. Blythe RA et al (2007) Nonequilibrium steady states of matrix product form: a solver's guide. *Math Theor* 40:R333–R441, doi:10.1088/1751-8113/40/46/R01
10. Bolay K (1998) Nichtlineare Phänomene in einem fluid-dynamischen Verkehrsmodell. Diploma Thesis, Stuttgart University
11. Boyce KE, Shields TJ, Silcock GWH (1999) Toward the Characterization of Building Occupancies for Fire Safety Engineering: Capabilities of Disabled People Moving Horizontally and on an Incline. *Fire Technol* 35:51–67
12. Bryan JL (1995) Behavioral response to fire and smoke. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy, p 263
13. Burstedde C, Klauck K, Schadschneider A, Zittartz J (2001) Simulation of pedestrian dynamics using a two-dimensional cellular automaton. *Physica A* 295:507–525
14. Burstedde C, Kirchner A, Klauck K, Schadschneider A, Zittartz J (2002) Cellular automaton approach to pedestrian dynamics – applications. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 87–98
15. Chakrabarti J, Dzubiella J, Löwen H (2004) Reentrance effect in the lane formation of driven colloids. *Phys Rev E* 70:012401
16. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329(4–6):199–329
17. Clarke L (2002) Panic: Myth or reality? *Contexts* 1(3):21–26
18. Coleman JS (1990) *Foundation of Social Theory*. Belknap, Cambridge, Chap 9
19. Daamen W (2004) *Modelling Passenger Flows in Public Transport Facilities*. Ph.D. thesis, Technical University of Delft
20. Daamen W, Hoogendoorn SP (2006) Flow-density relations for pedestrian traffic. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow 05*. Springer, Berlin, pp 315–322
21. Daamen W, Bovy PHL, Hoogendoorn SP (2002) Modelling pedestrians in transfer stations. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 59–73
22. de Gelder B, Snyder J, Greve D, Gerard G, Hadjikhani N (2004) Fear fosters flight: A mechanism for fear contagion when perceiving emotion expressed by a whole body. *Proc Natl Acad Sci* 101(47):16701–16706
23. Derrida B (1998) An exactly soluble non-equilibrium system: The asymmetric simple exclusion process. *Phys Rep* 301:65
24. Dieckmann D (1911) *Die Feuersicherheit in Theatern*. Jung, München (in German)
25. DiNenno PJ (ed) (2002) *SFPE Handbook of Fire Protection Engineering*, 3rd edn. National Fire Protection Association, Bethesda
26. DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) (1995) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy
27. Dogliani M (2002) An overview of present and under-development IMO's requirements concerning evacuation from ships. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 339–354
28. Dzubiella J, Hoffmann GP, Löwen H (2002) Lane formation in colloidal mixtures driven by an external field. *Phys Rev E* 65:021402
29. Ehm M, Linxweiler J (2004) Berechnungen von Evakuierungszeiten bei Sonderbauten mit dem Programm buildingExodus. Technical report, TU Braunschweig
30. El Yacoubi S, Chopard B, Bandini S (eds) (2006) *Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006*, Perpignan. Springer, Berlin
31. Federal Aviation Administration FAA (1990) Emergency evacuation – cfr sec. 25.803. Regulation CFR Sec. 25.803
32. Fischer H (1933) *Über die Leistungsfähigkeit von Türen, Gängen und Treppen bei ruhigem, dichtem Verkehr*. Dissertation, Technische Hochschule Dresden (in German)
33. Frantzi H (1996) Study of movement on stairs during evacuation using video analysing techniques. Technical report, Department of Fire Safety Engineering, Lund Institute of Technology
34. Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier-Stokes equation. *Phys Rev Lett* 56:1505
35. Fruin JJ (1971) *Pedestrian Planning and Design*. Metropolitan Association of Urban Designers and Environmental Planners, New York
36. Fruin JJ (1993) The causes and prevention of crowd disasters. In: Smith RA, Dickie JF (eds) *Engineering for Crowd Safety*. Amsterdam, Elsevier
37. Fujiyama T (2006) Collision avoidance of pedestrians on stairs. Technical report, Centre for Transport Studies. University College London, London
38. Fujiyama T, Tyler N (2004) An explicit study on walking speeds of pedestrians on stairs. In: 10th International Conference on Mobility and Transport for Elderly and Disabled People, Hamamatsu, Japan, May 2004
39. Fujiyama T, Tyler N (2004) Pedestrian Speeds on Stairs: An Initial Step for a Simulation Model. In: *Proceedings of 36th Universities Transport Studies Group Conference*, Life Science Centre, Newcastle upon Tyne, Jan 2004
40. Fukui M, Ishibashi Y (1999) Jamming transition in cellular automaton models for pedestrians on passageway. *J Phys Soc Jpn* 68:3738
41. Fukui M, Ishibashi Y (1999) Self-organized phase transitions in cellular automaton models for pedestrians. *J Phys Soc Jpn* 68:2861
42. Galbreath M (1969) Time of evacuation by stairs in high buildings. *Fire Res Note* 8, NRCC
43. Galea ER (2002) Simulating evacuation and circulation in planes, trains, buildings and ships using the EXODUS software. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin pp 203–226
44. Galea ER (ed) (2003) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London
45. Gipps PG, Marksjö B (1985) A micro-simulation model for pedestrian flows. *Math Comput Simul* 27:95–105
46. Graat E, Midden C, Bockholts P (1999) Complex evacuation; effects of motivation level and slope of stairs on emergency egress time in a sports stadium. *Saf Sci* 31:127–141
47. Grosshandler W, Sunder S, Snell J (2003) Building and fire safety investigation of the world trade center disaster. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 279–281

48. Hamacher HW, Tjandra SA (2002) Mathematical modelling of evacuation problems – a state of the art. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 227–266
49. Hankin BD, Wright RA (1958) Passenger flow in subways. *Oper Res Q* 9:81–88
50. Helbing D (1992) A fluid-dynamic model for the movement of pedestrians. *Complex Syst* 6:391–415
51. Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
52. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282–4286
53. Helbing D, Farkas I, Vicsek T (2000) Freezing by heating in a driven mesoscopic system. *Phys Rev Lett* 84:1240–1243
54. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407:487–490
55. Helbing D, Farkas I, Molnár P, Vicsek T (2002) Simulation of pedestrian crowds in normal and evacuation situations. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 21–58
56. Helbing D, Buzna L, Werner T (2003) Self-organized pedestrian crowd dynamics and design solutions. *Traffic Forum*, pp 2003–12
57. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transp Sci* 39:1–24
58. Helbing D, Johansson A, Al-Abideen HZ (2007) The dynamics of crowd disasters: An empirical study. *Phys Rev E* 75:046109
59. Helbing D, Johansson A, Al-Abideen HZ (2007) Crowd turbulence: the physics of crowd disasters. In: *The Fifth International Conference on Nonlinear Mechanics, ICNM-V*, Shanghai, pp 967–969
60. Henderson LF (1971) The statistics of crowd fluids. *Nature* 229:381–383
61. Henderson LF (1974) On the fluid mechanics of human crowd motion. *Transp Res* 8:509–515
62. Hoogendoorn SP (2003) Walker behaviour modelling by differential games. In: Emmerich H, Nestler B, Schreckenberg M (eds) *Interface and transport dynamics. Lecture notes in Computational Science and Engineering*, vol 32. Springer, Berlin, pp 275–294
63. Hoogendoorn SP, Bovy PHL (2003) Simulation of pedestrian flows by optimal control and differential games. *Optim Control Appl Meth* 24:153
64. Hoogendoorn SP, Daamen W (2005) Pedestrian behavior at bottlenecks. *Transp Sci* 39 2:0147–0159
65. Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic pedestrian wayfinding and dynamics modelling. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 123–154
66. Hoogendoorn SP, Daamen W, Bovy PHL (2003) Microscopic pedestrian traffic data collection and analysis by walking experiments: Behaviour at bottlenecks. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 89–100
67. Hoskin K (2004) Fire protection and evacuation procedures of stadia venues in new zealand. Master's thesis, University of Canterbury
68. Hughes RL (2000) The flow of large crowds of pedestrians. *Math Comput Simul* 53:367–370
69. Hughes RL (2002) A continuum theory for the flow of pedestrians. *Transp Res Part B* 36:507–535
70. International Maritime Organization (IMO) (2000) *International Code of Safety for High-Speed Craft*, 2000 (2000 HSC Code). Technical report, Resolution MSC 97(73)
71. International Organization for Standardization (2000) *ISO-TR-13387-8-1999 Fire safety engineering – part 8: Life safety – occupant behaviour, location and condition*. Technical report
72. Jian L, Lizhong Y, Daoling Z (2005) Simulation of bi-direction pedestrian movement in corridor. *Physica A* 354:619
73. Johnson NR (1987) Panic at “The Who Concert Stampede”: An Empirical Assessment. *Soc Probl* 34(4):362–373
74. Jungermann H, Göhlert C (2000) Emergency evacuation from double-deck aircraft. In: Cottam MP, Harvey DW, Pape RP, Tait J (eds) *Foresight and Precaution. Proceedings of ESREL 2000, SARS and SRA. Europe Annual Conference*, Rotterdam, pp 989–992
75. Kashiwagi T (ed) (1994) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London
76. Kaufman M (2007) *Lane Formation in Counterflow Situations of Pedestrian Traffic*. Master's thesis, Universität Duisburg-Essen
77. Keating JP (1982) The myth of panic. *Fire J* May:57–62
78. Kendik E (1983) Determination of the evacuation time pertinent to the projected area factor in the event of total evacuation of high-rise office buildings via staircases. *Fire Saf J* 5:223–232
79. Kendik E (1984) *Die Berechnung der Personenströme als Grundlage für die Bemessung von Gehwegen in Gebäuden und um Gebäude*. Ph.D. thesis, TU Wien
80. Kendik E (1986) Designing escape routes in buildings. *Fire Technol* 22:272–294
81. Kerner BS (2004) *The Physics of Traffic*. Springer, Heidelberg
82. Kirchner A (2003) *Modellierung und statistische Physik biologischer und sozialer Systeme*. Dissertation, Universität zu Köln
83. Kirchner A, Schadschneider A (2002) Simulation of evacuation processes using a bionics-inspired cellular automaton model for pedestrian dynamics. *Physica A* 312:260–276
84. Kirchner A, Namazi A, Nishinari K, Schadschneider A (2003) Role of conflicts in the floor field cellular automaton model for pedestrian dynamics. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, pp 51
85. Kirchner A, Nishinari K, Schadschneider A (2003) Friction effects and clogging in a cellular automaton model for pedestrian dynamics. *Phys Rev E* 67:056122
86. Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization effects and the influence of walking speed in cellular automata models for pedestrian dynamics. *J Stat Mech* 10:P10011
87. Klüpfel H (2003) *A Cellular Automaton Model for Crowd Movement and Egress Simulation*. Dissertation, University Duisburg-Essen
88. Klüpfel H (2006) The simulation of crowds at very large events. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow 05*. Springer, Berlin, pp 341–346
89. Klüpfel H, Meyer-König T, Wahle J, Schreckenberg M (2000) Microscopic simulation of evacuation processes on passen-

- ger ships. In: Bandini S, Worsch T (eds) *Theory and Practical Issues on Cellular Automata*. Springer, Berlin
90. Klüpfel H, Meyer-König T, Schreckenberg M (2001) Empirical data on an evacuation exercise in a movie theater. Technical report, University Duisburg-Essen
 91. Ko SY (2003) Comparison of evacuation times using Simulex and EvacuationNZ based on trial evacuations. Fire Engineering Research Report 03/9, University of Canterbury
 92. Kretz T (2007) *Pedestrian Traffic – Simulation and Experiments*. Dissertation, Universität Duisburg-Essen
 93. Kretz T, Grünebohm A, Schreckenberg M (2006) Experimental study of pedestrian flow through a bottleneck. *J Stat Mech* P10014
 94. Kretz T, Grünebohm A, Kaufmann M, Mazur F, Schreckenberg M (2006) Experimental study of pedestrian counterflow in a corridor. *J Stat Mech* P10001
 95. Kretz T, Grünebohm A, Keßel A, Klüpfel H, Meyer-König T, Schreckenberg M (2008) Upstairs walking speed distributions on a long stair. *Saf Sci* 46:72–78
 96. Kuligowski ED, Milke JA (2005) A performance-based egress analysis of a hotel building using two models. *J Fire Prot Eng* 15:287–305
 97. Kuligowski ED, Peacock RD (2005) A review of building evacuation models. Technical report 1471. National Institute of Standards and Technology, Gaithersburg
 98. Lakoba TI, Kaup DJ, Finkelstein NM (2005) Modifications of the Helbing-Molnár-Farkas-Vicsek social force model for pedestrian evolution. *Simulation* 81(5):339–352
 99. Lam WHK, Lee JYS, Chan KS, Goh PK (2003) A generalised function for modeling bi-directional flow effects on indoor walkways in Hong Kong. *Transp Res A: Policy Pract* 37:789–810
 100. Laur U, Jaakula H, Metsaveer J, Lehtola K, Livonen H, Karpinen T, Eksborg AL, Rosengren H, Noord O (1997) Final Report on the Capsizing on 28 September 1994 in the Baltic Sea of the Ro-Ro Passenger Vessel MV Estonia. Technical report. The Joint Accident Investigation Commission of Estonia, Finland and Sweden
 101. LeBon G (1895) *Lois Psychologiques De L'evolution Des Peuples*. Alcan, Paris
 102. Leutzbach W (1988) *Introduction to the Theory of Traffic Flow*. Springer, Berlin
 103. Lewin K (1951) *Field Theory in Social Science*. Harper, New York
 104. Lord J, Meacham B, Moore A, Fahy R, Proulx G (2005) Guide for evaluating the predictive capabilities of computer egress models. Technical report NIST GCR 06–886, NIST, Gaithersburg
 105. Lovas GG (1994) Modeling and simulation of pedestrian traffic flow. *Transp Res B* 28V:429
 106. Maniccam S (2003) Traffic jamming on hexagonal lattice. *Physica A* 321:653
 107. Maniccam S (2005) Effects of back step and update rule on congestion of mobile objects. *Physica A* 346:631
 108. Marconi S, Chopard B (2002) A multiparticle lattice gas automata model for a crowd. In: *Cellular Automata. Lecture Notes Computer Science*, vol 2493. Springer, Berlin, pp 231
 109. Mawson AR (2005) Understanding mass panic and other collective responses to threat and disaster. *Psychiatry* 68:95–113
 110. Melinek SJ, Booth S (1975) An analysis of evacuation times and the movement of crowds in buildings. Technical report CP 96/75, BRE
 111. Meyer-König T, Klüpfel H, Schreckenberg M (2001) A microscopic model for simulating mustering and evacuation processes onboard passenger ships. In: KH Dräger (ed) *Proceedings of the International Emergency Management Society Conference*. The International Emergency Management Society, Oslo
 112. Meyer-König T, Klüpfel H, Schreckenberg M (2002) Assessment and analysis of evacuation processes on passenger ships by microscopic simulation. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 297–302
 113. Mintz A (1951) Non-adaptive group behaviour. *J Abnorm Soc Psychol* 46:150–159
 114. Molnár P (1995) *Modellierung und Simulation der Dynamik von Fußgängerströmen*. Dissertation, Universität Stuttgart
 115. Mori M, Tsukaguchi H (1987) A new method for evaluation of level of service in pedestrian facilities. *Transp Res* 21A(3): 223–234
 116. Morrall JF, Ratnayake LL, Seneviratne PN (1991) Comparison of CBD pedestrian characteristics in Canada and Sri Lanka. In: *Transportation Research Record* 1294. TRB, National Research Council, Washington DC, pp 57–61
 117. MSA (1997) Report on Exercise Invicta. Technical report. Marine Safety Agency, Southampton
 118. MSC-Circ.1033. Interim guidelines for evacuation analyses for new and existing passenger ships. Technical report, International Maritime Organization, Marine Safety Committee, London, June, 6th 2002. MSC/Circ. 1033
 119. MSC-Circ.1166. Guidelines for a simplified evacuation analysis for high-speed passenger craft. Technical report, International Maritime Organisation, 2005
 120. Muir HC (1997) Airplane of the 21st century: Challenges in safety and survivability. International Conference on Aviation Safety and Security in the 21st Century, White House Commission on Aviation Safety and Security, Washington
 121. Muir HC, Bottomley DM, Marrison C (1996) Effects of motivation and cabin configuration on emergency aircraft evacuation behavior and rates of egress. *Int J Aviat Psychol* 6(1):57–77
 122. Müller K (1981) Zur Gestaltung und Bemessung von Fluchtwegen für die Evakuierung von Personen aus Bauwerken auf der Grundlage von Modellversuchen. Dissertation, Technische Hochschule Magdeburg
 123. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 3:65–70; to be continued in 4/1966
 124. Müller W (1966) Die Beurteilung von Treppen als Rückzugsweg in mehrgeschossigen Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 4:93–96; continuation from 3/1966
 125. Müller W (1968) Die Überschneidung der Verkehrsströme bei dem Berechnen der Räumungszeit von Gebäuden. *Unser Brandschutz – Wissenschaftlich-Technische Beil* 4:87–92
 126. Müller W (1970) Untersuchung über zulässige Räumungszeiten und die Bemessung von Rückzugswegen in Gebäuden. Habilitation, TU Dresden, Dresden

127. Muramatsu M, Nagatani T (2000) Jamming transition in two-dimensional pedestrian traffic. *Physica A* 275:281–291
128. Muramatsu M, Nagatani T (2000) Jamming transition of pedestrian traffic at crossing with open boundary conditions. *Physica A* 286:377–390
129. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487–498
130. Argebau (2005) MVStättV – Erläuterungen: Musterverordnung über den Bau und Betrieb von Versammlungsstätten. Erläuterungen, Juni 2005
131. Nagai R, Nagatani T (2006) Jamming transition in counter flow of slender particles on square lattice. *Physica A* 366:503
132. Nagai R, Fukamachi M, Nagatani T (2006) Evacuation of crawlers and walkers from corridor through an exit. *Physica A* 367:449–460
133. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2:2221
134. Nakayama A, Hasebe K, Sugiyama Y (2005) Instability of pedestrian flow and phase structure in a two-dimensional optimal velocity model. *Phys Rev E* 71:036121
135. Navin PD, Wheeler RJ (1969) Pedestrian flow characteristics. *Traffic Eng* 39:31–36
136. Nelson HE, Mowrer FW (2002) Emergency movement. In: DiNenno PJ (ed) *SFPE Handbook of Fire Protection Engineering*, 3rd edn. National Fire Protection Association, Bethesda, p 367
137. National Fire Protection Association (2007) NFPA 130: Standard for Fixed Guideway Transit and Passenger Rail Systems.
138. Norwegian Ministry of Justice and Police (2000) The High-Speed Craft MS Sleipner Disaster, 26 November 1999. Official Norwegian Reports 2000:31, Oslo
139. Oeding D (1963) Verkehrsbelastung und Dimensionierung von Gehwegen und anderen Anlagen des Fußgängerverkehrs. Forschungsbericht, vol 22. Technische Hochschule Braunschweig
140. O'Flaherty CA, Parkinson MH (1972) Movement in a city centre footway. *Traffic Eng Control*, p 434
141. Okazaki S, Matsushita S (1993) A study of simulation model for pedestrian movement with evacuation and queuing. In: Smith RA, Dickie JF (eds) *Proceedings International Conference Engineering Crowd Safety*. Elsevier, Amsterdam, pp 271
142. Older SJ (1968) Movement of pedestrians on footways in shopping streets. *Traffic Eng Control* 10:160–163
143. Owen M, Galea ER, Lawrence PJ, Filippidis L (1998) AASK – aircraft accident statistics and knowledge: a database of human experience in evacuation, derived from aviation accident reports. *Aero J* 102:353–363
144. Pauls JL (1971) Evacuation drill held in the b. c. hydro building, 26 June 1969. Building Research Note 80, National Republican Congressional Committee
145. Pauls JL (1995) Movement of people. In: DiNenno PJ, Beyler CL, Custer RLP, Walton WD, Watts JM, Drysdale D, Hall JR (eds) *SFPE Handbook of Fire Protection Engineering*, 2nd edn. National Fire Protection Association, Quincy, p 263
146. Pauls JL, Fruin JJ, Zupan JM (2007) Minimum stair width for evacuation, overtaking movement and counterflow – technical bases and suggestions for the past, present and future. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 57–69
147. Paulsen T, Soma H, Schneider V, Wiklund J, Lovas G (1995) Evaluation of simulation models of evacuating from complex spaces. SINTEF Report STF75 A95020. SINTEF, Trondheim
148. Polus A, Joseph JL, Ushpiz A (1983) Pedestrian flow and level of service. *J Transp Eng* 109(1):46–56
149. Popkov V, Schütz G (1999) Steady-state selection in driven diffusive systems with open boundaries. *Europhys Lett* 48:257
150. Predtechenskii VM, Milinskii AI (1969) Planning for foot traffic flow in buildings. Amerind Publishing, New Dehli, 1978. Translation of: Proektirovanie Zhdaniis Uchetom Organizatsii Dvizheniya Lyuddskikh Potokov, Stroiizdat Publishers, Moscow
151. Predtechenskii WM, Milinski AI (1971) Personenströme in Gebäuden – Berechnungsmethoden für die Modellierung. Müller, Köln-Braunsfeld
152. Predtechenskii WM, Cholschtschewnikow WW, Völkel H (1972) Vereinfachte Berechnung der Umformung von Personenströmen auf Wegabschnitten mit begrenzter Länge. *Unser Brandschutz Wissenschaftlich-Technische Beil* 6:90–94
153. Purser DA, Bensilium M (2001) Quantification of behaviour for engineering design standards and escape time calculations. *Saf Sci* 38(2):158–182
154. Pushkarev B, Zupan JM (1975) Capacity of walkways. *Transp Res Rec* 538:1–15
155. Quarantelli EL (1960) Images of withdrawal behavior in disasters: Some basic misconceptions. *Soc Probl* 8:63–79
156. Quarantelli EL (2001) The sociology of panic. In: Smelser NJ, Baltes PB (eds) *International Encyclopedia of the Social and Behavioral Sciences*. Pergamon, New York, pp 11020–11030
157. Revi A, Singh AK (2007) Cyclone and storm surge, pedestrian evacuation and emergency response in India. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 119–130
158. Rex M, Löwen H (2007) Lane formation in oppositely charged colloids driven by an electric field: Chaining and two-dimensional crystallization. *Phys Rev E* 75:051402
159. Rickert M, Nagel K, Schreckenberg M, Latour A (1996) Two lane traffic simulations using cellular automata. *Physica A* 231:534
160. Rogsch C (2005) Vergleichende Untersuchungen zur dynamischen Simulation von Personenströmen. Technical report JUEL-4185. Forschungszentrum Jülich
161. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) How reliable are commercial software-tools for evacuation calculation? In: *Interflam 2007 – Conference Proceedings*. Inter-science Communication Ltd, Greenwich, London, pp 235–245
162. Rogsch C, Klingsch W, Seyfried A, Weigel H (2007) Prediction accuracy of evacuation times for high-rise buildings and simple geometries by using different software-tools. In *Traffic and Granular Flow 2007*. Preprint
163. Roitman MJ (1966) Die Evakuierung von Menschen aus Bauwerken. Staatsverlag der Deutschen Demokratischen Republik
164. Rothman DH, Zaleski S (1994) Lattice-gas models of phase separation: Interfaces, phase transitions, and multiphase flow. *Rev Mod Phys* 66:1417
165. Rothman DH, Zaleski S (1997) *Lattice-Gas Cellular Automata*. Cambridge University Press, Cambridge
166. Saloma C, Perez GJ (2007) Herding in real escape panic. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M

- (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 471–479
167. Schadschneider A (2002) Cellular automaton approach to pedestrian dynamics – theory. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, pp 75–86
 168. Schelajew J, Schelajewa E, Semjonow N (2000) Nikolaus II. Der letzte russische Zar. Bechtermünz, Augsburg
 169. Schneider U, Kath K, Oswald M, Kirchberger H (2006) Evakuierung und Verhalten von Personen im Brandfall unter spezieller Berücksichtigung von schienengebundenen Fahrzeugen. Technical report 12, TU Wien
 170. Schreckenberg M, Sharma SD (eds) (2007) *Pedestrian and Evacuation Dynamics*. Springer, Berlin
 171. Schultz M, Lehmann S, Fricke H (2007) A discrete microscopic model for pedestrian dynamics to manage emergency situations in airport terminals. In: Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin, pp 389–395
 172. Schütz GM (2001) Exactly solvable models for many-body systems. In: Domb C, Lebowitz JL (eds) *Phase Transitions and Critical Phenomena*, vol 19. Academic Press, Amsterdam
 173. Seeger PG, John R (1978) Untersuchung der Räumungsabläufe in Gebäuden als Grundlage für die Ausbildung von Rettungswegen, Teil III: Reale Räumungsversuche. Technical report T395. Forschungsstelle für Brandschutztechnik an der Universität Karlsruhe (TH)
 174. Seyfried A, Steffen B, Klingsch W, Boltes M (2005) The fundamental diagram of pedestrian movement revisited. *J Stat Mech* P10002
 175. Seyfried A, Steffen B, Lippert T (2006) Basics of modelling the pedestrian flow. *Physica A* 368:232–238
 176. Seyfried A, Rupprecht T, Passon O, Steffen B, Klingsch W, Boltes M (2007) Capacity estimation for emergency exits and bottlenecks. In: *Interflam 2007 – Conference Proceedings*. Interscience Communication Ltd, Greenwich, London
 177. Shestopal VO, Grubits SJ (1994) Evacuation model for merging traffic flows in multi-room and multi-storey buildings. In: Kashiwagi T (ed) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 625–632
 178. Sime JD (1990) The Concept of Panic. In: Canter D (ed) *Fires and Human Behaviour*, vol 1. Wiley, London, pp 63–81
 179. Smelser NJ (1962) *Theory of Collective Behavior*. Free Press, New York
 180. Still KG (2001) *Crowd Dynamics*, Ph.D. thesis, University of Warwick
 181. Tajima Y, Nagatani T (2002) Clogging transition of pedestrian flow in t-shaped channel. *Physica A* 303:239–250
 182. Taylor PM (1990) *The Hillsborough Stadium Disaster: Inquiry Final Report*. Technical report, Great Britain Home Office
 183. Templer J (1992) *The Staircase*. MIT Press, Cambridge
 184. Thompson PA, Marchant EW (1994) Simulex; developing new computer modelling techniques for evaluation. In: Kashiwagi T (ed) *Fire Safety Science – 4th international Symposium Proceedings*. Interscience, West Yard House, Guildford. The International Association for Fire Safety Science. Grove, London, pp 613–624
 185. Togawa K (1955) Study on fire escapes basing on the observation of multitude currents. Report of the building research institute. Ministry of Construction, Japan (in Japanese)
 186. Tsuji Y (2003) Numerical simulation of pedestrian flow at high densities. In: Galea ER (ed) *Pedestrian and Evacuation Dynamics 2003*. CMS Press, London, p 27
 187. Tubbs JS, Meacham B (2007) *Egress Design Solutions – A Guide to Evacuation and Crowd Management Planning*. Wiley, New Jersey
 188. Virkler MR, Elayadath S (1994) Pedestrian density characteristics and shockwaves. In: Akcelik R (ed) *Proceedings of the Second International Symposium on Highway Capacity*, vol 2. Australian Road Research Board, Sydney, pp 671–684
 189. Waldau N (2002) *Massenpanik in Gebäuden*. Diploma thesis, Technische Universität Wien
 190. Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2006) *Pedestrian and Evacuation Dynamics 2005*. Springer, Berlin
 191. Weckman LS, Mannikkö S (1999) Evacuation of a theatre: Exercise vs calculations. *Fire Mater* 23:357–361
 192. Weidmann U (1993) *Transporttechnik der Fußgänger – Transporttechnische Eigenschaften des Fußgängerverkehrs (Literaturauswertung)*. Schriftenreihe des IVT 90, ETH Zürich, 3 1993. Zweite, ergänzte Auflage (in German)
 193. Weifeng F, Lizhong Y, Weicheng F (2003) Simulation of bi-directional pedestrian movement using a cellular automata model. *Physica A* 321:633–640
 194. Wolf DE, Grassberger P (eds) (1996) *Friction, Arching, Contact Dynamics*. World Scientific, Singapore
 195. Yamamoto K, Kokubo S, Nishinari K (2006) New approach for pedestrian dynamics by real-coded cellular automata (rca). In: El Yacoubi S, Chopard B, Bandini S (eds) *Cellular Automata – 7th International Conference on Cellular Automata for Research and Industry, ACRI 2006, Perpignan*. Springer, Berlin, pp 728–731
 196. Yamamoto K, Kokubo S, Nishinari K (2007) Simulation for pedestrian dynamics by real-coded cellular automata (rca). *Physica A* 379:654
 197. Yamori K (1998) Going with the flow: Micro-macro dynamics in the macrobehavioral patterns of pedestrian crowds. *Psychol Rev* 105(3):530–557

Books and Reviews

- Chopard B, Droz M (1998) *Cellular automaton modeling of physical systems*. Cambridge University Press, Cambridge
- Chowdhury D, Nishinari K, Santen L, Schadschneider A (2008) *Stochastic transport in complex systems: From molecules to vehicles*. Elsevier, Amsterdam
- DiNenno PJ (ed) (2002) *SFPE Handbook of Fire Protection Engineering*. National Fire Protection Association, Quincy
- Galea ER (ed) (2003) *Pedestrian and Evacuation Dynamics '03*. CMS Press, London
- Ped-Net collaboration. Webpage www.ped-net.org (including discussion forum)
- Predtechenskii VM, Milinskii AI (1978) *Planing for foot traffic flow in buildings*. Amerint Publishing, New Delhi
- Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and Granular Flow '05*. Springer, Berlin (see also previous issues of this conference series)

- Tubbs JS, Meacham BJ (2007) *Egress Design Solution – A Guide to Evacuation and Crowd Management Planning*. Wiley, New Jersey
- Waldau N, Gattermann P, Knoflacher H, Schreckenberg M (eds) (2007) *Pedestrian and Evacuation Dynamics '05*. Springer, Berlin

Evolutionary Game Theory

WILLIAM H. SANDHOLM

Department of Economics, University of Wisconsin, Madison, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Normal Form Games
 Static Notions of Evolutionary Stability
 Population Games
 Revision Protocols
 Deterministic Dynamics
 Stochastic Dynamics
 Local Interaction
 Applications
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Deterministic evolutionary dynamic A deterministic evolutionary dynamic is a rule for assigning population games to ordinary differential equations describing the evolution of behavior in the game. Deterministic evolutionary dynamics can be derived from *revision protocols*, which describe choices (in economic settings) or births and deaths (in biological settings) on an agent-by-agent basis.

Evolutionarily stable strategy (ESS) In a symmetric normal form game, an evolutionarily stable strategy is a (possibly mixed) strategy with the following property: a population in which all members play this strategy is resistant to invasion by a small group of mutants who play an alternative mixed strategy.

Normal form game A normal form game is a strategic interaction in which each of n players chooses a strategy and then receives a payoff that depends on all agents' choices of strategy. In a *symmetric two-player normal form game*, the two players choose from

the same set of strategies, and payoffs only depend on own and opponent's choices, not on a player's identity.

Population game A population game is a strategic interaction among one or more large populations of agents. Each agent's payoff depends on his own choice of strategy and the distribution of others' choices of strategies. One can generate a population game from a normal form game by introducing random matching; however, many population games of interest, including congestion games, do not take this form.

Replicator dynamic The replicator dynamic is a fundamental deterministic evolutionary dynamic for games. Under this dynamic, the percentage growth rate of the mass of agents using each strategy is proportional to the excess of the strategy's payoff over the population's average payoff. The replicator dynamic can be interpreted biologically as a model of natural selection, and economically as a model of imitation.

Revision protocol A revision protocol describes both the timing and the results of agents' decisions about how to behave in a repeated strategic interaction. Revision protocols are used to derive both deterministic and stochastic evolutionary dynamics for games.

Stochastically stable state In Game-theoretic models of stochastic evolution in games are often described by irreducible Markov processes. In these models, a population state is stochastically stable if it retains positive weight in the process's stationary distribution as the level of noise in agents' choices approaches zero, or as the population size approaches infinity.

Definition of the Subject

Evolutionary game theory studies the behavior of large populations of agents who repeatedly engage in strategic interactions. Changes in behavior in these populations are driven either by natural selection via differences in birth and death rates, or by the application of myopic decision rules by individual agents.

The birth of evolutionary game theory is marked by the publication of a series of papers by mathematical biologist John Maynard Smith [137,138,140]. Maynard Smith adapted the methods of traditional game theory [151,215], which were created to model the behavior of rational economic agents, to the context of biological natural selection. He proposed his notion of an *evolutionarily stable strategy (ESS)* as a way of explaining the existence of ritualized animal conflict.

Maynard Smith's equilibrium concept was provided with an explicit dynamic foundation through a differential

equation model introduced by Taylor and Jonker [205]. Schuster and Sigmund [189], following Dawkins [58], dubbed this model the *replicator dynamic*, and recognized the close links between this game-theoretic dynamic and dynamics studied much earlier in population ecology [132,214] and population genetics [73]. By the 1980s, evolutionary game theory was a well-developed and firmly established modeling framework in biology [106].

Towards the end of this period, economists realized the value of the evolutionary approach to game theory in social science contexts, both as a method of providing foundations for the equilibrium concepts of traditional game theory, and as a tool for selecting among equilibria in games that admit more than one. Especially in its early stages, work by economists in evolutionary game theory hewed closely to the interpretation set out by biologists, with the notion of ESS and the replicator dynamic understood as modeling natural selection in populations of agents genetically programmed to behave in specific ways. But it soon became clear that models of essentially the same form could be used to study the behavior of populations of active decision makers [50,76,133,149,167,191]. Indeed, the two approaches sometimes lead to identical models: the replicator dynamic itself can be understood not only as a model of natural selection, but also as one of imitation of successful opponents [35,188,216].

While the majority of work in evolutionary game theory has been undertaken by biologists and economists, closely related models have been applied to questions in a variety of fields, including transportation science [143, 150,173,175,177,197], computer science [72,173,177], and sociology [34,62,126,225,226]. Some paradigms from evolutionary game theory are close relatives of certain models from physics, and so have attracted the attention of workers in this field [141,201,202,203]. All told, evolutionary game theory provides a common ground for workers from a wide range of disciplines.

Introduction

This article offers a broad survey of the theory of evolution in games. Section “[Normal Form Games](#)” introduces normal form games, a simple and commonly studied model of strategic interaction. Section “[Static Notions of Evolutionary Stability](#)” presents the notion of an evolutionarily stable strategy, a static definition of stability proposed for this normal form context.

Section “[Population Games](#)” defines population games, a general model of strategic interaction in large populations. Section “[Revision Protocols](#)” offers the notion of a revision protocol, an individual-level description

of behavior used to define the population-level processes of central concern.

Most of the article concentrates on these population-level processes: Section “[Deterministic Dynamics](#)” considers deterministic differential equation models of game dynamics; Section “[Stochastic Dynamics](#)” studies stochastic models of evolution based on Markov processes; and Sect. “[Local Interaction](#)” presents deterministic and stochastic models of local interaction. Section “[Applications](#)” records a range of applications of evolutionary game theory, and Sect. “[Future Directions](#)” suggests directions for future research. Finally, Sect. “[Bibliography](#)” offers an extensive list of primary references.

Normal Form Games

In this section, we introduce a very simple model of strategic interaction: the symmetric two-player normal form game. We then define some of the standard solution concepts used to analyze this model, and provide some examples of games and their equilibria. With this background in place, we turn in subsequent sections to evolutionary analysis of behavior in games.

In a *symmetric two-player normal form game*, each of the two players chooses a (pure) strategy from the finite set S , which we write generically as $S = \{1, \dots, n\}$. The game’s *payoffs* are described by the matrix $A \in \mathbf{R}^{n \times n}$. Entry A_{ij} is the payoff a player obtains when he chooses strategy i and his opponent chooses strategy j ; this payoff does not depend on whether the player in question is called player 1 or player 2.

The fundamental solution concept of noncooperative game theory is Nash equilibrium [151]. We say that the pure strategy $i \in S$ is a *symmetric Nash equilibrium* of A if

$$A_{ii} \geq A_{ji} \quad \text{for all } j \in S. \quad (1)$$

Thus, if his opponent chooses a symmetric Nash equilibrium strategy i , a player can do no better than to choose i himself.

A stronger requirement on strategy i demands that it be superior to all other strategies regardless of the opponent’s choice:

$$A_{ik} > A_{jk} \quad \text{for all } j, k \in S. \quad (2)$$

When condition (2) holds, we say that strategy i is *strictly dominant* in A .

Example 1 The game below, with strategies C (“cooperate”) and D (“defect”), is an instance of a *Prisoner’s*

Dilemma:

	C	D
C	2	0
D	3	1

(To interpret this game, note that $A_{CD} = 0$ is the payoff to cooperating when one's opponent defects.) Since $1 > 0$, defecting is a symmetric Nash equilibrium of this game. In fact, since $3 > 2$ and $1 > 0$, defecting is even a strictly dominant strategy. But since $2 > 1$, both players are better off when both cooperate than when both defect.

In many instances, it is natural to allow players to choose *mixed* (or *randomized*) strategies. When a player chooses mixed strategy from the simplex $X = \{x \in \mathbb{R}_+^n : \sum_{i \in S} x_i = 1\}$, his behavior is stochastic: he commits to playing pure strategy $i \in S$ with probability x_i .

When either player makes a randomized choice, we evaluate payoffs by taking expectations: a player choosing mixed strategy x against an opponent choosing mixed strategy y garners an expected payoff of

$$x' Ay = \sum_{i \in S} \sum_{j \in S} x_i A_{ij} y_j. \quad (3)$$

In biological contexts, payoffs are *fitnesses*, and represent levels of reproductive success relative to some baseline level; Eq. (3) reflects the idea that in a large population, expected reproductive success is what matters. In economic contexts, payoffs are *utilities*: a numerical representation of players' preferences under which Eq. (3) captures players' choices between uncertain outcomes [215].

The notion of Nash equilibrium extends easily to allow for mixed strategies. Mixed strategy x is a *symmetric Nash equilibrium* of A if

$$x' Ax \geq y' Ax \quad \text{for all } y \in X. \quad (4)$$

In words, x is a symmetric Nash equilibrium if its expected payoff against itself is at least as high as the expected payoff obtainable by any other strategy y against x . Note that we can represent the pure strategy $i \in S$ using the mixed strategy $e_i \in X$, the i th standard basis vector in \mathbb{R}^n . If we do so, then definition (4) restricted to such strategies is equivalent to definition (1).

We illustrate these ideas with a few examples.

Example 2 Consider the *Stag Hunt* game:

	H	S
H	h	h
S	0	s

Each player in the Stag Hunt game chooses between hunting hare (H) and hunting stag (S). A player who hunts hare always catches one, obtaining a payoff of $h > 0$. But hunting stag is only successful if both players do so, in which case each obtains a payoff of $s > h$. Hunting stag is potentially more profitable than hunting hare, but requires a coordinated effort.

In the Stag Hunt game, H and S (or, equivalently, e_H and e_S) are symmetric pure Nash equilibria. This game also has a symmetric mixed Nash equilibrium, namely $x^* = (x_H^*, x_S^*) = (\frac{s-h}{s}, \frac{h}{s})$. If a player's opponent chooses this mixed strategy, the player's expected payoff is h whether he chooses H, S, or any mixture between the two; in particular, x^* is a best response against itself.

To distinguish between the two pure equilibria, we might focus on the one that is *payoff dominant*, in that it achieves the higher joint payoff. Alternatively, we can concentrate on the *risk dominant* equilibrium [89], which utilizes the strategy preferred by a player who thinks his opponent is equally likely to choose either option (that is, against an opponent playing mixed strategy $(x_H, x_S) = (\frac{1}{2}, \frac{1}{2})$). In the present case, since $s > h$, equilibrium S is payoff dominant. Which strategy is risk dominant depends on further information about payoffs. If $s > 2h$, then S is risk dominant. But if $s < 2h$, H is risk dominant: evidently, payoff dominance and risk dominance need not agree.

Example 3 In the *Hawk–Dove* game [139], the two players are animals contesting a resource of value $v > 0$. The players choose between two strategies: display (D) or escalate (E). If both display, the resource is split; if one escalates and the other displays, the escalator claims the entire resource; if both escalate, then each player is equally likely to claim the entire resource or to be injured, suffering a cost of $c > v$ in the latter case.

The payoff matrix for the Hawk–Dove game is therefore

	D	E
D	$\frac{1}{2}v$	0
E	v	$\frac{1}{2}(v - c)$

This game has no symmetric Nash equilibrium in pure strategies. It does, however, admit the symmetric mixed equilibrium $x^* = (x_D^*, x_E^*) = (\frac{c-v}{c}, \frac{v}{c})$. (In fact, it can be shown that every symmetric normal form game admits at least one symmetric mixed Nash equilibrium [151].)

In this example, our focus on symmetric behavior may seem odd: rather than randomizing symmetrically, it seems more natural for players to follow an asymmetric Nash equilibrium in which one player escalates and the

other displays. But the symmetric equilibrium is the most relevant one for understanding natural selection in populations whose members are randomly matched in pairwise contests – see Sect. “Static Notions of Evolutionary Stability”.

Example 4 Consider the class of *Rock–Paper–Scissors* games:

	R	P	S
R	0	− <i>l</i>	<i>w</i>
P	<i>w</i>	0	− <i>l</i>
S	− <i>l</i>	<i>w</i>	0

Here $w > 0$ is the benefit of winning the match and $l > 0$ the cost of losing; ties are worth 0 to both players. We call this game *good RPS* if $w > l$, so that the benefit of winning the match exceeds the cost of losing, *standard RPS* if $w = l$, and *bad RPS* if $w < l$. Regardless of the values of w and l , the unique symmetric Nash equilibrium of this game, $x^* = (x_R^*, x_P^*, x_S^*) = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$, requires uniform randomization over the three strategies.

Static Notions of Evolutionary Stability

In introducing game-theoretic ideas to the study of animal behavior, Maynard Smith advanced this fundamental principle: that the evolutionary success of (the genes underlying) a given behavioral trait can depend on the prevalences of all traits. It follows that natural selection among the traits can be modeled as random matching of animals to play normal form games [137,138,139,140]. Working in this vein, Maynard Smith offered a stability concept for populations of animals sharing a common behavioral trait – that of playing a particular mixed strategy in the game at hand. Maynard Smith’s concept of evolutionary stability, influenced by the work of Hamilton [87] on the evolution of sex ratios, defines such a population as stable if it is resistant to invasion by a small group of mutants carrying a different trait.

Suppose that a large population of animals is randomly matched to play the symmetric normal form game A . We call mixed strategy $x \in X$ an *evolutionarily stable strategy* (ESS) if

$$x'A((1-\varepsilon)x + \varepsilon y) > y'A((1-\varepsilon)x + \varepsilon y) \\ \text{for all } \varepsilon \leq \bar{\varepsilon}(y) \text{ and } y \neq x. \quad (5)$$

To interpret condition (5), imagine that a population of animals programmed to play mixed strategy x is invaded by a group of mutants programmed to play the alternative mixed strategy y . Equation (5) requires that regardless of the choice of y , an incumbent’s expected payoff from

a random match in the post-entry population exceeds that of a mutant so long as the size of the invading group is sufficiently small.

The definition of ESS above can also be expressed as a combination of two conditions:

$$x'Ax \geq y'Ax \quad \text{for all } y \in X; \quad (4)$$

$$\text{For all } y \neq x, [x'Ax = y'Ax] \\ \text{implies that } [x'Ay > y'Ay]. \quad (6)$$

Condition (4) is familiar: it requires that the incumbent strategy x be a best response to itself, and so is none other than our definition of symmetric Nash equilibrium. Condition (6) requires that if a mutant strategy y is an alternative best response against the incumbent strategy x , then the incumbent earns a higher payoff against the mutant than the mutant earns against itself.

A less demanding notion of stability can be obtained by allowing the incumbent and the mutant in condition (6) to perform equally well against the mutant:

$$\text{For all } y \in X, [x'Ax = y'Ax] \\ \text{implies that } [x'Ay \geq y'Ay]. \quad (7)$$

If x satisfies conditions (4) and (7), it is called a *neutrally stable strategy* (NSS) [139].

Let us apply these stability notions to the games introduced in the previous section. Since every ESS and NSS must be a Nash equilibrium, we need only consider whether the Nash equilibria of these games satisfy the additional stability conditions, (6) and (7).

Example 5 In the Prisoner’s Dilemma game (Example 1), the dominant strategy D is an ESS.

Example 6 In the Stag Hunt game (Example 2), each pure Nash equilibrium is an ESS. But the mixed equilibrium $(x_H^*, x_S^*) = (\frac{s-h}{s}, \frac{h}{s})$ is not an ESS: if mutants playing either pure strategy enter the population, they earn a higher payoff than the incumbents in the post-entry population.

Example 7 In the Hawk–Dove game (Example 3), the mixed equilibrium $(x_D^*, x_E^*) = (\frac{c-v}{c}, \frac{v}{c})$ is an ESS. Maynard Smith used this and other examples to explain the existence of ritualized fighting in animals. While an animal who escalates always obtains the resource when matched with an animal who merely displays, a population of escalators is unstable: it can be invaded by a group of mutants who display, or who merely escalate less often.

Example 8 In Rock–Paper–Scissors games (Example 4), whether the mixed equilibrium $x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ is evolutionarily stable depends on the relative payoffs to winning

and losing a match. In good RPS ($w > l$), x^* is an ESS; in standard RPS ($w = l$), x^* is a NSS but not an ESS, while in bad RPS ($w < l$), x^* is neither an ESS nor an NSS. The last case shows that neither evolutionary nor neutrally stable strategies need exist in a given game.

The definition of an evolutionarily stable strategy has been extended to cover a wide range of strategic settings, and has been generalized in a variety of directions. Prominent among these developments are set-valued versions of ESS: in rough terms, these concepts consider a set of mixed strategies $Y \subset X$ to be stable if the no population playing a strategy in the set can be invaded successfully by a population of mutants playing a strategy outside the set. [95] provides a thorough survey of the first 15 years of research on ESS and related notions of stability; key references on set-valued evolutionary solution concepts include [15,199,206].

Maynard Smith's notion of ESS attempts to capture the dynamic process of natural selection using a static definition. The advantage of this approach is that his definition is often easy to check in applications. Still, more convincing models of natural selection should be explicitly dynamic models, building on techniques from the theories of dynamical systems and stochastic processes. Indeed, this thoroughgoing approach can help us understand whether and when the ESS concept captures the notion of robustness to invasion in a satisfactory way.

The remainder of this article concerns explicitly dynamic models of behavior. In addition to being dynamic rather than static, these models will differ from the one considered in this section in two other important ways as well. First, rather than looking at populations whose members all play a particular mixed strategy, the dynamic models consider populations in which different members play different pure strategies. Second, instead of maintaining a purely biological point of view, our dynamic models will be equally well-suited to studying behavior in animal and human populations.

Population Games

Population games provide a simple and general framework for studying strategic interactions in large populations whose members play pure strategies. The simplest population games are generated by random matching in normal form games, but the population game framework allows for interactions of a more intricate nature.

We focus here on games played by a single population (i. e., games in which all agents play equivalent roles). We suppose that there is a unit mass of agents, each of whom chooses a pure strategy from the set $S = \{1, \dots, n\}$. The

aggregate behavior of these agents is described by a *population state* $x \in X$, with x_j representing the proportion of agents choosing pure strategy j . We identify a *population game* with a continuous vector-valued payoff function $F: X \rightarrow \mathbf{R}^n$. The scalar $F_i(x)$ represents the payoff to strategy i when the population state is x .

Population state x^* is a *Nash equilibrium* of F if no agent can improve his payoff by unilaterally switching strategies. More explicitly, x^* is a Nash equilibrium if

$$x_i^* > 0 \quad \text{implies that} \quad F_i(x) \geq F_j(x) \quad \text{for all } j \in S. \quad (8)$$

Example 9 Suppose that the unit mass of agents are randomly matched to play the symmetric normal form game A . At population state x , the (expected) payoff to strategy i is the linear function $F_i(x) = \sum_{j \in S} A_{ij}x_j$; the payoffs to all strategies can be expressed concisely as $F(x) = Ax$. It is easy to verify that x^* is a Nash equilibrium of the population game F if and only if x^* is a symmetric Nash equilibrium of the symmetric normal form game A .

While population games generated by random matching are especially simple, many games that arise in applications are not of this form. In the biology literature, games outside the random matching paradigm are known as *playing the field* models [139].

Example 10 Consider the following model of highway congestion [17,143,166,173]. A pair of towns, Home and Work, are connected by a network of *links*. To commute from Home to Work, an agent must choose a *path* $i \in S$ connecting the two towns. The payoff the agent obtains is the negation of the delay on the path he takes. The delay on the path is the sum of the delays on its constituent links, while the delay on a link is a function of the number of agents who use that link.

Population games embodying this description are known as a *congestion games*. To define a congestion game, let Φ be the collection of links in the highway network. Each strategy $i \in S$ is a route from Home to Work, and so is identified with a set of links $\Phi_i \subseteq \Phi$. Each link ϕ is assigned a *cost function* $c_\phi: \mathbf{R}_+ \rightarrow \mathbf{R}$, whose argument is link ϕ 's *utilization level* u_ϕ :

$$u_\phi(x) = \sum_{i \in \rho(\phi)} x_i, \quad \text{where } \rho(\phi) = \{i \in S: \phi \in \Phi_i\}$$

The payoff of choosing route i is the negation of the total delays on the links in this route:

$$F_i(x) = - \sum_{\phi \in \Phi_i} c_\phi(u_\phi(x)).$$

Since driving on a link increases the delays experienced by other drivers on that link (i.e., since highway congestion involves *negative externalities*), cost functions in models of highway congestion are increasing; they are typically convex as well. Congestion games can also be used to model positive externalities, like the choice between different technological standards; in this case, the cost functions are decreasing in the utilization levels.

Revision Protocols

We now introduce foundations for our models of evolutionary dynamics. These foundations are built on the notion of a revision protocol, which describes both the timing and results of agents' myopic decisions about how to continue playing the game at hand [24,35,96,175,217]. Revision protocols will be used to derive both the deterministic dynamics studied in Sect. "Deterministic Dynamics" and the stochastic dynamics studied in Sect. "Stochastic Dynamics"; similar ideas underlie the local interaction models introduced in Sect. "Local Interaction".

Definition

Formally, a *revision protocol* is a map $\rho: \mathbf{R}^n \times X \rightarrow \mathbf{R}_+^{n \times n}$ that takes the payoff vectors π and population states x as arguments, and returns nonnegative matrices as outputs. For reasons to be made clear below, scalar $\rho_{ij}(\pi, x)$ is called the *conditional switch rate* from strategy i to strategy j .

To move from this notion to an explicit model of evolution, let us consider a population consisting of $N < \infty$ members. (A number of the analyzes to follow will consider the limit of the present model as the population size N approaches infinity – see Sects. "Mean Dynamics", "Deterministic Approximation", and "Stochastic Stability via Large Population Limits".) In this case, the set of feasible social states is the finite set $\mathcal{X}^N = X \cap \frac{1}{N}\mathbf{Z}^n = \{x \in X: Nx \in \mathbf{Z}^n\}$, a grid embedded in the simplex X .

A revision protocol ρ , a population game F , and a population size N define a continuous-time evolutionary process – a Markov process $\{X_t^N\}$ – on the finite state space \mathcal{X}^N . A one-size-fits-all description of this process is as follows. Each agent in the society is equipped with a "stochastic alarm clock". The times between rings of an agent's clock are independent, each with a rate R exponential distribution. The ringing of a clock signals the arrival of a revision opportunity for the clock's owner. If an agent playing strategy $i \in S$ receives a revision opportunity, he switches to strategy $j \neq i$ with probability ρ_{ij}/R . If a switch occurs, the population state changes accordingly,

from the old state x to a new state y that accounts for the agent's change in strategy.

While this interpretation of the evolutionary process can be applied to any revision protocol, simpler interpretations are sometimes available for protocols with additional structure. The examples to follow illustrate this point.

Examples

Imitation Protocols and Natural Selection Protocols

In economic contexts, revision protocols of the form

$$\rho_{ij}(\pi, x) = x_j \hat{\rho}_{ij}(\pi, x) \quad (9)$$

are called *imitation protocols* [35,96,216]. These protocols can be given a very simple interpretation: when an agent receives a revision opportunity, he chooses an opponent at random and observes her strategy. If our agent is playing strategy i and the opponent strategy j , the agent switches from i to j with probability proportional to $\hat{\rho}_{ij}$. Notice that the value of the population share x_j is not something the agent need know; this term in (9) accounts for the agent's observing a randomly chosen opponent.

Example 11 Suppose that after selecting an opponent, the agent imitates the opponent only if the opponent's payoff is higher than his own, doing so in this case with probability proportional to the payoff difference:

$$\rho_{ij}(\pi, x) = x_j [\pi_j - \pi_i]_+.$$

This protocol is known as *pairwise proportional imitation* [188].

Protocols of form (9) also appear in biological contexts, [144], [153,158], where in these cases we refer to them as *natural selection protocols*. The biological interpretation of (9) supposes that each agent is programmed to play a single pure strategy. An agent who receives a revision opportunity dies, and is replaced through asexual reproduction. The reproducing agent is a strategy j player with probability $\rho_{ij}(\pi, x) = x_j \hat{\rho}_{ij}(\pi, x)$, which is proportional both to the number of strategy j players and to some function of the prevalences and fitnesses of all strategies. Note that this interpretation requires the restriction

$$\sum_{j \in S} \rho_{ij}(\pi, x) \equiv 1.$$

Example 12 Suppose that payoffs are always positive, and let

$$\rho_{ij}(\pi, x) = \frac{x_j \pi_j}{\sum_{k \in S} x_k \pi_k}. \quad (10)$$

Understood as a natural selection protocol, (10) says that the probability that the reproducing agent is a strategy j player is proportional to $x_j \pi_j$, the aggregate fitness of strategy j players.

In economic contexts, we can interpret (10) as an imitative protocol based on repeated sampling. When an agent's clock rings he chooses an opponent at random. If the opponent is playing strategy j , the agent imitates him with probability proportional to π_j . If the agent does not imitate this opponent, he draws a new opponent at random and repeats the procedure.

Direct Evaluation Protocols In the previous examples, only strategies currently in use have any chance of being chosen by a revising agent (or of being the programmed strategy of the newborn agent). Under other protocols, agents' choices are not mediated through the population's current behavior, except indirectly via the effect of behavior on payoffs. These *direct evaluation protocols* require agents to directly evaluate the payoffs of the strategies they consider, rather than to indirectly evaluate them as under an imitative procedure.

Example 13 Suppose that choices are made according to the *logit choice rule*:

$$\rho_{ij}(\pi, x) = \frac{\exp(\eta^{-1} \pi_j)}{\sum_{k \in S} \exp(\eta^{-1} \pi_k)}. \quad (11)$$

The interpretation of this protocol is simple. Revision opportunities arrive at unit rate. When an opportunity is received by an i player, he switches to strategy j with probability $\rho_{ij}(\pi, x)$, which is proportional to an exponential function of strategy j 's payoffs. The parameter $\eta > 0$ is called the *noise level*. If η is large, choice probabilities under the logit rule are nearly uniform. But if η is near zero, choices are optimal with probability close to one, at least when the difference between the best and second best payoff is not too small.

Additional examples of revision protocols can be found in the next section, and one can construct new revision protocols by taking linear combinations of old ones; see [183] for further discussion.

Deterministic Dynamics

Although antecedents of this approach date back to the early work of Brown and von Neumann [45], the use of differential equations to model evolution in games took root with the introduction of the replicator dynamic by Taylor and Jonker [205], and remains an vibrant area of re-

search; Hofbauer and Sigmund [108] and Sandholm [183] offer recent surveys. In this section, we derive a deterministic model of evolution: the *mean dynamic* generated by a revision protocol and a population game. We study this deterministic model from various angles, focusing in particular on local stability of rest points, global convergence to equilibrium, and nonconvergent limit behavior.

While the bulk of the literature on deterministic evolutionary dynamics is consistent with the approach we take here, we should mention that other specifications exist, including discrete time dynamics [5,59,131,218], and dynamics for games with continuous strategy sets [41,42,77,100,159,160] and for Bayesian population games [62,70,179]. Also, deterministic dynamics for extensive form games introduce new conceptual issues; see [28,30,51,53,55] and the monograph of Cressman [54].

Mean Dynamics

As described earlier in Sect. "Definition", a revision protocol ρ , a population game F , and a population size N define a Markov process $\{X_t^N\}$ on the finite state space \mathcal{X}^N . We now derive a deterministic process – the *mean dynamic* – that describes the expected motion of $\{X_t^N\}$. In Sect. "Deterministic Approximation", we will describe formally the sense in which this deterministic process provides a very good approximation of the behavior of the stochastic process $\{X_t^N\}$, at least over finite time horizons and for large population sizes. But having noted this result, we will focus in this section on the deterministic process itself.

To compute the expected increment of $\{X_t^N\}$ over the next dt time units, recall first that each of the N agents receives revision opportunities via a rate R exponential distribution, and so expects to receive Rdt opportunities during the next dt time units. If the current state is x , the expected number of revision opportunities received by agents currently playing strategy i is approximately $Nx_i Rdt$. Since an i player who receives a revision opportunity switches to strategy j with probability ρ_{ij}/R , the expected number of such switches during the next dt time units is approximately $Nx_i \rho_{ij} dt$. Therefore, the expected change in the number of agents choosing strategy i during the next dt time units is approximately

$$N \left(\sum_{j \in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j \in S} \rho_{ij}(F(x), x) \right) dt. \quad (12)$$

Dividing expression (12) by N and eliminating the time differential dt yields a differential equation for the rate of

change in the *proportion* of agents choosing strategy i :

$$\dot{x}_i = \sum_{j \in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j \in S} \rho_{ij}(F(x), x). \quad (\text{M})$$

Equation (M) is the *mean dynamic* (or *mean field*) generated by revision protocol ρ in population game F . The first term in (M) captures the inflow of agents to strategy i from other strategies, while the second captures the outflow of agents to other strategies from strategy i .

Examples

We now describe some examples of mean dynamics, starting with ones generated by the revision protocols from Sect. “Examples”. To do so, we let

$$\bar{F}(x) = \sum_{i \in S} x_i F_i(x)$$

denote the *average payoff* obtained by the members of the population, and define the *excess payoff* to strategy i ,

$$\hat{F}_i(x) = F_i(x) - \bar{F}(x),$$

to be the difference between strategy i 's payoff and the population's average payoff.

Example 14 In Example 11, we introduced the pairwise proportional imitation protocol $\rho_{ij}(\pi, x) = x_j[\pi_j - \pi_i]_+$. This protocol generates the mean dynamic

$$\dot{x}_i = x_i \hat{F}_i(x). \quad (13)$$

Equation (13) is the *replicator dynamic* [205], the best-known dynamic in evolutionary game theory. Under this dynamic, the percentage growth rate \dot{x}_i/x_i of each strategy currently in use is equal to that strategy's current excess payoff; unused strategies always remain so. There are a variety of revision protocols other than pairwise proportional imitation that generate the replicator dynamic as their mean dynamics; see [35,96,108,217].

Example 15 In Example 12, we assumed that payoffs are always positive, and introduced the protocol $\rho_{ij} \propto x_j \pi_j$, which we interpreted both as a model of biological natural selection and as a model of imitation with repeated sampling. The resulting mean dynamic,

$$\dot{x}_i = \frac{x_i F_i(x)}{\sum_{k \in S} x_k F_k(x)} - x_i = \frac{x_i \hat{F}_i(x)}{\bar{F}(x)}, \quad (14)$$

is the *Maynard Smith replicator dynamic* [139]. This dynamic only differs from the standard replicator dynamic

(13) by a change of speed, with motion under (14) being relatively fast when average payoffs are relatively low. (In multipopulation models, the two dynamics are less similar, and convergence under one does not imply convergence under the other – see [183,216].)

Example 16 In Example 13 we introduced the logit choice rule $\rho_{ij}(\pi, x) \propto \exp(\eta^{-1} \pi_j)$. The corresponding mean dynamic,

$$\dot{x}_i = \frac{\exp(\eta^{-1} F_i(x))}{\sum_{k \in S} \exp(\eta^{-1} F_k(x))} - x_i, \quad (15)$$

is called the *logit dynamic* [82].

If we take the noise level η to zero, then the probability with which a revising agent chooses the best response approaches one whenever the best response is unique. At such points, the logit dynamic approaches the *best response dynamic* [84]:

$$\dot{x} \in B^F(x) - x, \quad (16)$$

where

$$B^F(x) = \operatorname{argmax}_{y \in X} y' F(x)$$

defines the (*mixed*) *best response correspondence* for game F . Note that unlike the other dynamics we consider here, (16) is defined not by an ordinary differential equation, but by a differential inclusion, a formulation proposed in [97].

Example 17 Consider the protocol

$$\rho_{ij}(\pi, x) = \left[\pi_j - \sum_{k \in S} x_k \pi_k \right]_+.$$

When an agent's clock rings, he chooses a strategy at random; if that strategy's payoff is above average, the agent switches to it with probability proportional to its excess payoff. The resulting mean dynamic,

$$\dot{x}_i BM = [\hat{F}_i(x)]_+ - x_i \sum_{k \in S} [\hat{F}_k(x)]_+,$$

is called the *Brown–von Neumann–Nash (BNN) dynamic* [45]; see also [98,176,194,200,217].

Example 18

Consider the revision protocol

$$\rho_{ij}(\pi, x) = [\pi_j - \pi_i]_+.$$

When an agent's clock rings, he selects a strategy at random. If the new strategy's payoff is higher than his cur-

Evolutionary Game Theory, Table 1
Five basic deterministic dynamics

Revision Protocol	Mean Dynamic	Name and source
$\rho_{ij} = x_j[\pi_j - \pi_i]_+$	$\dot{x}_i = x_i \hat{F}_i(x)$	Replicator [205]
$\rho_{ij} = \frac{\exp(\eta^{-1} \pi_j)}{\sum_{k \in S} \exp(\eta^{-1} \pi_k)}$	$\dot{x}_i = \frac{\exp(\eta^{-1} F_i(x))}{\sum_{k \in S} \exp(\eta^{-1} F_k(x))} - x_i$	Logit [82]
$\rho_{ij} = 1_{\{j = \arg \max_{k \in S} \pi_k\}}$	$\dot{x} \in B^F(x) - x$	Best response [84]
$\rho_{ij} = [\pi_j - \sum_{k \in S} x_k \pi_k]_+$	$\dot{x}_i = [\hat{F}_i(x)]_+ - x_i \sum_{j \in S} [F_j(x)]_+$	BNN [45]
$\rho_{ij} = [\pi_j - \pi_i]_+$	$\dot{x}_i = \sum_{j \in S} x_j [F_i(x) - F_j(x)]_+ - x_i \sum_{j \in S} [F_j(x) - F_i(x)]_+$	Smith [197]

rent strategy's payoff, he switches strategies with probability proportional to the difference between the two payoffs. The resulting mean dynamic,

$$\dot{x}_i = \sum_{j \in S} x_j [F_i(x) - F_j(x)]_+ - x_i \sum_{j \in S} [F_j(x) - F_i(x)]_+, \quad (17)$$

is called the *Smith dynamic* [197]; see also [178].

We summarize these examples of revision protocols and mean dynamics in Table 1.

Figure 1 presents phase diagrams for the five basic dynamics when the population is randomly matched to play standard Rock–Paper–Scissors (Example 4). In the phase diagrams, colors represent speed of motion: within each diagram, motion is fastest in the red regions and slowest in the blue ones.

The phase diagram of the replicator dynamic reveals closed orbits around the unique Nash equilibrium $x^* = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Since this dynamic is based on imitation (or on reproduction), each face and each vertex of the simplex X is an invariant set: a strategy initially absent from the population will never subsequently appear.

The other four dynamics pictured are based on direct evaluation, allowing agents to select strategies that are currently unused. In these cases, the Nash equilibrium is the sole rest point, and attracts solutions from all initial conditions. (In the case of the logit dynamic, the rest point happens to coincide with the Nash equilibrium only because of the symmetry of the game; see [101, 104].) Under the logit and best response dynamics, solution trajectories quickly change direction and then accelerate when the best response to the population state changes; under the BNN and especially the Smith dynamic, solutions approach the Nash equilibrium in a less angular fashion.

Evolutionary Justification of Nash Equilibrium

One of the goals of evolutionary game theory is to justify the prediction of Nash equilibrium play. For this justification to be convincing, it must be based on a model that makes only mild assumptions about agents' knowledge about one another's behavior. This sentiment can be captured by introducing two desiderata for revision protocols:

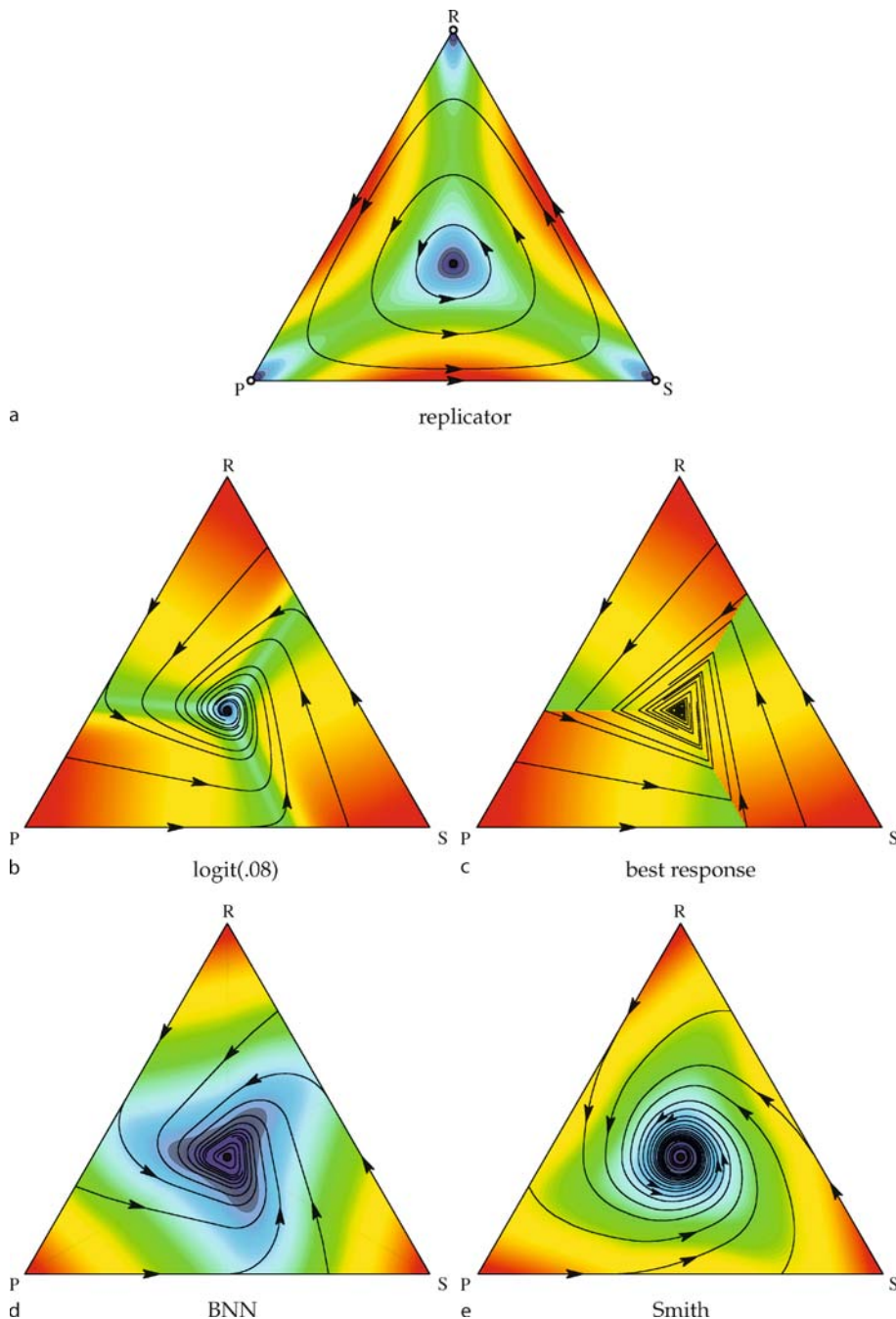
- (C) *Continuity*: ρ is Lipschitz continuous.
 (SD) *Scarcity of data*: ρ_{ij} only depends on π_i, π_j , and x_j .

Continuity (C) asks that revision protocols depend continuously on their inputs, so that small changes in aggregate behavior do not lead to large changes in players' responses. *Scarcity of data* (SD) demands that the conditional switch rate from strategy i to strategy j only depend on the payoffs of these two strategies, so that agents need only know those facts that are most germane to the decision at hand [183]. (The dependence of ρ_{ij} on x_j is included to allow for dynamics based on imitation.) Protocols that respect these two properties do not make unrealistic demands on the amount of information that agents in an evolutionary model possess.

Our two remaining desiderata impose restrictions on mean dynamics $\dot{x} = V^F(x)$, linking the evolution of aggregate behavior to incentives in the underlying game.

- (NS) *Nash stationarity*:
 $V^F(x) = \mathbf{0}$ if and only if $x \in NE(F)$.
 (PC) *Positive correlation*:
 $V^F(x) \neq \mathbf{0}$ implies that $V^F(x)' F(x) > 0$.

Nash stationarity (NS) is a restriction on stationary states: it asks that the rest points of the mean dynamic be pre-



Evolutionary Game Theory, Figure 1

Five basic deterministic dynamics in standard Rock–Paper–Scissors. Colors represent speeds: *red* is fastest, *blue* is slowest

cisely the Nash equilibria of the game being played. *Positive correlation* (PC) is a restriction on disequilibrium adjustment: it requires that away from rest points, strategies' growth rates be positively correlated with their pay-

offs. Condition (PC) is among the weakest of the many conditions linking growth rates of evolutionary dynamics and payoffs in the underlying game; for alternatives, see [76,110,149,162,170,173,200].

Evolutionary Game Theory, Table 2

Families of deterministic evolutionary dynamics and their properties; yes* indicates that a weaker or alternate form of the property is satisfied

Dynamic	Family	(C)	(SD)	(NS)	(PC)
Replicator	Imitation	yes	yes	no	yes
Best response		no	yes*	yes*	yes*
Logit	Perturbed best response	yes	yes*	no	no
BNN	Excess payoff	yes	no	yes	yes
Smith	Pairwise comparison	yes	yes	yes	yes

In Table 2, we report how the the five basic dynamics fare under the four criteria above. For the purposes of justifying the Nash prediction, the most important row in the table is the last one, which reveals that the Smith dynamic satisfies all four desiderata at once: while the revision protocol for the Smith dynamic (see Example 18) requires only limited information on the part of the agents who employ it, this information is enough to ensure that rest points of the dynamic and Nash equilibria coincide.

In fact, the dynamics introduced above can be viewed as members of families of dynamics that are based on similar revision protocols and that have similar qualitative properties. For instance, the Smith dynamic is a member of the family of *pairwise comparison* dynamics [178], under which agents only switch to strategies that outperform their current choice. For this reason, the exact functional forms of the previous examples are not essential to establishing the properties noted above.

In interpreting these results, it is important to remember that Nash stationarity only concerns the rest points of a dynamic; it says nothing about whether a dynamic will converge to Nash equilibrium from an arbitrary initial state. The question of convergence is addressed in Sects. “Global Convergence” and “Nonconvergence”. There we will see that in some classes of games, general guarantees of convergence can be obtained, but that there are some games in which no reasonable dynamic converges to equilibrium.

Local Stability

Before turning to the global behavior of evolutionary dynamics, we address the question of local stability. As we noted at the onset, an original motivation for introducing game dynamics was to provide an explicitly dynamic foundation for Maynard Smith’s notion of ESS [205]. Some of the earliest papers on evolutionary game dynamics [105,224] established that being an ESS is a sufficient condition for asymptotically stability under the replicator dynamic, but that it is not a necessary condition. It is cu-

rious that this connection obtains despite the fact that ESS is a stability condition for a population whose members all play the same mixed strategy, while (the usual version of) the replicator dynamic looks at populations of agents choosing among different pure strategies.

In fact, the implications of ESS for local stability are not limited to the replicator dynamic. Suppose that the symmetric normal form game A admits a symmetric Nash equilibrium that places positive probability on each strategy in S . One can show that this equilibrium is an ESS if and only if the payoff matrix A is negative definite with respect to the tangent space of the simplex:

$$z'Az < 0 \quad \text{for all } z \in TX = \left\{ \hat{z} \in \mathbf{R}^n : \sum_{i \in S} \hat{z}_i = 0 \right\}. \quad (18)$$

Condition (18) and its generalizations imply local stability of equilibrium not only under the replicator dynamic, but also under a wide range of other evolutionary dynamics: see [52,98,99,102,111,179] for further details.

The papers cited above use linearization and Lyapunov function arguments to establish local stability. An alternative approach to local stability analysis, via index theory, allows one to establish restrictions on the stability properties of all rest points at once – see [60].

Global Convergence

While analyses of local stability reveal whether a population will return to equilibrium after a small disturbance, they do not tell us whether the population will approach equilibrium from an arbitrary disequilibrium state. To establish such global convergence results, we must restrict attention to classes of games defined by certain interesting payoff structures. These structures appear in applications, lending strong support for the Nash prediction in the settings where they arise.

Potential Games A *potential game* [17,106,143,166,173,181] is a game that admits a *potential function*: a scalar val-

ued function whose gradient describes the game's payoffs. In a full potential game $F: \mathbf{R}_+^n \rightarrow \mathbf{R}^n$ (see [181]), all information about incentives is captured by the potential function $f: \mathbf{R}_+^n \rightarrow \mathbf{R}$, in the sense that

$$\nabla f(x) = F(x) \quad \text{for all } x \in \mathbf{R}_+^n. \quad (19)$$

If F is smooth, then it is a full potential game if and only if it satisfies *full externality symmetry*:

$$\frac{\partial F_i}{\partial x_j}(x) = \frac{\partial F_j}{\partial x_i}(x) \quad \text{for all } i, j \in S \text{ and } x \in \mathbf{R}_+^n. \quad (20)$$

That is, the effect on the payoff to strategy i of adding new strategy j players always equals the effect on the payoff to strategy j of adding new strategy i players.

Example 19 Suppose a single population is randomly matched to play the symmetric normal form game $A \in \mathbf{R}^{n \times n}$, generating the population game $F(x) = Ax$. We say that A exhibits *common interests* if the two players in a match always receive the same payoff. This means that $A_{ij} = A_{ji}$ for all i and j , or, equivalently, that the matrix A is symmetric. Since $DF(x) = A$, this is precisely what we need for F to be a full potential game. The full potential function for F is $f(x) = \frac{1}{2}x'Ax$, which is one-half of the average payoff function $\bar{F}(x) = \sum_{i \in S} x_i F_i(x) = x'Ax$. The common interest assumption defines a fundamental model from population genetics, this assumption reflects the shared fate of two genes that inhabit the same organism [73,106,107].

Example 20 In Example 10, we introduced congestion games, a basic model of network congestion. To see that these games are potential games, observe that an agent taking path $j \in S$ affects the payoffs of agents choosing path $i \in S$ through the marginal increases in congestion on the links $\phi \in \Phi_i \cap \Phi_j$ that the two paths have in common. But since the marginal effect of an agent taking path i on the payoffs of agents choosing path j is identical, full externality symmetry (20) holds:

$$\frac{\partial F_i}{\partial x_j}(x) = - \sum_{\phi \in \Phi_i \cap \Phi_j} c'_\phi(u_\phi(x)) = \frac{\partial F_j}{\partial x_i}(x).$$

In congestion games, the potential function takes the form

$$f(x) = - \sum_{\phi \in \Phi} \int_0^{u_\phi(x)} c_\phi(z) dz,$$

and so is typically unrelated to aggregate payoffs,

$$\bar{F}(x) = \sum_{i \in S} x_i F_i(x) = - \sum_{\phi \in \Phi} u_\phi(x) c_\phi(u_\phi(x)).$$

However, potential is proportional to aggregate payoffs if the cost functions c_ϕ are all monomials of the same degree [56,173].

Population state x is a Nash equilibrium of the potential game F if and only if it satisfies the Kuhn–Tucker first order conditions for maximizing the potential function f on the simplex X [17,173]. Furthermore, it is simple to verify that any dynamic $\dot{x} = V^F(x)$ satisfying positive correlation (PC) ascends the potential function:

$$\frac{d}{dt} f(x_t) = \nabla f(x_t)' \dot{x}_t = F(x_t)' V^F(x_t) \geq 0.$$

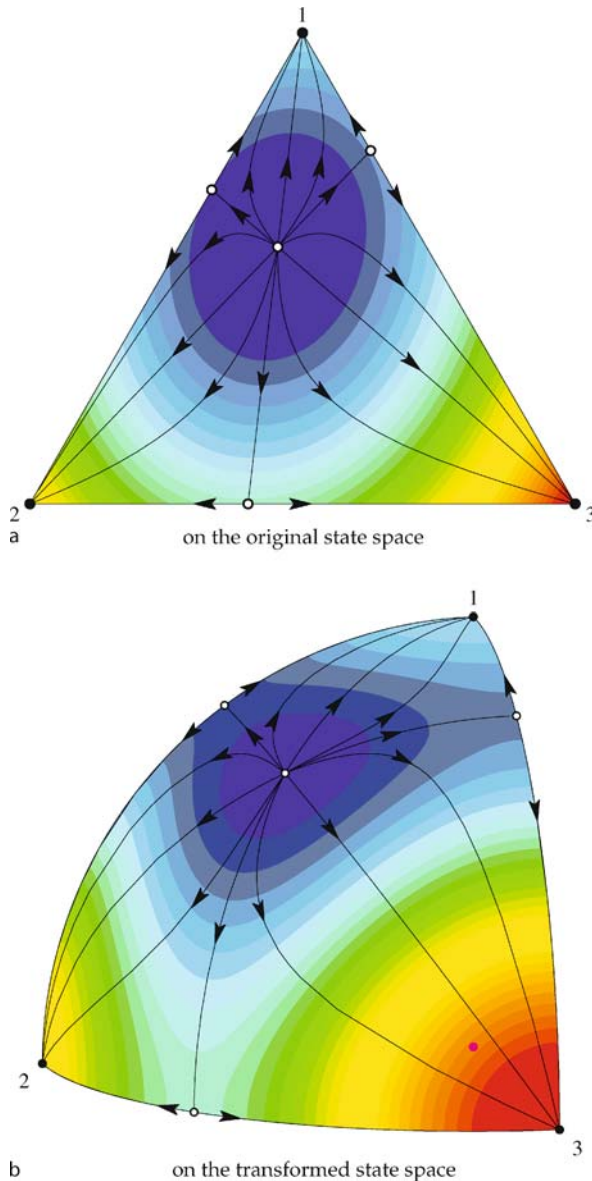
It then follows from classical results on Lyapunov functions that any dynamic satisfying positive correlation (PC) converges to a connected set of rest points. If the dynamic also satisfies Nash stationarity (NS), these sets consist entirely of Nash equilibria. Thus, in potential games, very mild conditions on agents' adjustment rules are sufficient to justify the prediction of Nash equilibrium play.

In the case of the replicator dynamic, one can say more. On the interior of the simplex X , the replicator dynamic for the potential game F is a *gradient system* for the potential function f (i.e., it always ascends f in the direction of maximum increase). However, this is only true after one introduces an appropriate Riemannian metric on X [123,192]. An equivalent statement of this result, due to [2], is that the replicator dynamic is the gradient system for f under the usual Euclidean metric if we stretch the state space X onto the radius 2 sphere. This stretching is accomplished using the *Akin transformation* $H_i(x) = 2\sqrt{x_i}$, which emphasizes changes in the use of rare strategies relative to changes in the use of common ones [2,4,185]. (There is also a dynamic that generates the gradient system for f on X under the usual metric: the so-called *projection dynamic* [130,150,185].)

Example 21 Consider evolution in 123 Coordination:

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 3 & 0 & 0 & 3 \end{array}.$$

Figure 2a presents a phase diagram of the replicator dynamic on its natural state space X , drawn atop of a contour plot of the potential function $f(x) = \frac{1}{2}((x_1)^2 + 2(x_2)^2 + 3(x_3)^2)$. Evidently, all solution trajectories ascend this function and converge to one of the seven symmetric Nash equilibria, with trajectories from all but a measure



Evolutionary Game Theory, Figure 2

The replicator dynamic in 123 Coordination. Colors represent the value of the game's potential function

zero set of initial conditions converging to one of the three pure equilibria.

Figure 2b presents another phase diagram for the replicator dynamic, this time after the solution trajectories and the potential function have been transported to the surface of the radius 2 sphere using the Akin transformation. In this case, solutions cross the level sets of the potential function orthogonally, moving in the direction that increases potential most quickly.

Stable Games A population game F is a *stable game* [102] if

$$(y - x)'(F(y) - F(x)) \leq 0 \quad \text{for all } x, y \in X. \quad (21)$$

If the inequality in (21) always holds strictly, then F is a *strictly stable game*.

If F is smooth, then F is a stable game if and only if it satisfies *self-defeating externalities*:

$$z' DF(x)z \leq 0 \quad \text{for all } z \in TX \text{ and } x \in X, \quad (22)$$

where $DF(x)$ is the derivative of $F: X \rightarrow \mathbb{R}^n$ at x . This condition requires that the improvements in the payoffs of strategies to which revising agents are switching are always exceeded by the improvements in the payoffs of strategies which revising agents are abandoning.

Example 22 The symmetric normal form game A is *symmetric zero-sum* if A is skew-symmetric (i. e., if $A = -A'$), so that the payoffs of the matched players always sum to zero. (An example is provided by the standard Rock–Paper–Scissors game (Example 4).) Under this assumption, $z'Az = 0$ for all $z \in \mathbb{R}^n$; thus, the population game generated by random matching in A , $F(x) = Ax$, is a stable game that is not strictly stable.

Example 23 Suppose that A satisfies the interior ESS condition (18). Then (22) holds strictly, so $F(x) = Ax$ is a strictly stable game. Examples satisfying this condition include the Hawk–Dove game (Example 3) and any good Rock–Paper–Scissors game (Example 4).

Example 24 A *war of attrition* [33] is a symmetric normal form game in which strategies represent amounts of time committed to waiting for a scarce resource. If the two players choose times i and $j > i$, then the j player obtains the resource, worth v , while both players pay a cost of c_i ; once the first player leaves, the other seizes the resource immediately. If both players choose time i , the resource is split, so payoffs are $\frac{v}{2} - c_i$ each. It can be shown that for any resource value $v \in \mathbb{R}$ and any increasing cost vector $c \in \mathbb{R}^n$, random matching in a war of attrition generates a stable game [102].

The flavor of the self-defeating externalities condition (22) suggests that obedience of incentives will push the population toward some “central” equilibrium state. In fact, the set of Nash equilibria of a stable game is always convex, and in the case of strictly stable games, equilibrium is unique. Moreover, it can be shown that the replicator dynamic converges to Nash equilibrium from all interior initial conditions in any strictly stable

Evolutionary Game Theory, Table 3

Lyapunov functions for five basic deterministic dynamics in stable games

Dynamic	Lyapunov function for stable games
Replicator	$H_{x^*}(x) = \sum_{i \in S(x^*)} x_i^* \log \frac{x_i^*}{x_i}$
Logit	$\tilde{G}(x) = \max_{y \in \text{int}(X)} (y' \hat{F}(x) - \eta \sum_{i \in S} y_i \log y_i) + \eta \sum_{i \in S} x_i \log x_i$
Best response	$G(x) = \max_{i \in S} \hat{F}_i(x)$
BNN	$\Gamma(x) = \frac{1}{2} \sum_{i \in S} [\hat{F}_i(x)]_+^2$
Smith	$\Psi(x) = \frac{1}{2} \sum_{i \in S} \sum_{j \in S} x_i [F_j(x) - F_i(x)]_+^2$

game [4,105,224], and that the direct evaluation dynamics introduced above converge to Nash equilibrium from all initial conditions in all stable games, strictly stable or not [98,102,104,197]. In each case, the proof of convergence is based on the construction of a Lyapunov function that solutions of the relevant dynamic descend. The Lyapunov functions for the five basic dynamics are presented in Table 3.

Interestingly, the convergence results for direct evaluation dynamics are not restricted to the dynamics listed in Table 3, but extend to other dynamics in the same families (cf Table 2). But compared to the conditions for convergence in potential games, the conditions for convergence in stable games demand additional structure on the adjustment process [102].

Perturbed Best Response Dynamics in Supermodular Games Supermodular games are defined by the property that higher choices by one's opponents (with respect to the natural ordering on $S = \{1, \dots, n\}$) make one's own higher strategies look relatively more desirable. Let the matrix $\Sigma \in \mathbf{R}^{(n-1) \times n}$ satisfy $\Sigma_{ij} = 1$ if $j > i$ and $\Sigma_{ij} = 0$ otherwise, so that $\Sigma x \in \mathbf{R}^{n-1}$ is the "decumulative distribution function" corresponding to the "density function" x . The population game F is a *supermodular game* if it exhibits *strategic complementarities*:

If $\Sigma y \geq \Sigma x$, then

$$F_{i+1}(y) - F_i(y) \geq F_{i+1}(x) - F_i(x) \quad \text{for all } i < n \text{ and } x \in X. \quad (23)$$

If F is smooth, condition (23) is equivalent to

$$\frac{\partial(F_{i+1} - F_i)}{\partial(e_{j+1} - e_j)}(x) \geq 0 \quad \text{for all } i, j < n \text{ and } x \in X. \quad (24)$$

Example 25 Consider this model of *search with positive externalities*. A population of agents choose levels of search effort in $S = \{1, \dots, n\}$. The payoff to choosing effort i is

$$F_i(x) = m(i) b(a(x)) - c(i),$$

where $a(x) = \sum_{k \leq n} kx_k$ is the aggregate search effort, b is some increasing benefit function, m is an increasing multiplier function, and c is an arbitrary cost function. Notice that the benefits from searching are increasing in both own search effort and in the aggregate search effort. It is easy to check that F is a supermodular game.

Complementarity condition (23) implies that the agents' best response correspondence is monotone in the stochastic dominance order, which in turn ensures the existence of minimal and maximal Nash equilibria [207]. One can take advantage of the monotonicity of best responses in studying evolutionary dynamics by appealing to the theory of monotone dynamical systems [196]. To do so, one needs to focus on dynamics that respect the monotonicity of best responses and that also are smooth, so that the theory of monotone dynamics can be applied. It turns out that the logit dynamic satisfies these criteria; so does any perturbed best response dynamic defined in terms of stochastic payoff perturbations. In supermodular games, these dynamics define cooperative differential equations; consequently, solutions of these dynamics from almost every initial condition converge to an approximate Nash equilibrium [104].

Imitation Dynamics in Dominance Solvable Games

Suppose that in the population game F , strategy i is a strictly dominated by strategy j : $F_i(x) < F_j(x)$ for all $x \in X$. Consider the evolution of behavior under the repli-

cator dynamic (13). Since for this dynamic we have

$$\begin{aligned}\frac{d}{dt} \frac{x_i}{x_j} &= \frac{\dot{x}_i x_j - \dot{x}_j x_i}{(x_j)^2} \\ &= \frac{x_i \hat{F}_i(x) x_j - x_j \hat{F}_j(x) x_i}{(x_j)^2} \\ &= \frac{x_i}{x_j} (\hat{F}_i(x) - \hat{F}_j(x)),\end{aligned}$$

solutions from every interior initial condition converge to the face of the simplex where the dominated strategy is unplayed [3]. It follows that the replicator dynamic converges in games with a strictly dominant strategy, and by iterating this argument, one can show that this dynamic converges to equilibrium in any game that can be solved by iterative deletion of strictly dominated strategies. In fact, this argument is not specific to the replicator dynamic, but can be shown to apply to a range of dynamics based on imitation [110,170]. Even in games which are not dominance solvable, arguments of a similar flavor can be used to restrict the long run behavior of imitative dynamics to better-reply closed sets [162]; see Sect. “Convergence to Equilibria and to Better-Reply Closed Sets” for a related discussion.

While the analysis here has focused on imitative dynamics, it is natural to expect that elimination of dominated strategies will extend to any reasonable evolutionary dynamic. But we will see in Sect. “Survival of Dominated Strategies” that this is not the case: the elimination of dominated strategies that obtains under imitative dynamics is the exception, not the rule.

Nonconvergence

The previous section revealed that when certain global structural conditions on payoffs are satisfied, one can establish global convergence to equilibrium under various classes of evolutionary dynamics. Of course, if these conditions are not met, convergence cannot be guaranteed. In this section, we offer examples to illustrate some of the possibilities for nonconvergent limit behavior.

Conservative Properties of the Replicator Dynamic in Zero-Sum Games In Sect. “Stable Games”, we noted that in strictly stable games, the replicator dynamic converges to Nash equilibrium from all interior initial conditions. To prove this, one shows that interior solutions descend the function

$$H_{x^*}(x) = \sum_{i \in S(x^*)} x_i^* \log \frac{x_i}{x_i^*},$$

until converging to its minimizer, the unique Nash equilibrium x^* .

Now, random matching in a symmetric zero-sum game generates a population game that is stable, but not strictly stable (Example 22). In this case, for each interior Nash equilibrium x^* , the function H_{x^*} is a constant of motion for the replicator dynamic: its value is fixed along every interior solution trajectory.

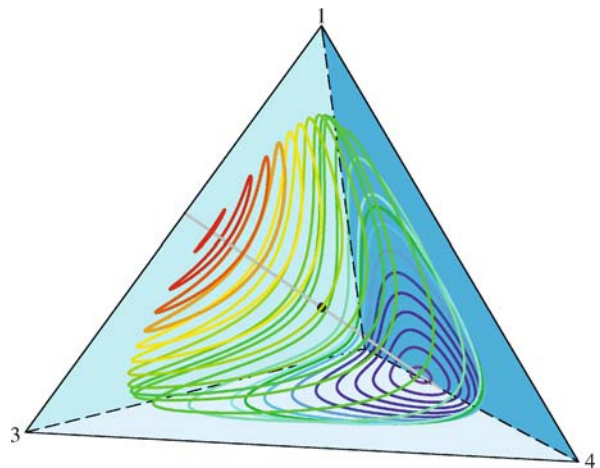
Example 26 Suppose that agents are randomly matched to play the symmetric zero-sum game A , given by

	1	2	3	4
1	0	-1	0	1
2	1	0	-1	0
3	0	1	0	-1
4	-1	0	1	0

The Nash equilibria of $F(x) = Ax$ are the points on the line segment NE connecting states $(\frac{1}{2}, 0, \frac{1}{2}, 0)$ and $(0, \frac{1}{2}, 0, \frac{1}{2})$, a segment that passes through the barycenter $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$. Figure 3 shows solutions to the replicator dynamic that lie on the level set $H_{x^*}(x) = .58$. Evidently, each of these solutions forms a closed orbit.

Although solution trajectories of the replicator dynamic do not converge in zero-sum games, it can be proved that the time average of each solution trajectory converges to Nash equilibrium [190].

The existence of a constant of motion is not the only conservative property enjoyed by replicator dynamics for symmetric zero-sum games: these dynamics are also vol-



Evolutionary Game Theory, Figure 3

Solutions of the replicator dynamic in a zero-sum game. The solutions pictured lie on the level set $H_{x^*}(x) = .58$

ume preserving after an appropriate change of speed or change of measure [5,96].

Games with Nonconvergent Dynamics The conservative properties described in the previous section have been established only for the replicator dynamic (and its distant relative, the projection dynamic [185]). Inspired by Shapley [193], many researchers have sought to construct games in which large classes of evolutionary dynamics fail to converge to equilibrium.

Example 27 Suppose that players are randomly matched to play the following symmetric normal form game [107,109]:

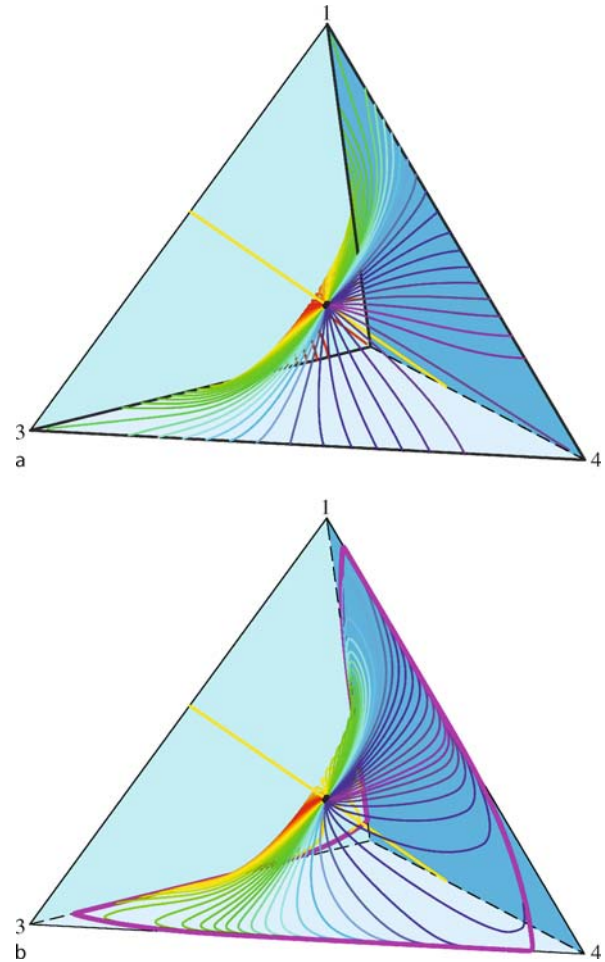
	1	2	3	4
1	0	0	-1	ε
2	ε	0	0	-1
3	-1	ε	0	0
4	0	-1	ε	0

When $\varepsilon = 0$, the payoff matrix $A^\varepsilon = A^0$ is symmetric, so F^0 is a potential game with potential function $f(x) = \frac{1}{2}x'A^0x = -x_1x_3 - x_2x_4$. The function f attains its minimum of $-\frac{1}{4}$ at states $v = (\frac{1}{2}, 0, \frac{1}{2}, 0)$ and $w = (0, \frac{1}{2}, 0, \frac{1}{2})$, has a saddle point with value $-\frac{1}{8}$ at the Nash equilibrium $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, and attains its maximum of 0 along the closed path of Nash equilibria γ consisting of edges e_1e_2 , e_2e_3 , e_3e_4 , and e_4e_1 .

Let $\dot{x} = V^F(x)$ be an evolutionary dynamic that satisfies Nash stationarity (NS) and positive correlation (PC), and that is based on a revision protocol that is continuous (C). If we apply this dynamic to game F^0 , then the foregoing discussion implies that all solutions to $\dot{x} = V^{F^0}(x)$ whose initial conditions ξ satisfy $f(\xi) > -\frac{1}{8}$ converge to γ . The Smith dynamic for F^0 is illustrated in Fig. 4a.

Now consider the same dynamic for the game F^ε , where $\varepsilon > 0$. By continuity (C), the attractor γ of V^{F^0} continues to an attractor γ^ε of V^{F^ε} whose basin of attraction approximates that of γ under $\dot{x} = V^{F^0}(x)$ (Fig. 4b). But since the unique Nash equilibrium of F^ε is the barycenter x^* , it follows that solutions from most initial conditions converge to an attractor far from any Nash equilibrium.

Other examples of games in which many dynamics fail to converge include monocyclic games [22,83,97,106], Mismatching Pennies [91,116], and the hypnodisk game [103]. These examples demonstrate that there is no evolutionary dynamic that converges to Nash equilibrium regardless of the game at hand. This suggests that in general, analyses of long run behavior should not restrict attention to equilibria alone.



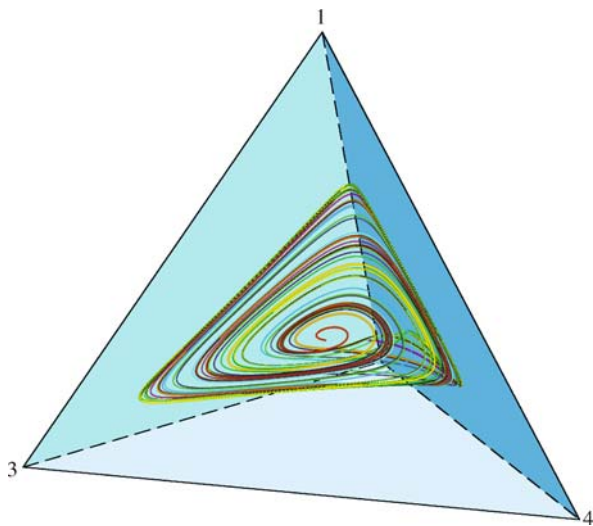
Evolutionary Game Theory, Figure 4

Solutions of the Smith dynamic in **a** the potential game F^0 ; **b** the perturbed potential game F^ε , $\varepsilon = \frac{1}{10}$

Chaotic Dynamics We have seen that deterministic evolutionary game dynamics can follow closed orbits and approach limit cycles. We now show that they also can behave chaotically.

Example 28 Consider evolution under the replicator dynamic when agents are randomly matched to play the symmetric normal form game below [13,195], whose lone interior Nash equilibrium is the barycenter $x^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$:

	1	2	3	4
1	0	-12	0	22
2	20	0	0	-10
3	-21	-4	0	35
4	10	-2	2	0



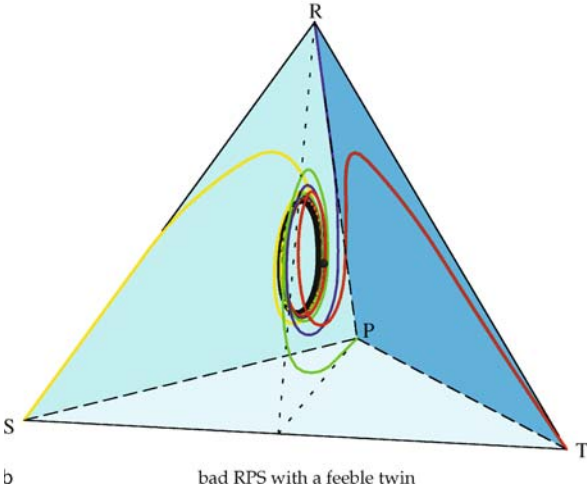
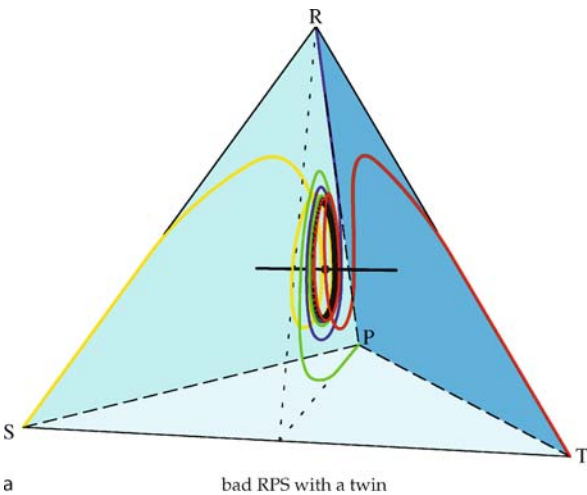
Evolutionary Game Theory, Figure 5
Chaotic behavior under the replicator dynamic

Figure 5 presents a solution to the replicator dynamic for this game from initial condition $x_0 = (.24, .26, .25, .25)$. This solution spirals clockwise about x^* . Near the right-most point of each circuit, where the value of x_3 gets close to zero, solutions sometimes proceed along an “outside” path on which the value of x_3 surpasses .6. But they sometimes follow an “inside” path on which x_3 remains below .4, and at other times do something in between. Which of these alternatives occurs is difficult to predict from approximate information about the previous behavior of the system.

While the game in Example 28 has a complicated payoff structure, in multipopulation contexts one can find chaotic evolutionary dynamics in very simple games [187].

Survival of Dominated Strategies In Sect. “Imitation Dynamics in Dominance Solvable Games”, we saw that dynamics based on imitation eliminate strictly dominated strategies along solutions from interior initial conditions. While this result seems unsurprising, it is actually extremely fragile: [25,103] prove that dynamics that satisfy continuity (C), Nash stationarity (NS), and positive correlation (PC) and that are not based exclusively on imitation must fail to eliminate strictly dominated strategies in some games. Thus, evolutionary support for a basic rationality criterion is more tenuous than the results for imitative dynamics suggest.

Example 29 Figure 6a presents the Smith dynamic for “bad RPS with a twin”:



Evolutionary Game Theory, Figure 6
The Smith dynamic in two games

	R	P	S	T
R	0	-2	1	1
P	1	0	-2	-2
S	-2	1	0	0
T	-2	1	0	0

The Nash equilibria of this game are the states on line segment $NE = \{x^* \in X: x^* = (\frac{1}{3}, \frac{1}{3}, c, \frac{1}{3} - c)\}$, which is a repellor under the Smith dynamic. Under this dynamic, strategies gain players at rates that depend on their payoffs, but lose players at rates proportional to their current usage levels. It follows that when the dynamics are not at rest, the proportions of players choosing strategies 3 and 4 become equal, so that the dynamic approaches the plane $P = \{x \in X: x_3 = x_4\}$ on which the twins receive equal

weight. Since the usual three-strategy version of bad RPS, exhibits cycling solutions here on the plane P approach a closed orbit away from any Nash equilibrium.

Figure 6b presents the Smith dynamic in “bad RPS with a feeble twin”,

	R	P	S	T
R	0	-2	1	1
P	1	0	-2	-2
S	-2	1	0	0
T	$-2 - \epsilon$	$1 - \epsilon$	$-\epsilon$	$-\epsilon$

with $\epsilon = \frac{1}{10}$. Evidently, the attractor from Fig. 6a moves slightly to the left, reflecting the fact that the payoff to Twin has gone down. But since the new attractor is in the interior of X , the strictly dominated strategy Twin is always played with probabilities bounded far away from zero.

Stochastic Dynamics

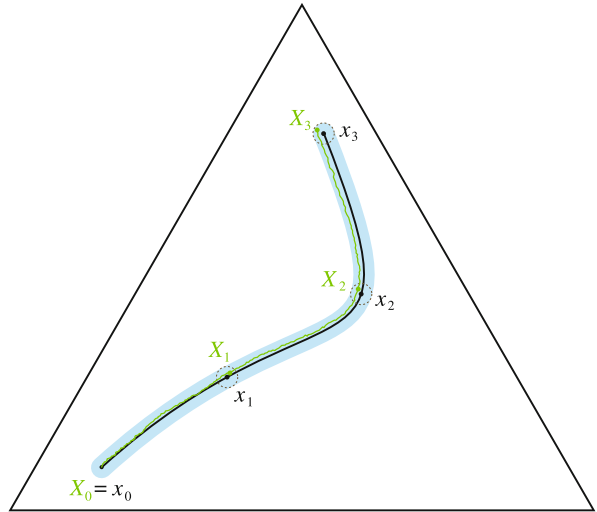
In Sect. “Revision Protocols” we defined the stochastic evolutionary process $\{X_t^N\}$ in terms of a simple model of myopic individual choice. We then turned to the study of deterministic dynamics, which we claimed could be used to approximate the stochastic process $\{X_t^N\}$ over finite time spans and for large population sizes. In this section, we turn our attention to the stochastic process $\{X_t^N\}$ itself. After offering a formal version of the deterministic approximation result, we investigate the long run behavior of $\{X_t^N\}$, focusing on the questions of convergence to equilibrium and selection among multiple stable equilibria.

Deterministic Approximation

In Sect. “Revision Protocols”, we defined the Markovian evolutionary process $\{X_t^N\}$ from a revision protocol ρ , a population game F , and a finite population size N . In Sect. “Mean Dynamics”, we argued that the expected motion of this process is captured by the mean dynamic

$$\dot{x}_i = V_i^F(x) = \sum_{j \in S} x_j \rho_{ji}(F(x), x) - x_i \sum_{j \in S} \rho_{ij}(F(x), x). \quad (\text{M})$$

The basic link between the Markov process $\{X_t^N\}$ and its mean dynamic (M) is provided by Kurtz’s Theorem [127], variations and extensions of which have been offered in a number of game-theoretic contexts [24,29,43,44,175,204]. Consider the sequence of Markov processes $\{\{X_t^N\}_{t \geq 0}\}_{N=N_0}^\infty$, supposing that the initial conditions X_0^N



Evolutionary Game Theory, Figure 7
Deterministic approximation of the Markov process $\{X_t^N\}$

converge to $x_0 \in X$. Let $\{x_t\}_{t \geq 0}$ be the solution to the mean dynamic (M) starting from x_0 . Kurtz’s Theorem tells us that for each finite time horizon $T < \infty$ and error bound $\epsilon > 0$, we have that

$$\lim_{N \rightarrow \infty} P \left(\sup_{t \in [0, T]} |X_t^N - x_t| < \epsilon \right) = 1. \quad (25)$$

Thus, when the population size N is large, nearly all sample paths of the Markov process $\{X_t^N\}$ stay within ϵ of a solution of the mean dynamic (M) through time T . By choosing N large enough, we can ensure that with probability close to one, X_t^N and x_t differ by no more than ϵ for all times t between 0 and T (Fig. 7).

The intuition for this result comes from the law of large numbers. At each revision opportunity, the increment in the process $\{X_t^N\}$ is stochastic. Still, at most population states the expected number of revision opportunities that arrive during the brief time interval $I = [t, t + dt]$ is large – in particular, of order Ndt . Since each opportunity leads to an increment of the state of size $\frac{1}{N}$, the size of the overall change in the state during time interval I is of order dt . Thus, during this interval there are a large number of revision opportunities, each following nearly the same transition probabilities, and hence having nearly the same expected increments. The law of large numbers therefore suggests that the change in $\{X_t^N\}$ during this interval should be almost completely determined by the expected motion of $\{X_t^N\}$, as described by the mean dynamic (M).

Convergence to Equilibria and to Better-Reply Closed Sets

Stochastic models of evolution can also be used to address directly the question of convergence to equilibrium [61,78,117,118,125,143,172,219]. Suppose that a society of agents is randomly matched to play an (asymmetric) normal form game that is *weakly acyclic in better replies*: from each strategy profile, there exists a sequence of profitable unilateral deviations leading to a Nash equilibrium. If agents switch to strategies that do at least as well as their current one against the choices of random samples of opponents, then the society will eventually escape any better-response cycle, ultimately settling upon a Nash equilibrium.

Importantly, many classes of normal form games are weakly acyclic in better replies: these include potential games, dominance solvable games, certain supermodular games, and certain *aggregative games*, in which each agent's payoffs only depend on opponents' behavior through a scalar aggregate statistic. Thus, in all of these cases, simple stochastic better-reply procedures are certain to lead to Nash equilibrium play.

Outside these classes of games, one can narrow down the possibilities for long run behavior by looking at *better-reply closed sets*: that is, subsets of the set of strategy profiles that cannot be escaped without a player switching to an inferior strategy (cf. [16,162]). Stochastic better-reply procedures must lead to a cluster of population states corresponding to a better-reply closed set; once the society enters such a cluster, it never departs.

Stochastic Stability and Equilibrium Selection

To this point, we used stochastic evolutionary dynamics to provide foundations for deterministic dynamics and to address the question of convergence to equilibrium. But stochastic evolutionary dynamics introduce an entirely new possibility: that of obtaining unique long-run predictions of play, even in games with multiple locally stable equilibria. This form of analysis, which we consider next, was pioneered by Foster and Young [74], Kandori, Mailath, and Rob [119], and Young [219], building on mathematical techniques due to Freidlin and Wentzell [75].

Stochastic Stability To minimize notation, let us describe the evolution of behavior using a discrete-time Markov chain $\{X_k^{N,\varepsilon}\}_{k=0}^{\infty}$ on \mathcal{X}^N , where the parameter $\varepsilon > 0$ represents the level of “noise” in agents' decision procedures. The noise ensures that the Markov chain is irreducible and aperiodic: any state in \mathcal{X}^N can be reached

from any other, and there is positive probability that a period passes without a change in the state.

Under these conditions, the Markov chain $\{X_k^{N,\varepsilon}\}$ admits a unique *stationary distribution*, $\mu^{N,\varepsilon}$, a measure on the state space \mathcal{X}^N that is invariant under the Markov chain:

$$\sum_{x \in \mathcal{X}^N} \mu^{N,\varepsilon}(x) P(X_{k+1}^{N,\varepsilon} = y \mid X_k^{N,\varepsilon} = x) = \mu^{N,\varepsilon}(y) \quad \text{for all } y \in \mathcal{X}^N.$$

The stationary distribution describes the long run behavior of the process $\{X_t^{N,\varepsilon}\}$ in two distinct ways. First, $\mu^{N,\varepsilon}$ is the *limiting distribution* of $\{X_t^{N,\varepsilon}\}$:

$$\lim_{k \rightarrow \infty} P(X_k^{N,\varepsilon} = y \mid X_0^{N,\varepsilon} = x) = \mu^{N,\varepsilon}(y) \quad \text{for all } x, y \in \mathcal{X}^N.$$

Second, $\mu^{N,\varepsilon}$ almost surely describes the *limiting empirical distribution* of $\{X_t^{N,\varepsilon}\}$:

$$P\left(\lim_{K \rightarrow \infty} \frac{1}{K} \sum_{k=0}^{K-1} 1_{\{X_k^{N,\varepsilon} \in A\}} = \mu^{N,\varepsilon}(A)\right) = 1 \quad \text{for any } A \subseteq \mathcal{X}^N.$$

Thus, if most of the mass in the stationary distribution $\mu^{N,\varepsilon}$ were placed on a single state, then this state would provide a unique prediction of long run behavior.

With this motivation, consider a sequence of Markov chains $\{\{X_k^{N,\varepsilon}\}_{k=0}^{\infty}\}_{\varepsilon \in (0,\bar{\varepsilon})}$ parametrized by noise levels ε that approach zero. Population state $x \in \mathcal{X}^N$ is said to be *stochastically stable* if it retains positive weight in the stationary distributions of these Markov chains as ε becomes arbitrarily small:

$$\lim_{\varepsilon \rightarrow 0} \mu^{N,\varepsilon}(x) > 0.$$

When the stochastically stable state is unique, it offers a unique prediction of play that is relevant over sufficiently long time spans.

Bernoulli Arrivals and Mutations Following the approach of many early contributors to the literature, let us consider a model of stochastic evolution based on *Bernoulli arrivals of revision opportunities* and *best responses with mutations*. The former assumption means that during each discrete time period, each agent has probability $\theta \in (0, 1]$ of receiving an opportunity to update his strategy. This assumption differs than the one we proposed in Sect. “[Revision Protocols](#)”; the key new implication is

that all agents may receive revision opportunities simultaneously. (Models that assume this directly generate similar results.) The latter assumption posits that when an agent receives a revision opportunity, he plays a best response to the current strategy distribution with probability $1 - \varepsilon$, and chooses a strategy at random with probability ε .

Example 30 Suppose that a population of N agents is randomly matched to play the Stag Hunt game (Example 2):

	H	S
H	h	h
S	0	s

Since $s > h > 0$, hunting hare and hunting stag are both symmetric pure equilibria; the game also admits the symmetric mixed equilibrium $x^* = (x_H^*, x_S^*) = (\frac{s-h}{s}, \frac{h}{s})$.

If more than fraction x_H^* of the agents hunt hare, then hare is the unique best response, while if more than fraction x_S^* of the agents hunt stag, then stag is the unique best response. Thus, under any deterministic dynamic that respects payoffs, the mixed equilibrium x^* divides the state space into two basins of attraction, one for each of the two pure equilibria.

Now consider our stochastic evolutionary process. If the noise level ε is small, this process typically behaves like a deterministic process, moving quickly toward one of the two pure states, $e_H = (1, 0)$ or $e_S = (0, 1)$, and remaining there for some time. But since the process is ergodic, it will eventually leave the pure state it reaches first, and in fact will switch from one pure state to the other infinitely often.

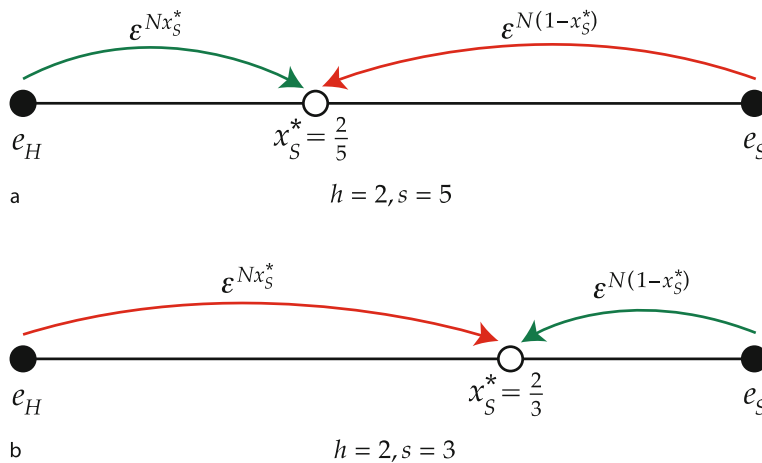
To determine the stochastically stable state, we must compute and compare the “improbabilities” of these transitions. If the current state is e_H , a transition to e_S re-

quires mutations to cause roughly Nx_S^* agents to switch to the suboptimal strategy S, sending the population into the basin of attraction of e_S ; the probability of this event is of order $\varepsilon^{Nx_S^*}$. Similarly, to transit from e_S to e_H , mutations must cause roughly $Nx_H^* = N(1 - x_S^*)$ to switch from S to H; this probability of this event is of order $\varepsilon^{N(1-x_S^*)}$.

Which of these rare events is more likely ones depends on whether x_S^* is greater than or less than $\frac{1}{2}$. If $s > 2h$, so that $x_S^* < \frac{1}{2}$, then $\varepsilon^{Nx_S^*}$ is much smaller than $\varepsilon^{N(1-x_S^*)}$ when ε is small; thus, state e_S is stochastically stable (Fig. 8a). If instead $s < 2h$, so that $x_S^* > \frac{1}{2}$, then $\varepsilon^{N(1-x_S^*)} < \varepsilon^{Nx_S^*}$, so e_H is stochastically stable (Fig. 8b).

These calculations show that *risk dominance* – being the optimal response against a uniformly randomizing opponent – drives stochastic stability 2×2 games. In particular, when $s < 2h$, so that risk dominance and payoff dominance disagree, stochastic stability favors the former over the latter.

This example illustrates how under Bernoulli arrivals and mutations, stochastic stability analysis is based on *mutation counting*: that is, on determining how many simultaneous mutations are required to move from each equilibrium into the basin of attraction of each other equilibrium. In games with more than two strategies, completing the argument becomes more complicated than in the example above: the analysis, typically based on the tree-analysis techniques of [75,219], requires one to account for the relative difficulties of transitions between all pairs of equilibria. [68] develops a streamlined method of computing the stochastically stable state based on radius-coradius calculations; while this approach is not always sufficiently fine



Evolutionary Game Theory, Figure 8

Equilibrium selection via mutation counting in Stag Hunt games

to yield a complete analysis, in the cases where it works it can be considerably simpler to apply than the tree-analysis method.

These techniques have been employed successfully to variety of classes of games, including pure coordination games, supermodular games, games satisfying “bandwagon” properties, and games with equilibria that satisfy generalizations of risk dominance [68,120,121,134]. A closely related literature uses stochastic stability as a basis for evaluating traditional solution concepts for extensive form games [90,115,122,128,152,168,169].

A number of authors have shown that variations on the Bernoulli arrivals and mutations model can lead to different equilibrium selection results. For instance, [165,211] show that if choices are determined from the payoffs from a single round of matching (rather than from expected payoffs), the payoff dominant equilibrium rather than the risk dominant equilibrium is selected. If choices depend on strategies’ relative performances rather than their absolute performances, then long run behavior need not resemble a Nash equilibrium at all [26,161,171,198]. Finally, if the probability of mutation depends on the current population state, then any recurrent set of the unperturbed process (e. g., any pure equilibrium of a coordination game) can be selected in the long run if the mutation rates are specified in an appropriate way [27]. This last result suggests that mistake probabilities should be provided with an explicit foundation, a topic we take up in Sect. “Poisson Arrivals and Payoff Noise”.

Another important criticism of the stochastic stability literature concerns the length of time needed for its predictions to become relevant [31,67]. If the population size N is large and the mutation rate ε is small, then the probability ε^{cN} that a transition between equilibria occurs during given period is miniscule; the waiting time between transitions is thus enormous. Indeed, if the mutation rate falls over time, or if the population size grows over time, then ergodicity may fail, abrogating equilibrium selection entirely [163,186]. These analyses suggest that except in applications with very long time horizons, the unique predictions generated by analyses of stochastic stability may be inappropriate, and that modelers would do better to focus on history-dependent predictions of the sort provided by deterministic models. At the same time, there are frameworks in which stochastic stability becomes relevant much more quickly. The most important of these are local interaction models, which we discuss in Sect. “Local Interaction”.

Poisson Arrivals and Payoff Noise Combining the assumption of Bernoulli arrivals of revision opportunities

with that of best responses with mutations creates a model in which the probabilities of transitions between equilibria are easy to compute: one can focus on events in which large numbers of agents switch to a suboptimal strategy at once, each doing so with the same probability. But the simplicity of this argument also highlights the potency of the assumptions behind it.

An appealing alternative approach is to model stochastic evolution using *Poisson arrivals of revision opportunities* and *payoff noise* [29,31,38,39,63,135,145,209,210,222]. (One can achieve similar effects by looking at models defined in terms of stochastic differential equations; see [18,48,74,79,113].) By allowing revision opportunities to arrive in continuous time, as we did in Sect. “Revision Protocols”, we ensure that agents do not receive opportunities simultaneously, ruling out the simultaneous mass revisions that drive the Bernoulli arrival model. (One can accomplish the same end using a discrete time model by assuming that one agent updates during each period; the resulting process is a random time change away from the Poisson arrivals model.)

Under Poisson arrivals, transitions between equilibria occur gradually, as the population works its way out of basins of attraction one agent at a time. In this context, the mutation assumption becomes particularly potent, ensuring that the probabilities of suboptimal choices do not vary with their payoff consequences. Under the alternative assumption of payoff noise, one supposes that agents play best responses to payoffs that are subject to random perturbations drawn from a fixed multivariate distribution. In this case, suboptimal choices are much more likely near basin boundaries, where the payoffs of second-best strategies are not much less than those of optimal ones, than they are at stable equilibria, where payoff differences are larger.

Evidently, assuming Poisson arrivals and payoff noise means that stochastic stability cannot be assessed by way of mutation counting. To determine the unlikelihood of escaping from an equilibrium’s basin of attraction, one must not only account for the “width” of the basin of attraction (i. e., the number of suboptimal choices needed to escape it), but also for its “depth” (the unlikelihood of each of these choices). In two-strategy games this is not difficult to accomplish: in this case the evolutionary process is a birth-and-death chain, and its stationary distribution can be expressed using an explicit formula. Beyond this case, one can employ the Freidlin and Wentzell [75] machinery, although doing so tends to be computationally demanding.

This computational burden is less in models that retain Poisson arrivals, but replace perturbed optimization with decision rules based on imitation and mutation [80].

Because agents imitate successful opponents, the population spends the vast majority of periods on the edges of the simplex, implying that the probabilities of transitions between vertices can be determined using birth-and-death chain methods [158]. As a consequence, one can reduce the problem of finding the stochastically stable state in an n strategy coordination game to that of computing the limiting stationary distribution of an n state Markov chain.

Stochastic Stability via Large Population Limits The approach to stochastic stability followed thus far relies on small noise limits: that is, on evaluating the limit of the stationary distributions $\mu^{N,\varepsilon}$ as the noise level ε approaches zero. Binmore and Samuelson [29] argue that in the contexts where evolutionary models are appropriate, the amount of noise in agents decisions is not negligible, so that taking the low noise limit may not be desirable. At the same time, evolutionary models are intended to describe behavior in large populations, suggesting an alternative approach: that of evaluating the limit of the stationary distributions $\mu^{N,\varepsilon}$ as the population size N grows large.

In one respect, this approach complicates the analysis. When N is fixed and ε varies, each stationary distribution $\mu^{N,\varepsilon}$ is a measure on the fixed state space $\mathcal{X}^N = \{x \in X : Nx \in \mathbb{Z}^n\}$. But when ε is fixed and N varies, the state space \mathcal{X}^N varies as well, and one must introduce notions of weak convergence of probability measures in order to define stochastic stability.

But in other respects taking large population limits can make analysis simpler. We saw in Sect. “[Deterministic Approximation](#)” that by taking the large population limit, we can approximate the finite-horizon sample paths of the stochastic evolutionary process $\{X_t^{N,\varepsilon}\}$ by solutions to the mean dynamic (M). Now we are concerned with infinite horizon behavior, but it is still reasonable to hope that the large population limit will again reduce some of our computations to a calculus problems.

As one might expect, this approach is easiest to follow in the two-strategy case, where for each fixed population size N , the evolutionary process $\{X_t^{N,\varepsilon}\}$ is a birth-and-death chain. When one takes the large population limit, the formulas for waiting times and for the stationary distribution can be evaluated using integral approximations [24,29,39,222]. Indeed, the approximations so obtained take an appealing simple form [182].

The analysis becomes more complicated beyond the two-strategy case, but certain models have proved amenable to analysis. For instance [80], characterizes large population stochastic stability in models based on imitation and mutation. Imitation ensures that the population spends nearly all periods on the edges of the simplex X ,

and the large population limit makes evaluating the probabilities of transitions along these edges relatively simple.

If one supposes that agents play best responses to noisy payoffs, then one must account directly for the behavior of the process $\{X_t^{N,\varepsilon}\}$ in the interior of the simplex. One possibility is to combine the deterministic approximation results from Sect. “[Deterministic Approximation](#)” with techniques from the theory of stochastic approximation [20,21] to show that the large N limiting stationary distribution is concentrated on attractors of the mean dynamic. By combining this idea with convergence results for deterministic dynamics from Sect. “[Global Convergence](#)”, Ref. [104] shows that the limiting stationary distribution must be concentrated around equilibrium states in potential games, stable games, and supermodular games.

The results in [104] do not address the question of equilibrium selection. However, for the specific case of logit evolution in potential games, a complete characterization of the large population limit of the process $\{X_t^{N,\varepsilon}\}$ has been obtained [23]. By combining deterministic approximation results, which describe the usual behavior of the process within basins of attraction, with a large deviations analysis, which characterizes the rare escapes from basins of attraction, one can obtain a precise asymptotic formula for the large N limiting stationary distribution. This formula accounts both for the typical procession of the process along solutions of the mean dynamic, and for the rare sojourns of the process against this deterministic flow.

Local Interaction

All of the game dynamics considered so far have been based implicitly on the assumption of *global interaction*: each agent’s payoffs depend directly on all agents’ actions. In many contexts, one expects to the contrary that interactions will be local in nature: for instance, agents may live in fixed locations and interact only with neighbors. In addition to providing a natural fit for these applications, *local interaction* models respond to some of the criticisms of the stochastic stability literature. At the same time, once one moves beyond relatively simple cases, local interaction models become exceptionally complicated, and so lend themselves to methods of analysis very different from those considered thus far.

Stochastic Stability and Equilibrium Selection Revisited

In Sect. “[Stochastic Stability and Equilibrium Selection](#)”, we saw the prediction of risk dominant equilibrium play provided by stochastic stability models is subverted by the waiting-time critique: namely, that the length of time re-

quired before this equilibrium is reached may be extremely long. Ellison [67,68] shows that if interactions are local, then selection of the risk dominant equilibrium persists, and waiting times are no longer an issue.

Example 31 In the simplest local interaction model, a population of N agents are located at N distinct positions around a circle. During each period of play, each agent plays the Stag Hunt game (Examples 2 and 30) with his two nearest neighbors, following the same action against both of his opponents. If we suppose that $s \in (h, 2h)$, so that hunting hare is the risk dominant strategy, then by definition, an agent whose neighbors play different strategies finds it optimal to choose H himself.

Now suppose that there are Bernoulli arrivals of revision opportunities, and that decisions are based on best responses and rare mutations. To move from the all S state to the all H state, it is enough that a single agent mutates S to H . This one mutation begins a chain reaction: the mutating agent's neighbors respond optimally by switching to H themselves; they are followed in this by their own neighbors; and the contagion continues until all agents choose H . Since a single mutation is always enough to spur the transition from all S to all H , the expected wait before this transition is small, even when the population is large.

In contrast, the transition back from all H to all S is extremely unlikely. Even if all but one of the agents simultaneously mutate to S , the contagion process described above will return the population to the all- H state. Thus, while the transition from all- S to all- H occurs quickly, the reverse transition takes even longer than in the global interaction setting.

The local interaction approach to equilibrium selection has been advanced in a variety of directions: by allowing agents to choose their locations [69], or to pay a cost to choose different strategies against different opponents [86], and by basing agents' decisions on the attainment of aspiration levels [11], or on imitation of successful opponents [9,10]. A portion of this literature initiated by Blume develops connections between local interaction models in evolutionary game theory with models from statistical mechanics [36,37,38,124,141]. These models provide a point of departure for research on complex spatial dynamics in games, which we consider next.

Complex Spatial Dynamics

The local interaction models described above address the questions of convergence to equilibrium and selection among multiple equilibria. In the cases where convergence and selection results obtain, behavior in these models is

relatively simple, as most periods are spent with most agents coordinating on a single strategy. A distinct branch of the literature on evolution and local interaction focuses on cases with complex dynamics, where instead of settling quickly into a homogeneous, static configuration, behavior remains in flux, with multiple strategies coexisting for long periods of time.

Example 32 Cooperating is a dominated strategy in the Prisoner's Dilemma, and is not played in equilibrium in finitely repeated versions of this game. Nevertheless, a pair of Prisoner's Dilemma tournaments conducted by Axelrod [14] were won by the strategy Tit-for-Tat, which cooperates against cooperative opponents and defects against defectors. Axelrod's work spawned a vast literature aiming to understand the persistence of individually irrational but socially beneficial behavior.

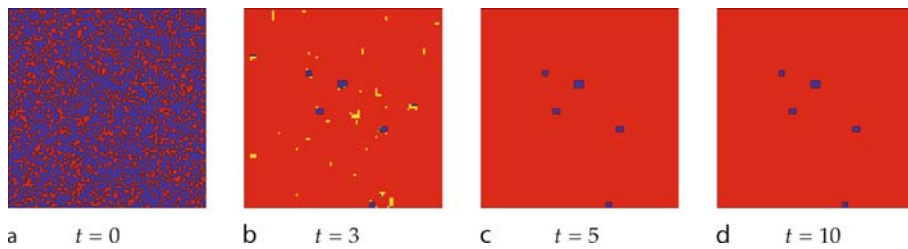
To address this question, Nowak and May [153,154,155,156,157] consider a population of agents who are repeatedly matched to play the Prisoner's Dilemma

$$\begin{array}{cc} & \begin{array}{cc} C & D \end{array} \\ \begin{array}{c} C \\ D \end{array} & \begin{array}{|cc|} \hline 1 & -\varepsilon \\ g & 0 \\ \hline \end{array} \end{array},$$

where the greedy payoff g exceeds 1 and $\varepsilon > 0$ is small. The agents are positioned on a two-dimensional grid. During each period, each agent plays the Prisoner's Dilemma with the eight agents in his (Moore) neighborhood. In the simplest version of the model, all agents simultaneously update their strategies at the end of each period. If an agent's total payoff that period is as high as that of any of neighbor, he continues to play the same strategy; otherwise, he switches to the strategy of the neighbor who obtained the highest payoff.

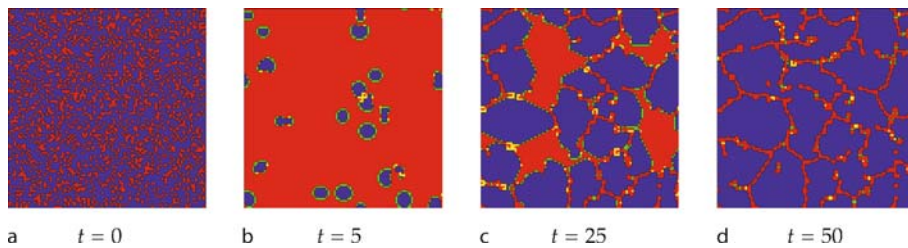
Since defecting is a dominant strategy in the Prisoner's Dilemma, one might expect the local interaction process to converge to a state at which all agents defect, as would be the case in nearly any model of global interaction. But while an agent is always better off defecting himself, he also is better off the more of his neighbors cooperate; and since evolution is based on imitation, cooperators tend to have more cooperators as neighbors than do defectors.

In Figs. 9–11, we present snapshots of the local interaction process for choices of the greedy payoff g from each of three distinct parameter regions. If $g > \frac{5}{3}$ (Fig. 9), the process quickly converges to a configuration containing a few rectangular islands of cooperators in a sea of defectors; the exact configuration depending on the initial conditions. If instead $g < \frac{8}{5}$ (Fig. 10), the process moves towards a configuration in which agents other than those in a “web” of defectors cooperate. But for $g \in (\frac{8}{5}, \frac{5}{3})$ (Fig. 11), the sys-



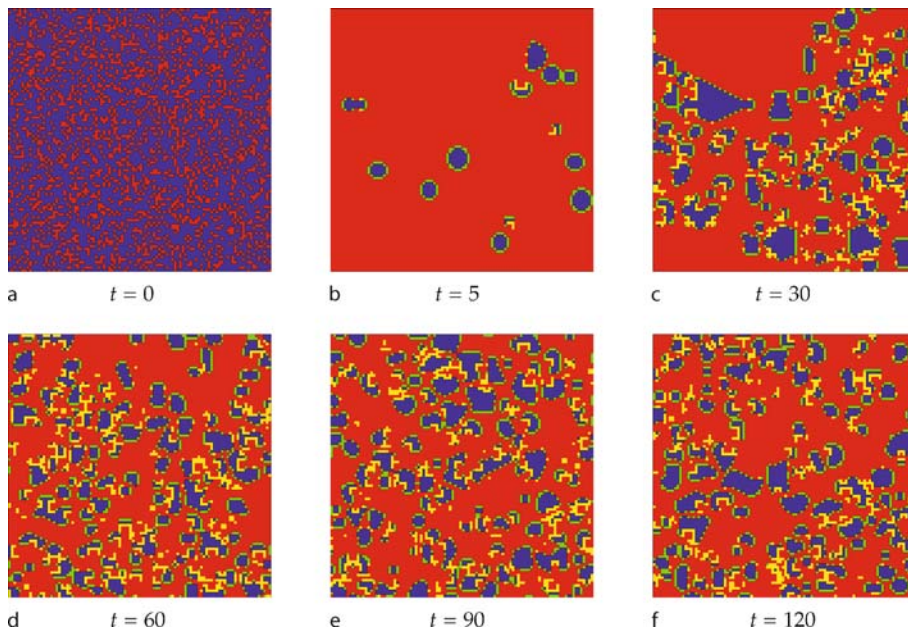
Evolutionary Game Theory, Figure 9

Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.7$. In Figs. 9–11, agents are arrayed on a 100×100 grid with periodic boundaries (i.e., a torus). Initial conditions are random with 75% cooperators and 25% defectors. Agents update simultaneously, imitating the neighbor who earned the highest payoff. *Blue* cells represent cooperators who also cooperated last period, *green* cells represent new cooperators; *red* cells represent defectors who also defected last period, *yellow* cells represent new defectors. (Figs. 9–11 created using VirtualLabs [92])



Evolutionary Game Theory, Figure 10

Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.55$



Evolutionary Game Theory, Figure 11

Local interaction in a Prisoner's Dilemma; greedy payoff $g = 1.65$

tem evolves in a complicated fashion, with clusters of cooperators and of defectors forming, expanding, disappearing, and reforming. But while the configuration of behavior never stabilizes, the proportion of cooperators appears to settle down to about .30.

The specification of the dynamics considered above, based on simultaneous updating and certain imitation of the most successful neighbor, presents a relatively favorable environment for cooperative behavior. Nevertheless, under Poisson arrivals of revision opportunities, or probabilistic decision rules, or both, cooperation can persist for very long periods of time for values of g significantly larger than 1 [154,155].

The literature on complex spatial dynamics in evolutionary game models is large and rapidly growing, with the evolution of behavior in the spatial Prisoners' Dilemma being the single most-studied environment. While analyses are typically based on simulations, analytical results have been obtained in some relatively simple settings [71,94].

Recent work on complex spatial dynamics has considered games with three or more strategies, including Rock-Paper-Scissors games, as well as public good contribution games and Prisoner's Dilemmas with voluntary participation. Introducing more than two strategies can lead to qualitatively novel dynamic phenomena, including large-scale spatial cycles and traveling waves [93,202,203]. In addition to simulations, the analysis of complex spatial dynamics is often based on approximation techniques from non-equilibrium statistical physics, and much of the research on these dynamics has appeared in the physics literature. [201] offers a comprehensive survey of work on this topic.

Applications

Evolutionary game theory was created with biological applications squarely in mind. In the prehistory of the field, Fisher [73] and Hamilton [87] used game-theoretic ideas to understand the evolution of sex ratios. Maynard Smith [137,138,139,140] introduced his definition of ESS as a way of understanding ritualized animal conflicts. Since these early contributions, evolutionary game theory has been used to study a diverse array of biological questions, including mate choice, parental investment, parent-offspring conflict, social foraging, and predator-prey systems. For overviews of research on these and other topics in biology, see [65,88].

The early development of evolutionary game theory in economics was motivated primarily by theoretical concerns: the justification of traditional game-theoretic solu-

tion concepts, and the development of methods for equilibrium selection in games with multiple stable equilibria. More recently, evolutionary game theory has been applied to concrete economic environments, in some instances as a means of contending with equilibrium selection problems, and in others to obtain an explicitly dynamic model of the phenomena of interest. Of course, these applications are most successful when the behavioral assumptions that underlie the evolutionary approach are appropriate, and when the time horizon needed for the results to become relevant corresponds to the one germane to the application at hand.

Topics in economics theoretical studied using the methods of evolutionary game theory range from behavior in markets [1,6,7,8,12,19,64,112,129,212], to bargaining and hold-up problems [32,46,57,66,164,208,220,221,222], to externality and implementation problems [47,49,136,174,177,180], to questions of public good provision and collective action [146,147,148]. The techniques described here are being applied with increasing frequency to problems of broader social science interest, including residential segregation [40,62,142,222,223,225,226] and cultural evolution [34,126], and to the study of behavior in transportation and computer networks [72,143,150,173,175,177,197]. A proliferating branch of research extends the approaches described in this article to address the evolution of structure and behavior in social networks; a number of recent books [85,114,213] offer detailed treatments of work in this domain.

Future Directions

Evolutionary game theory is a maturing field; many basic theoretical issues are well understood, but many difficult questions remain. It is tempting to say that stochastic and local interaction models offer the more open terrain for further explorations. But while it is true that we know less about these models than about deterministic evolutionary dynamics, even our knowledge of the latter is limited: while dynamics on one and two dimensional state spaces, and for games satisfying a few interesting structural assumptions, are well-understood, the dynamics of behavior in the vast majority of many-strategy games are not.

The prospects for further applications of the tools of evolutionary game theory are brighter still. In economics, and in other social sciences, the analysis of mathematical models has too often been synonymous with the computation and evaluation of equilibrium behavior. The questions of whether and how equilibrium will come to be are often ignored, and the possibility of long-term disequilibrium behavior left unmentioned. For settings in which its

assumptions are tenable, evolutionary game theory offers a host of techniques for modeling the dynamics of economic behavior. The exploitation of the possibilities for a deeper understanding of human social interactions has hardly begun.

Acknowledgments

The figures in Sects. “Deterministic Dynamics” and “Local Interaction” were created using Dynamo [184] and VirtualLabs [92], respectively. I am grateful to Caltech for its hospitality as I completed this article, and I gratefully acknowledge financial support under NSF Grant SES-0617753.

Bibliography

- Agastya M (2004) Stochastic stability in a double auction. *Games Econ Behav* 48:203–222
- Akin E (1979) *The geometry of population genetics*. Springer, Berlin
- Akin E (1980) Domination or equilibrium. *Math Biosci* 50: 239–250
- Akin E (1990) The differential geometry of population genetics and evolutionary games. In: Lessard S (ed) *Mathematical and statistical developments of evolutionary theory*. Kluwer, Dordrecht, pp 1–93
- Akin E, Losert V (1984) Evolutionary dynamics of zero-sum games. *J Math Biol* 20:231–258
- Alós-Ferrer C (2005) The evolutionary stability of perfectly competitive behavior. *Econ Theory* 26:497–516
- Alós-Ferrer C, Ania AB, Schenk-Hoppé KR (2000) An evolutionary model of Bertrand oligopoly. *Games Econ Behav* 33:1–19
- Alós-Ferrer C, Kirchsteiger G, Walzl M (2006) On the evolution of market institutions: The platform design paradox. Unpublished manuscript, University of Konstanz
- Alós-Ferrer C, Weidenholzer S (2006) Contagion and efficiency. *J Econ Theory* forthcoming, University of Konstanz and University of Vienna
- Alós-Ferrer C, Weidenholzer S (2006) Imitation, local interactions, and efficiency. *Econ Lett* 93:163–168
- Anderlini L, Ianni A (1996) Path dependence and learning from neighbors. *Games Econ Behav* 13:141–177
- Ania AB, Tröger T, Wambach A (2002) An evolutionary analysis of insurance markets with adverse selection. *Games Econ Behav* 40:153–184
- Arneodo A, Couillet P, Tresser C (1980) Occurrence of strange attractors in three-dimensional Volterra equations. *Phys Lett* 79A:259–263
- Axelrod R (1984) *The evolution of cooperation*. Basic Books, New York
- Balkenborg D, Schlag KH (2001) Evolutionarily stable sets. *Int J Game Theory* 29:571–595
- Basu K, Weibull JW (1991) Strategy sets closed under rational behavior. *Econ Lett* 36:141–146
- Beckmann M, McGuire CB, Winsten CB (1956) *Studies in the economics of transportation*. Yale University Press, New Haven
- Beggs AW (2002) Stochastic evolution with slow learning. *Econ Theory* 19:379–405
- Ben-Shoham A, Serrano R, Volij O (2004) The evolution of exchange. *J Econ Theory* 114:310–328
- Benaïm M (1998) Recursive algorithms, urn processes, and the chaining number of chain recurrent sets. *Ergod Theory Dyn Syst* 18:53–87
- Benaïm M, Hirsch MW (1999) On stochastic approximation algorithms with constant step size whose average is cooperative. *Ann Appl Probab* 30:850–869
- Benaïm M, Hofbauer J, Hopkins E (2006) Learning in games with unstable equilibria. Unpublished manuscript, Université de Neuchâtel, University of Vienna and University of Edinburgh
- Benaïm M, Sandholm WH (2007) Logit evolution in potential games: Reversibility, rates of convergence, large deviations, and equilibrium selection. Unpublished manuscript, Université de Neuchâtel and University of Wisconsin
- Benaïm M, Weibull JW (2003) Deterministic approximation of stochastic evolution in games. *Econometrica* 71:873–903
- Berger U, Hofbauer J (2006) Irrational behavior in the Brown-von Neumann-Nash dynamics. *Games Econ Behav* 56:1–6
- Bergin J, Bernhardt D (2004) Comparative learning dynamics. *Int Econ Rev* 45:431–465
- Bergin J, Lipman BL (1996) Evolution with state-dependent mutations. *Econometrica* 64:943–956
- Binmore K, Gale J, Samuelson L (1995) Learning to be imperfect: The ultimatum game. *Games Econ Behav* 8:56–90
- Binmore K, Samuelson L (1997) Muddling through: Noisy equilibrium selection. *J Econ Theory* 74:235–265
- Binmore K, Samuelson L (1999) Evolutionary drift and equilibrium selection. *Rev Econ Stud* 66:363–393
- Binmore K, Samuelson L, Vaughan R (1995) Musical chairs: Modeling noisy evolution. *Games Econ Behav* 11:1–35
- Binmore K, Samuelson L, Peyton Young H (2003) Equilibrium selection in bargaining models. *Games Econ Behav* 45:296–328
- Bishop DT, Cannings C (1978) A generalised war of attrition. *J Theor Biol* 70:85–124
- Bisin A, Verdier T (2001) The economics of cultural transmission and the dynamics of preferences. *J Econ Theory* 97:298–319
- Björnerstedt J, Weibull JW (1996) Nash equilibrium and evolution by imitation. In: Arrow KJ et al. (eds) *The Rational Foundations of Economic Behavior*. St. Martin's Press, New York, pp 155–181
- Blume LE (1993) The statistical mechanics of strategic interaction. *Games Econ Behav* 5:387–424
- Blume LE (1995) The statistical mechanics of best response strategy revision. *Games Econ Behav* 11:111–145
- Blume LE (1997) Population games. In: Arthur WB, Durlauf SN, Lane DA (eds) *The economy as an evolving complex system II*. Addison-Wesley, Reading pp 425–460
- Blume LE (2003) How noise matters. *Games Econ Behav* 44:251–271
- Bøg M (2006) Is segregation robust? Unpublished manuscript, Stockholm School of Economics
- Bomze IM (1990) Dynamical aspects of evolutionary stability. *Monatshefte Mathematik* 110:189–206
- Bomze IM (1991) Cross entropy minimization in uninhabitable states of complex populations. *J Math Biol* 30:73–87

43. Börgers T, Sarin R (1997) Learning through reinforcement and the replicator dynamics. *J Econ Theory* 77:1–14
44. Boylan RT (1995) Continuous approximation of dynamical systems with randomly matched individuals. *J Econ Theory* 66:615–625
45. Brown GW, von Neumann J (1950) Solutions of games by differential equations. In: Kuhn HW, Tucker AW (eds) *Contributions to the theory of games I*, volume 24 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, pp 73–79
46. Burke MA, Peyton Young H (2001) Competition and custom in economic contracts: A case study of Illinois agriculture. *Am Econ Rev* 91:559–573
47. Cabrales A (1999) Adaptive dynamics and the implementation problem with complete information. *J Econ Theory* 86:159–184
48. Cabrales A (2000) Stochastic replicator dynamics. *Int Econ Rev* 41:451–481
49. Cabrales A, Ponti G (2000) Implementation, elimination of weakly dominated strategies and evolutionary dynamics. *Rev Econ Dyn* 3:247–282
50. Crawford VP (1991) An “evolutionary” interpretation of Van Huyck, Battalio, and Beil’s experimental results on coordination. *Games Econ Behav* 3:25–59
51. Cressman R (1996) Evolutionary stability in the finitely repeated prisoner’s dilemma game. *J Econ Theory* 68:234–248
52. Cressman R (1997) Local stability of smooth selection dynamics for normal form games. *Math Soc Sci* 34:1–19
53. Cressman R (2000) Subgame monotonicity in extensive form evolutionary games. *Games Econ Behav* 32:183–205
54. Cressman R (2003) *Evolutionary dynamics and extensive form games*. MIT Press, Cambridge
55. Cressman R, Schlag KH (1998) On the dynamic (in)stability of backwards induction. *J Econ Theory* 83:260–285
56. Dafermos S, Sparrow FT (1969) The traffic assignment problem for a general network. *J Res Nat Bureau Stand B* 73:91–118
57. Dawid H, Bentley MacLeod W (2008) Hold-up and the evolution of investment and bargaining norms. *Games Econ Behav* forthcoming 62:26–52
58. Dawkins R (1976) *The selfish gene*. Oxford University Press, Oxford
59. Dekel E, Scotchmer S (1992) On the evolution of optimizing behavior. *J Econ Theory* 57:392–407
60. Demichelis S, Ritzberger K (2003) From evolutionary to strategic stability. *J Econ Theory* 113:51–75
61. Dindoš M, Mezzetti C (2006) Better-reply dynamics and global convergence to Nash equilibrium in aggregative games. *Games Econ Behav* 54:261–292
62. Dokumacı E, Sandholm WH (2007) Schelling redux: An evolutionary model of residential segregation. Unpublished manuscript, University of Wisconsin
63. Dokumacı E, Sandholm WH (2007) Stochastic evolution with perturbed payoffs and rapid play. Unpublished manuscript, University of Wisconsin
64. Droste E, Hommes, Tuinstra J (2002) Endogenous fluctuations under evolutionary pressure in Cournot competition. *Games Econ Behav* 40:232–269
65. Dugatkin LA, Reeve HK (eds) (1998) *Game theory and animal behavior*. Oxford University Press, Oxford
66. Ellingsen T, Robles J (2002) Does evolution solve the hold-up problem? *Games Econ Behav* 39:28–53
67. Ellison G (1993) Learning, local interaction, and coordination. *Econometrica* 61:1047–1071
68. Ellison G (2000) Basins of attraction, long run equilibria, and the speed of step-bystep evolution. *Rev Econ Stud* 67:17–45
69. Ely JC (2002) Local conventions. *Adv Econ Theory* 2:1(30)
70. Ely JC, Sandholm WH (2005) Evolution in Bayesian games I: Theory. *Games Econ Behav* 53:83–109
71. Eshel I, Samuelson L, Shaked A (1998) Altruists, egoists, and hooligans in a local interaction model. *Am Econ Rev* 88:157–179
72. Fischer S, Vöcking B (2006) On the evolution of selfish routing. Unpublished manuscript, RWTH Aachen
73. Fisher RA (1930) *The genetical theory of natural selection*. Clarendon Press, Oxford
74. Foster DP, Peyton Young H (1990) Stochastic evolutionary game dynamics. *Theor Popul Biol* 38:219–232 also in *Corrigendum* 51:77–78 (1997)
75. Freidlin MI, Wentzell AD (1998) *Random perturbations of dynamical systems*, 2nd edn. Springer, New York
76. Friedman D (1991) Evolutionary games in economics. *Econometrica* 59:637–666
77. Friedman D, Yellin J (1997) Evolving landscapes for population games. Unpublished manuscript, UC Santa Cruz
78. Friedman JW, Mezzetti C (2001) Learning in games by random sampling. *J Econ Theory* 98:55–84
79. Fudenberg D, Harris C (1992) Evolutionary dynamics with aggregate shocks. *J Econ Theory* 57:420–441
80. Fudenberg D, Imhof LA (2006) Imitation processes with small mutations. *J Econ Theory* 131:251–262
81. Fudenberg D, Imhof LA (2008) Monotone imitation dynamics in large populations. *J Econ Theory* 140:229–245
82. Fudenberg D, Levine DK (1998) *Theory of learning in games*. MIT Press, Cambridge
83. Gaunersdorfer A, Hofbauer J (1995) Fictitious play, shapley polygons, and the replicator equation. *Games Econ Behav* 11:279–303
84. Gilboa I, Matsui A (1991) Social stability and equilibrium. *Econometrica* 59:859–867
85. Goyal S (2007) *Connections: An introduction to the economics of networks*. Princeton University Press, Princeton
86. Goyal S, Janssen MCW (1997) Non-exclusive conventions and social coordination. *J Econ Theory* 77:34–57
87. Hamilton WD (1967) Extraordinary sex ratios. *Science* 156:477–488
88. Hammerstein P, Selten R (1994) Game theory and evolutionary biology. In: Aumann RJ, Hart S (eds) *Handbook of Game Theory*. vol 2, chap 28, Elsevier, Amsterdam, pp 929–993
89. Harsanyi JC, Selten R (1988) *A General Theory of equilibrium selection in games*. MIT Press, Cambridge
90. Hart S (2002) Evolutionary dynamics and backward induction. *Games Econ Behav* 41:227–264
91. Hart S, Mas-Colell A (2003) Uncoupled dynamics do not lead to Nash equilibrium. *Am Econ Rev* 93:1830–1836
92. Hauert C (2007) Virtual Labs in evolutionary game theory. Software <http://www.univie.ac.at/virtuallabs>. Accessed 31 Dec 2007
93. Hauert C, De Monte S, Hofbauer J, Sigmund K (2002) Volunteering as Red Queen mechanism for cooperation in public goods games. *Science* 296:1129–1132

94. Herz AVM (1994) Collective phenomena in spatially extended evolutionary games. *J Theor Biol* 169:65–87
95. Hines WGS (1987) Evolutionary stable strategies: A review of basic theory. *Theor Popul Biol* 31:195–272
96. Hofbauer J (1995) Imitation dynamics for games. Unpublished manuscript, University of Vienna
97. Hofbauer J (1995) Stability for the best response dynamics. Unpublished manuscript, University of Vienna
98. Hofbauer J (2000) From Nash and Brown to Maynard Smith: Equilibria, dynamics and ESS. *Selection* 1:81–88
99. Hofbauer J, Hopkins E (2005) Learning in perturbed asymmetric games. *Games Econ Behav* 52:133–152
100. Hofbauer J, Oechssler J, Riedel F (2005) Brown-von Neumann-Nash dynamics: The continuous strategy case. Unpublished manuscript, University of Vienna
101. Hofbauer J, Sandholm WH (2002) On the global convergence of stochastic fictitious play. *Econometrica* 70:2265–2294
102. Hofbauer J, Sandholm WH (2006) Stable games. Unpublished manuscript, University of Vienna and University of Wisconsin
103. Hofbauer J, Sandholm WH (2006) Survival of dominated strategies under evolutionary dynamics. Unpublished manuscript, University of Vienna and University of Wisconsin
104. Hofbauer J, Sandholm WH (2007) Evolution in games with randomly disturbed payoffs. *J Econ Theory* 132:47–69
105. Hofbauer J, Schuster P, Sigmund K (1979) A note on evolutionarily stable strategies and game dynamics. *J Theor Biol* 81:609–612
106. Hofbauer J, Sigmund K (1988) Theory of evolution and dynamical systems. Cambridge University Press, Cambridge
107. Hofbauer J, Sigmund K (1998) Evolutionary games and population dynamics. Cambridge University Press, Cambridge
108. Hofbauer J, Sigmund K (2003) Evolutionary game dynamics. *Bull Am Math Soc (New Series)* 40:479–519
109. Hofbauer J, Swinkels JM (1996) A universal Shapley example. Unpublished manuscript, University of Vienna and Northwestern University
110. Hofbauer J, Weibull JW (1996) Evolutionary selection against dominated strategies. *J Econ Theory* 71:558–573
111. Hopkins E (1999) A note on best response dynamics. *Games Econ Behav* 29:138–150
112. Hopkins E, Seymour RM (2002) The stability of price dispersion under seller and consumer learning. *Int Econ Rev* 43:1157–1190
113. Imhof LA (2005) The long-run behavior of the stochastic replicator dynamics. *Ann Appl Probab* 15:1019–1045
114. Jackson MO Social and economic networks. Princeton University Press, Princeton, forthcoming
115. Jacobsen HJ, Jensen M, Sloth B (2001) Evolutionary learning in signalling games. *Games Econ Behav* 34:34–63
116. Jordan JS (1993) Three problems in learning mixed-strategy Nash equilibria. *Games Econ Behav* 5:368–386
117. Josephson J (2008) Stochastic better reply dynamics in finite games. *Econ Theory*, 35:381–389
118. Josephson J, Matros A (2004) Stochastic imitation in finite games. *Games Econ Behav* 49:244–259
119. Kandori M, Mailath GJ, Rob R (1993) Learning, mutation, and long run equilibria in games. *Econometrica* 61:29–56
120. Kandori M, Rob R (1995) Evolution of equilibria in the long run: A general theory and applications. *J Econ Theory* 65:383–414
121. Kandori M, Rob R (1998) Bandwagon effects and long run technology choice. *Games Econ Behav* 22:84–120
122. Kim Y-G, Sobel J (1995) An evolutionary approach to pre-play communication. *Econometrica* 63:1181–1193
123. Kimura M (1958) On the change of population fitness by natural selection. *Heredity* 12:145–167
124. Kosfeld M (2002) Stochastic strategy adjustment in coordination games. *Econ Theory* 20:321–339
125. Kukushkin NS (2004) Best response dynamics in finite games with additive aggregation. *Games Econ Behav* 48:94–110
126. Kuran T, Sandholm WH (2008) Cultural integration and its discontents. *Rev Economic Stud* 75:201–228
127. Kurtz TG (1970) Solutions of ordinary differential equations as limits of pure jump Markov processes. *J Appl Probab* 7:49–58
128. Kuzmics C (2004) Stochastic evolutionary stability in extensive form games of perfect information. *Games Econ Behav* 48:321–336
129. Lahkar R (2007) The dynamic instability of dispersed price equilibria. Unpublished manuscript, University College London
130. Lahkar R, Sandholm WH The projection dynamic and the geometry of population games. *Games Econ Behav*, forthcoming
131. Losert V, Akin E (1983) Dynamics of games and genes: Discrete versus continuous time. *J Math Biol* 17:241–251
132. Lotka AJ (1920) Undamped oscillation derived from the law of mass action. *J Am Chem Soc* 42:1595–1598
133. Mailath GJ (1992) Introduction: Symposium on evolutionary game theory. *J Econ Theory* 57:259–277
134. Maruta T (1997) On the relationship between risk-dominance and stochastic stability. *Games Econ Behav* 19:221–234
135. Maruta T (2002) Binary games with state dependent stochastic choice. *J Econ Theory* 103:351–376
136. Mathevet L (2007) Supermodular Bayesian implementation: Learning and incentive design. Unpublished manuscript, Caltech
137. Maynard Smith J (1972) Game theory and the evolution of fighting. In: Maynard Smith J On Evolution. Edinburgh University Press, Edinburgh, pp 8–28
138. Maynard Smith J (1974) The theory of games and the evolution of animal conflicts. *J Theor Biol* 47:209–221
139. Maynard Smith J (1982) Evolution and the theory of games. Cambridge University Press, Cambridge
140. Maynard Smith J, Price GR (1973) The logic of animal conflict. *Nature* 246:15–18
141. Miękisz J (2004) Statistical mechanics of spatial evolutionary games. *J Phys A* 37:9891–9906
142. Möbius MM (2000) The formation of ghettos as a local interaction phenomenon. Unpublished manuscript, MIT
143. Monderer D, Shapley LS (1996) Potential games. *Games Econ Behav* 14:124–143
144. Moran PAP (1962) The statistical processes of evolutionary theory. Clarendon Press, Oxford
145. Myatt DP, Wallace CC (2003) A multinomial probit model of stochastic evolution. *J Econ Theory* 113:286–301
146. Myatt DP, Wallace CC (2007) An evolutionary justification for thresholds in collective-action problems. Unpublished manuscript, Oxford University
147. Myatt DP, Wallace CC (2008) An evolutionary analysis of the volunteer's dilemma. *Games Econ Behav* 62:67–76

148. Myatt DP, Wallace CC (2008) When does one bad apple spoil the barrel? An evolutionary analysis of collective action. *Rev Econ Stud* 75:499–527
149. Nachbar JH (1990) “Evolutionary” selection dynamics in games: Convergence and limit properties. *Int J Game Theory* 19:59–89
150. Nagurney A, Zhang D (1997) Projected dynamical systems in the formulation, stability analysis and computation of fixed demand traffic network equilibria. *Transp Sci* 31:147–158
151. Nash JF (1951) Non-cooperative games. *Ann Math* 54:287–295
152. Nöldeke G, Samuelson L (1993) An evolutionary analysis of backward and forward induction. *Games Econ Behav* 5:425–454
153. Nowak MA (2006) *Evolutionary dynamics: Exploring the equations of life*. Belknap/Harvard, Cambridge
154. Nowak MA, Bonhoeffer S, May RM (1994) More spatial games. *Int J Bifurc Chaos* 4:33–56
155. Nowak MA, Bonhoeffer S, May RM (1994) Spatial games and the maintenance of cooperation. *Proc Nat Acad Sci* 91:4877–4881
156. Nowak MA, May RM (1992) Evolutionary games and spatial chaos. *Nature* 359:826–829
157. Nowak MA, May RM (1993) The spatial dilemmas of evolution. *Int J Bifurc Chaos* 3:35–78
158. Nowak MA, Sasaki A, Taylor C, Fudenberg D (2004) Emergence of cooperation and evolutionary stability in finite populations. *Nature* 428:646–650
159. Oechssler J, Riedel F (2001) Evolutionary dynamics on infinite strategy spaces. *Econ Theory* 17:141–162
160. Oechssler J, Riedel F (2002) On the dynamic foundation of evolutionary stability in continuous models. *J Econ Theory* 107:141–162
161. Rhode P, Stegeman M (1996) A comment on “learning, mutation, and long run equilibria in games”. *Econometrica* 64:443–449
162. Ritzberger K, Weibull JW (1995) Evolutionary selection in normal form games. *Econometrica* 63:1371–1399
163. Robles J (1998) Evolution with changing mutation rates. *J Econ Theory* 79:207–223
164. Robles J (2008) Evolution, bargaining and time preferences. *Econ Theory* 35:19–36
165. Robson A, Vega-Redondo F (1996) Efficient equilibrium selection in evolutionary games with random matching. *J Econ Theory* 70:65–92
166. Rosenthal RW (1973) A class of games possessing pure strategy Nash equilibria. *Int J Game Theory* 2:65–67
167. Samuelson L (1988) Evolutionary foundations of solution concepts for finite, two-player, normal-form games. In: Vardi MY (ed) *Proc. of the Second Conference on Theoretical Aspects of Reasoning About Knowledge* (Pacific Grove, CA, 1988), Morgan Kaufmann Publishers, Los Altos, pp 211–225
168. Samuelson L (1994) Stochastic stability in games with alternative best replies. *J Econ Theory* 64:35–65
169. Samuelson L (1997) *Evolutionary games and equilibrium selection*. MIT Press, Cambridge
170. Samuelson L, Zhang J (1992) Evolutionary stability in asymmetric games. *J Econ Theory* 57:363–391
171. Sandholm WH (1998) Simple and clever decision rules in a model of evolution. *Econ Lett* 61:165–170
172. Sandholm WH (2001) Almost global convergence to p -dominant equilibrium. *Int J Game Theory* 30:107–116
173. Sandholm WH (2001) Potential games with continuous player sets. *J Econ Theory* 97:81–108
174. Sandholm WH (2002) Evolutionary implementation and congestion pricing. *Rev Econ Stud* 69:81–108
175. Sandholm WH (2003) Evolution and equilibrium under inexact information. *Games Econ Behav* 44:343–378
176. Sandholm WH (2005) Excess payoff dynamics and other well-behaved evolutionary dynamics. *J Econ Theory* 124:149–170
177. Sandholm WH (2005) Negative externalities and evolutionary implementation. *Rev Econ Stud* 72:885–915
178. Sandholm WH (2006) Pairwise comparison dynamics. Unpublished manuscript, University of Wisconsin
179. Sandholm WH (2007) Evolution in Bayesian games II: Stability of purified equilibria. *J Econ Theory* 136:641–667
180. Sandholm WH (2007) Pigouvian pricing and stochastic evolutionary implementation. *J Econ Theory* 132:367–382
181. Sandholm WH (2007) Large population potential games. Unpublished manuscript, University of Wisconsin
182. Sandholm WH (2007) Simple formulas for stationary distributions and stochastically stable states. *Games Econ Behav* 59:154–162
183. Sandholm WH *Population games and evolutionary dynamics*. MIT Press, Cambridge, forthcoming
184. Sandholm WH, Dokumaci E (2007) *Dynamo: Phase diagrams for evolutionary dynamics*. Software <http://www.ssc.wisc.edu/~whs/dynamo>
185. Sandholm WH, Dokumaci E, Lahkar R The projection dynamic and the replicator dynamic. *Games Econ Behav*, forthcoming
186. Sandholm WH, Pauzner A (1998) Evolution, population growth, and history dependence. *Games Econ Behav* 22:84–120
187. Sato Y, Akiyama E, Doyné Farmer J (2002) Chaos in learning a simple two-person game. *Proc Nat Acad Sci* 99:4748–4751
188. Schlag KH (1998) Why imitate, and if so, how? A boundedly rational approach to multi-armed bandits. *J Econ Theory* 78:130–156
189. Schuster P, Sigmund K (1983) Replicator dynamics. *J Theor Biol* 100:533–538
190. Schuster P, Sigmund K, Hofbauer J, Wolff R (1981) Selfregulation of behaviour in animal societies I: Symmetric contests. *Biol Cybern* 40:1–8
191. Selten R (1991) Evolution, learning, and economic behavior. *Games Econ Behav* 3:3–24
192. Shahshahani S (1979) A new mathematical framework for the study of linkage and selection. *Mem Am Math Soc* 211
193. Shapley LS (1964) Some topics in two person games. In: Dresher M, Shapley LS, Tucker AW (eds) *Advances in game theory*. vol 52 of *Annals of Mathematics Studies*. Princeton University Press, Princeton, pp 1–28
194. Skyrms B (1990) *The Dynamics of Rational Deliberation*. Harvard University Press, Cambridge
195. Skyrms B (1992) Chaos in game dynamics. *J Log Lang Inf* 1:111–130
196. Smith HL (1995) *Monotone Dynamical Systems: An introduction to the theory of competitive and cooperative systems*. American Mathematical Society, Providence, RI

197. Smith MJ (1984) The stability of a dynamic model of traffic assignment – an application of a method of Lyapunov. *Transp Sci* 18:245–252
198. Stegeman M, Rhode P (2004) Stochastic Darwinian equilibria in small and large populations. *Games Econ Behav* 49:171–214
199. Swinkels JM (1992) Evolutionary stability with equilibrium entrants. *J Econ Theory* 57:306–332
200. Swinkels JM (1993) Adjustment dynamics and rational play in games. *Games Econ Behav* 5:455–484
201. Szabó G, Fáth G (2007) Evolutionary games on graphs. *Phys Rep* 446:97–216
202. Szabó G, Hauert C (2002) Phase transitions and volunteering in spatial public goods games. *Phys Rev Lett* 89:11801(4)
203. Tainaka K-I (2001) Physics and ecology of rock-paper-scissors game. In: Marsland TA, Frank I (eds) *Computers and games, Second International Conference (Hamamatsu 2000)*, vol 2063 in *Lecture Notes in Computer Science*. Springer, Berlin, pp 384–395
204. Tanabe Y (2006) The propagation of chaos for interacting individuals in a large population. *Math Soc Sci* 51:125–152
205. Taylor PD, Jonker L (1978) Evolutionarily stable strategies and game dynamics. *Math Biosci* 40:145–156
206. Thomas B (1985) On evolutionarily stable sets. *J Math Biol* 22:105–115
207. Topkis D (1979) Equilibrium points in nonzero-sum n -person submodular games. *SIAM J Control Optim* 17:773–787
208. Tröger T (2002) Why sunk costs matter for bargaining outcomes: An evolutionary approach. *J Econ Theory* 102:28–53
209. Ui T (1998) Robustness of stochastic stability. Unpublished manuscript, Bank of Japan
210. van Damme E, Weibull JW (2002) Evolution in games with endogenous mistake probabilities. *J Econ Theory* 106:296–315
211. Vega-Redondo F (1996) *Evolution, games, and economic behaviour*. Oxford University Press, Oxford
212. Vega-Redondo F (1997) The evolution of Walrasian behavior. *Econometrica* 65:375–384
213. Vega-Redondo F (2007) *Complex social networks*. Cambridge University Press, Cambridge
214. Volterra V (1931) *Lecons sur la Theorie Mathematique de la Lutte pour la Vie*. Gauthier-Villars, Paris
215. von Neumann J, Morgenstern O (1944) *Theory of games and economic behavior*. Prentice-Hall, Princeton
216. Weibull JW (1995) *Evolutionary game theory*. MIT Press, Cambridge
217. Weibull JW (1996) The mass action interpretation. Excerpt from “The work of John Nash in game theory: Nobel Seminar, December 8, 1994”. *J Econ Theory* 69:165–171
218. Weissing FJ (1991) Evolutionary stability and dynamic stability in a class of evolutionary normal form games. In: Selten R (ed) *Game Equilibrium Models I*. Springer, Berlin, pp 29–97
219. Peyton Young H (1993) The evolution of conventions. *Econometrica* 61:57–84
220. Peyton Young H (1993) An evolutionary model of bargaining. *J Econ Theory* 59:145–168
221. Peyton Young H (1998) Conventional contracts. *Review Econ Stud* 65:773–792
222. Peyton Young H (1998) *Individual strategy and social structure*. Princeton University Press, Princeton
223. Peyton Young H (2001) The dynamics of conformity. In: Durlauf SN, Peyton Young H (eds) *Social dynamics*. Brookings Institution Press/MIT Press, Washington/Cambridge, pp 133–153
224. Zeeman EC (1980) Population dynamics from game theory. In: Nitecki Z, Robinson C (eds) *Global theory of dynamical systems* (Evanston, 1979). number 819 in *Lecture Notes in Mathematics*. Springer, Berlin, pp 472–497
225. Zhang J (2004) A dynamic model of residential segregation. *J Math Sociol* 28:147–170
226. Zhang J (2004) Residential segregation in an all-integrationist world. *J Econ Behav Organ* 24:533–550

Evolution of Culture, Memetics

FRANCIS HEYLIGHEN, KLAAS CHIELENS
Vrije Universiteit Brussel, Brussels, Belgium

Article Outline

Glossary
Definition of the Subject
Introduction
Defining the Meme
Dynamics of Meme Replication and Spread
Social Structures
Computer Simulations of Cultural Evolution
Selection Criteria for Memes
Parasitic Memes
Empirical Tests
Future Directions
Bibliography

Glossary

Culture The attitudes, beliefs, and behaviors that, for a certain group, define their general way of life and that they have taken over from others.

Cultural evolution The development of culture over time, as conceptualized through the mechanisms of variation and natural selection of cultural elements.

Replicator An information pattern that is able to make copies of itself, typically with the help of another system. Examples are genes, memes, and (computer) viruses.

Meme A cultural replicator; a unit of imitation or communication.

Memeplex (or meme complex) A collection of mutually supporting memes, which tend to replicate together.

Memetics The theoretical and empirical science that studies the replication, spread and evolution of memes.

Fitness The overall success rate of a replicator, as determined by its degree of adaptation to its environment, and the three requirements of longevity, fecundity and copying-fidelity.

Longevity The duration that an individual replicator survives.

Fecundity The speed of reproduction of a replicator, as measured by the number of copies made per time unit.

Copying-fidelity The degree to which a replicator is accurately reproduced.

Vertical transmission Transmission of traits (memes or genes) from parents to offspring.

Horizontal transmission Transmission of traits between individuals of the same generation.

Memotype A meme in the form of information held in an individual's memory.

Mediotype A meme as expressed in an external medium, such as a text, an artefact, a song, or a behavior.

Sociotype The group or community of individuals who hold a particular meme in their memory.

Definition of the Subject

Cultural traits are transmitted from person to person, similarly to genes or viruses. Cultural evolution therefore can be understood through the same basic mechanisms of reproduction, spread, variation, and natural selection that underlie biological evolution. This implies a shift from genes as units of biological information to a new type of units of cultural information: *memes*. The concept of meme can be defined as an information pattern, held in an individual's memory, which is capable of being copied to another individual's memory. Memetics can then be defined as the theoretical and empirical science that studies the replication, spread and evolution of memes. Memes differ in their degree of fitness, i. e. adaptedness to the socio-cultural environment in which they propagate. Fitter memes will be more successful in being communicated, "infecting" more individuals, thus spreading over a larger population. This biological analogy allows us to apply Darwinian concepts and theories to model cultural evolution.

Introduction

The transmission of cultural traits is a process that in many ways resembles the spread of an infectious disease: the carrier of a certain idea, behavior or attitude directly or indirectly communicates this idea to another person, who now also becomes a carrier, ready to "infect" further people. For example, after you heard your neighbor whistling a catchy tune a couple of times, you may well start whistling it your-

self, thus being ready to infect some more people with the tune. Similarly, after you hear your friends recommend a new electronic tool they have bought, you may well buy one yourself, and, if you like it, start recommending it to those acquaintances who do not know it yet. Thus, cultural traits can be seen as analogous to *mind viruses* [18,28], *idea viruses* [41] or *thought contagions* [59], which are reproduced from mind to mind via imitation or communication. A truly successful trait is one that spreads like an epidemic, infecting the whole of the population, in order to end up as a stable, endemic component of that population's culture. For example, the tune may become part of the repertoire of "evergreens" that everyone knows, and the tool may become as widespread as the mobile phone or color television.

This virus metaphor is attractive in that it suggests a new perspective and new methods, such as epidemiology [4], for studying the dynamics of culture. However, in order to turn it into a well-founded scientific theory, we need a deeper understanding of the underlying assumptions and implications of this analogy. For this, we can turn to the science that studies viruses and other self-reproducing systems: biology.

It is an old idea to see a correspondence between cultural and biological evolution, with cultural entities undergoing similar processes of variation, reproduction and natural selection as organisms or genes. Around the end of the 18th century Western linguists discovered the similarities between different languages. Sir William Jones gave birth to the field of language evolution studies, more specifically in the search for the origin of languages, and their "common descent" [72]. The German linguist August Schleicher attempted to recreate this common ancestor of languages, publishing tree-diagrams of languages as early as 1853, six years before Darwin published his *Origin of Species*. In an 1870 article one can already read: "How does a new style of architecture prevail? How, again, does fashion change? (...) or take language itself (...) it is the idea of 'natural selection' that was wanted" [65]. The American philosopher and psychologist William James [55] pointed out in a presentation to the Harvard Natural History Society that: "A remarkable parallel, ..., obtains between the facts of social evolution on the one hand, and of zoological evolution as expounded by Mr. Darwin on the other."

By the end of the 20th century, the parallel study of cultural and biological evolution got a new impetus with the introduction, by Richard Dawkins [27] (first edition 1976), of the concept of *meme* (for a review see [7]). A meme, named in analogy with a gene, is defined as a cultural replicator, i. e. an element of culture such as a tradition, belief, idea, melody, or fashion, that can be held in memory and

transmitted or copied to the memory of another person. The core idea of memetics is that memes differ in their degree of fitness, i. e. adaptedness to the socio-cultural environment in which they propagate [29,50]. Mutations and recombinations of existing ideas will produce a variety of memes that compete with each other for the attention of people. Fitter memes will be more successful in being communicated, “infecting” more individuals, thus spreading over a larger population. The resulting evolutionary dynamics is one of variation creating new meme variants, followed by natural selection retaining only the ones that are most fit. Thus, the Darwinian principle of the survival of the fittest can be seen to underlie cultural evolution as well as biological evolution [3,5,6,33,58].

The memetic perspective on culture is complementary to the traditional social science perspective, which focuses on the characteristics of the individuals and groups communicating rather than on the characteristics of the information being communicated. This does not imply a “memetic reductionism”, which would deny individual control over what you communicate. It just notes that in many cases the dynamics of information propagation and the ensuing evolution of culture can be modeled more simply from the “meme’s point of view” than by analyzing the conscious or unconscious intentions of the communicating agents.

Over the past thirty years, several models of cultural evolution have been proposed that study the propagation of memes or similarly defined cultural traits e.g. “culturgens” [58] or “mnemons” [21]. Most of those models are purely theoretical, proposing various conceptualizations, implications and speculations based on the memetic perspective e.g. [15,31,38,54,57]. Some studies are mathematical in nature, applying techniques from mathematical genetics or epidemiology to quantitatively estimate the spread of particular types of memes within a population e.g. [17,23,58,60]. Others are computational, simulating the transmission of knowledge or behaviors between software agents e.g. [12,19,39]. A few are observational case studies, where the spread of a particular cultural phenomenon, such as a chain letter, an urban legend, or a social stereotype, is investigated qualitatively or quantitatively e.g. [10,24,42,68].

However, in spite of these advances, the memetic perspective on culture is not very well developed yet, and remains controversial [2,3,34]. There are several reasons why memeticists have not yet been able to convince the bulk of social and cultural scientists of the soundness of their approach.

First, the analogy with the gene, and its embodiment as DNA, seems to indicate that a meme should have

a clear, well-delineated, stable structure. (Although one should note that natural selection was proposed by Darwin well before genes were postulated by Mendel, and a century before their structure was elucidated by Watson and Crick). Cultural entities, such as beliefs, ideas, fashions, and norms, on the other hand are typically ambiguous, difficult to delimit and constantly changing. Memetic models that are based on “hard”, explicitly defined units, therefore, only seem applicable to a very small subset of cultural phenomena, such as chain letters. However, the biological analogy does not imply such rigidity: unlike higher organisms, the genes of bacteria and viruses too are in a flux, constantly mutating and exchanging bits of DNA with other organisms, but that does not imply that they do not obey evolutionary principles.

A second criticism of the memetic approach is that people are not passive “vehicles” or “carriers” of ideas and beliefs, the way they may carry viruses. Individuals actively interpret the information they receive in the light of their existing knowledge and values, and on the basis of that may decide to reject, accept, or modify the information that is communicated to them. In other words, individuals and groups actively intervene in the formulation and propagation of culture. In that sense, cultural evolution is Lamarckian rather than purely Darwinian.

A final criticism is that memetic models have not yet been sufficiently subjected to empirical tests [25,34]. Part of the reason is that most memetic theories do not make sufficiently concrete predictions to be falsifiable by observation. Most of these theories remain very speculative—often hardly better than a form of “armchair philosophy”. Moreover, until now there simply have been very few empirical studies of how memes propagate, whether in the laboratory e.g. [62] or in real life e.g. [10], and even fewer links have been established between these observations and theoretical or mathematical models.

We will try to address these criticisms in the remainder of this article. First, we will discuss the issue of how to define a meme in an as accurate way as possible. Then we will review the process of transmission of memes between individuals, emphasizing the active role played by an individual’s cognitive structure. This will give us a basis to review the dynamics of memetic propagation across a population, and the mathematical and simulation models that have been used to study it. To introduce empirical tests, we will first discuss the criteria that determine the fitness of a meme, specifying which memes are most likely to spread. We will then summarize a few experiments and case studies in which the predictive value of such selection criteria was tested. Finally, we will discuss some potential future applications of memetic research.

Defining the Meme

Replicators

The original definition of a meme by Dawkins [27] was based on the concept of *replicator*. A replicator is a system that is able to make copies of itself, typically with the help of some other system. Examples include real and computer viruses, which need respectively a cell and a computer processor to make copies of themselves. The fundamental example discussed by Dawkins is the gene, the string of DNA that carries the information on how to make a protein, and that is copied with the help of the cellular machinery whenever a cell divides. A meme too is a replicator, as it is copied whenever information is transmitted from one individual to another via communication or imitation.

Because replicators can be reproduced in different quantities, they are subject to natural selection: the one that tends to produce the largest number of replicas over an extended time span will win the competition with less productive replicators. To succeed in this, according to Dawkins [27], a good replicator should exhibit the following characteristics:

- Longevity:** The longer any instance of the replicating pattern survives, the more copies can be made of it. A drawing made by etching lines in the sand is likely to be erased before anybody could have reproduced it.
- Fecundity:** The faster the rate of copying, the more the replicator will spread. An industrial printing press can churn out many more copies of a pamphlet than an office-copying machine.
- Copying-fidelity:** The more accurate or faithful the copy, the more will remain of the initial pattern after several rounds of copying. If a painting is reproduced by making photocopies from photocopies, the picture will quickly become unrecognizable.

Dawkins called memes the “new” replicators, in the sense that they appeared very recently compared to genes. The reason for this evolution is clear: the typically human ability of imitation, i. e. learning new ideas, knowledge or behavior by copying what another individual already learnt, provides a tremendous shortcut for the multiple experiences of trial-and-error that are otherwise necessary to discover a useful new behavior pattern [21]. While some other animals are capable of limited imitation—e. g. song-birds learn songs from each other, and apes can imitate simple behaviors [16]—this capability is best developed in humans [15]. This accounts for our ability to develop a culture that is passed on from generation to generation, thus accumulating ever more useful knowledge in the course

of its evolution. In that sense, memes can be seen to be responsible for the extremely fast development of human society and its subsequent dominance of the ecosystem.

Memes vs. Genes

When we compare the two most important replicators, genes and memes, we immediately notice a number of fundamental differences. Genes can only be transmitted from parent to offspring. Memes can in principle be transmitted between any two individuals. For genes to be transmitted, you need one generation. Memes can be transmitted in the span of minutes. Meme propagation is also much faster than gene spreading, because gene replication is restricted by the relatively small number of offspring a single parent can have, whereas the number of individuals that can take over a meme from a single individual is almost unlimited. Moreover, it seems much easier for memes to undergo variation, since the information in the nervous system is more plastic than that in the DNA, and since individuals can come into contact with many more different sources of novel memes. On the other hand, selection processes too are more efficient because of “vicarious” selection [21]: the carrier of a meme does not need to be killed in order to eliminate an inadequate meme; it suffices that he witnesses or hears about the troubles of another individual due to that same meme.

The conclusion is that cultural evolution will be several orders of magnitude faster and more efficient than genetic evolution. It should not surprise us then that during the last ten thousand years, humans have hardly changed on the genetic level, whereas their culture has undergone the most radical developments. In practice the superior *evolvability* of memes also means that in cases where genetic and memetic replicators are in competition, we would expect the memes to win, even though the genes would start with the advantage of a well-established, stable structure [15], as we will discuss further when reviewing computer simulations of such dual evolution.

While memes have a much higher fecundity than genes, their plasticity implies a much lower copying-fidelity: a message as received and understood by an individual will rarely be identical to the one that was expressed, as illustrated by the many misunderstandings and reinterpretations during communication. Yet, we should not conclude from this that effective communication is impossible: if you believed that, you would not be reading this article, hoping to assimilate the main ideas presented by its authors. The reason for such mixture of accurate transmission with creative reinterpretation is that, most fundamentally, humans are cognitive agents. This means that they

process incoming information depending on the knowledge they already have and the computing machinery they are endowed with, selectively retain some of that information in their memory, and selectively express some of that information to other agents. Generally, the transmission of information by an agent will change both the agent, who has learned something new, and the information, which will be affected by the knowledge the agent already had.

Therefore, a meme reaching an agent, if it is reproduced at all, will typically be transmitted in a changed form, possibly recombined with other information learned earlier. This explains why it is often so difficult to define or pinpoint an individual meme. In that sense, cultural evolution is Lamarckian: characteristics acquired during the lifetime of the meme's carrier can be transmitted to later carriers. Lamarckian evolution, while not being Darwinian in the strict sense, is still subject to the principle of natural selection: acquired characteristics too will be passed on selectively, depending on their fitness. Natural selection by definition will pick out the memes who survive this transmission process and are favorable among other memes thanks to possible variations they underwent in this process (although it is likely that most changes will not be beneficial to the favorability). Therefore, the fittest memes, such as certain songs, religious beliefs, scientific laws, or brand names, will have a stable, recognizable identity, even though they may differ in appearance, as exemplified by the many renditions of a song or joke. All such memes together define the culture shared by a community.

This identity will be reinforced by positive feedback that characterizes the non-linear interaction between meme and carrier: the more people encounter a particular version of a meme, the more they will tend to adapt their own version to this common prototype, the more commonly they will express this version, and thus the more people will encounter it. In this way, a variety of versions that are constantly being exchanged within the same group will tend to *converge* to a single, canonical version [8]. A newcomer to this group with a variant version will be extensively subjected to the accepted version, and is likely to eventually give in to this *conformist pressure* by adopting the majority version [17].

This *winner-takes-all* dynamics, where the initially most frequent variant comes to dominate all others, is elegantly illustrated by computer simulations of the evolution of language, in which many communicating individuals who use different words for the same concept quickly converge on a single word [69]. Similarly, most systems of ethics or religious belief tend to actively suppress any variant from their canonical version. This explains why in spite of the great variability of memes, we generally have

no problems determining whether an individual belongs to a certain religious or linguistic group [51]. Note that such non-linear reinforcement does not exist for genes, since genes are transmitted only once, from parent to offspring. Moreover, once a gene is given, it can no longer be affected by the presence of other versions in the population.

Another fundamental difference between memes and genes is that for memes there is no equivalent for the traditional distinction between *genotype* (the information carried by the genes and passed on to the next generation) and *phenotype* (the specific appearance of an organism as determined by genes and environmental influences). In biological evolution, the genotype is the site of evolutionary variation (since variations in the phenotype are not passed on during reproduction) and the phenotype the site of selection (since it is the organism as a whole that survives and reproduces, or is eliminated). In memetics, we can distinguish three levels:

1. The *memotype* denotes the information as held in an individual's memory;
2. The *mediotype* denotes that information as expressed in an external medium, such as a text, an artefact, a song, or a behavior;
3. The *sociotype* denotes the group or community of individuals who hold that information in their memory [15].

Variation and selection take place on all three levels. A memotype can vary or be eliminated (forgotten) while residing in an individual's brain. A mediotype can similarly mutate (e. g. via a printing error) or be lost, and a sociotype can change when new individuals are added to the group, who may introduce different memes, or be eliminated (as when an unsuccessful tribe dies out). In conclusion, the processes of variation and selection, while analogous at the deepest level, are much more complex for memes than for genes.

Delimiting the Memetic Unit

What are the elements that make up a meme? In order to analyze meme structure, we can use some concepts from cognitive science, the discipline that studies mental content. Perhaps the most popular unit used to represent knowledge in artificial intelligence is the *production rule*. It has the form "if condition, then action". The action leads in general to the activation of another condition or category. A production rule can thus be analyzed as a combination of even more primitive elements: two *concepts* or *categories* and a *connection* (the "then" part, which

makes the first category entail the second one). For example, a meme like “God is omnipotent” can be interpreted as “if a phenomenon is God (it belongs to the category of God-like entities), then that phenomenon is omnipotent (it belongs to the category of omnipotent entities)”.

Production rules are connected when the output condition (action) of the one matches the input condition of the other. This makes it possible to construct complex cognitive systems on the basis of simple rules. In memetics, such systems are called *meme complexes* or *memeplexes*. For example, a scientific theory or a religious system of belief may be represented as a collection of mutually connected propositions or production rules, such as “God is omnipotent”, “God is good”, “God punishes bad people”, “if you steal, you are bad”, etc. This collection of rules together determines a knowledge system that allows making inferences, such as “if you steal, God will punish you”. Even more concrete perceptual or behavioral memes, such as a tune, might be modeled in this way, as concatenations of production rules of the type “if C (musical note distinguished), then E (note produced and subsequently distinguished)”, “if E, then A”, and so on. (In fact, genetic information too can be modeled using networks of “if... then” productions: a DNA string is activated by the presence of certain proteins (condition) to which it responds by producing specific other proteins (action), see [56]).

Production rules—or at least a simplified, binary representation of them, called “classifiers”—can be used to build computer simulations of cognitive evolution, using genetic algorithms, i. e. algorithmically applied operators that perform the equivalents of mutation, recombination, and selection on the basis of “fitness” on such strings [53]. Although classifier models generally do not take into account distinct carriers, this looks like a promising road to study the evolution of memeplexes formally and computationally. As we will see later, though, simulations of cultural evolution are usually limited to the mutation and spread of simple memes, ignoring the cognitive structures and processes that support inferences and that create new meme(plex)es out of combinations of existing ones.

Even if we would model memes as connected sets of production rules, we still have the problem of how many production rules define a single meme(plex). If we call a religion or a scientific theory a meme, it is clear that this will encompass a very large number of interconnected rules. In practice it will be impossible to enumerate all rules, or to define sharp boundaries between the rules that belong to the meme and those that do not. For example, should you believe in the existence of Hell, the creation of the world in seven days, and the virginity of Mary to be called a Catholic?

A pragmatic criterion that can be used in this regard is to define a meme or memeplex as the smallest collection of propositions or memory items that tends to replicate as a whole cf. [73]. For example, a proposition like “God is omnipotent” on its own, without specification of God’s other characteristics, is much too abstract to be clearly understood or applied, and as such is unlikely to replicate well. However, in combination with a number of other propositions, like “God is good”, “God is the creator of the world”, etc., that flesh out, apply, and support this abstract idea, the package will make much more sense, and be more likely to be passed on to other individuals. Similarly, the first three notes of a melody are unlikely to be remembered as a unit, but the first eight, as in the beginning of Beethoven’s fifth symphony, may well be.

It remains that often we can add or subtract a few production rules (such as the virginity of Mary) from a memeplex without significantly changing its chances of replication. Therefore, in practice it will rarely be possible to determine the precise boundaries of a meme(plex). However, this should not detract us from considering memes while analyzing cultural evolution. Indeed, the same problem besets genetic models of biological evolution: as yet, it is in practice impossible to specify the exact combination of DNA codons that determine the gene for, say, fair skin, big ears or altruism. The biochemical definition of a gene as a string of DNA that codes for one protein is not very useful when studying evolution, since most practical functions require a combination of proteins, most proteins exhibit a combination of functions, and much of the DNA is non-coding, but therefore not necessarily useless, as it may contain control information that determines the activation of other DNA strings.

As Dawkins [27] notes, we do not need to know the constitutive elements or boundaries of a gene in order to explain the evolution of particular characteristics, such as altruism or fair skin, for which such a gene would be responsible. It is sufficient that we can distinguish the effects of that gene from the effects of its rival genes (alleles). If we can determine the fitness differences resulting from these effects, then we can make predictions about which type of genes will win the competition in a particular situation, and thus which characteristics the species is most likely to evolve. For example, knowing that people with lighter skin need less sunlight to produce sufficient vitamin D, we can predict that in Northern regions natural selection will favor genes for light skin over genes for dark skin—whatever DNA codons make up these respective genes.

The same applies to memes. If, for example, we observe that one meme (say Catholicism) induces its carriers to have more children than its competitors (say Anglican-

ism), and that the children tend to take over their memes from their parents, then, *all other things being equal*, we can predict that after sufficient time this meme will dominate in the population. This prediction does not require any explicit definition of the meme of Catholicism, but only the ability to distinguish it from its competitors. Of course, in practice it is never the case that all other things are equal, but that is the predicament of all scientific modeling: we must always simplify, and ignore potentially important influences. The question is to do that as wisely as possible, and to maximally include relevant variables without making the model too complex.

Dynamics of Meme Replication and Spread

To be replicated, a meme must pass successfully through four subsequent stages:

1. *Assimilation* by an individual, who thereby becomes a carrier or *host* of the meme;
2. *Retention* in that individual's memory;
3. *Expression* by the individual in language, behavior, or another form that can be perceived by others;
4. *Transmission* of the thus created message or mediotype to one or more other individuals.

This last stage is followed again by stage 1, thus closing the replication loop. At each stage there is selection, meaning that some memes will be eliminated. Let us look in more detail at the mechanisms governing these four stages.

Assimilation

A successful meme must be able to “infect” a new host, that is, enter into its memory, and thus acquire its *memotype* form. Let us assume that a meme is presented to a potential new host. “Presented” means either that the individual encounters an existing mediotype form of a meme, or that he or she independently discovers the meme, by observation of outside phenomena or by thought, i. e. recombination of existing cognitive elements. To be assimilated, the presented meme must be respectively *noticed*, *understood* and *accepted* by the host. Noticing requires that the mediotype be sufficiently salient to attract the host's attention. Understanding means that the host recognizes the meme as something that fits in with his or her cognitive system. The mind is not a blank slate on which any idea can be impressed. To be understood, a new idea or phenomenon must connect to cognitive structures that are already available to the individual. Finally, a host that has understood a new idea must also be willing to believe it or to take it seriously. For example, although you are likely to under-

stand the proposition that your car was built by little green men from Mars, you are unlikely to accept that proposition without very strong evidence. Therefore, you will in general not memorize it, and the meme will not manage to infect you.

Retention

The second stage of memetic replication is the retention of the meme in memory. The longer the meme stays, the more opportunities it will have to spread further by infecting other hosts. This is Dawkins's [27] *longevity* characteristic for replicators.

Just like assimilation, retention is characterized by strong selection, which few memes will survive. Indeed, most of the things we hear, see or understand during the day are not stored in memory for longer than a few hours. Although you may have clearly assimilated the news that the national party won the Swaziland elections with 54% of the votes, you are unlikely to remember this a week later—unless you live in Swaziland, perhaps. Retention will depend on how important the idea is to you, and how often it is repeated, either by recurrent encounter or by internal rehearsal.

Expression

To be communicated to other individuals, a meme must emerge from its storage as memory pattern or memotype and enter its mediotype phase, i. e. assume a physical shape that can be perceived by others. This process may be called “expression”. The most obvious medium for expression is speech. Other common means are text, pictures, behavior, and the creation of artifacts such as tools, buildings or works of art. Expression does not require the conscious decision of the host to communicate the meme. A meme can be expressed simply by the way somebody walks or manipulates an object, or by what he or she wears.

Some retained memes will never be expressed, for example because the host does not consider the meme interesting enough for others to know, uses it unconsciously without it showing up in his or her behavior, does not know how to express it, or wants to keep it secret. On the other hand, the host may be convinced that the meme is so important that it must be expressed again and again to everybody he or she meets.

Transmission

To reach another individual, an expression needs a physical carrier or medium that is sufficiently stable to transmit the expression without too much loss or deformation.

Speech, for example, uses sound to transmit an expression, while text will be transmitted through ink on paper or electrical impulses in a wire. The expression will take the form of a physical signal, modulating the carrier into a specific shape—the mediotype—from which the original meme can be re-derived. For example, mediotypes can be books, photographs, artifacts or CD-ROMs.

Selection at the transmission stage happens through either elimination of certain memes, when the mediotype is destroyed or gets corrupted before it is perceived by another individual, or through differential multiplication, when the mediotype is reproduced into many copies. For example, a manuscript may be put into the shredder or turned into a book that is printed in millions of copies. Especially since the emergence of mass media and mass manufacturing, the transmission stage is the one where the contrast between successful and unsuccessful memes is largest, and where selection can have the largest impact.

Meme Fitness

The overall survival rate of a meme m can be expressed as the meme *fitness* $F(m)$, which measures the expected number $N(m)$ of memes at the next time step or generation $t + 1$ divided by the average number of memes at the present time t . This fitness can be expressed in a simplified model as the product of the survival/multiplication rates for each of the four stages, respectively assimilation A , retention R , expression E and transmission T :

$$F(m) \equiv \frac{N(m, t+1)}{N(m, t)} = A(m).P(m).E(m).T(m)$$

A denotes the proportion of mediotypes encountered by the host that are assimilated. R represents the proportion of these assimilated memes that are retained in memory. Therefore, $A \leq 1$, $R \leq 1$. E is the number of times a retained meme is expressed by the host. T is the number of potential new hosts reached by a copy of the expression. Unlike A and R , E and T do not have an upper bound, although E is likely to be more restricted than T . Note that F is zero as soon as one of its components (A , R , E , T) is zero. This expresses the fact that a meme must successfully pass through *all* four stages in order to replicate. Also note that for a meme to spread ($F > 1$), you must have at least $E > 1$ or $T > 1$.

Dynamics of Spread

From the standard definition of fitness F , we can derive the rate of growth for the number $N(t)$ of meme copies at time t . This determines the speed with which the meme

spreads through the population of carriers:

$$\frac{dN}{dt} \cong \frac{N(t+1) - N(t)}{1} = (F - 1).N$$

This results in a traditional exponential growth if $F > 1$, exponential decay (and eventual extinction) if $F < 1$, and stability if $F = 1$. This model is too simple if the population is finite. In that case, we need to take into account the total size of the population of potential carriers K , which functions as the “carrying capacity” of the socio-cultural environment in which a meme proliferates. The increase in the number $N(t)$ of memes can be represented by the following Verhulst type of equation:

$$\frac{dN}{dt} = (F - 1).N \left(1 - \frac{N}{K}\right).$$

This equation expresses the fact that the growth in meme number (dN) is in first instance proportional to the number (N) that is already there—since more memes produce more copies of themselves—, but eventually limited by the number K of potential hosts in the population, so that growth becomes zero when the population reaches this limit ($N = K$). The function $N(t)$ that is the solution to this differential equation is the logistic function with its characteristic *sigmoid* (S-like) shape.

Interactions Between Memes

The dynamics of a single growing meme population $N(m)$ could be extended to several interacting memes $N_i = N(m_i)$. Here we should add an interaction term A_{ij} which describes the strength of the influence of meme i on meme j . This influence can be positive ($A_{ij} > 0$), which means that an increase in i produces an increase in j , i. e. i helps j to grow. A negative influence ($A_{ij} < 0$) means that the growth of i suppresses the growth of j . A neutral relation ($A_{ij} = 0$) means that the spread of the one does not influence the spread of the other. This applies to memes from independent domains, such as “God exists” and “apples are healthy”. If we now consider the reciprocal influence (A_{ji}), we can distinguish the following specific types of interaction:

- $A_{ij} > 0$, $A_{ji} > 0$: the memetic species i and j can be seen as *mutualists*, that help each other to spread, e. g. by reinforcing each others’ message. An example could be “God is good” and “God is great”.
- $A_{ij} < 0$, $A_{ji} < 0$: i and j are rivals or *competitors* [12]: an increase in the one produces a decrease in the other. Examples are “God is good” and “God does not exist”.

- $A_{ij} > 0$, $A_{ji} < 0$: i and j stand in a *predator-prey* type of relationship, i.e. i grows at the expense of j . This may happen when i (e.g. relativity theory) is a more advanced version of j (e.g. Newtonian mechanics), so that carriers of j would quickly convert to i , but non-carriers of j would be more difficult to convince of i 's value.

The overall dynamics can be represented by the following system of non-linear differential \sum_j equations:

$$\frac{dN_i}{dt} = N_i \left(\sum_j A_{ij} N_j + B_i \right)$$

A_{ii} , the influence of meme i on itself will here normally always be negative and equal to $(1 - F_i)/K$, while $B_i = F_i - 1$, as in the previous equation for a single meme. Such dynamical models quickly become very complex to solve, but are not fundamentally different from traditional growth and competition models used in population biology, epidemiology, or studies of the diffusion of innovations [12]. However, they do not take into account the dependence of a meme on its carrier, nor the specific communication channels between carriers.

Social Structures

One way to make the model more realistic without adding too much complication is to consider the structure of the social space in which the potential carriers of a meme reside. Here we make the additional assumption of continuity, namely that a meme cannot jump from one carrier to another without there being some form of proximity or relationship between the carriers.

Horizontal Transmission and the Evolution of Cooperation

The simplest form of relationship is the one between parents and their offspring. Parent-to-child transmission (or more generally transmission between generations) is called *vertical transmission* [23]. Memes belonging to domains such as religion, language, ethics, and general culture are commonly transmitted in this way. This form of propagation is analogous to the transmission of genes. Therefore “vertical” models of cultural evolution find results similar to those of biological evolution. This means that vertically transmitted memes, such as established religions, will typically reinforce or elaborate genetically transmitted behavioral patterns and thus directly contribute to biological fitness [26].

The same does not apply to *horizontally transmitted* culture, i.e. memes exchanged between members of the same generation [23]. Here what is good for a meme (e.g. slavish imitation of fads and fashions) is not necessarily good for the biological individual or gene pool, since genes and memes are subjected to different kinds of natural selection. This may promote the evolution of parasitic memes that are deleterious to their carriers, as we will discuss further.

However, in addition to the fact that it spreads new information more quickly, horizontal transmission also offers another benefit that vertical transmission lacks. A classic problem in biological evolution is the *evolution of cooperation* [27,51]: given that genes are selected to promote their own good, with a disregard or even hostility toward any rivals that compete for their scarce resources, how can we explain cooperative or altruistic behavior where an individual invests more in helping another than in his or her own good? In the animal world, cases of altruism, such as among social insects, are usually explained via *kin selection*: individuals will help others as long as these are related to them, i.e. share their genes. In human society, however, people often help strangers that are totally unrelated. The initially proposed explanation of *group selection*, namely that groups of individuals that help each other survive better than groups of selfish individuals, has the shortcoming that, within altruist groups, it are the selfish profiteers that do best, and thus spread their genes most [27].

Horizontal transmission of cooperation norms solves this problem, since the members of a cultural group are all *memetically related* to each other, sharing their memes rather than their genes. Therefore, cultural kin selection will extend to all members of the group [36]. This entails a selective pressure for memes to support the fitness of the whole group of their carriers, e.g. by promoting cooperation. Moreover, selfish profiteers will not be able to undermine the cooperation produced by such altruism-promoting memes because of conformist pressure [17,51], or what we have called “winner-takes-all”: when one meme establishes a majority position it will eventually get imposed on *all* members of the group, thus suppressing the appearance of selfish dissidents—or at least not allowing them to make any converts and thus spread their memes. This cultural solution to the cooperation paradox in biological evolution appears to have been developed more or less independently by different meme theorists [15,17,36,47].

Topologies of Communication

Horizontal transmission will generally follow existing social or geographical topologies. This can be modeled in

two different ways:

1. Individuals are situated in a *space* (typically a two-dimensional plane, or its discrete equivalent, a two-dimensional lattice of cells);
2. Individuals are considered as nodes in a (social) *network*, which are connected by ties of acquaintance or trust.

The basic assumption in these models is that memes diffuse *continuously* across the space or network. This means that, in first instance, communications are considered to be *local*, i. e. agents exchange memes only with their direct neighbors in the space or social network. The neighbor can pass on the meme to its neighbors, and so on, so that a meme eventually may spread across the whole population.

When a population consists of different clusters or local communities, that have little communication with each other, this will typically lead to different cultures establishing themselves in different communities [8,17]. The reason is that intense communication within each community will produce a “winner-takes-all” dynamics where by chance or local adaptation one of several variant memes becomes dominant. Memes from other communities, however, will only rarely be encountered, so that they will generally not receive enough reinforcement to displace the established memes.

Recent research in complex networks, including social networks, has shown that such networks commonly have a *scale-free structure* [1]. This means that a few agents, the so-called “hubs” of the network, have a great many social ties, while most agents only have a few links. The implication for cultural diffusion is that memes hosted by “hub” agents will have a disproportionately large effect, and are much more likely to spread widely. A similar effect has been observed in the spread of sexually transmitted diseases, such as AIDS, where the infection of a few hubs in the network (in this case individuals with a large number of sexual partners) may make the difference between a large-scale epidemic and a few isolated infections. This observation has provided inspiration to researchers in “viral” marketing, who look for methods to make publicity for a brand or product by creating a “buzz”, i. e. a positive message about their product that is propagated via word-of-mouth [63,64]. Their strategies focus on identifying and targeting the “opinion-leaders” within a community, i. e. those central individuals that many know and tend to imitate.

Although it is in principle possible to make analytical models of the propagation of memes across space or across networks, calculating the precise spread in a realistic envi-

ronment is far too difficult. Therefore, these processes are typically explored via multi-agent computer simulations.

Computer Simulations of Cultural Evolution

Cultural transmission of rules, norms or information is a common ingredient in many social simulations (e. g. [8,20,32,37,45]), that are based on an “artificial society” of interacting software agents [35]. However, such memetic propagation is often added merely as one of the many assumptions within a complicated model of a specific type of socio-cultural evolution, such as the evolution of a shared vocabulary [69] or of cooperation norms [45]. There have been relatively few simulations that have explored cultural evolution in the broadest sense. We will now discuss some typical examples that illustrate the wider issue.

Probably the first explicitly memetic simulation, *Meme and Variations*, was made by Gabora [39] (first written 1992). The assumptions underlying this, and related simulations of cultural diffusion e. g. [9,30] are the following: agents search the best solution for a particular problem. They can either find a solution on their own through trial-and-error, or they can take over a solution from another agent, by observing the solutions each of their neighbors has found and imitating the best one. The result of the simulation is that the agents collectively find the best solutions if they partially imitate others, partially explore individually. If they only imitate, there is no creativity and the best solution cannot be improved. If they only explore individually, lots of search is needed to merely rediscover what was already known elsewhere. In the ideal situation, which is achieved by trying out different parameter values for the simulation until one has found the optimal mix of innovation and imitation, good solutions will spread very quickly throughout the population, but this without preventing the discovery of even better solutions by certain agents.

This simulation investigated the relative effectiveness of, and interaction between, *individual learning* and *cultural diffusion*. An older classic simulation [52] investigated the relative effectiveness of, and interaction between, *individual learning* and *genetic evolution*. Inspired by this work, Best [14] studied the three-way interactions between individual learning, genetic evolution, and cultural evolution. In Best’s simulation, agents can acquire knowledge that allows them to maximize their fitness in three ways: 1) by inheriting it, possibly with variations, from their parents (vertical, genetic transmission); 2) by copying it from another, fitter agent (horizontal, cultural transmission); 3) by individually discovering it via trial-and-error. The sim-

ulation showed that cultural transmission, just like individual learning, can enhance genetic evolution, accelerating its convergence to the optimal solution. Moreover, cultural transmission appeared superior to individual learning in that it produced convergence more quickly.

Best [14] also examined the situation in which cultural and genetic evolution pursue opposite goals, and found that in this case genetic evolution normally wins the competition. However, Bull, Holland and Blackmore [19] further investigated this situation by allowing cultural evolution to be much more rapid than genetic evolution, as is normally the case. They found that under these conditions memetic effects are stronger than genetic effects, and the only way genes can still keep some control over the process is by evolving mechanisms to filter out particularly harmful memes.

These simulations of cultural evolution are still rather simplistic, in the sense that agents literally copy any knowledge exhibited by a fitter agent. In practice, individuals do not a priori know which individual is fitter, and when they receive a message, this information will interact with the knowledge they already had. Van Overwalle and Heylighen [71] have proposed a more realistic simulation model in which agents do not just copy a message, but actively “reinterpret” it, based on their previous experience. Agents are modeled as simple neural networks that learn from experience. A message corresponds to a pattern of activation over the nodes in such a network, and communication to the spread of that activation from agent to agent via variable inter-agent connections. The strength of the connection between two agents represents the degree of trust of the one in the information received from the other. This trust is learned on the basis of the degree to which information received previously from that agent is confirmed by own knowledge. Unlike most multi-agent simulations, the Van Overwalle and Heylighen [71] model is supported by solid empirical evidence, in that it manages to accurately reproduce the results of several classic communication experiments, including the Lyons and Kashima [61,62] study of meme transmission that we will discuss in a later section.

Selection Criteria for Memes

Since mathematical models and computer simulations of meme spread necessarily have to make plenty of simplifying assumptions, and cannot incorporate all the specific social, psychological, linguistic and cultural factors that influence the propagation of a meme, they are not very useful in predicting which concrete memes will be successful and which will not. Yet, such predictions are neces-

sary if we want to arrive at an empirically testable theory, which can be applied to practical problems. One way to arrive at a more practical, predictive model is to formulate general selection criteria that distinguish fitter memes from less fit ones. All other things being equal, *a meme that scores better on one of these criteria is predicted to become more numerous* than a meme that scores worse. This is a falsifiable hypothesis that can be tested through experiments or observations. It suffices to operationalize the criteria so that satisfaction of a criterion can be objectively measured.

Many authors have proposed criteria for memetic success, and a few (e.g. [22,44,48,49,50]) have compiled lists of such criteria. Since these proposals, while related, are all different, we need to examine more clearly what is needed for a good list of criteria. First, such criteria should be formulated to be as much as possible independent or non-overlapping, so that a piece of information can vary along one dimension of evaluation without varying along the others. Second, without becoming too restrictive, they should be defined as precisely, concretely and unambiguously as possible, so that different observers using these criteria can come to the same conclusions.

To illustrate the importance of these methodological considerations, let us review some proposed criteria, and point to their shortcomings. For example, one might naively propose that fit memes should be *attractive* to their receivers. While this is true in a general sense, it helps us very little in operationalizing meme fitness, as we cannot say what makes a meme attractive without becoming much more explicit about its properties. A somewhat more sophisticated hypothesis may propose that good memes should be *communicable* [68]. Again, this is obviously correct, but communicability has so many different aspects, depending on the meme itself, its audience, the used medium, etc., that we might as well say that it simply should be a good meme. A more specific criterion, used e.g. by [46], is *plausibility*. The problem here is that people may use very different procedures to estimate plausibility, e.g. by looking at the source of the information, the available evidence for it, or their own previous experience.

We will here summarize the criteria proposed by Heylighen [48,49,50], which are based on an extensive review of relevant cognitive, social and communicative mechanisms. At the most abstract level, there are three classes of entities that information depends on: the object that it refers to, the subject who assimilates and remembers it, and the communication process that is used to transmit it between subjects. These determine three categories of selection criteria, *objective*, *subjective* and *intersubjective*:

Objective Criteria

Distinctiveness Information that refers to something precise, distinct or detailed, can be confirmed more easily by observation.

Invariance Information that remains valid over a wide range of contexts or situations, is more stable and broadly applicable.

Evidence Information that is supported by independent observations, is more reliable.

Subjective Criteria

Utility Information that is valuable or useful to its carrier is more likely to be remembered and passed on.

Affectivity Information that provokes strong emotions is more likely to be remembered and passed on: this typically stimulates instinctive reactions, such as fear, desire or disgust [46].

Coherence The better information fits in with the knowledge that individuals already have, the more easily they will understand and accept it [70]

Simplicity Short, simple messages are easier to assimilate, remember and transmit.

Novelty Information that is unexpected will attract more attention.

Repetition Repeated exposure to the same message helps it to be assimilated and retained.

Intersubjective Criteria

Publicity The more effort an individual puts into spreading a message, the more people will receive it.

Formality Messages formulated explicitly and unambiguously are less likely to be misinterpreted.

Expressivity Information must be easy to express in a given language or medium.

Authority An authoritative, trustworthy source of the information makes it more likely to be accepted.

Conformity Information confirmed by many people is more easily accepted [17].

Collective utility Information, if adopted by a group, may help the group to function better, and therefore to grow or function as a model for others. Examples are standards, linguistic conventions, and traffic rules.

Parasitic Memes

Mememes being communicated undergo natural selection. Some mememes are transmitted easily, thus reaching a large number of people, while others are rejected, misunderstood, forgotten or otherwise eliminated from circulation.

This means that the mememes best adapted to the underlying cognitive and communicative processes will spread farthest. We may assume that our brain, general culture, and social structures have evolved so as to maximize the fitness of society and its members. This means that they should be good at assimilating useful mememes, and at rejecting bad ones, e. g. [15]. For example, we know that we should not accept things without evidence, and that some sources are more reliable than others. Insofar that these socio-cognitive guidelines, as exemplified by the above list of selection criteria, efficiently filter out poor-quality information, successful mememes will also increase the fitness of their carriers.

However, since no system is foolproof, these mechanisms will not always be reliable. This leaves a niche for mememes to evolve that propagate well, apparently satisfying the criteria that people intuitively use, but without delivering any benefit to their carriers. We may call such mememes *selfish* [47] or *parasitic* cf. [26], as they free ride on the effort invested by individuals to gather and communicate useful information. Such information parasites succeed by faking the criteria that we use to recognize high-quality information. This is similar to the way many biological organisms mimic other phenomena, such as viruses that mimic the cell's own DNA, so that they are reproduced for free by the cellular machinery. Mememes have therefore been described as "mind viruses" [18,28], since they similarly exploit our cognitive machinery to get themselves replicated.

There are plenty of examples of such parasitic mememes. Perhaps the most studied from a memetic point of view are *chain letters*, whose only purpose is to have themselves replicated and sent to as many people as possible [11,42]. A more modern variant are *virus hoaxes* [24,25], i. e. email messages that warn the receivers for a non-existent type of computer virus, and urge them to pass on this warning to as many people as possible. Probably the most dangerous information parasites are certain *religious cults* [26], which indoctrinate their followers to make as many converts as possible, while isolating them from alternative sources of information, so that they tend to develop a view of reality that is so distorted that it may end fatally, as in the mass suicides of the Heaven's Gate cult. *Pseudosciences* too can be dangerous, parading as solid scientific theories, but asserting statements that at best are not supported by the facts, like in astrology, at worst fatally wrong, like in certain quack cures for cancer. Somewhat more benign are *urban legends* and various *rumors* and *fads*, which tend to spread in waves, being passed on from person to person but without any authoritative source or real evidence. Bangerter and Heath [10] have tracked the evolution of

one such legend, the *Mozart effect* (i. e. the unfounded belief that babies listening to classical music become more intelligent), starting from its source: a scientific experiment that merely found that after listening to music adults temporarily scored better on certain tests—perhaps simply because they were more relaxed.

Several memeticists, e. g. [18,28,60], have investigated the many tricks that memes use to appear more acceptable than they deserve to be, including the following. *Self-justification* means that the components of a memplex mutually justify each other, but without independent support. An example is: “God exists because the Bible says so”, and “You should believe the Bible because it is the word of God”. *Self-reinforcement* means that a meme stimulates its host to rehearse itself, e. g. by repeated study, meditation, prayer, etc. *Intolerance* means that a meme indoctrinates its host to a priori reject any potentially competing memes. *Proselytism* occurs when a meme urges its host to maximally spread the meme to other hosts, like in the case of chain letters. *Baiting* occurs when a meme promises its carriers a reward if they only accept and spread the meme. All these tactics are common in religious cults, which promise their adherents that they will go to Heaven if they believe the teachings, pray, and spread the word, while they will burn in Hell if they dare to doubt [28].

Parasitic memes have been the subject and inspiration of most empirical approaches to memetics, since their spread is relatively easy to track, and since they are prime illustrations of the way in which cultural evolution is fundamentally different from genetic evolution.

Empirical Tests

Memetics has often been challenged and has known some very virulent critics. One of the main criticisms is that there are no empirical data to back up the theories that were put forth, and in that sense memetics is merely a way of thinking rather than a scientific discipline [34,40,43]. This criticism is to some degree justifiable. The lack of a universally accepted meme definition and the vagueness of meme boundaries cf. [2] indeed make empirical studies less evident. Yet, there have already been a few empirical studies of meme propagation in different conditions, both in the laboratory, e. g. [61,62], and in real life, e. g. [11,24,42]. More generally, we must note that memetics is an approach that illuminates important aspects of culture, society and communication that more traditional approaches, such as sociology, psychology or history, tend to neglect. Empirical tests cannot confirm or falsify this perspective as a whole, but merely specific implementations of it. This is analogous to the observation that ex-

periments in psychology or sociology can test particular theories within their field, but not the field as such.

Within the memetics field, one simple way to test specific theories is by considering the memetic selection criteria they imply [25]. We can then measure the apparent success rate of different memes, and examine its correlation with the degree to which the memes fulfill the proposed selection criteria. Heath, Bell and Sternberg [46] applied this approach to investigate the criteria of *affectivity* (which they call *emotional selection*) and *plausibility*. From the affects, they focused on disgust because this is a relatively simple emotion whose strength is easy to measure. When comparing different urban legends that contained an element of disgust (e. g. the story of a man who discovers a dead rat in the cola bottle he has just been drinking from), they found that the more disgusting variations typically had spread more widely than the less disgusting ones. The same applied to more plausible variations. Chieffens [24] used a similar method to examine the spread of virus hoaxes. He used both expert and non-expert surveys to score different hoaxes on different criteria, including novelty, simplicity, utility, authority and proselytism. In this case, the spread of the hoax correlated most strongly with its *novelty*. Schaller, Conway and Tanchuk [68] examined the correlation between the *communicability* of traits, i. e. the probability with which subjects would speak about a particular trait, and the frequency of these traits. They found that for traits used to describe different ethnic groups, the most communicable traits are also the ones that are most widely spread and persistent in society. Pocklington and Best [66] and [12,13] used automatic text analysis to measure how often a certain discussion subject was mentioned in a particular discussion group on the net. They found evidence for memetic competition between mutually *incoherent* subjects, meaning that an increase in the one correlated with a decrease in the other.

While these studies merely observed existing patterns of spread, Lyons and Kashima [61,62] performed a laboratory experiment in which they deliberately produced a memetic transmission chain. They created a fictional story and asked the participants in the experiment to subsequently tell the story from person 1 to person 2, from 2 to 3, and so on, like in the game of “telephone” or “Chinese Whispers.” The story involved a fictitious tribe, the Jamayans, about which all participants had received some background information. After several experiments, a statistical analysis of the story elements that remained at the end of the transmission chain (i. e. as reported by the last person to hear the story) found a number of systematic selection effects. These seem to confirm four of the previously mentioned criteria:

1. *Coherence*: Elements inconsistent with the background information were more likely to be left out;
2. *Novelty*: Elements that the speakers assumed were already known by the listeners were more likely to be left out;
3. *Simplicity*: Details or embellishments that did not affect the story line tended to be left out;
4. *Conformity*: When the participants were told that the majority of them believed that the Jamayans were, e. g., peaceful, they were more likely to leave out elements inconsistent with this fact than if they thought that this was only a minority opinion.

Experiments on memes “in the wild” are more difficult to control than in such a laboratory situation. Because of ethical concerns it is hardly possible to release a well-doctored meme upon the world to observe how it evolves. Another problem is that it is hard to track memes once they are released. Even when feedback is asked or if mechanical trackers are included in the meme (for example, images included in an email message that are automatically downloaded from a controlled server each time the message is opened), it is hard to see whether the meme has already been changed by the time these devices are triggered.

In conclusion, empirical memetics research remains in its infancy. However, experiments have already shown that one can quantitatively confirm or falsify a number of non-trivial memetic predictions. It is to be hoped that many more studies along these lines will be carried out. As long as memetics has not been thoroughly investigated empirically, it is likely to remain a theoretical niche framework used only for description rather than prediction.

Future Directions

The theory of memetics and cultural evolution holds out great promises for a better understanding, anticipation and control of fundamental social problems that depend on the propagation of ideas and behaviors. This will require more extensive empirical observations and tests, more detailed computer models of the interaction between memes and their hosts, and a better conceptualization of what a meme precisely is. A first basic result that should come out of such research is a concrete and reliable list of criteria that characterize successful memes, i. e. ideas or cultural traits that propagate widely and easily across large populations. This would allow us not only to recognize such memes, but to some degree to design or improve them.

The ability to create successful memes is the Holy Grail of marketing research [41], which is constantly on the look-out for techniques to create a “buzz” and have their

publicity message or brand name [63,64] become as widely known as possible. Another application of these principles lies in public education. For example, if the government makes a campaign to convince people to stop smoking, it would be very useful to have the campaign designed according to sound memetic principles. This should take into account both the characteristics of the message itself (e. g. being sufficiently simple and unambiguous), of the intended audience (e. g. being consistent with what the audience already believes, while being sufficiently novel to attract their attention), and of the way it is transmitted (e. g. having the meme expressed in a common medium by people considered trustworthy).

Memetic selection criteria can be applied not only positively, to help spread a beneficial idea, but negatively, to prevent or suppress harmful memes. Examples are the idea that it is cool to smoke, false rumors and scares that may promote panic or accentuate social prejudice, fundamentalist ideologies that incite hatred or terrorism, and dangerous superstitions, such as the belief that you can cure AIDS by having sex with a virgin. A better understanding of memetic dynamics may help us to understand how such mind viruses arise and spread. It may moreover help us to “immunize” the population by educating them about basic memetics, so as not to be misled by apparently plausible—but fundamentally misleading—cults, fads and superstitions [18,28].

Another basic result of future memetic research should be a complex dynamical model of the interactions between individuals, groups, and the memes they carry. This should allow making longer-term predictions about the interactions between different groups and subcultures within our globalizing society. A crucial issue in this regard is whether minority cultures will eventually be assimilated into the majority, or on the contrary become polarized, asserting their divergent habits and beliefs ever more forcefully [8,71]. Two concrete examples are minority languages, such as the Welsh still spoken in Britain, where there is a tendency for the subculture to be slowly erased by the majority culture, and the culture of Islamic immigrants in Europe, where there is a tendency towards polarization in the sense of increased radicalism. Neither complete assimilation nor polarization are desirable outcomes, but at first sight they seem like the most likely results of the “winner-takes-all” dynamics created by the pressure to conform to the group one has most contact with. A more detailed theory of cultural evolution may help us to find a middle way that preserves cultural diversity without exacerbating conflicts, and to pinpoint the crucial factors that can steer the dynamics in one direction rather than another.

Bibliography

Primary Literature

1. Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
2. Atran S (2001) The Trouble with memes: Inference versus imitation in cultural creation. *Hum Nat* 12:351–381
3. Aunger R (2001) *Darwinizing Culture: The Status of Memetics As a Science*. University Press, Oxford
4. Aunger R (2002) Exposure versus susceptibility in the epidemiology of 'everyday' beliefs. *J Cogn Cult* 2(2):113–154
5. Aunger R (2003) Cultural transmission and diffusion. In: Nadel L (ed) *Encyclopedia of Cognitive Science*. MacMillan, London
6. Aunger R (2003) The Electric Meme: A New Model of How We Think. Simon and Schuster, New York
7. Aunger R (2004) Memes. In: Kuper A, Kuper J (eds) *The Social Science Encyclopedia*, 3rd edn. Routledge, London
8. Axelrod R (1997) The Dissemination of Culture: A Model with Local Convergence and Global Polarization. *J Confl Resolut* 41:203–26
9. Baldassarre G, Parisi D (1999) Trial-and-Error Learning, Noise and Selection in Cultural Evolution: A Study Through Artificial Life Simulations. In: *Proceedings of the AISB'99 Symposium on Imitation in Animals and Artifacts*
10. Bangerter A, Heath C (2004) The Mozart effect: Tracking the evolution of a scientific legend. *Br J Soc Psychol* 43:605–623
11. Bennet C, Li M, Ma B (2003) Chain Letters & Evolutionary Histories. *Sci Am* 288(6):76–81
12. Best ML (1997) Models for interacting populations of memes: Competition and niche behavior. *J Memetics* 1(2) http://jom-emit.cfp.m.org/1997/vol1/best_ml.html
13. Best ML (1998) An ecology of text: Using text retrieval to study alive on the net. *J Artif Life* 3(4):261–287
14. Best ML (1999) How culture can guide evolution: An inquiry into gene/meme enhancement. *J Adapt Behav* 7(3/4):289–306
15. Blackmore S (2000) *The Meme Machine*. Oxford University Press, Oxford
16. Bonner JT (1980) *The Evolution of Culture in Animals*. Princeton University Press, Princeton
17. Boyd R, Richerson PJ (1985) *Culture and the evolutionary process*. Chicago University Press, Chicago
18. Brodie R (1996) *Virus of the mind: The new science of the meme*. Integral Press, Seattle
19. Bull L, Holland W, Blackmore S (2001) On meme-gene coevolution. *Artif Life* 6(3):227–235
20. Bura S (1994) MINIMEME: Of life and death in the Noosphere. In: Cliff D, Husbands P, Meyer JA Wilson SW (eds) *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, pp 479–486
21. Campbell DT (1974) Evolutionary Epistemology. In: Schilpp PA (ed) *The Philosophy of Karl Popper*. Open Court Publish, La Salle, pp 413–463
22. Castelfranchi C (2001) Towards a Cognitive Memetics: Socio-Cognitive Mechanisms for Memes Selection and Spreading. *J Memetics Evol Models Inf Transm* 5
23. Cavalli-Sforza LL, Feldman MW (1981) *Cultural transmission and evolution: A quantitative approach*. Princeton University Press, Princeton
24. Chielens K (2003) *The Viral Aspects of Language: A Quantitative Research of Memetic Selection Criteria*. Master's thesis, Vrije Universiteit Brussel. <http://memetics.chielens.net/master/thesis.pdf>
25. Chielens K, Heylighen F (2005) Operationalization of Meme Selection Criteria: procedures to empirically test memetic hypotheses. In: *Proceedings AISB 2005*
26. Cullen B (1999) Parasite Ecology and the Evolution of Religion. In: Heylighen F, Bollen J, Riegler A (eds) *The Evolution of Complexity*. Kluwer, Dordrecht
27. Dawkins R (1989) *The Selfish Gene*, 2nd edn. Oxford University Press, Oxford
28. Dawkins R (1993) Viruses of the mind. In: Dahlbom B (ed) *Dennett and his Critics: Demystifying the Mind*. Blackwell, USA, pp 12–27
29. de Jong M (1999) Survival of the institutionally fittest concepts. *J Memetics Evol Models Inf Transm* 3. http://jom-emit.cfp.m.org/1999/vol3/de_jong_m.html
30. Denaro D, Parisi D (1996) Cultural evolution in a population of neural networks. In: Marinaro M, Tagliaferri R (eds) *Neural nets: Wirm96*. Springer, New York
31. Dennett D (1995) *Darwin's dangerous idea*. Hammondsworth, Penguin
32. Doran J (1998) Simulating Collective Misbelief. *J Artif Soc Soc Simul* 1:1. <http://www.soc.surrey.ac.uk/JASSS/1/1/3.html>
33. Durham WH (1991) *Coevolution: Genes, culture and human diversity*. Stanford University Press, Stanford
34. Edmonds B (2002) Three Challenges for the Survival of Memetics. *J Memetics Evol Models Inf Transm* 6. http://jom-emit.cfp.m.org/2002/vol6/edmonds_b_letter.html
35. Epstein J, Axtell R (1996) *Growing Artificial Societies: Social Science from the Bottom Up*. MIT Press, London
36. Evers JR (1998) A justification of societal altruism according to the memetic application of Hamilton's rule. *Proc 16th Int Congress on Cybernetics. Namur: Association Internat de Cybernétique*, pp 437–442
37. Flentge F, Polani D, Uthmann T (2001) Modelling the Emergence of Possession Norms using Memes. *J Artif Soc Soc Simul* 4(4). <http://www.soc.surrey.ac.uk/JASSS/4/4/3.html>
38. Flinn MV, Alexander RD (1982) Culture theory: The developing synthesis from biology. *Hum Ecol* 10:383–400
39. Gabora L (1995) Meme and variations: A Computer Model of Cultural Evolution. In: Nadel L, Stein D (eds) *1993 Lectures in Complex Systems*. Addison-Wesley, Boston, pp 471–485
40. Gil-White FJ (2005) Common Misunderstandings of Memes (and Genes). The Promise and the Limits of the Genetic Analogy to Cultural Transmission Processes. In: Hurley S, Chater N (eds) *Perspectives on Imitation. From Neuroscience to Social Science*, vol 2. MIT Press, Cambridge, pp 317–338
41. Godin S (2002) *Unleashing the Ideavirus*. Simon and Schuster, London
42. Goodenough OR, Dawkins R (2002) The 'St Jude' mind virus. *Nature* 371:23–24
43. Greenberg M (2005) Goals versus Memes: Explanation in the Theory of Cultural Evolution. In: Hurley S, Chater N (ed) *Perspectives on Imitation. From Neuroscience to Social Science*, vol 2. MIT Press, Cambridge, pp 339–353
44. Hale-Evans R (1995) *Memetics: A Systems Metabiology working report*. <http://ron.ludism.org/memetics.html>

45. Hales D (1998) Selfish memes & selfless agents – Altruism in the swap shop. In: *Proceedings of the 3rd International Conference on Multi-Agent Systems*. IEEE Press, Los Gatos
46. Heath C, Bell C, Sternberg E (2001) Emotional Selection in Memes: The Case of Urban Legends. *J Pers Soc Psychol* 81(6):1028–1041
47. Heylighen F (1992) Selfish Memes and the Evolution of Cooperation. *J Ideas* 2(4):77–84
48. Heylighen F (1993) Selection C criteria for the Evolution of Knowledge. In: *Proc 13th Int Congress on Cybernetics Association Internat de Cybernétique*, Namur, pp 524–528
49. Heylighen F (1997) Objective, subjective and intersubjective selectors of knowledge. *Evol Cogn* 3(1):63–67
50. Heylighen F (1998) What makes a meme successful? Selection Criteria for Cultural Evolution. *Proc 16th Int Congress on Cybernetics. Association Internat de Cybernétique*, Namur
51. Heylighen F, Campbell DT (1995) Selection of Organization at the Social Level: obstacles and facilitators of metasystem transitions. *World Futures: J General Evol* 45:181–212
52. Hinton GE, Nowlan SJ (1987) How Learning Can Guide Evolution. *Complex Syst* 1:495–502
53. Holland JH, Holyoak KJ, Nisbett RE, Thagard PR (1986) *Induction: processes of inference, learning and discovery*. MIT Press, Cambridge
54. Hull DL (1982) The naked meme. In: Plotkin HC (ed) *Development and culture: Essays in evolutionary epistemology*. Chichester, Wiley, pp 272–327
55. James W (1880) *Great Men and their Environment*. <http://www.des.emory.edu/mfp/jgreatmen.html>. Accessed 28 Sept 2005
56. Kauffman SA (1993) *The Origins of Order: Self-Organization and Selection in Evolution*. Oxford University Press, New York
57. Lake MW (1998) Digging for memes: The role of material objects in cultural evolution. In: Renfrew C, Scarre C (eds) *Cognition and material culture: the archeology of symbolic storage*. McDonald Institute Monographs, Cambridge pp 77–88
58. Lumsden C, Wilson E (1981) *Genes, Mind, and Culture: the Co-evolutionary Process*. Harvard University Press, Cambridge
59. Lynch A (1996) *Thought contagion: How beliefs spread through society: The new science of memes*. Basic Books, New York
60. Lynch A (1998) Units, events and dynamics in memetic evolution. *J Memetics Evol Models Inf Transm* 2. http://www.cpm.mmu.ac.uk/jomemit/1998/vol2/lynch_a.html
61. Lyons A, Kashima Y (2001) The reproduction of culture: Communication processes tend to maintain cultural stereotypes. *Soc Cogn* 19:372–394
62. Lyons A, Kashima Y (2003) How Are Stereotypes Maintained Through Communication? *J Pers Soc Psychol* 85:989–1005
63. Marsden P (2002) Brand positioning: meme's the word. *Mark Intell Plan* 20(5):307–12
64. Marsden P, Kirby J (2005) *Connected Marketing: The Buzz, Viral and Word of Mouth Revolution*. Elsevier, London
65. Müller M (1870) The Science of Language. *Nature* 2:256–259
66. Pocklington R, Best ML (1997) Cultural evolution and units of selection in replicating text. *J Theor Biol* 188:79–87
67. Rogers EM (2003) *Diffusion of Innovations*, Fifth Edition. Free Press, New York
68. Schaller M, Conway LG III, Tanchuk TL (2002) Selective pressures on the once and future contents of ethnic stereotypes: Effects of the 'communicability' of traits. *J Pers Soc Psychol* 82:861–877
69. Steels L (1998) Synthesising the origins of language and meaning using co-evolution, selforganisation and level formation. In: Hurford J (ed) *Evolution of Human Language*. Edinburgh Univ Press, Edinburgh
70. Thagard P (1989) Explanatory Coherence. *Behav Brain Sci* 12:435–467
71. Van Overwalle F, Heylighen F (2006) Talking Nets: A Multi-Agent Connectionist Approach to Communication and Trust between Individuals. *Psychol Rev* 113:606–627
72. Van Wyhe J (2005) The Descent of words: evolutionary thinking 1780–1880. *Endeavour* 29(3):94–100
73. Wilkins JS (1998) What's in a meme? Reflections from the perspective of the history and philosophy of evolutionary biology. *J Memetics Evol Models Inf Transm* 2. http://jom-emit.cfm.org/1998/vol2/wilkins_js.html

Books and Reviews

- Benzon W (1996) Culture as an evolutionary arena. *J Soc Evol Syst* 19:321–362 <http://www.newsavanna.com/wlb/CE/Arena/Arena00.shtml>
- Blackmore S (2000) *The Meme Machine*. University Press, Oxford
- Blackmore S (2001) Evolution And Memes: The Human Brain As A Selective Imitation Device. *Cybern Syst* 32(1–2):225–255; <http://www.ingentaconnect.com/content/tandf/ucbs;jsessionid=1b1oec7s7li7s.henrietta>
- Boyd R, Richerson PJ (1985) *Culture and the evolutionary process*. Chicago University Press, Chicago
- Edmonds B (1998) On Modelling in Memetics. *J Memetics Evol Models Inf Transm* 2. http://jom-emit.cfm.org/1998/vol2/edmonds_b.html
- Fog A (1999) *Cultural Selection*. Kluwer, Dordrecht
- Heyes CM (1994) Imitation and culture: Longevity, fecundity and fidelity in social transmission. In: Galef BG, Mainardi M, Valsecchi P (eds) *Behavioural Aspects of Feeding*. Harwood, Chur, pp 271–287
- Jesiek BK (2003) *Betwixt the Popular and Academic: The Histories and Origins of Memetics*. Master Thesis, Virginia Tech
- Marsden P (1998) *Memetics and Social Contagion: Two Sides of the Same Coin*. *J Memetics Evol Models Inf Transm* 2. http://jom-emit.cfm.org/1998/vol2/marsden_p.html
- Moritz E (1990) *Memetic Science: I – General Introduction*. *J Ideas* 1:1–23

Evolution in Materio

SIMON HARDING¹, JULIAN F. MILLER²

¹ Department of Computer Science,
Memorial University, St. John's, Canada

² Department of Electronics, University of York,
Heslington, UK

Article Outline

Glossary

Definition of the Subject

Introduction

Evolutionary Algorithms

Evolution in Materio: Historical Background

Evolution in Materio: Defining Suitable Materials

Evolution in Materio Is Verified with Liquid Crystal

Evolution in Materio Using Liquid Crystal:

Implementational Details

The Computational Power of Materials

Future Directions

Bibliography

Glossary

Evolutionary algorithm A computer algorithm loosely inspired by Darwinian evolution.

Generate-and-test The process of generating a potential solution to a computational problem and testing it to see how good a solution it is. The idea behind it is that no human ingenuity is employed to make good solutions more likely.

Genotype A string of information that encodes a potential solution instance of a problem and allows its suitability to be assessed.

Evolution in materio The method of applying computer controlled evolution to manipulate or configure a physical system.

Liquid crystal Substances that have properties between those of a liquid and a crystal.

Definition of the Subject

Evolution in materio refers to the use of computers running search algorithms, called evolutionary algorithms, to find the values of variables that should be applied to material systems so that they carry out useful computation. Examples of such variables might be the location and magnitude of voltages that need to be applied to a particular physical system. Evolution in materio is a methodology for programming materials that utilizes physical effects that the human programmer need not be aware of. It is a general methodology for obtaining analogue computation that is specific to the desired problem domain. Although a form of this methodology was hinted at in the work of Gordon Pask in the 1950s it was not convincingly demonstrated until 1996 by Adrian Thompson, who showed that physical properties of a digital chip could be exploited by computer controlled evolution. This article describes the first demonstration that such a method can be used to obtain specific analogue computation in a non-silicon based physical material (liquid crystal). The work is important for a number of reasons. Firstly, it proposes a general method for building analogue computational devices. Secondly it explains how previously un-

known physical effects may be utilized to carry out computations. Thirdly, it presents a method that can be used to *discover* useful physical effects that can form the basis of future computational devices.

Introduction

Physical Computation

Classical computation is founded on a mathematical model of computation based on an abstract (but physically inspired) machine called a Turing Machine [1]. A Turing machine is a machine that can write or erase symbols on a possibly infinite one dimensional tape. Its actions are determined by a table of instructions that determine what the machine will write on the tape (by moving one square left or right) given its state (stored in a state register) and the symbol on the tape. Turing showed that the calculations that could be performed on such a machine accord with the notion of computation in mathematics. The Turing machine is an abstraction (partly because it uses a possibly infinite tape) and to this day it is still not understood what limitations or extensions to the computational power of Turing's model might be possible using real physical processes. Von Neumann and others at the Institute for Advanced Study at Princeton devised a design for a computer based on the ideas of Turing that has formed the foundation of modern computers. Modern computers are digital in operation. Although they are made of physical devices (i. e. transistors), computations are made on the basis of whether a voltage is above or below some threshold. Prior to the invention of digital computers there have been a variety of analogue computing machines. Some of these were purely mechanical (e.g. an abacus, a slide-rule, Charles Babbage's difference engine, Vannevar Bush's Differential Analyzer) but later computing machines were built using operational amplifiers [2].

There are many aspects of computation that were deliberately ignored by Turing in his model of computation. For instance, speed, programmability, parallelism, openness, adaptivity are not considered. The speed at which an operation can be performed is clearly an important issue since it would be of little use to have a machine that can calculate any computable function but takes an arbitrarily large amount of time to do so. Programmability is another issue that is of great importance. Writing programs directly in the form of instruction tables that could be used with a device based on a Turing is extremely tedious. This is why many high-level computer languages have been devised. The general issue of how to subdivide a computer program into a number of parallel executing processes so that the intended computation is carried out as quickly as

possible is still unsolved. Openness refers to systems that can interact with an external environment during their operation. Openness is exhibited strongly in biological systems where new resources can be added or removed either by an external agency or by the actions taken by the system itself. Adaptivity refers to the ability of systems to change their characteristics in response to an environment.

In addition to these aspects, the extent to which the underlying physics affects both the abstract notion of computation and its tractability has been brought to prominence through the discovery of quantum computation, where Deutsch pointed out that Turing machines implicitly use assumptions based on physics [3]. He also showed that through ‘quantum parallelism’ certain computations could be performed much more quickly than on classical computers. Other forms of physical computation that have recently been explored are: reaction-diffusion systems [4], DNA computing [5,6] and synthetic biology [7].

In the UK a number of Grand Challenges in computing research have been proposed [8], in particular ‘Journeys in Non-Classical Computation’ [9,10] seeks to explore, unify and generalize many diverse non-classical computational paradigms to produce a mature science of computation.

Toffoli argued that ‘Nothing Makes Sense in Computation Except in the Light of Evolution’ [11]. He argues firstly that a necessary but not sufficient condition for a computation to have taken place, is when a novel function is produced from a fixed and finite repertoire of components (i. e. logic gates, protein molecules). He suggests that a sufficient condition requires *intention*. That is to say, we cannot argue that computation has taken place unless a system has arisen for a higher purpose (this is why he insists on intention as being a prerequisite for computation). Otherwise, almost everything is carrying out some form of computation (which is not a helpful point of view). Thus a Turing machine does not carry out computations unless it has been programmed to do so, and since natural evolution constructs organisms that have an increased chance of survival (the higher ‘purpose’) we can regard them as carrying out computations. It is in this sense that Toffoli points to the fundamental role of evolution in the definition of a computation as it has provided animals with the ability to have intention.

This brings us to one of the fundamental questions in computation. How can we program a physical system to perform a particular computation? The dominant method used to answer this question has been to construct logic gates and from these build a von Neumann machine (i. e. a digital computer). The mechanism that has been used to devise a computer program to carry out a particular com-

putation is the familiar top-down design process, where ultimately the computation is represented using Boolean operations. According to Conrad this process leads us to pay “The Price of Programmability” [12], whereby in conventional programming and design we proceed by excluding many of the processes that may lead to us solving the problem at hand. Natural evolution does not do this. It is noteworthy that natural evolution has constructed systems of extraordinary sophistication, complexity and computational power. We argue that it is not possible to construct computational systems of such power using a conventional methodology and that complex software systems that directly utilize physical effects will require some form of search process akin to natural evolution together with a way of manipulating the properties of materials. We suggest that some form of evolution ought to be an appropriate methodology for arriving at *physical* systems that compute. In this chapter we discuss work that has adopted this methodology. We call it evolution in materio.

Evolutionary Algorithms

Firstly we propose that to overcome the limitations of a top-down design process, we should use a more unconstrained design technique that is more akin to a process of generate-and-test. However, a guided search method is also required that spends more time in areas of the search space that confer favorable traits for computation. One such approach is the use of evolutionary algorithms. These algorithms are inspired by the Darwinian concepts of survival of the fittest and the genetic inheritance of information. Using a computer, a population of randomly generated solutions is systematically tested, selected and modified until a solution has been found [13,14,15].

As in nature, a genetic algorithm optimizes a population of individuals by selecting the ones that are best suited to solving a problem and allowing their genetic make-up to propagate into future generations. It is typically guided only by the evolutionary process and often contains very limited domain specific knowledge. Although these algorithms are bio-inspired, it is important that any analogies drawn with nature are considered only as analogies.

Their lack of specialization for a problem makes genetic algorithms ideal search techniques where little is known about a problem. As long as a suitable representation is chosen along with a fitness function that allows for ease of movement around a search space, a GA can search vast problem spaces rapidly. Another feature of their behavior is that provided that the genetic representation chosen is sufficiently expressive the algorithm can explore potential solutions that are unconventional. A human de-

signer normally has a set of predefined rules and strategies that they adopt to solve a problem. These preconceptions may prevent trying a new method, and may prevent the designer using a better solution. A genetic algorithm does not necessarily require such domain knowledge. Evolutionary algorithms have been shown to be competitive or surpass human designed solutions in a number of different areas. The largest conference on evolutionary computation called GECCO has an annual session on evolutionary approaches that have produced human competitive scientific and technological results. Moreover the increase in computational power of computers makes such results increasingly more likely.

Many different versions of genetic algorithms exist. Variations in representations and genetic operators change the performance characteristics of the algorithm, and depending on the problem, people employ a variety of modifications of the basic algorithm. However, all the algorithms follow a similar basic set of steps.

Firstly the numbers or physical variables that are required to define a potential solution have to be identified and encoded into a data representation that can be manipulated inside a computer program. This is referred to as the encoding step. The representation chosen is of crucial importance as it is possible to inadvertently choose overly constrained representations which limits the portions of the space of potential solutions that will be considered by the evolutionary algorithm. Generally the encoded information is referred to as a genotype and genotypes are sometimes divided into a number of separate strings called chromosomes. Each entry in the chromosome string is an allele, and one or more of these make up a gene.

The second step is to create inside the computer a number of independently generated genotypes whose alleles have been chosen with uniform probability from the allowed set of values. This collection of genotypes is called a population.

In its most basic form, an individual genotype is a single chromosome made of 1s and 0s. However, it is also common to use integer and floating-point numbers if they are more appropriate for the task at hand. Combinations of different representations can also be used within the same chromosome, and that is the approach used in the work described in this article. Whatever representation is used, it should be able to adequately describe the individual and provide a mechanism where its characteristics can be transferred to future generations without loss of information.

Each of these individuals is then decoded into its phenotype, the outward, physical manifestation of the individual and tested to see how well the candidate solution solves

the problem at hand. This is usually returned as a number that is referred to as the *fitness* of the genotype. Typically it is this phase in a genetic algorithm that is the most time consuming.

The next stage is to select what genetic information will proceed to the next generation. In nature the fitness function and selection are essentially the same – individuals that are better suited to the environment survive to reproduce and pass on their genes. In the genetic algorithm a procedure is applied to determine what information gets to proceed.

Genetic algorithms are often generational – where all the old population is removed before moving to the next generation, in nature this process is much less algorithmic. However, to increase the continuity of information between generations, some versions of the algorithm use elitism, where the fittest individuals are always selected for promotion to the next generation. This ensures that good solutions are not lost from the population, but it may have the side effect of causing the genetic information in the population to converge too quickly so that the search stagnates on a sub-optimal solution.

To generate the next population, a procedure analogous to sexual reproduction occurs. For example, two individuals will be selected and they will then have their genetic information combined together to produce the genotype for the offspring. This process is called recombination or crossover. The genotype is split into sections at randomly selected points called crossover points. A “simple” GA has only one of these points, however it is possible to perform this operation at multiple points.

Sections of the two chromosomes are then put together to form a new individual. This individual shares some of the characteristics of both parents. There are many different ways to choose which members of the population to breed with each other, the aim in general is to try and ensure that fit individuals get to reproduce with other fit individuals. Individuals can be selected with a probability proportional to their relative fitness or selected through some form of tournament, which may choose two or more chromosomes at random from the population and select the fittest.

In natural recombination, errors occur when the DNA is split and combined together. Also, errors in the DNA of a cell can occur at any time under the influence of a mutagen, such as radiation, a virus or toxic chemical. The genetic algorithm also has mutations. A number of alleles are selected at random and modified in some way. For a binary GA, the bit may be flipped, in a real-numbered GA a random value may be added to or subtracted from the previous allele.

Although GAs often have both mutation and cross-over, it is possible to just use mutation. A mutation only approach has in some cases been demonstrated to work, and often crossover is seen as a macro mutation operator – effectively changing large sections of a chromosome.

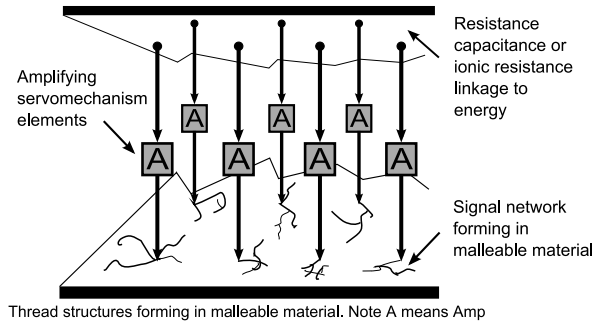
After the previous operations have been carried out, the new individuals in the population are then retested and their new fitness scores calculated. Eventually this process leads to an increase in the average fitness of the population, and so the population moves closer toward a solution. This cycle of test, select and reproduce is continued until a solution is found (or some other termination condition is reached), at which point the algorithm stops. The performance of a genetic algorithm is normally measured in terms of the number of evaluations required to find a solution of a given quality.

Evolution in Materio: Historical Background

It is arguable that ‘evolution in materio’ began in 1958 in the work of Gordon Pask who worked on experiments to grow neural structures using electrochemical assemblages [16,17,18,19]. Gordon Pask’s goal was to create a device sensitive to either sound or magnetic fields that could perform some form of signal processing – a kind of ear. He realized he needed a system that was rich in structural possibilities, and chose to use a metal solution. Using electric currents, wires can be made to self-assemble in an acidic aqueous metal-salt solution (e.g. ferrous sulphate). Changing the electric currents can alter the structure of these wires and their positions – the behavior of the system can be modified through external influence. Pask used an array of electrodes suspended in a dish containing the metal-salt solution, and by applying current (either transiently or a slowly changing source) was able to build iron wires that responded differently to two different frequencies of sound – 50 Hz and 100 Hz.

Pask had developed a system whereby he could manually train the wire formation in such a way that no complete specification had to be given – a complete paradigm shift from previous engineering techniques which would have dictated the position and behavior of every component in the system. His training technique relied on making changes to a set of resistors, and updating the values with given probabilities – in effect a test-randomly modify-test cycle. We would today recognize this algorithm as some form of evolutionary, hill climbing strategy – with the test stage as the fitness evaluation.

In 1996 Adrian Thompson started what we might call the modern era of evolution in materio. He was investigating whether it was possible to build working electronic



Evolution in Materio, Figure 1

Pask's experimental set up for growing dendritic wires in ferrous sulphate solution [17]

circuits using unconstrained evolution (effectively, generate-and-test) using a re-configurable electronic silicon chip called an Field Programmable Gate Array (FPGA). Carrying out evolution by defining configurations of actual hardware components is known as *intrinsic* evolution. This is quite possible using FPGAs which are devices that have a two-dimensional array of logic functions that a configuration bit string defines and connects together. Thompson had set himself the task of evolving a digital circuit that could discriminate between an applied 1 kHz or 10 kHz applied signal [20,21]. He found that computer controlled evolution of the configuring bit strings could relatively easily solve this problem. However, when he analyzed the successful circuits he found to his surprise that they worked by utilizing subtle electrical properties of the silicon. Despite painstaking analysis and simulation work he was unable to explain how, or what property was being utilized. This lack of knowledge of how the system works, of course, prevents humans from designing systems that are intended to exploit these subtle and complex physical characteristics. However, it does not prevent exploitation through artificial evolution. Since then a number of researchers have demonstrated the viability of intrinsic evolution in silicon devices [21,22,23,24,25,26,27].

The term *evolution in materio* was first coined by Miller and Downing [28]. They argued that the lesson that should be drawn from the work of [21] is that evolution may be used to exploit the properties of a wider range of materials than silicon.

In summary, evolution in materio can be described as:

Exploitation, using an unconstrained evolutionary algorithm, of the non-linear properties of a malleable or programmable material to perform a desired function by altering its physical or electrical configuration.

Evolution in materio is a subset of a research field known as evolvable hardware. It aims to exploit properties of physical systems with much fewer preconditions and constraints than is usual, and it deliberately tries to avoid paying Conrad's 'Price of Programmability'. However, to get access to physically rich systems, we may have to discard devices designed with human programming in mind. Such devices are often based on abstract idealizations of processes occurring in the physical world. For example, FPGAs are considered as digital, but they are fundamentally analogue devices that have been constrained to behave in certain, human understandable ways. This means that intrinsically complex physical processes are carefully manipulated to represent extremely simple effects (e.g. a rapid switch from one voltage level to another). Unconstrained evolution, as demonstrated by Thompson, allows for the analogue properties of such devices to be effectively utilized.

We would expect physically rich systems to exhibit non-linear properties – they will be complex systems. This is because physical systems generally have huge numbers of parts interacting in complex ways. Arguably, humans have difficulty working with complex systems, and the use of evolution enables us to potentially overcome these limitations when dealing with such systems.

When systems are abstracted, the relationship to the physical world becomes more distant. This is highly convenient for human designers who do not wish to understand, or work with, hidden or subtle properties of materials. Exploitation through evolution reduces the need for abstraction, as it appears evolution is capable of discovering and utilizing any physical effects it can find. The aim of this new methodology in computation is to evolve special purpose computational processors. By directly exploiting physical systems and processes, one should be able to build extremely fast and efficient computational devices. It is our view that computer controlled evolution is a universal methodology for doing this. Of course, von Neumann machines (i.e. digital computers) are *individually* universal and this is precisely what confers their great utility in modern technology, however this universality comes at a price. They ignore the rich computational possibilities of materials and try to create operations that are close to a mathematical abstraction. Evolution in materio is a universal methodology for producing specific, highly tuned computational devices.

It is important not to underestimate the real practical difficulties associated with using an unconstrained design process. Firstly the evolved behavior of the material may be extremely sensitive to the specific properties of the material sample, so each piece would require individual train-

ing. Thompson originally experienced this difficulty, however in later work he showed that it was possible to evolve the configuration of FPGAs so that they produced reliable behavior in a variety of environmental conditions [29].

Secondly, the evolutionary algorithm may utilize physical aspects of any part of the training set-up. Both of these difficulties have already been experienced [21,23]. A third problem can be thought of as “the wiring problem”. The means to supply huge amounts of configuration data to a tiny sample. This problem is a very fundamental one. It suggests that if we wish to exploit the full physical richness of materials we might have to allow the material to grow its own wires and be self-wiring. This has profound implications for intrinsic evolution as artificial hardware evolution requires complete reconfigurability, this implies that one would have to be able to “wipe-clean” the evolved wiring and start again with a new artificial genotype. This might be possible by using nanoparticles that assemble into nanowires. These considerations bring us to an important issue in evolution in materio. Namely, the problem of choosing a suitable materials that can be exploited by computer controlled evolution.

Evolution in Materio: Defining Suitable Materials

The obvious characteristic required by a candidate material is the ability to reconfigure it in some way. Liquid crystal, clay, salt solutions etc can be readily configured either electrically or mechanically; their physical state can be adjusted, and readjusted, by applying a signal or force. In contrast (excluding its electrical properties) the physical properties of an FPGA would remain unchanged during configuration. It is also desirable to bulk configure the system. It would be infeasible to configure every molecule in the material, so the material should support the ability to be reconfigured over large areas using a small amount of configuration.

The material needs to perform some form of transformation (or computation) on incident signals that we apply. To do this, the material will have to interfere with the incident signal and perform a modification to it. We will need to be able to observe this modification, in order to extract the result of the computation. To perform a non-trivial computation, the material should be capable of performing complex operations upon the signal. Such capabilities would be maximized if the system exhibited non-linear behavior when interacting with input signals.

In summary, we can say that for a material to be useful to evolution in materio it should have the following properties:

- Modify incident signals in observable ways.
- The components of a system (i. e. the molecules within a material) interact with each other locally such that non-linear effects occur at either the local or global levels.
- It is possible to configure the state of the material locally.
- It is possible to observe the state of the material – either as a whole or in one or more locations.
- For practical reasons we can state that the material should be reconfigurable, and that changes in state should be temporary or reversible.

Miller and Downing [28] identified a number of physical systems that have some, if not all, of these desirable properties. They identified liquid crystal as the most promising in this regard as it digitally writable, reconfigurable and works at a molecular level. Most interestingly, it is an example of mesoscopic organization. Some people have argued that it is within such systems that emergent, organized behavior can occur [30]. Liquid crystals also exhibit the phenomenon of self-assembly. They form a class of substances that are being designed and developed in a field of chemistry called Supramolecular Chemistry [31]. This is a new and exciting branch of chemistry that can be characterized as ‘the designed chemistry of the intermolecular bond’. Supramolecular chemicals are in a permanent process of being assembled and disassembled. It is interesting to consider that conceptually liquid crystals appear to sit on the ‘edge of chaos’ [32] in that they are fluids (chaotic) that can be ordered, under certain circumstances.

Liquid Crystal

Liquid crystal (LC) is commonly defined as a substance that can exist in a mesomorphic state [33,34]. Mesomorphic states have a degree of molecular order that lies between that of a solid crystal (long-range positional and orientational) and a liquid, gas or amorphous solid (no long-range order). In LC there is long-range orientational order but no long-range positional order.

LC tends to be transparent in the visible and near infrared and quite absorptive in UV. There are three distinct types of LC: lyotropic, polymeric and thermotropic. Lyotropic LC is obtained when an appropriate amount of material is dissolved in a solvent. Most commonly this is formed by water and amphiphilic molecules: molecules with a hydrophobic part (water insoluble) and a hydrophilic part (strongly interacting with water). Polymeric LC is basically a polymer version of the aromatic LC discussed. They are characterized by high viscosity and include vinyls and Kevlar. Thermotropic LC (TLC) is the

most common form and is widely used. TLC exhibit various liquid crystalline phases as a function of temperature. They can be depicted as rod-like molecules and interact with each other in distinctive ordered structures. TLC exists in three main forms: nematic, cholesteric and smectic. In nematic LC the molecules are positionally arranged randomly but they all share a common alignment axis. Cholesteric LC (or chiral nematic) is like nematic however they have a chiral orientation. In smectic LC there is typically a layered positionally disordered structure. The three types A, B and C are defined as follows. In type A the molecules are oriented in alignment with the natural physical axes (i.e normal to the glass container), however in type C the common molecular axes of orientation is at an angle to the container. LC molecules typically are dipolar. Thus the organization of the molecular dipoles give another order of symmetry to the LC. Normally the dipoles would be randomly oriented. However in some forms the natural molecular dipoles are aligned with one another. This gives rise to ferroelectric and ferrielectric forms.

There is a vast range of different types of liquid crystal. LC of different types can be mixed. LC can be doped (as in Dye-Doped LC) to alter their light absorption characteristics. Dye-Doped LC film has been made that is optically addressable and can undergo very large changes in refractive index [35]. There are Polymer-Dispersed Liquid Crystals, which can have tailored, electrically controlled light refractive properties. Another interesting form of LC being actively investigated is Discotic LC. These have the form of disordered stacks (1-dimensional fluids) of disc-shaped molecules on a two-dimensional lattice. Although discotic LC is an electrical insulator, it can be made to conduct by doping with oxidants [36]. The oxidants are incorporated into the fluid hydrocarbon chain matrix (between disks). LC is widely known as useful in electronic displays, however, there are in fact, many non-display applications too. There are many applications of LC (especially ferroelectric LC) to electrically controlled light modulation: phase modulation, optical correlation, optical interconnects and switches, wavelength filters, optical neural networks. In the latter case a ferroelectric LC is used to encode the weights in a neural network [37].

Conducting and Electroactive Polymers

Conducting polymer composites have been made that rapidly change their microwave reflection coefficient when an electric field is applied. When the field is removed, the composite reverts to its original state. Experiments have shown that the composite can change from one state to the other in the order of 100 ms [38]. Also, some poly-

mers exhibit electrochromism. These substances change their reflectance when a voltage is applied. This can be reversed by a change in voltage polarity [39]. Electroactive polymers [40] are polymers that change their volume with the application of an electric field. They are particularly interesting as voltage controlled artificial muscle. Organic semiconductors also look promising especially when some damage is introduced. Further details of electronic properties of polymers and organic crystals can be found in [41].

Voltage Controlled Colloids

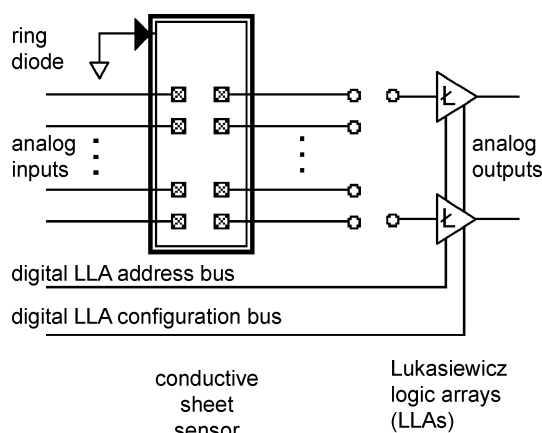
Colloids are suspensions of particles of sub-micron sizes in a liquid. The phase behavior of colloids is not fully understood. Simple colloids can self assemble into crystals, while multi-component suspensions can exhibit a rich variety of crystalline structures. There are also electrorheological fluids. These are suspensions of extremely fine non-conducting particles in an electrically insulating fluid. The viscosity of these fluids can be changed in a reversible way by large factors in response to an applied electric field in times of the order of milliseconds [42]. Also colloids can also be made in which the particles are charged making them easily manipulatable by suitable applied electric fields. Even if the particles are not charged they may be moved through the action of applied fields using a phenomenon known as dielectrophoresis which is the motion of polarized but electrically uncharged particles in nonuniform electric fields [43]. In work that echoes the methods of Pask nearly four decades ago, dielectrophoresis has been used to grow tiny gold wires through a process of self-assembly [44].

Langmuir–Blodgett Films

Langmuir–Blodgett films are molecular monolayers of organic material that can be transferred to a solid substrate [45]. They usually consist of hydrophilic heads and hydrophobic tails attached to the substrate. Multiple monolayers can be built and films can be built with very accurate and regular thicknesses. By arranging an electrode layer above the film it seems feasible that the local electronic properties of the layers could be altered. These systems look like feasible systems whose properties might be exploitable through computer controlled evolution of the voltages.

Kirchoff–Lukasiewicz Machines

Work by Mills [46,47] also demonstrates the use of materials in computation. He has designed an ‘Extended Analog Computer’ (EAC) that is a physical implementation of



Evolution in Materio, Figure 2
Kirchhoff–Lukasiewicz Machine

a Kirchhoff–Lukasiewicz Machine (KLM) [46]. The machines are composed of logical function units connected to a conductive media, typically a conductive polymer sheet. The logical units implement Lukasiewicz Logic – a type of multi-valued logic [47]. Figure 2 shows how the Lukasiewicz Logic Arrays (LLA) are connected to the conductive polymer. The LLA bridge areas of the sheet together. The logic units measure the current at one point, perform a transformation and then apply a current source to the other end of the bridge.

Computation is performed by applying current sinks and sources to the conductive polymer and reading the output from the LLAs. Different computations can be performed that are determined by the location of applied signals in the conducting sheet and the configuration of the LLAs. Hence, computation is performed by an interaction of the physics described by Kirchhoff's laws and the Lukasiewicz Logic units. Together they form a physical device that can solve certain kinds of partial differential equations. Using this form of analogue computation, a large number of these equations can be solved in nanoseconds – much faster than on a conventional computer. The speed of computation is dependent on materials used and how they are interfaced to digital computers, but it is expected that silicon implementations will be capable of finding tens of millions of solutions to the equations per second.

Examples of computation so far implemented in this system include robot control, control of a cyclotron beam [48], models of biological systems (including neural networks) [49] and radiosity based image rendering.

One of the most interesting feature of these devices is the programming method. It is very difficult to understand the actual processes used by the system to perform

computation, and until recently most of the reconfiguration has been done manually. This is difficult as the system is not amenable to traditional software development approaches. However, evolutionary algorithms can be used to automatically define the parameters of the LLAs and the placement of current sinks and sources. By defining a suitable fitness function, the configuration of the EAC can be evolved – which removes the need for human interaction and for knowledge of the underlying system.

Although it is clear that such KLMs are clearly using the physical properties of a material to perform computation, the physical state of the material is not reconfigured (i. e., programmed), only the currents in the sheet are changed.

Evolution in Materio Is Verified with Liquid Crystal

Harding [50] has verified Miller's intuition about the suitability of liquid crystal as an evolvable material by demonstrating that it is relatively easy to configure liquid crystal to perform various forms of computation.

In 2004, Harding constructed an analogue processor that utilizes the physical properties of liquid crystal for computation. He evolved the configuration of the liquid crystal to discriminate between two square waves of many different frequencies. This demonstrated, for the first time, that the principle of using computer-controlled evolution was a viable and powerful technique for using non-silicon materials for computation. The analogue processor consists of a passive liquid crystal display mounted on a reconfigurable circuit, known as an evolvable motherboard. The motherboard allows signals and configuration voltages to be routed to physical locations in the liquid crystal.

Harding has shown that many different devices can be evolved in liquid crystal including:

- **Tone discriminator.** A device was evolved in liquid crystal that could differentiate many different frequencies of square wave. The results were competitive, if not superior to those evolved in the FPGA.
- **Logic gates.** A variety of two input logic gates were evolved, showing that liquid crystal could behave in a digital fashion. This indicates that liquid crystal is capable of universal computation.
- **Robot controller.** An obstacle avoidance system for a simple exploratory robot was evolved. The results were highly competitive, with solutions taking fewer evaluations to find compared to other work on evolved robot controllers.

One of the surprising findings in this work has been that it turns out to be relatively easy to evolve the configura-

tion of liquid crystal to solve tasks; i. e., only 40 generations of a modest population of configurations are required to evolve a very good frequency discriminator, compared to the thousands of generations required to evolve a similar circuit on an FPGA. This work has shown that evolving such devices in liquid crystal is easier than when using conventional components, such as FPGAs. The work is a clear demonstration that evolutionary design can produce solutions that are beyond the scope of human design.

Evolution in Materio Using Liquid Crystal: Implementational Details

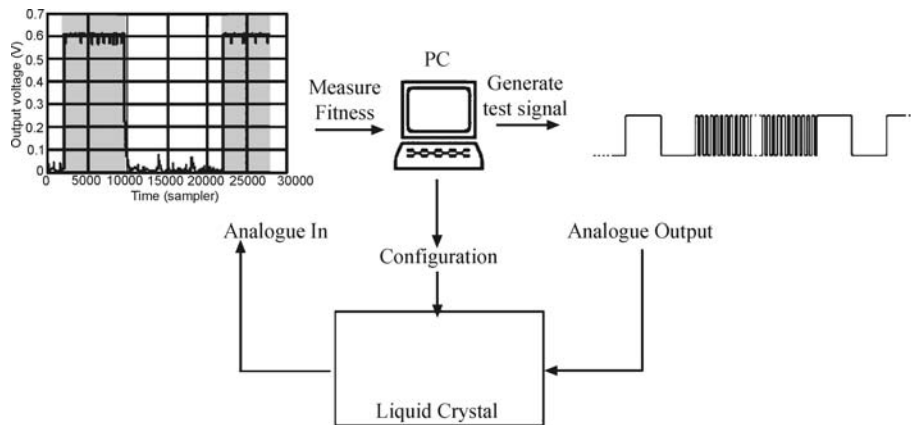
An evolvable motherboard (EM) [23] is a circuit that can be used to investigate intrinsic evolution. The EM is a reconfigurable circuit that rewires a circuit under computer control. Previous EMs have been used to evolve circuits containing electronic components [23,51] – however they can also be used to evolve in materio by replacing the standard components with a candidate material.

An EM is connected to an Evolvatron. This is essentially a PC that is used to control the evolutionary processes. The Evolvatron also has digital and analog I/O, and can be used to provide test signals and record the response of the material under evolution.

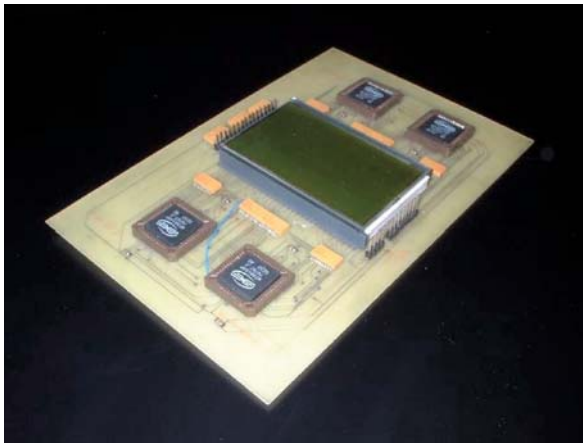
The Liquid Crystal Evolvable Motherboard (LCEM) is a circuit that uses four cross-switch matrix devices to dynamically configure the circuits connecting to the liquid crystal. The switches are used to wire the 64 connections on the LCD to one of 8 external connections. The external connections are: input voltages, grounding, signals and connections to measurement devices. Each of the external connectors can be wired to any of the connections to the LCD.

The external connections of the LCEM are connected to the Evolvatron's analogue inputs and outputs. One connection was assigned for the incident signal, one for measurement and the other for fixed voltages. The value of the fixed voltages is determined by the evolutionary algorithm, but is constant throughout each evaluation.

In these experiments the liquid crystal glass sandwich was removed from the display controller it was originally mounted on, and placed on the LCEM. The display has a large number of connections (in excess of 200), however because of PCB manufacturing constraints we are limited in the size of connection we can make, and hence the number of connections. The LCD is therefore roughly positioned over the pads on the PCB, with many of the PCB pads touching more than one of the connectors on the LCD. This means that we are applying configuration voltages to several areas of LC at the same time.



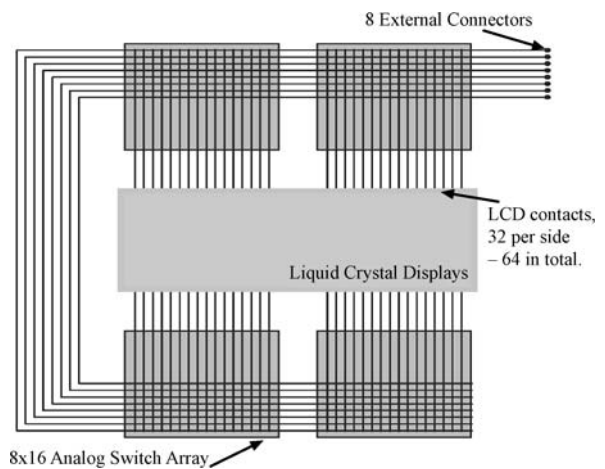
Evolution in Materio, Figure 3
Equipment configuration



Evolution in Materio, Figure 4
The LCEM

Unfortunately neither the internal structure nor the electrical characteristics of the LCD are known. This raises the possibility that a configuration may be applied that would damage the device. The wires inside the LCD are made of an extremely thin material that could easily be burnt out if too much current flows through them. To guard against this, each connection to the LCD is made through a 4.7 Kohm resistor in order to provide protection against short circuits and to help limit the current in the LCD. The current supplied to the LCD is limited to 100 mA. The software controlling the evolution is also responsible for avoiding configurations that may endanger the device (such as short circuits).

It is important to note that other than the control circuitry for the switch arrays there are no other active components on the motherboard – only analog switches, smoothing capacitors, resistors and the LCD are present.



Evolution in Materio, Figure 5
Schematic of LCEM

Stability and Repeatability Issues

When the liquid crystal display is observed while solving a problem it is seen that some regions of the liquid display go dark indicating that the local molecular direction has been changed. This means that the configuration of the liquid crystal is changing while signals are being applied. To draw an analogy with circuit design, the incident signals would be changing component values or changing the circuit topology, which would have an affect on the behavior of the system. This is likely to be detrimental to the measured performance of the circuit. When a solution is evolved, the fitness function automatically measures its stability over the period of the evaluation. Changes made by the incident signals can be considered part of the genotype-phenotype mapping. Solutions that cannot cope with

their initial configurations being altered will achieve a low score. However, the fitness function cannot measure the behavior beyond the end of the evaluation time. Therein lies the difficulty, in evolution in materio long term stability cannot be guaranteed.

Another issue concerns repeatability. When a configuration is applied to the liquid crystal the molecules are unlikely go back to exactly where they were when this configuration was tried previously. Assuming, that there is a strong correlation between genotype and phenotype, then it is likely that evolution will cope with this extra noise. However, if evolved devices are to be useful one needs to be sure that previously evolved devices will function in the same way as they did when originally evolved.

In [27] it is noted that the behavior of circuits evolved intrinsically can be influenced by previous configurations – therefore their behavior (and hence fitness) is dependent not only on the currently evaluated individual's configuration but on those that came before. It is worth noting that this is precisely what happens in natural evolution. For example, in a circuit, capacitors may still hold charge from a previously tested circuit. This charge would then affect the circuits operation, however if the circuit was tested again with no stored charge a different behavior would be expected and a different fitness score would be obtained. Not only does this affect the ability to evolve circuits, but would mean that some circuits are not valid. Without the influence of the previously evaluated circuits the current solution may not function as expected. It is expected that such problems will have analogies in evolution in materio. The configurations are likely to be highly sensitive to initial conditions (i. e. conditions introduced by previous configurations).

Dealing with Environmental Issues

A major problem when working with intrinsic evolution is separating out the computation allegedly being carried out by the target device, and that actually done by the material being used. For example, whilst trying to evolve an oscillator Bird and Layzell discovered that evolution was using part of the circuit for a radio antenna, and picking up emissions from the environment [22]. Layzell also found that evolved circuits were sensitive to whether or not a soldering iron was plugged in (not even switched on) in another part of the room [23]!

An evolved device is not useful if it is highly sensitive to its environment in unpredictable ways, and it will not always be clear what environmental effects the system is

using. It would be unfortunate to evolve a device for use in a space craft, only to find out it fails to work once out of range of a local radio tower!

To minimize these risks, we will need to check the operation of evolved systems under different conditions. We will need to test the behavior of a device using a different set up in a different location. It will be important to know if a particular configuration only works with one particular sample of a given material.

The Computational Power of Materials

In [52], Lloyd argued that the theoretical computing power of a kilogram of material is far more than is possible with a kilogram of traditional computer. He notes that computers are subject to the laws of physics, and that these laws place limits on the maximum speed they can operate and the amount of information it can process. Lloyd shows that if we were fully able to exploit a material, we would get an enormous increase in computing power. For example, with 1 kg of matter we should be able to perform roughly 5×10^{50} operations per second, and store 10^{31} bits. Amazingly, contemporary quantum computers do operate near these theoretical limits [52].

A small amount of material also contains a large number of components (regardless of whether we consider the molecular or atomic scale). This leads to some interesting thoughts. If we can exploit materials at this level, we would be able to do a vast amount of computation in a small volume. A small size also hints at low power consumption, as less energy has to be spent to perform an operation. Many components also provide a mechanism for reliability through redundancy. A particularly interesting observation, especially when considered in terms of non Von-Neumann computation, is the massive parallelism we may be able to achieve. The reason that systems such as quantum, DNA and chemical computation can operate so quickly is that many operations are performed at the same time. A programmable material might be capable of performing vast numbers of tasks simultaneously, and therefore provide a computational advantage.

In commercial terms, small is often synonymous with low cost. It may be possible to construct devices using cheaply available materials. Reliability may not be an issue, as the systems could be evolved to be massively fault tolerant using their intrinsic redundancy. Evolution is capable of producing novel designs. Koza has already rediscovered circuits that infringe on recent patents, and his genetic programming method has 'invented' brand new circuit designs [53]. Evolving in materio could produce many novel designs, and indeed given the infancy of pro-

grammable materials all designs may be unique and hence patentable.

Future Directions

The work described here concerning liquid crystal computational devices is at an early stage. We have merely demonstrated that it is possible to evolve configurations of voltages that allow a material to perform desired computations. Any application that ensues from this work is unlikely to be a replacement for a simple electronic circuit. We can design and build those very successfully. What we have difficulty with is building complex, fault tolerant systems for performing complex computation. It appears that nature managed to do this. It used a simple process of a repetitive test and modify, and it did this in a universe of unimaginable physical complexity. If nature can exploit the physical properties of a material and its surroundings through evolution, then so should we.

There are many important issues that remain to be addressed. Although we have made some suggestions about materials worthy of investigation, it is at present unclear which materials are most suitable. An experimental platform needs to be constructed that allows many materials to be tried and investigated. The use of microelectrode arrays in a small volume container would allow this. This would also have the virtue of allowing internal signals in the materials to be inspected and potentially understood.

We need materials that are rapidly configurable. They must not be fragile and sensitive to minute changes in physical setup. They must be capable of maintaining themselves in a stable configuration. The materials should be complex and allow us to carry out difficult computations more easily than conventional means. One would like materials that can be packaged into small volumes. The materials should be relatively easily interfaced with. So far, material systems have been configured by applying a constant configuration pattern, however this may not be appropriate for all systems. It may be necessary to put the physical system under some form of responsive control, in order to program and then keep the behavior stable.

We may or may not know if a particular material can be used to perform some form of computation. However, we can treat our material as a “black box”, and using evolution as a search technique, automatically discover what, if any, computations our black box can perform. The first step is to build an interface that will allow us to communicate with a material. Then we will use evolution to find a configuration we can apply using this platform, and then attempt to find a mapping from a given problem to an input suitable for that material, and a mapping from the ma-

terials response to an output. If this is done correctly, we might be automatically able to tell if a material can perform computation, and then classify the computation.

When we evolve in materio, using mappings evolved in software, how can we tell when the material is giving us any real benefit? The lesson of evolution in materio has been that the evolved systems can be very difficult to analyze, and the principal obstacle to the analysis is the problem of separating out the computational role that each component plays in the evolved system. These issues are by no means just a problem for evolution in materio. They may be an inherent part of complex evolved systems. Certainly the understanding of biological systems are providing immense challenges to scientists.

The single most important aspect that suggests that evolution in materio has a future is that natural evolution has produced immensely sophisticated material computational systems. It would seem foolish to ignore this and merely try to construct computational devices that operate according to one paradigm of computation (i. e. Turing). Oddly enough, it is precisely the sophistication of the latter that allows us to attempt the former.

Bibliography

Primary Literature

1. Turing AM (1936) On computable numbers, with an application to the entscheidungsproblem. *Proc Lond Math Soc* 42(2):230–265
2. Bissell C (2004) A great disappearing act: the electronic analogue computer. In: *IEEE Conference on the History of Electronics*, 28–30 June
3. Deutsch D (1985) Quantum theory, the church–turing principle and the universal quantum computer. *Proc Royal Soc Lond A* 400:97–117
4. Adamatzky A, Costello BDL, Asai T (2005) *Reaction-Diffusion Computers*. Elsevier, Amsterdam
5. Adleman LM (1994) Molecular computation of solutions to combinatorial problems. *Science* 266(11):1021–1024
6. Amos M (2005) *Theoretical and Experimental DNA Computation*. Springer, Berlin
7. Weiss R, Basu S, Hooshangi S, Kalmbach A, Karig D, Mehreja R, Netravali I (2003) Genetic circuit building blocks for cellular computation, communications, and signal processing. *Nat Comput* 2(1):47–84
8. UK Computing Research Committee (2005) Grand challenges in computer research. http://www.ukcrc.org.uk/grand_challenges/
9. Stepney S, Braunstein SL, Clark JA, Tyrrell A, Adamatzky A, Smith RE, Addis T, Johnson C, Timmis J, Welch P, Milner R, Partridge D (2005) Journeys in non-classical computation I: A grand challenge for computing research. *Int J Parallel Emerg Distrib Syst* 20(1):5–19
10. Stepney S, Braunstein S, Clark J, Tyrrell A, Adamatzky A, Smith R, Addis T, Johnson C, Timmis J, Welch P, Milner R, Partridge D

- (2006) Journeys in non-classical computation II: Initial journeys and waypoints. *Int J Parallel, Emerg Distrib Syst* 21(2):97–125
11. Toffoli T (2005) Nothing makes sense in computing except in the light of evolution. *Int J Unconv Comput* 1(1):3–29
 12. Conrad M (1988) The price of programmability. *The Universal Turing Machine*. pp 285–307
 13. Goldberg D (1989) *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley, Reading, Massachusetts
 14. Holland J (1992) *Adaptation in Natural and Artificial Systems*. 2nd edn. MIT Press, Cambridge
 15. Mitchell M (1996) *An introduction to genetic algorithms*. MIT Press, Cambridge, MA, USA
 16. Pask G (1958) Physical analogues to the growth of a concept. In: *Mechanization of Thought Processes*. Symposium 10, National Physical Laboratory, pp 765–794
 17. Pask G (1959) The natural history of networks. In: *Proceedings of International Tracts*. In: *Computer Science and Technology and their Application*. vol 2, pp 232–263
 18. Cariani P (1993) To evolve an ear: epistemological implications of gordon pask's electrochemical devices. *Syst Res* 3:19–33
 19. Pickering A (2002) Cybernetics and the mangle: Ashby, beer and pask. *Soc Stud Sci* 32:413–437
 20. Thompson A, Harvey I, Husbands P (1996) Unconstrained evolution and hard consequences. In: Sanchez E, Tomassini M (eds) *Towards Evolvable Hardware: The evolutionary engineering approach*. LNCS, vol 1062. Springer, Berlin, pp 136–165
 21. Thompson A (1996) An evolved circuit, intrinsic in silicon, entwined with physics. *ICES* 390–405
 22. Bird J, Layzell P (2002) The evolved radio and its implications for modelling the evolution of novel sensors. In: *Proceedings of Congress on Evolutionary Computation*, pp 1836–1841
 23. Layzell P (1998) A new research tool for intrinsic hardware evolution. *Proceedings of The Second International Conference on Evolvable Systems: From Biology to Hardware*. LNCS, vol 1478. Springer, Berlin, pp 47–56
 24. Linden DS, Altschuler EE (2001) A system for evolving antennas in-situ. In: *3rd NASA / DoD Workshop on Evolvable Hardware*. IEEE Computer Society, pp 249–255
 25. Linden DS, Altschuler EE (1999) Evolving wire antennas using genetic algorithms: A review. In: *1st NASA / DoD Workshop on Evolvable Hardware*. IEEE Computer Society, pp 225–232
 26. Stoica A, Zebulum RS, Guo X, Keymeulen D, Ferguson MI, Duong V (2003) Silicon validation of evolution-designed circuits. In: *Proceedings. NASA/DoD Conference on Evolvable Hardware*, pp 21–25
 27. Stoica A, Zebulum RS, Keymeulen D (2000) Mixtrinsic evolution. In: *Proceedings of the Third International Conference on Evolvable Systems: From Biology to Hardware (ICES2000)*. Lecture Notes in Computer Science, vol 1801. Springer, Berlin, pp 208–217
 28. Miller JF, Downing K (2002) Evolution in materio: Looking beyond the silicon box. In: *Proceedings of NASA/DoD Evolvable Hardware Workshop*, pp 167–176
 29. Thompson A (1998) On the automatic design of robust electronics through artificial evolution. In: Sipper M, Mange D, Pérez-Uribe A (eds) *Evolvable Systems: From Biology to Hardware*, vol 1478. Springer, New York, pp 13–24
 30. Laughlin RB, Pines D, Schmalian J, Stojkovic BP, Wolynes P (2000) The middle way. *Proc Natl Acad Sci* 97(1):32–37
 31. Lindoy LF, Atkinson IM (2000) *Self-assembly in Supramolecular Systems*. Royal Society of Chemistry
 32. Langton C (1991) Computation at the edge of chaos: Phase transitions and emergent computation. In: *Emergent Computation*, pp 12–37. MIT Press
 33. Demus D, Goodby JW, Gray GW, Spiess HW, Villi V (eds) (1998) *Handbook of Liquid Crystals*, vol 4. Wiley-VCH, ISBN 3-527-29502-X, pp 2180
 34. Khoo IC (1995) *Liquid Crystals: physical properties and nonlinear optical phenomena*. Wiley
 35. Khoo IC, Slussarenko S, Guenther BD, Shih MY, Chen P, Wood WV (1998) Optically induced space-charge fields, dc voltage, and extraordinarily large nonlinearity in dye-doped nematic liquid crystals. *Opt Lett* 23(4):253–255
 36. Chandrasekhar S (1998) Columnar, discotic nematic and lamellar liquid crystals: Their structure and physical properties. In: *Handbook of Liquid Crystals*, vol 2B. Wiley-VCH pp 749–780
 37. Crossland WA, Wilkinson TD (1998) Nondisplay applications of liquid crystals. In: *Handbook of Liquid Crystals*, vol 1. Wiley-VCH, pp 763–822
 38. Wright PV, Chambers B, Barnes A, Lees K, Despotakis A (2000) Progress in smart microwave materials and structures. *Smart Mater Struct* 9:272–279
 39. Mortimer RJ (1997) Electrochromic materials. *Chem Soc Rev* 26:147–156
 40. Bar-Cohen Y (2001) *Electroactive Polymer (EAP) Actuators as Artificial Muscles – Reality, Potential and Challenges*. SPIE Press
 41. Pope M, Swenberg CE (1999) *Electronic Processes of Organic Crystals and Polymers*. Oxford University Press, Oxford
 42. Hao T (2005) *Electrorheological Fluids: The Non-aqueous Suspensions*. Elsevier Science
 43. Khusid B, Activos A (1996) Effects of interparticle electric interactions on dielectrophoresis in colloidal suspensions. *Phys Rev E* 54(5):5428–5435
 44. Khusid B, Activos A (2001) Hermanson KD, Lumsdon SO, Williams JP, Kaler EW, Velev OD. *Science* 294:1082–1086
 45. Petty MC (1996) *Langmuir–Blodgett Films: An Introduction*. Cambridge University Press, Cambridge
 46. Mills JW (1995) *Polymer processors*. Technical Report TR580, Department of Computer Science, University of Indiana
 47. Mills JW, Beavers MG, Daffinger CA (1989) Lukasiewicz logic arrays. Technical Report TR296, Department of Computer Science, University of Indiana
 48. Mills JW (1995) Programmable vlsi extended analog computer for cyclotron beam control. Technical Report TR441, Department of Computer Science, University of Indiana
 49. Mills JW (1995) The continuous retina: Image processing with a single sensor artificial neural field network. Technical Report TR443, Department of Computer Science, University of Indiana
 50. Harding S, Miller JF (2004) Evolution in materio: A tone discriminator in liquid crystal. In: *Proceedings of the Congress on Evolutionary Computation 2004 (CEC'2004)*, vol 2, pp 1800–1807
 51. Crooks J (2002) *Evolvable analogue hardware*. Meng project report, The University Of York
 52. Lloyd S (2000) Ultimate physical limits to computation. *Nature* 406:1047–1054
 53. Koza JR (1999) Human-competitive machine intelligence by means of genetic algorithms. In: Booker L, Forrest S, Mitchell M, Riolo R (eds) *Festschrift in honor of John H Holland*. Center for the Study of Complex Systems, Ann Arbor, pp 15–22

Books and Reviews

- Analog Computer Museum and History Center: Analog Computer Reading List. <http://dcoward.best.vwh.net/analog/readlist.htm>
- Bringsjord S (2001) In Computation, Parallel is Nothing, Physical Everything. *Minds and Machines* 11(1)
- Feynman RP (2000) Feynman Lectures on Computation. Perseus Books Group
- Fifer S (1961) *Analogue computation: theory, techniques, and applications*. McGraw-Hill, New York
- Greenwood GW, Tyrrell AM (2006) *Introduction to Evolvable Hardware: A Practical Guide for Designing Self-Adaptive Systems*. Wiley-IEEE Press
- Hey AJG (ed) (2002) *Feynman and Computation*. Westview Press
- Penrose R (1989) *The Emperor's New Mind, Concerning Computers, Minds, and the Laws of Physics*. Oxford University, Oxford
- Piccinini G: The Physical Church-Turing Thesis: Modest or Bold? <http://www.umsl.edu/~piccinini/CTModestorBold5.htm>
- Raichman N, Ben-Jacob N, Segev R (2003) Evolvable Hardware: Genetic Search in a Physical Realm. *Phys A* 326:265–285
- Sekanina L (2004) *Evolvable Components: From Theory to Hardware Implementations*, 1st edn. Springer, Heidelberg
- Siegelmann HT (1999) *Neural Networks and Analog Computation, Beyond the Turing Limits*. Birkhauser, Boston
- Sienko T, Adamatzky A, Rambidi N, Conrad M (2003) *Molecular Computing*. MIT Press
- Thompson A (1999) *Hardware Evolution: Automatic Design of Electronic Circuits in Reconfigurable Hardware by Artificial Evolution*, 1st edn. Springer, Heidelberg

Evolving Cellular Automata

MARTIN CENEK, MELANIE MITCHELL
Computer Science Department,
Portland State University, Portland, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Cellular Automata
Computation in CAs
Evolving Cellular Automata with Genetic Algorithms
Previous Work on Evolving CAs
Coevolution
Other Applications
Future Directions
Acknowledgments
Bibliography

Glossary

Cellular automaton (CA) Discrete-space and discrete-time spatially extended lattice of cells connected in a regular pattern. Each cell stores its state and a state-

transition function. At each time step, each cell applies the transition function to update its state based on its local neighborhood of cell states. The update of the system is performed in synchronous steps – i. e., all cells update simultaneously.

Cellular programming A variation of genetic algorithms designed to simultaneously evolve state transition rules and local neighborhood connection topologies for non-homogeneous cellular automata.

Coevolution An extension to the genetic algorithm in which candidate solutions and their “environment” (typically test cases) are evolved simultaneously.

Density classification A computational task for binary CAs: the desired behavior for the CA is to iterate to an all-1s configuration if the initial configuration has a majority of cells in state 1, and to an all-0s configuration otherwise.

Genetic algorithm (GA) A stochastic search method inspired by the Darwinian model of evolution. A population of candidate solutions is evolved by reproduction with variation, followed by selection, for a number of generations.

Genetic programming A variation of genetic algorithms that evolves genetic trees.

Genetic tree Tree-like representation of a transition function, used by genetic programming algorithm.

Lookup table (LUT) Fixed-length table representation of a transition function.

Neighborhood Pattern of connectivity specifying to which other cells each cell is connected.

Non-homogeneous cellular automaton A CA in which each cell can have its own distinct transition function and local neighborhood connection pattern.

Ordering A computational task for one-dimensional binary CAs with fixed boundaries: The desired behavior is for the CA to iterate to a final configuration in which all initial 0 states migrate to the left-hand side of the lattice and all initial 1 states migrate to the right-hand side of the lattice.

Particle Periodic, temporally coherent boundary between two regular domains in a set of successive CA configurations. Particles can be interpreted as carrying information about the neighboring domains. Collisions between particles can be interpreted as the processing of information, with the resulting information carried by new particles formed by the collision.

Regular domain Region defined by a set of successive CA configurations that can be described by a simple regular language.

Synchronization A computational task for binary CAs: the desired behavior for the CA is to iterate to a tem-

poral oscillation between two configurations: all cells have state 1 and all cells have state 0s.

Transition function Maps a local neighborhood of cell states to an update state for the center cell of that neighborhood.

Definition of the Subject

Evolving cellular automata refers to the application of evolutionary computation methods to evolve cellular automata transition rules. This has been used as one approach to automatically “programming” cellular automata to perform desired computations, and as an approach to model the evolution of collective behavior in complex systems.

Introduction

In recent years, the theory and application of cellular automata (CAs) has experienced a renaissance, due to advances in the related fields of reconfigurable hardware, sensor networks, and molecular-scale computing systems. In particular, architectures similar to CAs can be used to construct physical devices such as field configurable gate arrays for electronics, networks of robots for environmental sensing and nano-devices embedded in interconnect fabric used for fault tolerant nanoscale computing. Such devices consist of networks of simple components that communicate locally without centralized control. Two major areas of research on such networks are (1) *programming* – how to construct and configure the locally connected components such that they will collectively perform a desired task; and (2) *computation theory* – what types of tasks are such networks able to perform efficiently, and how does the configuration of components affect the computational capability of these networks?

This article describes research into one particular automatic programming method: the use of genetic algorithms (GAs) to evolve cellular automata to perform desired tasks. We survey some of the leading approaches to evolving CAs with GAs, and discuss some of the open problems in this area.

Cellular Automata

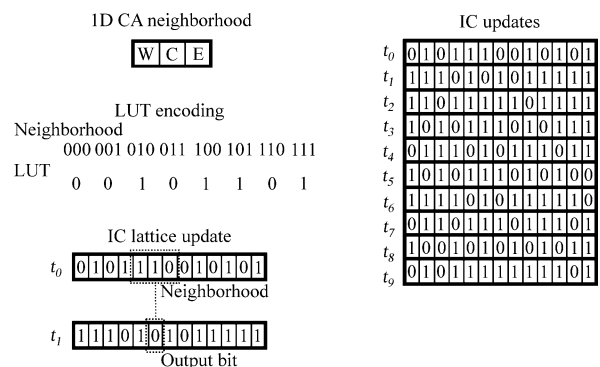
A *cellular automaton* (CA) is a spatially extended lattice of locally connected simple processors (cells). CAs can be used both to model physical systems and to perform parallel distributed computations.

In a CA, each cell maintains a discrete state and a transition function that maps the cell’s current state to its next state. This function is often represented as a lookup table

(LUT). The LUT stores all possible configurations of a cell’s local neighborhood, which consists of its own current state and the state of its neighboring cells. Change of state is performed in discrete time steps: the entire lattice is updated synchronously. There are many possible definitions of a *neighborhood*, but here we will define a neighborhood as the cell to be updated and the cells adjacent to it at a distance of radius r . The number of entries in the LUT will be s^N , where s is the number of possible states and N is the total number of cells in the neighborhood: $(2r + 1)^d$ for a square shaped neighborhood in a d -dimensional lattice, also known as a *Moore neighborhood*. CAs typically are given *periodic boundary conditions*, which treat the lattice as a torus.

To transform a cell’s state, the values of the cell’s state and those of its neighbors are encoded as a lookup index to the LUT that stores a value representing the cell’s new state (Fig. 1: left) [8,16,59]. For the scope of this article, we will focus on homogeneous binary CAs, which means that all cells in the CAs have the same LUT and each cell has one of two possible states, $s \in \{0, 1\}$. Figure 1 shows the mechanism of updates in a homogeneous one-dimensional two-state CA with a neighborhood radius $r = 1$.

CAs were invented in the 1940s by Stanislaw Ulam and John von Neumann. Ulam used CAs as a mathematical abstraction to study the growth of crystals, and von Neumann used them as an abstraction of a physical system with the concepts of a cell, state and transition function in order to study the logic of self-reproducing sys-



Evolving Cellular Automata, Figure 1

Left top: A one-dimensional neighborhood of three cells (radius 1): Center cell, West neighbor, and East neighbor. **Left middle:** A sample look-up table in which all possible neighborhood configurations are listed, along with the update state for the center cell in each neighborhood. **Left bottom:** Mechanism of update in a one dimensional binary CA of length 13: t_0 is the initial configuration at time 0 and t_1 is the initial configuration at next time step. **Right:** The sequence of synchronous updates starting at the initial state t_0 and ending at state t_9

tems [8,11,55]. Von Neumann's seminal work on CAs had great significance. Science after the industrial revolution was primarily concerned with energy, force and motion, but the concept of CAs shifted the focus to information processing, organization, programming, and most importantly, control [8]. The universal computational ability of CAs was realized early on, but harnessing this power continues to intrigue scientists [8,11,32,55].

Computation in CAs

In the early 1970s John Conway published a description of his deceptively simple Game of Life CA [18]. Conway proved that the Game of Life, like von Neumann's self-reproducing automaton, has the power of a universal Turing machine: any program that can be run on a Turing machine can be simulated by the Game of Life with the appropriate initial configuration of states. This initial configuration (IC) encodes both the input and the program to be run on that input. It is interesting that so simple a CA as the Game of Life (as well as even simpler CAs – see chapter 11 in [60]) has the power of a universal computer. However, the actual application of CAs as universal computers is, in general, impractical due to the difficulty of encoding a given program and input as an IC, as well as very long simulation times.

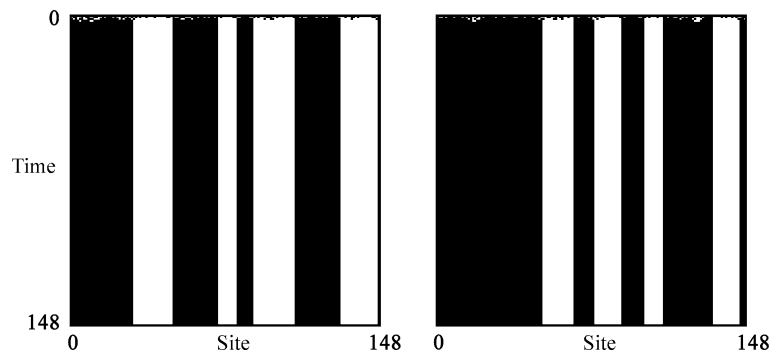
An alternative use of CAs as computers is to design a CA to perform a particular computational task. In this case, the initial configuration is the input to the program, the transition function corresponds to the program performing the specific task, and some set of final configurations is interpreted as the output of the computation. The intermediate configurations comprise the actual computation being done.

Examples of tasks for which CAs have been designed include location management in mobile computing networks [50], classification of initial configuration densities [38], pseudo-random number generation [51], multi-agent synchronization [47], image processing [26], simulation of growth patterns of material microstructures [5], chemical reactions [35], and pedestrian dynamics [45].

The problem of designing a CA to perform a task requires defining a cell's local neighborhood and boundary conditions, and constructing a transition function for cells that produces the desired input-output mapping. Given a CA's states, neighborhood radius, boundary conditions, and initial configuration, it is the LUT values that must be set by the "programmer" so that the computation will be performed correctly over all inputs.

In order to study the application of genetic algorithms to designing CAs, substantial experimentation has been done using the *density classification* (or *majority classification*) task. Here, "density" refers to the fraction of 1s in the initial configuration. In this task, a binary-state CA must iterate to an all-1s configuration if the initial configuration has a majority of cells in state 1, and iterate to an all-0s configuration otherwise. The maximum time allowed for completing this computation is a function of the lattice size.

One "naïve" solution for designing the LUT for this task would be *local majority voting*: set the output bit to 1 for all neighborhood configurations with a majority of 1s, and 0 otherwise. Figure 2 gives two space-time diagrams illustrating the behavior of this LUT in a one-dimensional binary CA with $N = 149$, and $r = 3$, where N denotes the number of cells in the lattice, and r is the neighborhood radius.



Evolving Cellular Automata, Figure 2

Two space-time diagrams illustrating the behavior of the "naïve" local majority voting rule, with lattice size $N = 149$, neighborhood radius $r = 3$, and number of time steps $M = 149$. *Left*: initial configuration has a majority of 0s. *Right*: initial configuration has a majority of 1s. Individual cells are colored black for state 1 and white for state 0. (Reprinted from [37] with permission of the author.)

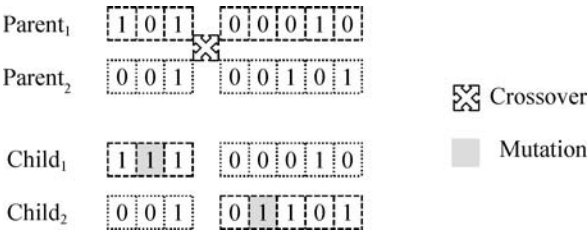
Each diagram shows an initial configuration of 149 cells (horizontal) iterating over 149 time steps (vertical, down the page). The left-hand diagram has an initial configuration with a majority of 0 (white) cells, and the right-hand diagram has an initial configuration with a majority of 1 (black) cells. In neither case does the CA produce the “correct” global behavior: an all-0s configuration for the left diagram and an all-1s configuration for the right diagram. This illustrates the general fact that human intuition often fails when trying to capture emergent collective behavior by manipulating individual bits in the lookup table that reflect the settings of the local neighborhood.

Evolving Cellular Automata with Genetic Algorithms

Genetic Algorithms (GAs) are a group of stochastic search algorithms, inspired by the Darwinian model of evolution, that have proved successful for solving various difficult problems [3,4,36].

A GA works as follows: (1) A population of individuals (“chromosomes”) representing candidate solutions to a given problem is initially generated at random. (2) The fitness of each individual is calculated as a function of its quality as a solution. (3) The fittest individuals are then selected to be the parents of a new generation of candidate solutions. Offspring are created from parents via copying, random mutation, and crossover. Once a new generation of individuals is created, the process returns to step two. This entire process is iterated for some number of generations, and the result is (hopefully) one or more highly fit individuals that are good solutions to the given problem.

GAs have been used by a number of groups to evolve LUTs for binary CAs [2,10,14,15,34,43,47,51]. The individuals in the GA population are LUTs, typically encoded as binary strings. Figure 3 shows a mechanism of encoding LUTs from a particular neighborhood configuration. For



Evolving Cellular Automata, Figure 4
Reproduction applied to *Parent*₁ and *Parent*₂ producing *Child*₁ and *Child*₂. The one-point crossover is performed at a randomly selected crossover point (bit 3) and a mutation is performed on bits 2 and 5 in *Child*₁ and *Child*₂ respectively

example: the decimal value for the neighborhood 11010 is 26. The updated value for the neighborhood’s center cell 11010 is retrieved from the 26th position in the LUT, updating cell’s value to 1.

The fitness of a LUT is a measure of how well the corresponding CA performs a given task after a fixed number of time steps, starting from a number of *test* initial configurations. For example, given the density classification task, the fitness of a LUT is calculated by running the corresponding CA on some number *k* of random initial configurations, and returning the fraction of those *k* on which the CA produces the correct final configuration (all 1s for initial configurations with majority 1s, all 0s otherwise). The set of random test ICs is typically regenerated at each generation.

For LUTs represented as bit strings, crossover is applied to two parents by randomly selecting a crossover point, so that each child inherits one segment of bits from each parent. Next, each child is subject to a mutation, where the genome’s individual bits are subject to a bit complement with a very low probability. An example of the reproduction process is illustrated in Fig. 4 for a lookup

1D CA neighborhood r=2



Neighborhood values:

00000	00001	00010	00011	00100	00101	11010	11011	11100	11101	11110	11111
LUT Offset: 0	1	2	3	4	5	26	27	28	29	30	31

LUT (length $2^{(2r+1)}$):

0	1	0	0	1	1	1	1	0	0	1	1
---	---	---	---	---	---	-----	-----	---	---	---	---	---	---

Evolving Cellular Automata, Figure 3

Lookup table encoding for 1D CA with neighborhood $r = 2$. All permutations of neighborhood values are encoded as an offset to the LUT. The LUT bit represents a new value for the center cell of the neighborhood. The binary string (LUT) encodes an individual’s chromosome used by evolution

table representation of $r = 1$. Here, one of two children is chosen for survival at random and placed in an offspring population. This process is repeated until the offspring population is filled. Before a new evolutionary cycle begins, the newly created population of offspring replaces the previous population of parents.

Previous Work on Evolving CAs

Von Neumann's self-reproducing automaton was the first construction that showed that CAs can perform universal computation [55], meaning that the CAs are capable, in principle, of performing any desired computation. However, in general it was unknown how to effectively "program" CAs to perform computations or what information processing dynamics CAs could best use to accomplish a task. In the 1980s and 1990s, a number of researchers attempted to determine how the generic dynamical behavior of a CA might be related to its ability to perform computations [18,19,21,33,58]. In particular, Langton defined a parameter on CA LUTs, λ , that he claimed correlated with computational ability. In Langton's work, λ is a function of the state-update values in the LUT; for binary CAs, λ is defined as the fraction of 1s in the state-update values.

Computation at the Edge of Chaos

Packard [40] was the first to use a genetic algorithm to evolve CA LUTs in order to test the hypothesis that LUTs with a critical value of λ will have maximal computational capability. Langton had shown that generic CA behavior seemed to undergo a sequence of phase transitions – from simple to "complex" to chaotic – as λ was varied. Both Langton and Packard believed that the "complex" region was necessary for non-trivial computation in CAs, thus the phrase "computation at the edge of chaos" was coined [33,40]. Packard's experiments indicated that CAs evolved by GAs to perform the density classification task indeed tended to exhibit critical λ values. However, this conclusion was not replicated in later work [38]. Correlations between λ (or other statistics of LUTs) and computational capability in CAs have been hinted at in further work, but have not been definitively established. A major problem is the difficulty of quantifying "computational capability" in CAs beyond the general (and not very practical) capability of universal computation.

Computation via CA "Particles"

While Mitchell, Hraber, and Crutchfield were not able to replicate Packard's results on λ , they were able to show

that genetic algorithms can indeed evolve CAs to perform computations [38]. Using earlier work by Hanson and Crutchfield on characterizing computation in CAs [20,21], Das, Mitchell and Crutchfield gave an information-processing interpretation of the dynamics exhibited by the evolved CAs in terms of *regular domains* and *particles* [21]. This work was extended by Das, Crutchfield, Mitchell, and Hanson [14] and Hordijk, Crutchfield and Mitchell [24].

In particular these groups showed that when regular domains – patterns described by simple regular languages – are filtered out of CA space-time behavior, the boundaries between these domains become forefront and can be interpreted as information-carrying "particles". These particles can characterize non-trivial computation carried out by CAs [15,21].

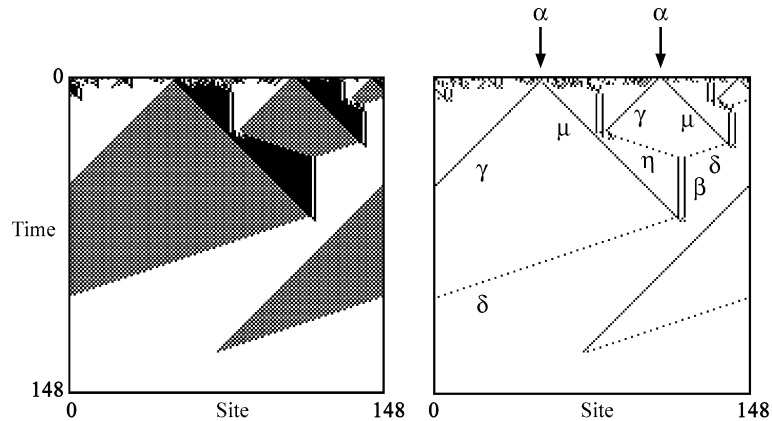
The information-carrying role of particles becomes clear when applied to CAs evolved by the GA for the density classification task. Figure 5, left, shows typical behavior of the best CAs evolved by the GA. The CA contains three regular domains: all white (0^*), all black (1^*), and checkerboard ($(01)^*$). Figure 5, right, shows the particles remaining after the regular domains are filtered out. Each particle has an origin and velocity, and carries information about the neighboring regions [37]. Hordijk et al. [24] showed that a small set of particles and their interactions can explain the computational behavior (i. e., the fitness) of the evolved cellular automata. Crutchfield et al. [13] describe how the analysis of evolved CAs in terms of particles can also explain how the GA evolved CAs with high fitness.

Land and Belew [31] proved that no two-state homogeneous CA can perform the density classification task perfectly. However, the maximum possible performance for CAs on this task is not known.

The density classification task remains a popular benchmark for studying the evolution of CAs with GAs, since the task requires collective behavior: the decision about the global density of the IC is based on information only from each local neighborhood. Das et al. [14] also used GAs to evolve CAs to perform a global synchronization task, which requires that, starting from any initial configuration, all cells of the CA will synchronize their states (to all 1s or 0s) and in the next time step all cells must change state to the opposite value. Again, this behavior requires global coordination based on local communication. Das et al. showed that an analysis in terms of particles and their interactions was also possible for this task.

Genetic Programming

Andre et al. [2] applied genetic programming (GP), a variation of GAs, to the density classification task. GP method-



Evolving Cellular Automata, Figure 5

Analysis of a GA evolved CA for density classification task. *Left*: The original spacetime diagram containing particle strategies in a CA evolved by GA. The regions of regular domains are all white, all black, or have a checkerboard pattern. *Right*: Spacetime diagram after regular domains are filtered out. (Reprinted from [37] with permission of the author.)

ology also uses a population of evolving candidate solutions, and the principles of reproduction and survival are the same for both GP and GAs. The main difference between these two methods is the encoding of individuals in the population. Unlike the binary strings used in GAs, individuals in a GP population have tree structures, made up of *function* and *terminal* nodes. The *function* nodes (internal nodes) are operators from a pre-defined function set, and the *terminal* nodes (leaves) represent operands from a terminal set. The fitness value is obtained by evaluating the tree on a set of test initial configurations. The crossover operator is applied to two parents by swapping randomly selected sub-trees, and the mutation operation is performed on a single node by creating a new node or by changing its value (Fig. 6) [29,30].

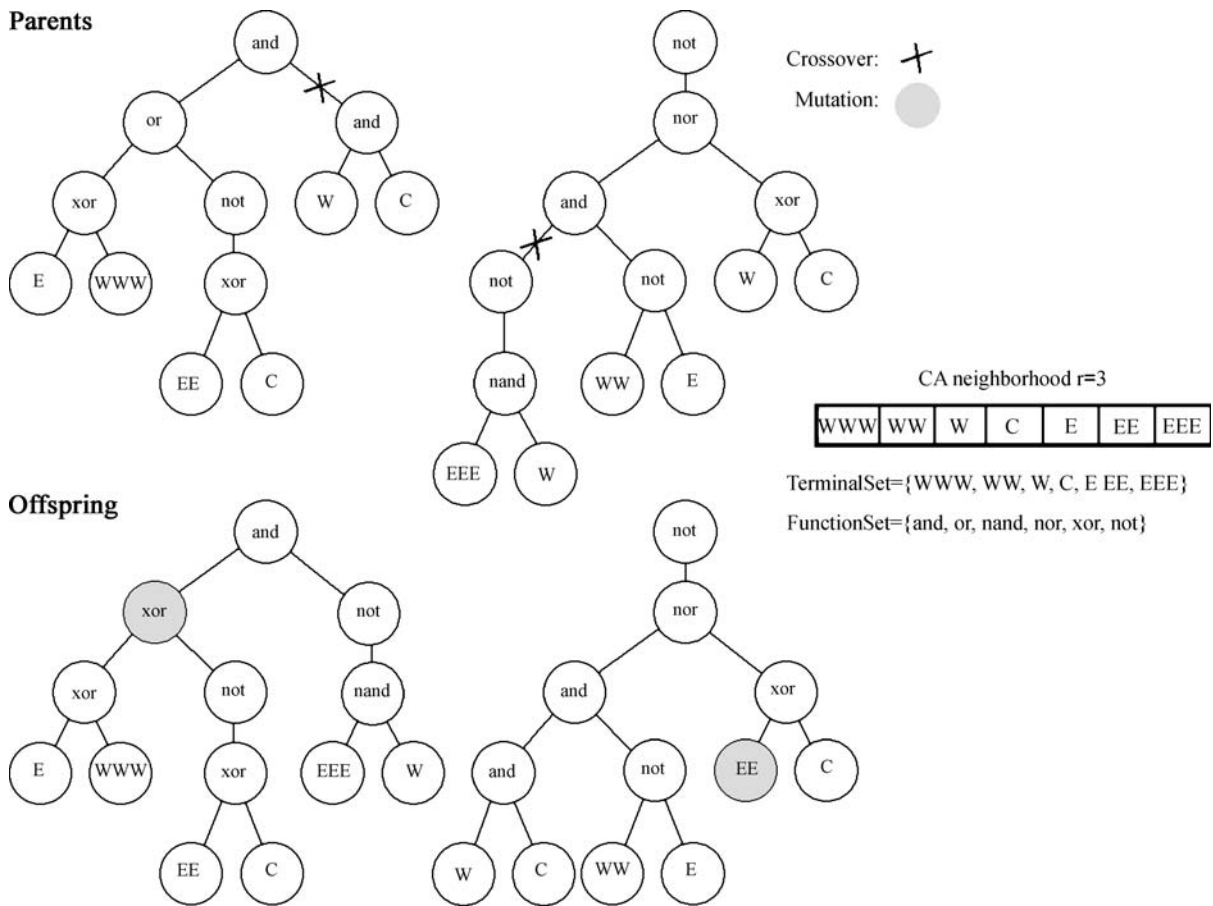
The GP algorithm evolved CAs whose performance is slightly higher than the performance of the best CAs evolved by a traditional GA.

Unlike traditional GAs that use crossover and mutation to evolve fixed length genome solutions, GP trees evolve to different sizes or shapes, and the subtrees can be substituted out and added to the function set as automatically defined functions. According to Andre et al., this allows GP to better explore the “regularities, symmetries, homogeneities, and modularities of the problem domain” [2]. The best-evolved CAs by GP revealed more complex particles and particle interactions than the CAs found by the EvCA group [13,24]. It is unclear whether the improved results were due to the GP representation or to the increased population sizes and computation time used by Andre et al.

Parallel Cellular Machines

The field of evolving CAs has grown in several directions. One important area is evolving non-homogeneous cellular automata [22,47,48,54]. Each cell of a non-homogeneous CA contains two independently evolving chromosomes. One represents the LUT for the cell (different cells can have different LUTs), and the second represents the neighborhood connections for the cell. Both the LUTs and the cell’s connectivity can be evolved at the same time. Since a task is performed by a collection of cells with different LUTs, there is no single best performing individual; the fitness is a measure of the collective behavior of the cells’ LUTs and their neighborhood assignments [46,48].

One of many tasks studied by Sipper was the global ordering task [47]. Here, the CA has fixed rather than periodic boundaries, so the “left” and “right” parts of the CA lattice are defined. The *ordering* in any given IC pattern will place all 0s on the left, followed by all 1s on the right. The initial density of the IC has to be preserved in the final configuration. Sipper designed a *cellular programming algorithm* to co-evolve multiple LUTs and their neighborhood topologies. Cellular programming carries out the same steps as the conventional GA (initialization, evaluation, reproduction, replacement), but each cell reproduces only with its local neighbors. The LUTs and connectivity chromosomes from the locally connected sites are the only potential parents for the reproduction and replacement of cell’s LUTs and the connectivity tables respectively. The cell’s limited connectivity results in genetically diverse population. If a current population has a cell with



Evolving Cellular Automata, Figure 6

An example of the encoding of individuals in a GP population, similar to one used in [2]. The function set here consists of the logical operators {and, or, not, nand, nor, and xor}. The terminal set represents the states of cells in a CA neighborhood, here {Center, East, West, EastOfEast, WestOfWest, EastOfEastOfEast, WestOfWestOfWest}. The figure shows the reproduction of *Parent*₁ and *Parent*₂ by crossover with subsequent mutation to produce *Child*₁ and *Child*₂

a high-fitness LUT, its LUT will not be directly inherited by a given cell unless they are connected. The connectivity chromosome causes spatial isolation that allows evolution to explore multiple CA rules as a part of a collective solution [47,48].

Sipper exhaustively tested all homogeneous CAs with $r = 1$ on the ordering task, and found that the best performing rule (rule 232) correctly ordered 71% of 1000 randomly generated ICs. The cellular programming algorithm evolved a non-homogeneous CA that outperformed the best homogeneous CA. The evolutionary search identified multiple rules that the non-homogeneous CA used as the components in the final solution. The rules composing the collective CA solution were classified as state preserving or repairing the incorrect ordering of the neighborhood bits. The untested hypothesis is that the cellular

programming algorithm can discover multiple important rules (partial traits) that compose more complex collective behavior.

Coevolution

Coevolution is an extension of the GA, introduced by Hillis [23], inspired by host-parasite coevolution in nature. The main idea is that randomly generated test cases will not continually challenge evolving candidate solutions. Coevolution solves this problem by evolving two populations – candidate solutions and test cases – also referred to as hosts and parasites. The hosts obtain high fitness by performing well on many of the parasites, whereas the parasites obtain high fitness by being difficult for the hosts. Simultaneously coevolving both populations engages hosts

and parasites in a mutual competition to achieve increasingly better results [7,17,56].

Successful applications of coevolutionary learning include discovery of minimal sorting networks, training artificial neural networks for robotics, function induction from data, and evolving game strategies [9,23,41,44,56,57]. Coevolution also improved upon GA results on evolving CA rules for density classification [28].

In the context of evolving CAs, the LUT candidate solutions are hosts, and the ICs are parasites. The fitness of a host is a fraction of correctly evaluated ICs from the parasite population. The fitness of a parasite is a function of the number of hosts that failed to correctly classify it.

Pagie et al. and Mitchell et al. among others, have found that embedding the host and parasite populations in a spatial grid, where hosts and parasites compete and evolve locally, significantly improves the performance of coevolution on evolving CAs [39,41,42,57].

Other Applications

The examples described in previous sections illustrate the power and versatility of genetic algorithms used to evolve desired collective behavior in CAs. The following are some additional examples of applications of CAs evolved by GAs.

CAs are most commonly used for modeling physical systems. CAs evolved by GAs modeled multi-phase fluid flow in porous material [61]. A 3D CA represented a pore model, and the GA evolved the permeability characteristics of the model to match the fluid flow pattern collected from the sample data. Another example is the modeling of physical properties of material microstructures [5]. An alternative definition of CAs (*effector automata*) represented a 2D cross-section of a material. The rule table specified the next location of the neighborhood's center cell. The results show that the GA evolved rules that reconstructed microstructures in the sample superalloy.

Network theory and topology studies for distributed sensor networks rely on connectivity and communication among its components. Evolved CAs for location management in mobile computing networks is an application in this field [50]. The cells in the mobile networks are mapped to CA cells where each cell is either a reporting or non-reporting cell. Subrata and Zomaya's study used three network datasets that assigned unique communication costs to each cell. A GA evolved the rules that designate each cell as reporting or not while minimizing the communication costs in the network. The results show that the GA found optimal or near optimal rules to determine which cells in a network are reporting. Sipper also hinted at ap-

plying his cellular programming algorithm to non-homogeneous CAs with non-standard topology to evolve network topology assignments [47].

Chopra and Bender applied GAs to evolve CAs to predict protein secondary structure [10]. The 1D CA with $r = 5$ represents interactions among local fragments of a protein chain. A GA evolved the weights for each of the neighboring fragments that determine the shape of the secondary protein structure. The algorithm achieved superior results in comparison with some other protein-secondary-structure prediction algorithms.

Built-In Self-Test (BIST) is a test method widely used in the design and production of hardware components. A combination of a *selfish gene algorithm* (a GA variant) and CAs were used to program the BIST architecture [12]. The individual CA cells correspond to the circuitry's input terminals, and the transition function serves as a test pattern generator. The GA identified CA rules that produce input test sequences that detect circuitry faults. The results achieved are comparable with previously proposed GA-based methods but with lower overhead.

Computer vision is a fast growing research area where CAs have been used for low-level image processing. The cellular programming algorithm has evolved non-homogeneous CAs to perform image thinning, finding and enhancing an object's rectangle boundaries, image shrinking, and edge detection [47].

Future Directions

Initial work on evolving two-dimensional CAs with GAs was done by Sipper [47] and Jiménez-Morales, Crutchfield, and Mitchell [27]. An extension of domain-particle analysis for 2D CAs is needed in order to analyze the information processing of CAs and to identify the epochs of innovations in evolutionary learning.

Spatially extended coevolution was successfully used to evolve high performance CAs for density classification. Parallel cellular machines also used spatial embedding of their components and found better performing CAs than the homogeneous CAs evolved by a traditional GA. The hypothesis is that spatially extended search techniques are successful more often than non-spatial techniques because spatial embedding enforces greater genetic diversity and, in the case of coevolution, more effective competition between hosts and parasites. This hypothesis deserves more detailed investigation.

Additional important research topics include the study of the error resiliency and the effect of noise on both the information processing in CAs and evolution of CAs. How successful is evolutionary learning in noisy environment?

What is the impact of failing CA components on information processing and evolutionary adaptation? Similarly, to make CAs more realistic as models of physical systems, evolving CAs with asynchronous cell updates is an important topic for future research. A number of groups have shown that CAs and similar decentralized spatially extended systems using asynchronous updates can have very different behavior from those using synchronous updates (e.g., [1,6,25,49,53]). An additional topic for future research is the effect of connectivity network structure on the behavior and computational capability of CAs. Some work along these lines has been done by Teuscher [52].

Acknowledgments

This work has been funded by the Center on Functional Engineered Nano Architectonics (FENA), through the Focus Center Research Program of the Semiconductor Industry Association.

Bibliography

- Alba E, Giacobini M, Tomassini M, Romero S (2002) Comparing synchronous and asynchronous cellular genetic algorithms. In: Guervos MJJ et al (eds) Parallel problem solving from nature. PPSN VII, Seventh International Conference. Springer, Berlin, pp 601–610
- Andre D, Bennett FH III, Koza JR (1996) Evolution of intricate long-distance communication signals in cellular automata using genetic programming. In: Artificial life V: Proceedings of the fifth international workshop on the synthesis and simulation of living systems. MIT Press, Cambridge
- Ashlock D (2006) Evolutionary computation for modeling and optimization. Springer, New York
- Back T (1996) Evolutionary algorithms in theory and practice. Oxford University Press, New York
- Basanta D, Bentley PJ, Miodownik MA, Holm EA (2004) Evolving cellular automata to grow microstructures. In: Genetic programming: 6th European Conference. EuroGP 2003, Essex, UK, April 14–16, 2003. Proceedings. Springer, Berlin, pp 77–130
- Bersini H, Detours V (2002) Asynchrony induces stability in cellular automata based models. In: Proceedings of the IVth conference on artificial life. MIT Press, Cambridge, pp 382–387
- Bucci A, Pollack JB (2002) Order-theoretic analysis of coevolution problems: Coevolutionary statics. In: GECCO 2002 Workshop on Understanding Coevolution: Theory and Analysis of Coevolutionary Algorithms, vol 1. Morgan Kaufmann, San Francisco, pp 229–235
- Burks A (1970) Essays on cellular automata. University of Illinois Press, Urban
- Cartlidge J, Bullock S (2004) Combating coevolutionary disengagement by reducing parasite virulence. *Evol Comput* 12(2):193–222
- Chopra P, Bender A (2006) Evolved cellular automata for protein secondary structure prediction imitate the determinants for folding observed in nature. *Silico Biol* 7(0007):87–93
- Codd EF (1968) Cellular automata. ACM Monograph series, New York
- Corno F, Reorda MS, Squillero G (2000) Exploiting the selfish gene algorithm for evolving cellular automata. *IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00)* 06:6577
- Crutchfield JP, Mitchell M, Das R (2003) The evolutionary design of collective computation in cellular automata. In: Crutchfield JP, Schuster PK (eds) *Evolutionary Dynamics – Exploring the Interplay of Selection, Neutrality, Accident, and Function*. Oxford University Press, New York, pp 361–411
- Das R, Crutchfield JP, Mitchell M, Hanson JE (1995) Evolving globally synchronized cellular automata. In: Eshelman L (ed) *Proceedings of the sixth international conference on genetic algorithms*. Morgan Kaufmann, San Francisco, pp 336–343
- Das R, Mitchell M, Crutchfield JP (1994) A genetic algorithm discovers particle-based computation in cellular automata. In: Davidor Y, Schwefel HP, Männer R (eds) *Parallel Problem Solving from Nature-III*. Springer, Berlin, pp 344–353
- Farmer JD, Toffoli T, Wolfram S (1984) Cellular automata: Proceedings of an interdisciplinary workshop. Elsevier Science, Los Alamos
- Funes P, Sklar E, Juille H, Pollack J (1998) Animal-animat coevolution: Using the animal population as fitness function. In: Pfeiffer R, Blumberg B, Wilson JA, Meyer S (eds) *From animals to animats 5: Proceedings of the fifth international conference on simulation of adaptive behavior*. MIT Press, Cambridge, pp 525–533
- Gardner M (1970) Mathematical games: The fantastic combinations of John Conway's new solitaire game "Life". *Sci Am* 223:120–123
- Grassberger P (1983) Chaos and diffusion in deterministic cellular automata. *Physica D* 10(1–2):52–58
- Hanson JE (1993) Computational mechanics of cellular automata. Ph D Thesis, University of California at Berkeley
- Hanson JE, Crutchfield JP (1992) The attractor-basin portrait of a cellular automaton. *J Stat Phys* 66:1415–1462
- Hartman H, Vichniac GY (1986) Inhomogeneous cellular automata (inca). In: Bienenstock E, Fogelman F, Weisbuch G (eds) *Disordered Systems and Biological Organization*, vol F20. Springer, Berlin, pp 53–57
- Hillis WD (1990) Co-evolving parasites improve simulated evolution as an optimization procedure. *Physica D* 42:228–234
- Hordijk W, Crutchfield JP, Mitchell M (1996) Embedded-particle computation in evolved cellular automata. In: Toffoli T, Bialek M, Leão J (eds) *Physics and Computation 1996*. New England Complex Systems Institute, Cambridge, pp 153–158
- Huberman BA, Glance NS (1993) Evolutionary games and computer simulations. *Proc Natl Acad Sci* 90:7716–7718
- Ikebe M, Amemiya Y (2001) VMoS cellular-automaton circuit for picture processing. In: Miki T (ed) *Brainware: Bio-inspired architectures and its hardware implementation*, vol 6 of FLSI Soft Computing, chapter 6. World Scientific, Singapore, pp 135–162
- Jiménez-Morales F, Crutchfield JP, Mitchell M (2001) Evolving two-dimensional cellular automata to perform density classification: A report on work in progress. *Parallel Comput* 27(5):571–585
- Juillé H, Pollack JB (1998) Coevolutionary learning: A case study. In: *Proceedings of the fifteenth international conference on machine learning (ICML-98)*. Morgan Kaufmann, San Francisco, pp 24–26

29. Koza JR (1992) Genetic programming: On the programming of computers by means of natural selection. MIT Press, Cambridge
30. Koza JR (1994) Genetic programming II: Automatic discovery of reusable programs. MIT Press, Cambridge
31. Land M, Belew RK (1995) No perfect two-state cellular automata for density classification exists. *Phys Rev Lett* 74(25): 5148–5150
32. Langton C (1986) Studying artificial life with cellular automata. *Physica D* 10D:120
33. Langton C (1990) Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D* 42:12–37
34. Lohn JD, Reggia JA (1997) Automatic discovery of self-replicating structures in cellular automata. *IEEE Trans Evol Comput* 1(3):165–178
35. Madore BF, Freedman WL (1983) Computer simulations of the Belousov-Zhabotinsky reaction. *Science* 222:615–616
36. Mitchell M (1996) An introduction to genetic algorithms. MIT Press, Cambridge
37. Mitchell M (1998) Computation in cellular automata: A selected review. In: Gramss T, Bornholdt S, Gross M, Mitchell M, Pellizzari T (eds) *Nonstandard Computation*. VCH, Weinheim, pp 95–140
38. Mitchell M, Hraber PT, Crutchfield JP (1993) Revisiting the edge of chaos: Evolving cellular automata to perform computations. *Complex Syst* 7:89–130
39. Mitchell M, Thomure MD, Williams NL (2006) The role of space in the success of coevolutionary learning. In: Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (eds) *Artificial life X: Proceedings of the tenth international conference on the simulation and synthesis of living systems*. MIT Press, Cambridge, pp 118–124
40. Packard NH (1988) Adaptation toward the edge of chaos. In: Kelso JAS, Mandell AJ, Shlesinger M (eds) *Dynamic patterns in complex systems*. World Scientific, Singapore, pp 293–301
41. Pagie L, Hogeweg P (1997) Evolutionary consequences of coevolving targets. *Evol Comput* 5(4):401–418
42. Pagie L, Mitchell M (2002) A comparison of evolutionary and coevolutionary search. *Int J Comput Intell Appl* 2(1):53–69
43. Reynaga R, Amthauer E (2003) Two-dimensional cellular automata of radius one for density classification task $\rho = \frac{1}{2}$. *Pattern Recogn Lett* 24(15):2849–2856
44. Rosin C, Belew R (1997) New methods for competitive coevolution. *Evol Comput* 5(1):1–29
45. Schadschneider A (2001) Cellular automaton approach to pedestrian dynamics – theory. In: *Pedestrian and evacuation dynamics*. Springer, Berlin, pp 75–86
46. Sipper M (1994) Non-uniform cellular automata: Evolution in rule space and formation of complex structures. In: Brooks RA, Maes P (eds) *Artificial life IV*. MIT Press, Cambridge, pp 394–399
47. Sipper M (1997) Evolution of parallel cellular machines: The cellular programming approach. Springer, Heidelberg
48. Sipper M, Ruppert E (1997) Co-evolving architectures for cellular machines. *Physica D* 99:428–441
49. Sipper M, Tomassini M, Capcarrere M (1997) Evolving asynchronous and scalable non-uniform cellular automata. In: *Proceedings of the international conference on artificial neural networks and genetic algorithms (ICANNGA97)*. Springer, Vienna, pp 382–387
50. Subrata R, Zomaya AY (2003) Evolving cellular automata for location management in mobile computing networks. *IEEE Trans Parallel Distrib Syst* 14(1):13–26
51. Tan SK, Guan SU (2007) Evolving cellular automata to generate nonlinear sequences with desirable properties. *Appl Soft Comput* 7(3):1131–1134
52. Teuscher C (2006) On irregular interconnect fabrics for self-assembled nanoscale electronics. In: Tyrrell AM, Haddow PC, Torresen J (eds) *2nd IEEE international workshop on defect and fault tolerant nanoscale architectures, NANOARCH'06*. Lecture Notes in Computer Science, vol 2602. ACM Press, New York, pp 60–67
53. Teuscher C, Capcarrere MS (2003) On fireflies, cellular systems, and evolvable. In: Tyrrell AM, Haddow PC, Torresen J (eds) *Evolvable systems: From biology to hardware*. Proceedings of the 5th international conference, ICES2003. Lecture Notes in Computer Science, vol 2602. Springer, Berlin, pp 1–12
54. Vichniac GY, Tamayo P, Hartman H (1986) Annealed and quenched inhomogeneous cellular automata. *J Stat Phys* 45:875–883
55. von Neumann J (1966) *Theory of Self-Reproducing Automata*. University of Illinois Press, Champaign
56. Wiegand PR, Sarma J (2004) Spatial embedding and loss of gradient in cooperative coevolutionary algorithms. *Parallel Probl Solving Nat* 1:912–921
57. Williams N, Mitchell M (2005) Investigating the success of spatial coevolution. In: *Proceedings of the 2005 conference on genetic and evolutionary computation*. Washington DC, pp 523–530
58. Wolfram S (1984) Universality and complexity in cellular automata. *Physica D* 10D:1
59. Wolfram S (1986) *Theory and application of cellular automata*. World Scientific Publishing, Singapore
60. Wolfram S (2002) *A new kind of science*. Wolfram Media, Champaign
61. Yu T, Lee S (2002) Evolving cellular automata to model fluid flow in porous media. In: *2002 Nasa/DoD conference on evolvable hardware (EH '02)*. IEEE Computer Society, Los Alamitos, pp 210

Evolving Fuzzy Systems

PLAMEN ANGELOV

Intelligent Systems Research Laboratory, Digital Signal Processing Research Group, Communication Systems Department, Lancaster University, Lancaster, UK

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Evolving Clustering](#)

[Evolving TS Fuzzy Systems](#)

[Evolving Fuzzy Classifiers](#)

[Evolving Fuzzy Controllers](#)

Application Case Studies
 Future Directions
 Acknowledgments
 Bibliography

Glossary

Evolving system In the context of this article the term ‘evolving’ is used in the sense of the self-development of a system (in terms of both its structure and parameters) based on the stream of data coming to the system on-line and in real-time from the environment and the system itself. The system is assumed to be mathematically described by a set of fuzzy rules of the form:

Ruleⁱ :

$$\begin{aligned} & \text{IF } (Input_1 \text{ is close to prototype}_1^i) \\ & \text{AND } \dots \text{ AND } (Input_n \text{ is close to prototype}_n^i) \quad (1) \\ & \text{THEN } (Output^i = \overline{Inputs}^T \text{ ConseqParams}) \end{aligned}$$

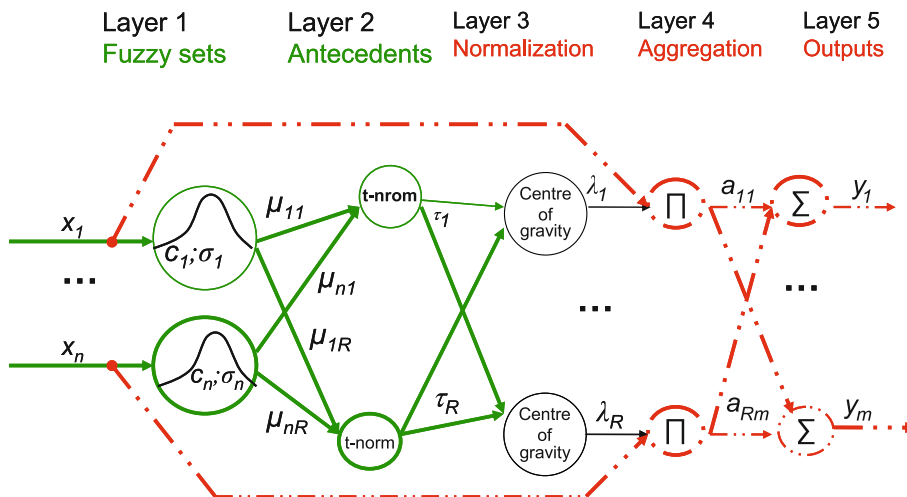
In this sense, this definition strictly follows the meaning of the English word “evolving” as described in [34], p. 294, namely “unfolding; developing; being developed, naturally and gradually”. Contrast this to the definition of “evolutionary” in the same source, which is “development of more complicated forms of life (plants, animals) from earlier and simpler forms”. The terms evolutionary or genetic are also associated with such phenomena (respectively operators that mimic these) as chromosome crossover, mutation, selection and reproduction, parents and off-springs [32]. Evolving

(fuzzy and neuro-fuzzy) systems do not deal with such phenomena. They rather consider a *gradual development* of the underlying (fuzzy or neuro-fuzzy) system structure.

Fuzzy system structure Structure of a fuzzy (or neuro-fuzzy) system is constituted of a set of fuzzy rules (1). Each fuzzy rule is composed of antecedent (IF) and consequents (THEN) parts. They are linguistically expressed. The antecedent part consists of a number of fuzzy sets that are linked with fuzzy logic aggregators such as conjunction, disjunction, more rarely, negation [43]. In the above example, a conjunction (logical AND) is used. It can be mathematically described by so-called t-norms or t-conorms between membership functions. The most popular membership functions are Gaussian, triangular, trapezoidal [73]. The consequent part of the fuzzy rules in the so-called Takagi–Sugeno (TS) form is represented by mathematical functions (usually linear). The structure of the TS fuzzy system can also be represented as a neural network with a specific (five layer) composition (Fig. 1). Therefore, these systems are also called neuro-fuzzy (NF).

The number of fuzzy rules and inputs (which in case of classification problems are also called features or attributes) is also a part of the structure.

The first layer consists of neurons corresponding to the membership functions of a particular fuzzy set. This layer takes the inputs, x and gives as output the degree, μ to which these fuzzy descriptors are satisfied. The second layer represents the antecedent parts



Evolving Fuzzy Systems, Figure 1
 Structure of the (neuro-fuzzy) system of TS type

of the fuzzy rules. It takes as inputs the membership function values and gives as output the firing level of the i th rule, τ_i . The third layer of the network takes as inputs the firing levels of the respective rule, τ_i and gives as output the normalized firing level, λ_i as “center of gravity” [43] of τ_i . As an alternative one can use the “winner takes all” operator. This operator is used usually in classification, while the “center of gravity” is preferred for time-series prediction and general system modeling and control. The fourth layer aggregates the antecedent and the consequent part that represents the local sub-systems (singletons or hyper planes). Finally, the last 5th layer forms the total output of the NF system. It performs a weighed summation of local sub-systems.

Fuzzy system parameters Parameters of the NF system of TS type include the center, c and spread, σ of the Gaussians or parameters of the triangular (or trapezoidal) membership functions. An example of a Gaussian type membership function can be given as:

$$\mu = e^{-\frac{1}{2}\left(\frac{d}{r}\right)^2} \quad (2)$$

where d denotes distance between a data sample (point in the data space) and a prototype/cluster center (focal point of a fuzzy set); r is the radius of the cluster (spread of the membership function).

Note that the distance can be represented by Euclidean (the most typical example), Mahalanobis [33], cosine etc. forms.

These parameters are associated with the antecedent part of the system. Consequent part parameters are coefficients of the (usually) linear functions, singleton coefficients or coefficients of more complex functions (e.g. exponential) if such ones are used.

$$y_i = a_{i0} + a_{i1}x_1 + \dots + a_{in}x_n \quad (3)$$

where a denotes parameters of the consequent part; x denote the inputs (features); i is the index of the i th fuzzy rule; n is the number (dimensionality) of the inputs (features).

Potential Potential is a mathematical measure of the data density. It is calculated at a data point, z and represents numerically the accumulated proximity (*density*) of the data surrounding this data point. It resembles the probability distribution used in so-called Parzen windows [33] and is described in [26,72] by a Gaussian-like function:

$$P(z) = e^{-\frac{1}{2r}\bar{\sigma}^2} \quad (4)$$

where $z = [x, y]$ denotes the joint (input/output) vector;

$\bar{\sigma}_k^2 = \frac{1}{k-1} \sum_{i=1}^{k-1} d^2(z_k, z_i)$ is the variance of the data in terms of the cluster center.

In [3,9] the Cauchy function is used which has the same properties as the Gaussian but is suitable for recursive calculations.

$$P(z) = \frac{1}{1 + \bar{\sigma}^2} \quad (5)$$

Age of a cluster or fuzzy rule The age of the (evolving) cluster is defined as the accumulated time of appearance of the samples that form the cluster which support that fuzzy rule.

$$A^i = k - \frac{\sum_{l=1}^{S_k^i} k_l}{S_k^i} \quad (6)$$

where k denotes the current time instant; S_k^i denotes the support of the cluster that is the number of data samples (points) that are in the zone of influence of the cluster (formed by its radius). It is derived by simple counting of data samples (points) at the moment of their arrival (when they are first read) and assigned to the nearest cluster [10].

The values of A vary from 0 to k and the derivative of A in respect to time is always less or equal to 1 [17]. An “old” cluster (fuzzy rule) has not been updated recently. A “young” cluster (fuzzy rule) is one that has predominantly new samples or recent ones. The (first and second) derivatives of the *age* are very informative and useful for detection of data “*shift*” and “*drift*” [17].

Definition of the Subject

Evolving Fuzzy Systems (EFS) are a class of Fuzzy Rule-based (FRB) and Neuro-Fuzzy (NF) systems that have both their parameters and underlying structure self-adapting, self-developing, self-learning from the data in on-line mode and, possibly, in real-time. The concept was conceived at the beginning of this century [2,5]. Parallel investigations have led to similar developments in neural networks (NN) [41,42]. EFS have the significant advantage compared to the evolving NN of being linguistically tractable and transparent. EFS have been instrumental in the emergence of new branches of *evolving* clustering algorithms [3], *evolving* classifiers [16,51], *evolving* time-series predictors [9,47], *evolving* fuzzy controllers [4], *evolving* fault detectors [30] etc. Over the last years EFS has demonstrated a wide range of applications spanning robotics [76]

and defense [24] to biomedical [70] and industrial process [29] data processing in real-time, new generations of self-calibrating, self-adapting sensors [52,53], speech [37] and image processing [56] etc. EFS have the potential to revolutionize such areas as autonomous systems [66], intelligent sensors [45], early cancer detection and diagnosis; they are instrumental in raising the so-called machine intelligence quotient [74] by developing systems that self-adapt in real-time to the dynamically changing environment and to internal changes that occur in the system itself (e. g. wearing, contamination, performance degradation, faults etc.). Although the terms *intelligent* and *artificial intelligence* have been used often during the last several decades the technical systems that claim to have such features are in reality far from true intelligence. One of the main reasons is that true intelligence is evolving, it is not fixed. EFS are the first mathematical constructs that combine the approximate reasoning typical for humans represented by the fuzzy inference with the dynamically evolving structure and respective formal mathematically sound learning mechanisms to implement it.

Introduction

Fuzzy Sets and Fuzzy Logic were introduced by Lotfi Zadeh in 1965 in his seminal paper [71]. During the last decade of the previous century there was an increase of the various applications of fuzzy logic-based systems mainly due to the introduction of fuzzy logic controllers (FLC) by Ebrahim Mamdani in 1975 [54], the introduction of the fuzzily blended linear systems construct called Takagi-Sugeno (TS) fuzzy systems in 1985 [65], and the theoretical proof that FRB systems are universal approximators (that is any arbitrary non-linear function in the $[0; 1]$ range can be asymptotically approximated by a FRB system [68]). Historically, the FRB systems were first being designed based entirely or predominantly on human expert knowledge [54,71]. This offers advantages and was a novel technique at that time for incorporating uncertain, subjective information, preferences, experience, intuition, which are difficult or impossible to be described otherwise. However, it poses enormous difficulties for the process of designing and routine use of these systems, especially in real industrial environments and in on-line and real-time modes. TS fuzzy systems made possible the development of efficient algorithms for their design not only in off-line, but also in on-line mode [14]. This is facilitated by their dual nature – they combine a fuzzy linguistic premise (antecedent) part with a functional (usually linear) consequent part [65]. With the invention of the concept of EFS [2,5] the problem of the design was completely

automated and data-driven. This means, EFS systems self-develop their model, respectively system structure as well as adapt their parameters “from scratch” on the fly using experimental data and efficient recursive learning mechanisms. Human expert knowledge is not compulsory, not limiting, not essential (especially if it is difficult to obtain in real-time). This does not necessarily mean that such knowledge is prohibited or not possible to be used. On the contrary, the concept of EFS makes possible the use of such knowledge in initialization stages, even during the learning process itself, but this is not essential, it is optional. Examples of EFS are intelligent sensors for oil refineries [52,53], autonomous self-localization algorithms used by mobile robots [75,76], smart agents for machine health monitoring and prognosis in the car industry [30], smart systems for automatic classification of images in CD production process [51] etc. This is a new promising area of research and new applications in different branches of industry are emerging.

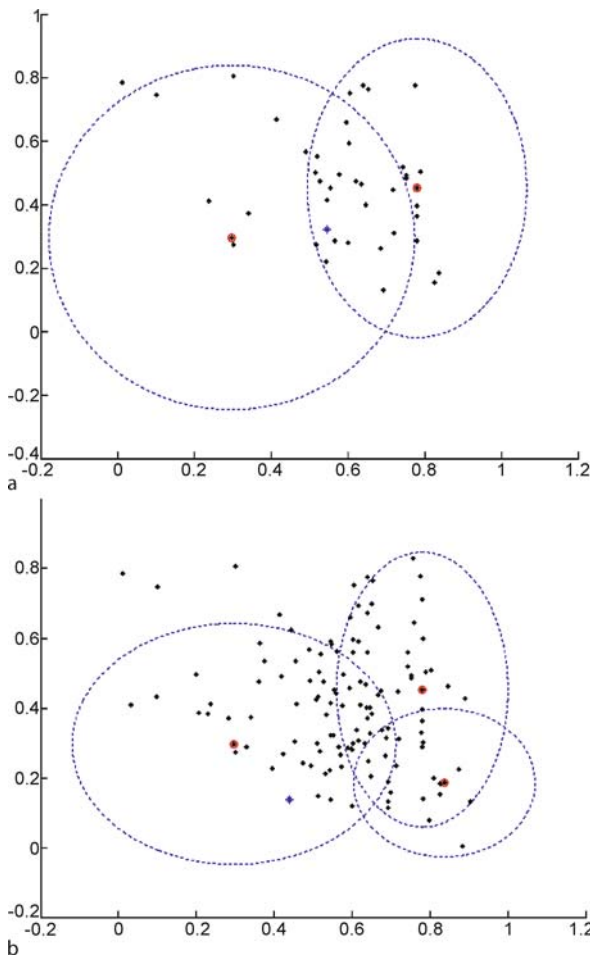
Evolving Clustering

Data Clustering and, Fuzzy Clustering in particular, are methods for grouping the data based on their similarity, density in the data space and proximity. Partitioning of the data into clusters can be done off-line (using a batch set of data, performing iterative computations over this set, minimizing certain criteria/cost function) or on-line, incrementally. Examples of incremental clustering approaches are self-organizing maps (SOM) conceived by Teuvo Kohonen in the early 1980s [44], adaptive resonance theory (ART) by Stephen Grossberg conceived in the same period [25] etc. Clustering is a type of unsupervised learning technique where the correct examples are not provided. Usually, the number of clusters is pre-specified, e.g. in SOM the number of nodes of the map is pre-defined; the number of neighbors, k in the k -nearest neighbor clustering method [33] is also supposed to be provided, the number C in the fuzzy c -means (FCM) fuzzy clustering algorithm by Jim Bezdek should also be provided [22]. Usually these approaches rely on a threshold and are very sensitive to the specific values of this threshold. Most of the existing approaches are also mean-based (i. e. they use the mean of all data or mean of groups of data). The problem is that the mean is a virtual (non-existing and possibly infeasible) point in the data space.

In contrast, the evolving clustering method eClustering conceived in the last decade [3] does not need the number of clusters, the threshold or any other parameter to be pre-specified. It is parameter-free and starts “from scratch” to cluster the data based on their density distri-

bution alone. It is based on the recursive calculation of the potential (5). eClustering is prototype-based (some of the data points are used as prototypes of cluster centers). The procedure of the evolving clustering approach starts from scratch assuming that the first available data point is a center of a cluster. This assumption is temporary and if a priori knowledge exists the procedure can start with an initial set of cluster centers that will be further refined. The coordinates of the first cluster center are formed from the coordinates of the first data point ($z_1^* \leftarrow z_1$). Its potential is set to the ideal value, $P_1(z_1) \rightarrow 1$. Starting from the **next** data point which is read in real-time, the following steps are performed for each new data point:

- calculate its potential, $P_k(z_k)$;



Evolving Fuzzy Systems, Figure 2

The Evolving Clustering method applied to data concerning NOx emissions; **a** top plot – after 43 samples are read (after 43 s because the sampling rate is 1 sample/second or 1 Hz); **b** bottom plot – after 124 samples are read (after 124 s)

- update the potential of the existing cluster centers (because their potential has been affected by adding a new data point);
- compare the potential of the new data point with the potential of the previously existing centers. On the basis of this comparison and the membership of the existing clusters one of the following actions is taken: (**add** a new cluster center based on the new data point) **OR** (**remove** the cluster that describes well the new point which brings an increment to the potential) **AND** (**replace** it with a cluster formed around the new point) **OR** (**ignore** (do not change the cluster structure)).

The process is illustrated in Fig. 2 for the data of NOx emissions from a car exhaust [13].

One can see that the clustering evolves (number of clusters increases from two to three, their position and their radius has changed). Note that in this experiment only two normalized to the range [0; 1] inputs (features), namely the engine output torque in N/m, x_1 and pressure in the second cylinder in Pa, x_2 are used. For more details on eClustering, please, consult the papers from the bibliography, and especially [2,3,9,16].

Evolving TS Fuzzy Systems

TS fuzzy systems [as illustrated in Fig. 1 and described in a very general form in Eq. (1)] were first introduced in 1985 [65] in the form:

$$\mathfrak{R}_i: \text{IF } (x_1 \text{ is } A_{i1}) \text{ AND } \dots \text{ AND } (x_n \text{ is } A_{in}) \quad (7) \\ \text{THEN } (y_i = a_{i0} + a_{i1}x_1 + \dots + a_{in}x_n)$$

where A_i denotes the i th fuzzy rule ($i = [1, R]$); R is the number of fuzzy rules; \mathbf{x} is the input vector; $\mathbf{x} = [x_1, x_2, \dots, x_n]^T$; A_{ij} denotes the antecedent fuzzy sets, $j \in \{1, n\}$; y_i is the output of the i th linear subsystem; a_{il} are its parameters, $l \in \{0, n\}$.

The structure of the TS system (number of fuzzy rules), antecedent part of the rules, number of inputs etc., are supposed to be known and are fixed. The data may be provided to the TS system in off-line or in on-line manner. Different data-driven techniques were developed to identify the best in terms of certain (local or global) error minimization criteria such as using a (recursive) least squares technique [65], using genetic algorithms [7,62] etc. The overall output is found to be a weighted sum of local outputs produced by each fuzzy rule:

$$y = \sum_{i=1}^R \lambda^i y^i. \quad (8)$$

Where the weights, λ represent the normalized firing level of the respective fuzzy rule and can be determined by:

$$\lambda^i = \frac{\sum_{j=1}^n \mu_j^i(x_j)}{\sum_{l=1}^R \sum_{j=1}^n \mu_j^l(x_j)} . \quad (9)$$

In a vector form the above equations can be represented as:

$$y = \psi^T \theta \quad (10)$$

where $\psi = [\lambda_1 x_e^T, \lambda_2 x_e^T, \dots, \lambda_R x_e^T]^T$ is the vector of weighted extended inputs; $x_e^T = [1, x^T]$;

$\theta = [\pi_1^T, \pi_2^T, \dots, \pi_R^T]^T$ is the vector of parameters;

$$\pi^i = \begin{bmatrix} \alpha_{01}^i & \alpha_{02}^i & \dots & \alpha_{0m}^i \\ \alpha_{11}^i & \alpha_{12}^i & \dots & \alpha_{1m}^i \\ \dots & \dots & \dots & \dots \\ \alpha_{n1}^i & \alpha_{n2}^i & \dots & \alpha_{nm}^i \end{bmatrix}$$

are the parameters of the m local linear sub-systems.

The assumption that the TS fuzzy system structure has to be known a priori was for the first time questioned in [2,5] and ultimately in [9] with the proposal of evolving TS (eTS) systems. In [15] a further extension of the eTS system was proposed, namely that they can have many outputs. In this way, the multi-input-multi-output eTS systems were introduced (MIMO-eTS).

eTS is a very flexible and powerful tool for time-series prediction, prognosis, modeling non-stationary phenomena, intelligent sensors etc. The algorithm for its learning from streaming data in real-time has two basic phases, which can both be performed very quickly (in one time step between the arrival of two data samples – the current one and the next one). The learning mechanism proposed in [9] is computationally very efficient because it is fully recursive. The two phases include:

- Data space partitioning and based on this forming and update of the fuzzy rule-base structure;
- Learning parameters of the consequent part of the fuzzy rules.

Note that the partitioning of the data space serves in eTS identification a different purpose compared to the purpose of data space partitioning in eClustering. In eTS there are outputs and the aim is to find such (perhaps overlapping) clustering of the input-output joint data space that fragments the input-output relationship into locally valid simpler (possibly linear) dependences. In eClustering the aim is to cluster the input data space into distinctive

regions. Other than that, the first phase of the eTS model identification is the same as the procedure in the eClustering method described above.

The second phase of the learning is parameter identification. It can be performed using a fuzzily weighted version [9] of the well-known recursive least squares (RLS) method [50]. One can perform either local (11) or global (12) identification by minimizing different cost functions [9]:

$$J_L = \sum_{i=1}^R (Y - X^T \pi_i)^T \Lambda_i (Y - X^T \pi_i) \quad (11)$$

$$J_G = (Y - \Psi^T \theta)^T (Y - \Psi^T \theta) . \quad (12)$$

In one of the cases (when a local cost function is used) the result will be a better approximation locally of the overall non-linear function by the local linear sub-models. The pay-off is, however, a poorer overall approximation. This is, however, compensated by a simpler and computationally more efficient procedure (if we use a locally valid cost function the covariance matrices of much smaller size can be used and they require much less memory space and time to perform computations) [9].

Evolving Fuzzy Classifiers

Classification is a problem that has been well studied and a large number of conventional approaches exist to address this problem [33]. Most of them, however, are designed to operate in batch mode and do not change their structure on-line (do not capture new patterns that may be present in the streaming data once the classifier is built). Off-line pre-trained classifiers may be good for certain scenarios, but they need to be redesigned or retrained if the circumstances change. There are also so-called incremental (or on-line) classifiers which work on a “sample-by-sample” basis and only require the features of that sample plus a small amount of aggregated information (a rule-base, a small number of variables needed for recursive calculations). They do not require all the history of the data stream (all previously seen data samples). Sometimes they are also called one-pass (each sample is processed only once at a time and is then discarded from the memory).

FRB systems have been successfully applied to a range of classification tasks including, but not limited to, decision making, fault detection, pattern recognition, image processing [46]. FRB systems have become one of the alternative frameworks for classifier design together with the more established Bayesian classifiers, decision trees [33], neural network-based classifiers [57], and support-vector machines (SVM) [67]. The task of the classifier is to map

the set of features of the sample data onto the set of class labels. A particular advantage of the FRB classifiers is that they are linguistic in form while being also proven universal approximators [68].

In the framework of the concept of *evolving fuzzy systems* a family of evolving fuzzy classifiers, eClass was proposed in [16,17,51]. The first type of evolving fuzzy classifiers, eClass0 has the typical structure of a fuzzy classifier [46] that differs from structure (1) by the consequent part only:

$$\begin{aligned} R^i: & \text{ IF } (Feature_1 \text{ is close to } prototype_1^i) \\ & \text{ AND } \dots \text{ AND } (Feature_n \text{ is close to } prototype_n^i) \\ & \text{ THEN } (ClassLabel^i) \end{aligned} \quad (13)$$

The output of eClass0, in the same way as typical fuzzy classifiers [46] provides the label of the class (0, 1 etc.) directly. In this sense, it is not a TS fuzzy system, but is closer to the Mamdani-type fuzzy systems [54]. The main difference of the eClass0 from the typical classifiers [46] is its ability to evolve, to expand the set of fuzzy rules that enables it to capture new data patterns, to adapt to possibly changing characteristics of the streaming data [16,17]. The inference in eClass0 is produced using the so-called “winner takes all” rule [33,46]:

$$Label = Label^{i^*}; i^* = \arg \max_{i=1}^R \left(\sum_{j=1}^n \mu_j^i(x_j) \right). \quad (14)$$

It is much easier and faster to build and train eClass0 in real-time, but the results of the classification can be further significantly improved if the classifier structure is assumed to be of TS-type. eClass is designed for on-line applications with an evolving (self-developing) FRB structure. The antecedents of the FRB are formed from the data stream around highly descriptive focal points (prototypes) per class in the input-output space. The features vector, \mathbf{x} is augmented with the class label, L to form a joint input-output vector $\mathbf{z} = [\mathbf{x}^T, L]^T$. The eClustering algorithm is applied *per class* (not over all the data). In this way, information granules (primitive forms of knowledge) [36] are formed **in real-time** around descriptive data samples, represented linguistically by fuzzy sets. This on-line algorithm works similarly to adaptive control [19] and estimation [39] – in the period between two samples two phases are performed: (1) class prediction (classification); (2) classifier update or evolution. During the first phase the class label is not known and is being predicted; during the second phase, however, it is known and is used as supervisory

information to update the classifier (including its **structure evolution** as well as its parameters update).

An alternative structure of the fuzzy classifier, eClass1 is based on the TS-type fuzzy system which has a consequent part of functional type as described in (7). The architecture of eClass1 differs significantly from the architecture of eClass0 and the typical FRB [46]. It performs a regression over the features. Having in mind that the classification surface in a data stream changes dynamically the goal of the evolving fuzzy classifier eClass1 is to evolve a rule-base which takes these changes into account by adapting parameters of the FRB (spreads, consequent parameters) as well as the focal points and the size of the rule-base. The output of each rule is a real (not integer as in the typical fuzzy classifiers) value, which if normalized represents the possibility of a data sample to be of certain class [16,17]:

$$\bar{y}^i = \frac{y^i}{\sum_{i=1}^R y^i}. \quad (15)$$

The overall output of the classifier is then taken as a weighted average (not as winner takes all as in typical fuzzy classifiers) of the normalized outputs of each fuzzy rule:

$$y = \sum_{i=1}^R \frac{\sum_{j=1}^n \mu_j^i}{\sum_{l=1}^R \sum_{j=1}^n \mu_j^l} \bar{y}^i. \quad (16)$$

This output is then used to discriminate between the classes. If the problem has two classes (A and B) then the target values are, obviously, 0 for one of the two classes (e.g. Class A) and 1 for the other one (Class B) or vice versa. To discriminate in this case one can simply use a threshold of 0.5. All the outputs that are above 0.5 are being classified as Class B while all the outputs below 0.5 are classified as Class A or vice versa:

$$\begin{aligned} & \text{ IF } (y > 0.5) \\ & \text{ THEN } (Class \ A) \\ & \text{ ELSE } (Class \ B) \end{aligned} \quad (17)$$

When the problem has more than two classes, one can apply MIMO eTS where each of the K outputs corresponds to the possibility that a data sample belongs to a certain class (as discussed above). It is interesting to note that it is possible to use MIMO eTS for a two-class problem. In order to do this, one needs to have vectors that represent the target outputs, for example $\mathbf{y} = [1 \ 0]$ for Class A and

$y = [0 \ 1]$ for Class B or vice versa. In eClass1-MIMO the label is determined by the highest value of the discriminator, \bar{y}_l :

$$\text{Label} = \text{Label}^{i*}; i^* = \arg \max_{l=1}^K \bar{y}_l \quad (18)$$

where K denotes the number of classes.

Evolving Fuzzy Controllers

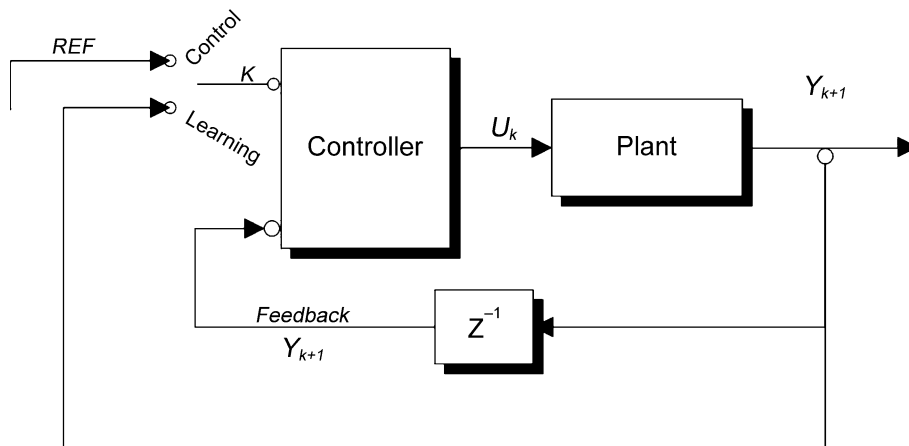
Fuzzy logic controllers have been applied in a range of industrial processes [48,59] around the world including in home appliances [21]. The structure of the controller, however, is often decided in an ad hoc manner [54,59] and parameters are tuned off-line using various numerical techniques such as genetic algorithms for example [27,32,63]. In reality, however, even if a validation test has been made beforehand, there is no guarantee that a controller designed in this way will perform satisfactorily if the object of control or its environment change [4]. The reasons could be aging, wearing, change in the mode of operation or development of a fault, seasonal changes in the environment etc. An effective mechanism for tackling such problems known from the classical control theory is adaptation [19]. It is well developed for the linear models and controllers [38], but not for a general (very often highly non-linear, complex and uncertain) case [2]. Adaptive control theory assumes a linear model with a fixed structure and applies to parameters only [19,38].

The concept of evolving fuzzy systems has been applied to the control problem in [2,4] in terms of self-developing the structure of a controller from experimental data in a data-driven manner based on the indirect adap-

tive learning scheme proposed initially by Psaltis [60] and developed further using NN by Anderson [1]. The indirect learning (IL) control scheme is based on the approximation of the inverse dynamics of the plant. The IL-based control scheme is a model-free concept. It feeds back the integrated (or delayed one-step back) output signal instead of feeding back the error between the plant output and the reference signal as represented in Fig. 3.

Figure 3 represents only the basic concept of the approach. It has two phases and the switching between them can be represented by an imaginary switch knob. When the imaginary knob, K is in position “1” the controller is used and we are in phase “Control”. When the imaginary knob is in position “2” the controller learns, self-develops and we are in phase “Learning”. During the supervisory learning phase, the true output signal (y_{k+1}) at the time-instant ($k + 1$) is fed back and the knob is in position “2”. The controller also receives a signal that is a delayed true output, y_k . The controller has as an output the value of the control signal, u_k . During the control phase (when the knob is in position “1”) the input is determined entirely based on the reference signal (ref) as an alternative to the predicted next step output, y_{k+1} . In this way, the controller already trained in the previous learning phases produces such a control signal (u_k), which brings the output of the plant at the next time step (y_{k+1}) close to the reference signal (ref).

The IL scheme was taken further in [2,4] by implementing the controller as an evolving FRB system of TS-type. The original works realized the controller as a NN that was trained *off-line* based on a batch set of training data for the control action and output triplets of the form $[y_k, y_{k+1}, u_k]$ for $k = 1, 2, \dots, N$; where N denotes the



Evolving Fuzzy Systems, Figure 3
Indirect learning-based control scheme

number of training data samples. However, learning techniques for NN are iterative and, therefore training of the NN-controller as described in [1,60] is performed *off-line*. Additionally, NN suffers from the important disadvantage in comparison to the FRB systems that they are not transparent. In [2,4] the basic scheme of IL control is taken further by adding a disturbance and by using eTS to realize the controller. This scheme was implemented on temperature-control problems.

Application Case Studies

Self-Calibrating Intelligent Sensors for Process Industries

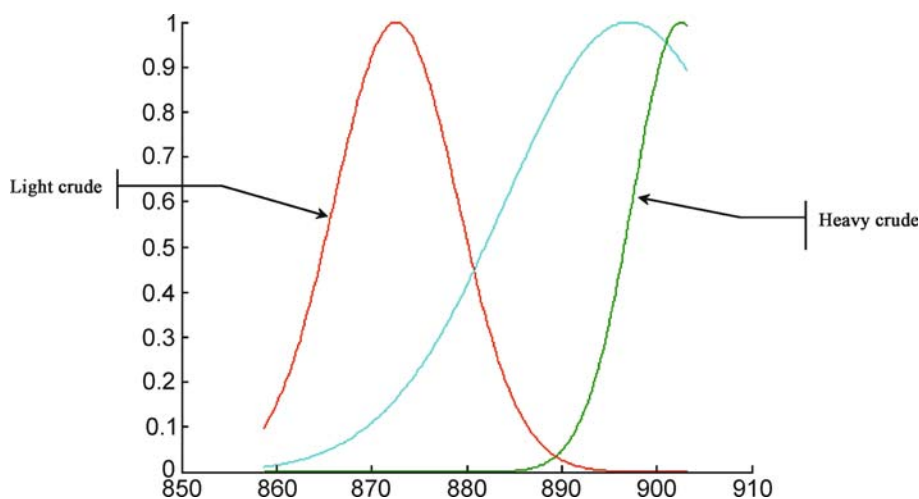
So-called intelligent or inferential sensors have been adopted by the process industries (chemical, petro-chemical, manufacturing) for several decades [31]. The main reason is that they provide accurate real time estimates of difficult to measure otherwise parameters or can replace (substitute) expensive measurements such as gas emissions, biomass, melt index, etc. They use as inputs the available (“hard”) sensors for easy to measure physical variables, such as temperatures, pressures, and flows which are also cheaper. The main disadvantage of the currently existing inferential or “soft” sensors is that significant efforts must be made based on batch sets of data to develop and maintain the mathematical models that support them (neural networks, statistical models etc.). Any process changes outside the conditions used for off-line model development can lead to significant performance

deterioration which, in turn, requires maintenance and recalibration.

Evolving Fuzzy Systems offer an effective opportunity to develop “soft” sensors that are more flexible, self-calibrating, and thus, more “intelligent” [45]. Several applications of EFS-based soft sensors, in particular for oil refineries [52,53] and propylene production [12] were reported. An important advantage of the evolving sensors is that they extract human-interpretable knowledge in the form of linguistic fuzzy rules. For example, Fig. 4 illustrates membership functions of the fuzzy sets (in values in the range [0;1] on the vertical axis) that describe the density of the crude, d in gram per liter (g/l) on the horizontal axis. The evolving fuzzy sensor (eSensor) implemented in the oil refinery at Santa Cruz, Tenerife, Spain predicts in real-time the temperature of the heavy naphtha (hn) T^{hn} , °C in degrees Celsius when it evaporates 95% liquid volume, according to the ASTM D86-04b standard based on real-time measurements of:

- The pressure of the tower, p , measured in kg/cm²g
- The amount of the product taking off, P , represented in %
- The density of the crude, d in g/l (illustrated in Fig. 4)
- Temperature of the column overhead, T^{co} in °C
- Temperature of the naphtha extraction, T^{ne} in °C

An expert in the area of oil refining processes can easily visually distinguish between the heavy crude and light crude represented by respective membership functions derived automatically from the data in real-time.



Evolving Fuzzy Systems, Figure 4

Fuzzy sets for different fractions of the crude that contribute to the different quality of the end product (naphtha in this case) can be extracted automatically in real-time from the data stream using eSensor

Adaptive Real-Time Classifiers (Image, Land-Mark, Robotic)

Presently the data that are to be processed in industry, in defense and other real-life applications are not only huge in volume, but very often they are in the form of data streams [28]. This requires not only precise classifiers, but also dynamically evolvable classifiers. For example, in mobile robotics, an autonomous vehicle produces a video stream while operating in a completely unknown environment that needs to be processed [75,76]. The Evolving Clustering method was used to automatically generate a fuzzy rule-base that describes the landmarks discovered without any prior learning, “on the fly” by a mobile robot Pioneer 3DX [58] exploring a completely unknown environment [76]. The mobile robot is using its on-board pan-tilt zoom camera to produce a video stream. The frames were grasped and processed by the eClustering algorithm based on a 3-dimensional color-vector (R, G, B). Fuzzy rules of the following form were extracted from the data automatically:

Note that the landmarks that were identified automatically represented real objects on the route of the mobile robot such as the underpass of Lancaster University from the example. They were identified to be distinctive from the surrounding background by eClustering. The fuzzy rule base (Fig. 5) has evolved “from scratch” based on the video information and the data distribution only.

Predictive Models (Air-Conditioning, Financial Time-Series, Benchmark Data Sets)

There are different techniques that can be used for predictive models, such as ARMAX models [50], neural networks [34] non-evolving (fixed structure) fuzzy rule-based models [65,73]. Evolving fuzzy systems, however, offer additionally the capability to have a predictor that evolves

following the dynamic changes of the data by gradually adapting not only its parameters, but also its structure. In [6] the problem of predicting the characteristic temperature difference across a coil in a heat exchanger of an air-conditioning unit installed in a real building in Iowa, USA is considered. The evolving fuzzy rule-based model develops its structure and parameters based on the data of the flow rate entering the coil, moisture content of the air entering the coil, temperature of the chilled water, and control signal to the valve as illustrated in Fig. 6. The model proved to work satisfactorily in all season conditions due to its ability to adapt to the changes in the environment (different seasons) as demonstrated in Fig. 6c. When pre-trained (based on 400 samples) and fixed as both structure and parameters the performance deteriorated unacceptably in changing seasonal conditions as seen in Fig. 6a. A partial re-training improved significantly the results (Fig. 6b), but still this was less valid when the season changed (at around sample 912) and the model structure evolution has stopped (at sample 1000). When, the model structure evolution continued uninterrupted the result was a satisfactory performance in all seasons as seen in Fig. 6c.

Fault Detection and Prognostics

Evolving clustering and eTS fuzzy systems were applied in Ford Motor Company to machine health monitoring and prognosis in [30]. The ability of the evolving clustering method to form new clusters that represent different operating modes of a machine was exploited and different types of faults (incipient or drastic) were automatically identified based on the difference in the cluster formation. A prediction of the direction of movement of the cluster centers was used for prediction of possible faults and of the end-of-life of the machine.

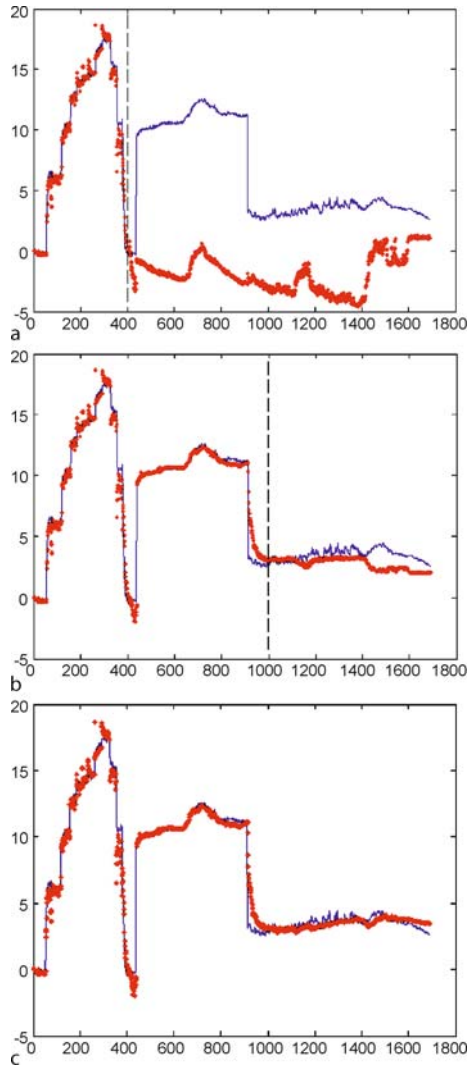
Rule_i: IF (μ^R_{il} is close to 123) AND... AND(μ^B_{in} is close to 84)



THEN the Frame_j is like the Landmark_i

Evolving Fuzzy Systems, Figure 5

Fuzzy rule describing a Landmark (underpass at Lancaster University campus) that was discovered automatically by eClustering using video streaming data



Evolving Fuzzy Systems, Figure 6

A predictive model of the characteristic temperature difference across a coil in a heat exchanger of an air-conditioning unit installed in a real building in Iowa, USA. **a** An off-line pre-trained model used in two different seasons (summer and spring); **b** Evolving FRB model trained and used during the summer (up to sample 1000) and then having its structure fixed during the spring season; **c** Evolving FRB model left to evolve (self-develop) during the whole period of usage (both spring and summer seasons)

Speech Signal Reconstruction

The ETS fuzzy system is used in [37] for error concealment in the next-generation Voice over Internet Protocol (VoIP) communication receivers. It is used in combination with parametric speech coders of analysis-by-synthesis-type. eTS MIMO [15] is used to predict the missing values of the linear spectral pairs (LSP) that will allow one

to reconstruct the lost in transmission packets. The eTS fuzzy model used ten inputs (current LSP parameter values) and ten outputs (predicted one step/20 milliseconds ahead LSP values). This research was a joint work between Lancaster University and Nokia-UK and aims the development of the next generation intelligent decoders at the receiver that will be able to conceal lost packets with a size of 80 to 160 ms without significant deterioration to the quality of service (QoS) in VoIP transmission.

Future Directions

The area of evolving fuzzy systems is in its infancy and has already demonstrated a remarkable success in addressing some of the most vibrating issues of the development, application and implementation of truly intelligent systems in a wide variety of branches of industry and real-life problems [11]. It opens the door for future developments that are related to the areas of autonomous systems, early cancer diagnosis and prognosis of the progression, even to the identification of structural changes in biological cells that correspond to the evolution of the disease. In the area of intelligent self-maintaining sensors the process industry can benefit from more flexible and smarter solutions. The problems that are yet to be addressed and can mark the future development of this vibrant area are; (1) collaboration aspects between two or more evolving fuzzy system-based intelligent systems (autonomous robots, intelligent sensors etc.); (2) further flexibility of the systems in terms of real-time self-analysis, optimal features and input selection, rule aggregation mechanism adaptation etc.; (3) even more flexible system structure architectures such as hierarchical, decentralized; (4) more robust learning algorithms that take care of missing data, different sampling intervals etc.

From a broader prospective, the future developments of this discipline will influence and are closely related to similar developments in the area of communication networks (self-adaptive networks [64]), self-validating soft sensors [61], autonomous aerial, ground-based, and underwater vehicles [20,40,49] etc. The area is closely related to the developments in the area of neural networks [34,48], so-called autonomous mental development [23] and cognitive psychology [55], mining data streams [28]. One can also expect more hardware implementations (the first hardware implementation of eClustering was reported in 2005 [8]). From the point of view of mathematical fundamentals and learning it is also closely related to adaptive filters theory [69] and the recent developments in particle filters [17] will certainly influence the future, more ef-

ficient techniques that will be developed in this emerging and highly potential branch of research.

Acknowledgments

The author would like to thank Mr. Xiaowei Zhou for his assistance in producing the illustrative material, Dr. Jose Macias Hernandez for kindly providing real data from the oil refinery CEPESA, Santa Cruz, Tenerife, Spain, Dr. Richard Buswell, Loughborough University and ASHRAE (RP-1020) for the real air conditioning data, and Dr. Edwin Lughofer from Johannes Kepler University of Linz, Austria for providing real data from car engines.

Bibliography

Primary Literature

- Andersen HC, Teng FC, Tsoi AC (1994) Single Net Indirect Learning Architecture. *IEEE Trans Neural Netw* 5:1003–1005
- Angelov P (2002) Evolving Rule-based Models: A Tool for Design of Flexible Adaptive Systems. Springer, Heidelberg
- Angelov P (2004) An Approach for Fuzzy Rule-base Adaptation using On-line Clustering. *Int J Approx Reason* 35(3): 275–289
- Angelov PP (2004) A Fuzzy Controller with Evolving Structure. *Inf Sci* 161:21–35
- Angelov P, Buswell R (2001) Evolving Rule-based Models: A Tool for Intelligent Adaptation. In: Proc of the Joint 9th IFSA World Congress and 20th NAFIPS Intern Conf, Vancouver, 25–28 July 2001. IEEE Press, USA, pp 1062–1066
- Angelov P, Buswell R (2002) Identification of Evolving Rule-based Models. *IEEE Trans Fuzzy Syst* 10(5):667–677
- Angelov P, Buswell R (2003) Automatic Generation of Fuzzy Rule-based Models from Data by Genetic Algorithms. *Inf Sci* 150(1/2):17–31
- Angelov P, Everett M (2005) EvoMap: On-Chip Implementation of Intelligent Information Modelling using EVolving MAPping. Lancaster University, Lancaster, pp 1–15
- Angelov P, Filev D (2004) An approach to on-line identification of evolving Takagi–Sugeno models. *IEEE Trans Syst Man Cybern part B Cybern* 34(1):484–498
- Angelov P, Filev D (2005) Simpl_eTS: A Simplified Method for Learning Evolving Takagi–Sugeno Fuzzy Models. In: Proc of The 2005 IEEE Intern. Conf. on Fuzzy Systems FUZZ-IEEE – 2005, Reno 2005, pp 1068–1073
- Angelov P, Filev D, Kasabov N, Cordon O (eds) (2006) Evolving Fuzzy Systems. Proc of the 2nd Int Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006. pp 1–350
- Angelov P, Kordon A, Zhou X (2008) Adaptive Inferential Sensors based on Evolving Fuzzy Models: An Industrial Case Study. *IEEE Trans Fuzzy Syst* (under review)
- Angelov P, Lughofer E, Klement PE (2005) Two Approaches for Data – Driven Design of Evolving Fuzzy Systems: eTS and FLEXFIS. In: Proc of The 2005 North American Fuzzy Information Processing Society, NAFIPS Annual Conference, Ann Arbor, June 2005, pp 31–35
- Angelov P, Victor J, Dourado A, Filev D (2004) On-line evolution of Takagi–Sugeno Fuzzy Models. In: Proc of the 2nd IFAC Workshop on Advanced Fuzzy and Neural Control, Oulu, 16–17 Sept 2004, pp 67–72
- Angelov P, Xydeas C, Filev D (2004) On-line Identification of MIMO Evolving Takagi–Sugeno Fuzzy Models. In: Proc of the Intern. Joint Conf. on Neural Networks and Intern. Conf. on Fuzzy Systems, IJCNN-FUZZ-IEEE, Budapest, 25–29 July 2004, pp 55–60
- Angelov P, Zhou X, Klawonn F (2007) Evolving Fuzzy Rule-based Classifiers. In: Proc of the First 2007 IEEE International Conference on Computational Intelligence Applications for Signal and Image Processing – a part of the IEEE Symposium Series on Computational Intelligence, SSCI-2007, Honolulu, 1–5 April 2007, pp 220–225
- Angelov P, Zhou X, Lughofer E, Filev D (2007) Architectures of Evolving Fuzzy Rule-based Classifiers. In: Proc of the 2007 IEEE International Conference on Systems, Man, and Cybernetics, Montreal, 7–10 Oct 2007, pp 2050–2055
- Arulampalam MS, Maskell S, Gordon N (2002) A Tutorial on Particle Filters for On-line Nonlinear Non-Gaussian Bayesian Tracking. *IEEE Trans Signal Process* 50(2):174–188
- Astroem KJ, Wittenmark B (1994) Adaptive Control. Prentice Hall, Upper Saddle River
- Azimi-Sadjadi MR, Yao D, Jamshidi AA, Dobeck GJ (2002) Underwater Target Classification in Changing Environments Using an Adaptive Feature Mapping. *IEEE Trans Neural Netw* 13(5):1099–1111
- Badami VV, Chbat NW (1998) Home appliances get smart. *IEEE Spectrum* 35(8):36–43
- Bezdek J (1974) Cluster Validity with Fuzzy Sets. *J Cybern* 3(3):58–71
- Bonarini A, Lazaric A, Restelli M, Vitali P (2006) Self-Development Framework for Reinforcement Learning Agents. In: Proc of the 5th Intern. Conf. on Development and Learning, ICDL-06 New Delhi
- Carline D, Angelov PP, Clifford R (2005) Agile Collaborative Autonomous Agents for Robust Underwater Classification Scenarios. In: the Proceedings of the Underwater Defense Technology Conference, Amsterdam, June 2005
- Carpenter GA, Grossberg S (2003) Adaptive Resonance Theory. In: Arbib MA (ed) *The Handbook of Brain Theory and Neural Networks*, 2nd edn. MIT Press, Cambridge, pp 87–90
- Chiu SL (1994) Fuzzy model identification based on cluster estimation. *J Intell Fuzzy Syst* 2:267–278
- Cordon O, Gomide F, Herrera F, Hoffmann F, Magdalena L (2004) Ten years of genetic fuzzy systems: Current framework and new trends. *Fuzzy Sets Syst* 141(1):5–31
- Domingos P, Hulten G (2001) Catching up with the data: Research issues in mining data streams. In: Proc of the Workshop on Research Issues in Data Mining and Knowledge Discovery, Santa Barbara
- Filev D, Larson T, Ma L (2000) Intelligent Control for Automotive Manufacturing – Rule-based Guided Adaptation. In: Proc of the IEEE Conference on Industrial Electronics, IECON-2000, Nagoya Oct 2000, pp 283–288
- Filev D, Tseng F (2006) Novelty detection-based Machine Health Prognostics. In: Proc of the 2006 Int Symposium on Evolving Fuzzy Systems. IEEE Press, USA, pp 193–199
- Fortuna L, Graziani S, Rizzo A, Xibilia MG (2007) Soft sensors for Monitoring and Control of Industrial Processes. Springer, London

32. Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
33. Hastie T, Tibshirani R, Friedman J (2001) The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, Heidelberg
34. Hornby AS (1974) Oxford Advance Learner's Dictionary. Oxford University Press, Oxford
35. Huang G-B, Saratchandran P, Sundarajan N (2005) A generalized growing and pruning RBF (GGAP – RBF) neural network for function approximation. *IEEE Trans Neural Netw* 16(1):57–67
36. Ishibuchi H, Nakashima T, Nii M (2004) Classification and Modeling with Linguistic Granules: Advanced Information Processing. Springer, Berlin
37. Jones E, Angelov P, Xydeas C (2006) Recovery of LSP Coefficients in VoIP Systems using Evolving Takagi–Sugeno Fuzzy MIMO Models. In: Proc of the 2006 Intern. Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006, pp 208–214
38. Kailath T, Sayed AH, Hassibi B (2000) Linear Estimation. Prentice Hall, Upper Saddle River
39. Kalman RE (1960) A New Approach to linear filtering and prediction problem. *Transactions of the American Society of Mechanical Engineering, ASME, Ser. D. J Basic Eng* 8:34–45
40. Kanakakis V, Valavanis KP, Tsourveloudis NC (2004) Fuzzy-Logic Based Navigation of Underwater Vehicles. *J Intell Robot Syst* 40:45–88
41. Kasabov N (2001) Evolving fuzzy neural networks for on-line supervised/unsupervised, knowledge-based learning. *IEEE Trans. on Systems, Man and Cybernetics – part B. Cybernetics* 31:902–918
42. Kasabov N, Song Q (2002) DENFIS: Dynamic Evolving Neural-Fuzzy Inference System and Its Application for Time-Series Prediction. *IEEE Trans Fuzzy Syst* 10(2):144–154
43. Klir G, Folger T (1988) Fuzzy Sets, Uncertainty and Information. Prentice Hall, Englewood Cliffs
44. Kohonen T (1995) Self-Organizing Maps. Series in Inf Sci, vol 30. Springer, Heidelberg
45. Kordon A (2006) Inferential Sensors as Potential Application Area of Intelligent Evolving Systems. 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 September 2006, key note presentation
46. Kuncheva L (2000) Fuzzy Classifiers. Physica, Heidelberg
47. Leng G, McGuinty TM, Prasad G (2005) An approach for on-line extraction of fuzzy rules using a self-organizing fuzzy neural network. *Fuzzy Sets Syst* 150(2):211–243
48. Lin F-J, Lin C-H, Shen P-H (2001) Self-constructing fuzzy neural network speed controller for permanent-magnet synchronous motor drives. *IEEE Trans Fuzzy Syst* 9(5):751–759
49. Liu PX, Meng MQ-X (2004) On-line Data-Driven Fuzzy Clustering with Applications to Real-time Robotic Tracking. *IEEE Trans Fuzzy Syst* 12(4):516–523
50. Ljung L (1987) System Identification: Theory for the User. Prentice-Hall, New Jersey
51. Lughofer E, Angelov P, Zhou X (2007) Evolving Single- and Multi-Model Fuzzy Classifiers with FLEXFIS-Class. In: Proc of the 2007 IEEE International Conference on Fuzzy Systems, London, 23–26 July 2007, pp 363–368
52. Macias J, Angelov P, Zhou X (2006) Predicting quality of the crude oil distillation using evolving Takagi–Sugeno fuzzy models. In: Proc of the 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006, pp 201–207
53. Macias-Hernandez JJ, Angelov P, Zhou X (2007) Soft Sensor for Predicting Crude Oil Distillation Side Streams using Takagi Sugeno Evolving Fuzzy Models. In: Proc of the 2007 IEEE Int Conf on Syst, Man, and Cybernetics, Montreal, 7–10 Oct. 2007, pp 3305–3310
54. Mamdani EH, Assilian S (1975) An Experiment in Linguistic Synthesis with a Fuzzy Logic Controller. *Int J Man-Mach Stud* 7:1–13
55. Massaro DW (1991) Integration versus Interactive Activation: The Joint Influence of Stimulus and Context in Perception. *Cogn Psychol* 23:558–614
56. Memon MA, Angelov P, Ahmed H (2006) An Approach to Real-Time Color-based Object Tracking. In: Proc 2006 International Symposium on Evolving Fuzzy Systems, Ambleside, 7–9 Sept 2006, pp 81–87
57. Nauck D, Kruse R (1997) A Neuro-fuzzy method to learn fuzzy classification rules from data. *Fuzzy Sets Syst* 89:277–288
58. Pioneer-3DX (2004) User Guide. ActiveMedia Robotics, Amherst
59. Procyk TJ, Mamdani EH (1979) A linguistic self-organizing process controller. *Automatica* 15:15–30
60. Psaltis D, Sideris A, Yamamura AA (1988) A Multilayered Neural Network Controller. *IEEE Trans Control Syst Manag* 8:17–21
61. Qin SJ, Yue H, Dunia R (1997) Self-validating inferential sensors with application to air emission monitoring. *Ind Eng Chem Res* 36:1675–1685
62. Setnes M, Roubos H (2000) Ga-fuzzy modeling and classification: complexity and performance. *IEEE Trans Fuzzy Syst* 8(5):509–522
63. Shimojima K, Fukuda T, Hashegawa Y (1995) Self-Tuning Modeling with Adaptive Membership Function, Rules, and Hierarchical Structure based on Genetic Algorithm. *Fuzzy Sets Syst* 71:295–309
64. Sifalakis M, Hutchison D (2004) From Active Networks to Cognitive Networks. In: Proc of the ICRC Dagstuhl Seminar 04411 on Service Management and Self-Organization in IP-based Networks. Waden, October 2004
65. Takagi T, Sugeno M (1985) Fuzzy identification of systems and its application to modeling and control. *IEEE Trans Syst Man Cybern B – Cybern* 15:116–132
66. Valavanis K (2006) Unmanned Vehicle Navigation and Control: A Fuzzy Logic Perspective. In: Proc of the 2006 International Symposium on Evolving Fuzzy Systems. Ambleside, 7–9 Sept. 2006, pp 200–207
67. Vapnik VN (1998) The Statistical Learning Theory. Springer, Berlin
68. Wang L-X (1992) Fuzzy Systems are Universal Approximators. In: Proc of the First IEEE International Conference on Fuzzy Systems, FUZZ-IEEE – 1992, San Diego, pp 1163–1170
69. Widrow B, Stearns S (1985) Adaptive Signal Processing. Prentice Hall, Englewood Cliffs
70. Xydeas C, Angelov P, Chiao S, Reoullas M (2006) Advances in EEG Signals Classification via Dependant HMM models and Evolving Fuzzy Classifiers. *Int J Comput Biol Medicine, special issue on Intell Technol Bio-Inform Medicine* 36(10):1064–1083
71. Yager R (2006) Learning Methods for Intelligent Evolving Systems. In: Proc 2006 International Symposium on Evolving Fuzzy Systems. Ambleside, 7–9 Sept. 2006, pp 3–7

72. Yager RR, Filev DP (1993) Learning of Fuzzy Rules by Mountain Clustering. In: Proc of the SPIE Conf. on Application of Fuzzy Logic Technology, Boston, pp 246–254
73. Yager RR, Filev DP (1994) Essentials of Fuzzy Modeling and Control. Wiley, New York
74. Zadeh LA (1993) Soft Computing. Introductory Lecture for the 1st European Congress on Fuzzy and Intelligent Technologies EUFIT'93, Aachen, pp vi–vii
75. Zhou X-W, Angelov P (2006) Real-Time joint Landmark Recognition and Classifier Generation by an Evolving Fuzzy System. In: Proc of the 2006 IEEE World Congress on Computational Intelligence, WCCI-2006, Vancouver, 16–21 July 2006, pp 6314–6321
76. Zhou X, Angelov P (2007) An approach to autonomous self-localization of a mobile robot in completely unknown environment using evolving fuzzy rule-based classifier. In: Proc of the First 2007 IEEE Int Symposium on Computational Intelligence Applications for Defense and Security – a part of the IEEE Symposium Series on Computational Intelligence, SSCI-2007, Honolulu, 1–5 April 2007, pp 131–138
- Juang C-F, Lin X-T (1999) A recurrent self-organizing neural fuzzy inference network. *IEEE Trans Neural Netw* 10:828–845
- Kasabov N (2006) Adaptation and Interaction in Dynamical Systems: Modelling and Rule Discovery Through Evolving Connectionist Systems. *Appl Soft Comput* 6(3):307–322
- Kasabov N (2006) Evolving connectionist systems: Brain-, gene-, and, quantum inspired computational intelligence. Springer, London
- Kasabov N, Chan Z, Song Q, Greer D (2005) Evolving neuro-fuzzy systems with evolutionary parameter self-optimisation. In: Do Adaptive Smart Systems exist? Series Study in Fuzziness, vol 173. Physica, Heidelberg
- Kim K, Baek J, Kim E, Park M (2005) TSK Fuzzy model based on-line identification. In: Proc of the 11th International Fuzzy Systems Association, IFSA World Congress, Beijing, pp 1435–1439
- Klinkenberg R, Joachims T (2000) Detection concept drift with support vector machines. Proc of the 7th International Conference on Machine Learning (ICML). Morgan Kaufman, Stanford University, pp 487–494
- Marin-Blazquez JG, Shen Q (2002) From approximative to descriptive fuzzy classifiers. *IEEE Trans Fuzzy Syst* 10(4):484–497
- Marshall MR, Song Q, Ma TM, MacDonell S, Kasabov N (2005) Evolving Connectionist System versus Algebraic Formulae for Prediction of Renal Function from Serum Creatinine. *Kidney Int* 6:1944–1954
- Ozawa S, Pang S, Kasabov N (2004) A Modified Incremental Principal Component Analysis for On-Line Learning of Feature Space and Classifier. *Lecture Notes in Artificial Intelligence LNAI*, vol 3157. Springer, Berlin, pp 231–240
- Pang S, Ozawa S, Kasabov N (2005) Incremental Linear Discriminant Analysis for Classification of Data Streams. *IEEE Trans Syst Man Cybern B – Cybern* 35(5):905–914
- Plat J (1991) A resource allocation network for function interpolation. *Neural Comput* 3(2):213–225
- Widmer G, Kubat M (1996) Learning in the presence of concept drift and hidden contexts. *Mach Learn* 23(1):69–101

Books and Reviews

- Angelov P, Xydeas C (2006) Fuzzy Systems Design: Direct and Indirect Approaches. *Int J Soft Comput*, special issue on New Trends in Fuzzy Modeling part I: Novel Approaches 10(9): 836–849
- Angelov P, Zhou X (2006) Evolving fuzzy systems from data streams in real-time. *Proc 2006 International Symposium on Evolving Fuzzy Systems*, Ambleside, 7–9 Sept. 2006, pp 29–35
- Bentley PJ (2000) Evolving Fuzzy Detectives: An Investigation into the Evolution of Fuzzy Rules. In: Suzuki, Roy, Ovasko, Furuhashi, Dote (eds) *Soft Computing in Industrial Applications*. Springer, London
- Chan Z, Kasabov N (2004) Evolutionary computation for on-line and off-line parameter tuning of evolving fuzzy neural networks. *Int J Comput Intell Appl* 4(3):309–319
- Fayyad UM, Piatetsky-Shapiro G, Smyth P (1996) *From Data Mining to Knowledge Discovery: An Overview*, Advances in Knowledge Discovery and Data Mining. MIT Press, Boston
- Fritzke B (1994) Growing cell structures – a self-organizing network for unsupervised and supervised learning. *Neural Netw* 7(9):1441–1460
- Futschik M, Reeve A, Kasabov N (2003) Evolving connectionist systems for knowledge discovery from gene expression data of cancer tissue. *Artif Intell Med* 28:165–189
- Hopner F, Klawonn F (2000) Obtaining interpretable fuzzy models from fuzzy clustering and fuzzy regression. In: Proc of the 4th Intern Conf on Knowledge-based Intelligent Engineering Systems (KES), Brighton, pp 162–165
- Huang G-B, Saratchandran P, Sundarajan N (2005) A generalized growing and pruning RBF (GGAP-RBF) neural network for function approximation. *IEEE Trans Neural Netw* 16(1):57–67
- Huang L, Song Q, Kasabov N (2005) Evolving Connectionist Systems Based Role Allocation of Robots for Soccer Playing. In: Proc of the Joint 2005 International Symposium on Intelligent Control and 13th Mediterranean Conference on Control and Automation, ISIC – MED – 2005, Limassol, 27–29 June 2005
- Jang JSR (1993) ANFIS: Adaptive Network-based Fuzzy Inference Systems. *IEEE Trans. on Syst Man Cybern B – Cybern* 23(3):665–685

Existence and Uniqueness of Solutions of Initial Value Problems

GIANNE DERKS
Department of Mathematics,
University of Surrey, Guildford, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Existence
Uniqueness
Continuous Dependence on Initial Conditions
Extended Concept of Differential Equation
Further Directions
Bibliography

Glossary

Ordinary differential equation An ordinary differential equation is a relation between a (vector) function $u: I \rightarrow \mathbb{R}^m$, (I an interval in \mathbb{R} , $m \in \mathbb{N}$) and its derivatives. A function u , which satisfies this relation, is called a solution.

Initial value problem An initial value problem is an ordinary differential equation with a prescribed value for the solution at one instance of its variable, often called the initial time. The initial value is the pair consisting of the initial time and the prescribed value of the solution.

Vector field For an ordinary differential equation of the form $\frac{du}{dt} = f(t, u)$, the function $f: \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ is called the vector field. Thus, at a solution, the vector field gives the tangent to the solution.

Flow map The flow map describes the solution of an ordinary differential equation for varying initial values. Hence, for an ordinary differential equation $\frac{du}{dt} = f(t, u)$, the flow map is a function $\Phi: \mathbb{R} \times \mathbb{R} \times \mathbb{R}^m \rightarrow \mathbb{R}^m$ such that $\Phi(t, t_0, u_0)$ is a solution of the ordinary differential equation, starting at $t = t_0$ in $u = u_0$.

Functions: bounded, continuous, uniformly Lipschitz continuous Let D be a connected set in \mathbb{R}^k with $k \in \mathbb{N}$ and let $f: D \rightarrow \mathbb{R}^m$ be a function on D :

- The function f is *bounded* if there is some $M > 0$ such that $|f(x)| \leq M$ for all $x \in D$.
- The function f is *continuous* on D if $\lim_{x \rightarrow x_0} f(x) = f(x_0)$, for every $x_0 \in D$. A continuous function on a bounded and closed set D is bounded. The function f is *equicontinuous* or *uniform continuous* if the convergence of the limits in the definition of continuity is uniform for all $x_0 \in D$, i. e., for every $\varepsilon > 0$, there is a $\delta > 0$, such that for all $x_0 \in D$ and all $x \in D$ with $|x - x_0| < \delta$ it holds that $|f(x) - f(x_0)| < \varepsilon$. A continuous function on a compact interval is equicontinuous.
- Let $D \subset \mathbb{R} \times \mathbb{R}^m$. The function $f: D \rightarrow \mathbb{R}^m$ is *uniformly Lipschitz continuous on D with respect to its second variable*, if f is continuous on D and there exists some constant $L > 0$ such that

$$|f(t, u) - f(t, v)| \leq L|u - v|, \\ \text{for all } (t, u), (t, v) \in D.$$

The constant L is called the *Lipschitz constant*.

Pointwise and uniform convergence A sequence of functions $\{u_n\}$ with $u_n: I \rightarrow \mathbb{R}^m$ is Pointwise con-

vergent if $\lim_{n \rightarrow \infty} u_n(t)$ exists for every $t \in I$. The sequence of functions $\{u_n\}$ is uniform convergent with limit function $u: I \rightarrow \mathbb{R}^m$ if $\lim_{n \rightarrow \infty} \sup\{|u - u_n| \mid t \in I\} = 0$. A sequence of pointwise convergent, equicontinuous functions is uniform convergent and the limit function is equicontinuous.

Notation

\dot{u}	Derivative of u , i. e., $\frac{du}{dt}$
$u^{(k)}$	The k th derivative of u , i. e., $\frac{d^k u}{dt^k}$
$I_a(t_0)$	The closed interval $[t_0, t_0 + a]$
$B_b(u_0)$	The closed ball with radius b about u_0 in \mathbb{R}^m , i. e., $B_b(u_0) := \{u \in \mathbb{R}^m \mid u - u_0 \leq b\}$
$\ u\ _\infty$	The supremum norm for a bounded function $u: I \rightarrow \mathbb{R}^m$, i. e., $\ u\ _\infty = \sup\{ u(t) \mid t \in I\}$

Definition of the Subject

Many problems in physics, engineering, biology, economics, etc., can be modeled as relations between observables or states and their derivatives, hence as differential equations. When only derivatives with respect to one variable play a role, the differential equation is called an ordinary differential equation. The field of differential equations has a long history, starting with Newton and Leibniz in the seventeenth century. In the beginning of the study of differential equations, the focus is on finding explicit solutions as the emphasis is on solving the underlying physical problems. But soon one starts to wonder: If a starting point for a solution of a differential equation is given, does the solution always exist? And if such a solution exists, how long does it exist and is there only one such solution? These are the questions of existence and uniqueness of solutions of initial value problems. The first existence result is given in the middle of the nineteenth century by Cauchy. At the end of the nineteenth century and the beginning of the twentieth century, substantial progress is made on the existence and uniqueness of solutions of initial value problems and currently the heart of the topic is quite well understood. But there are many open questions as soon as one considers delay equations, functional differential equations, partial differential equations or stochastic differential equations. Another area of intensive current research, which uses the existence and uniqueness of differential equations, is the area of finite- and infinite-dimensional dynamical systems.

Introduction

As indicated before, an initial value problem is the problem of finding a solution of an ordinary differential equation with a given initial condition. To be precise, let

D be an open, connected set in $\mathbb{R} \times \mathbb{R}^m$. Given are a function $f: D \rightarrow \mathbb{R}^m$ and a point $(t_0, u_0) \in D$. The initial value problem is the problem of finding an interval $I \subset \mathbb{R}$ and a function $u: I \rightarrow \mathbb{R}^m$ such that $t_0 \in I$, $\{(t, u(t)) \mid t \in I\} \subset D$ and

$$\frac{du}{dt} = f(t, u(t)), \quad t \in I, \quad \text{with } u(t_0) = u_0. \quad (1)$$

The function f is called the vector field and the point (t_0, u_0) the initial value. One often writes the derivative as $\frac{du}{dt} = \dot{u}$. If f is continuous, then the initial value problem (1) is equivalent to finding a function $u: I \rightarrow \mathbb{R}^m$ such that

$$u(t) = u_0 + \int_{t_0}^t f(\tau, u(\tau)) d\tau, \quad \text{for } t \in I. \quad (2)$$

If f is continuous and a solution exists, then clearly this solution is continuously differentiable on I .

It might seem restrictive to only consider first-order systems; however, most higher-order equations can be written as a system of first-order equations. Consider for example the n th order differential equation for $v: I \rightarrow \mathbb{R}$

$$v^{(n)}(t) = F(t, v, \dot{v}, \dots, v^{(n-1)}),$$

with initial conditions

$$v^{(k)}(t_0) = v_k, \quad k = 0, \dots, n-1.$$

This can be written as a first-order system by using the vector function $u: I \rightarrow \mathbb{R}^m$ defined as $u = (v, \dot{v}, \dots, v^{(n-1)})$. The equivalent initial value problem for u is

$$\dot{u} = f(t, u) = \begin{pmatrix} u_2 \\ \vdots \\ u_n \\ F(t, u_1, u_2, \dots, u_n) \end{pmatrix},$$

$$\text{with } u(t_0) = \begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix}.$$

Obviously, this system is not a unique representation, there are many other ways of obtaining first-order systems from the n th-order problem.

In the beginning of the study of differential equations, the focus is on finding explicit solutions. The first existence theorem is by Cauchy [3] in the middle of the nineteenth century and an initial value problem is also called a *Cauchy problem*. At the end of the nineteenth century

substantial progress is made on the existence of solutions of an initial value problem when Peano [20] shows that if f is continuous, then a solution exists near the initial value (t_0, u_0) , i. e., there is local existence of solutions. Global existence of solutions of initial value problems needs more than smoothness, as is illustrated by the following example.

Example 1 Consider the initial value problem on $D = \mathbb{R} \times \mathbb{R}$ given as

$$\dot{u} = u^2 \quad \text{and} \quad u(0) = u_0$$

for some $u_0 > 0$. As can be verified easily, the solution is $u(t) = u_0/(1 - tu_0)$, for $t \in (-\infty, 1/u_0)$. This solution cannot be extended for $t \geq 1/u_0$, even though the vector field $f(t, u) = u^2$ is infinitely differentiable on the full domain. As can be seen from later theorems, the lack of global existence is related to the fact that the vector field f is unbounded.

Once existence of solutions of initial value problems is known, the next question is if such a solution is unique. As can be seen from the following example, continuity of f is not sufficient for uniqueness of solutions.

Example 2 Consider the following initial value problem on $D = \mathbb{R} \times \mathbb{R}$:

$$\dot{u} = |u|^\alpha \quad \text{and} \quad u(0) = 0.$$

If $0 < \alpha < 1$, then there is an infinite number of solutions. Two obvious solutions for $t \in [0, \infty)$ are $\bar{u}(t) = 0$ and $\hat{u}(t) = ((1 - \alpha)t)^{\frac{1}{1-\alpha}}$. But these solutions are members of a large family of solutions. For any $c \geq 0$, the following functions are solutions for $I = \mathbb{R}$:

$$u_c(t) = \begin{cases} 0, & t < c \\ ((1 - \alpha)(t - c))^{\frac{1}{1-\alpha}}, & t \geq c. \end{cases}$$

A sufficient condition for uniqueness of solutions of the initial value problem is uniform Lipschitz continuity of the vector field f in its second variable u . Although the ideas behind this theorem go back to Cauchy and Lipschitz, Picard [21] and Lindelöf [18] are usually credited with this result. They used the so-called method of successive approximations to prove the uniqueness result; see Sect. “[Uniqueness](#)” for details.

Having determined existence and uniqueness, the next issue is the relation between a solution and its initial value. This is the topic of Sect. “[Continuous Dependence on Initial Conditions](#)”. The concept of a flow map is first introduced. Roughly speaking, the flow map is the solution,

with the initial value as an extra variable. If the vector field is uniformly Lipschitz continuous in its second variable, then the flow map is continuous. Thus, a small change in the initial value, will only give a small change in the solution (locally).

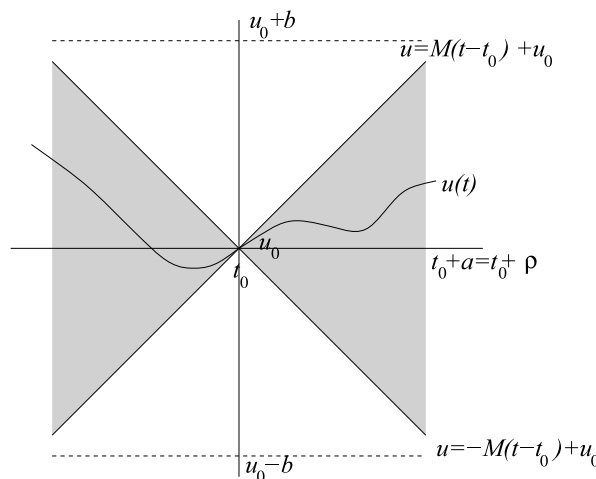
The concept of an initial value problem can be extended to noncontinuous vector fields f and existence theorems can be shown if the vector fields satisfy the so-called Carathéodory conditions. More details can be found in Sect. “[Extended Concept of Differential Equation](#)”. This chapter ends with a discussion of initial value problems in more general differential equations and the use of existence and uniqueness of solutions of initial value problems in dynamical systems.

Existence

Cauchy [3] seems to have been the first one to publish an existence result for differential equations, using an approximation of solutions by joined line segments. This work was extended considerably by Peano [20] in 1890.

Theorem 1 (Cauchy–Peano Theorem) *Let $a, b > 0$, define $I_a(t_0) := [t_0, t_0 + a]$ and $B_b(u_0) := \{u \in \mathbb{R}^m \mid |u - u_0| \leq b\}$ and assume that the cylinder $S := I_a(t_0) \times B_b(u_0) \subset D$. Let the vector field f be a continuous function on the cylinder S . This implies that f is bounded on S , say, $M = \max\{|f(t, u)| \mid (t, u) \in S\}$.*

Define the parameter $\rho = \min\left(a, \frac{b}{M}\right)$. Then there exists a solution $u(t)$, with $t_0 \leq t \leq t_0 + \rho$, which solves the initial value problem (1).



The theorem can also be phrased for an interval $[t_0 - \rho, t_0]$ or $[t_0 - \rho, t_0 + \rho]$. The parameter ρ gives a time interval such that the solution $(t, u(t))$ is guaranteed to be within S . From the integral formulation of the initial value problem (2), it follows that

$$|u(t) - u_0| \leq \int_{t_0}^t |f(\tau, u(\tau))| d\tau \leq M(t - t_0).$$

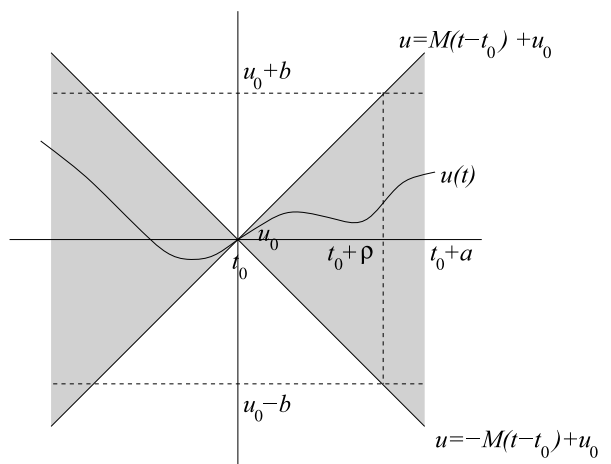
To guarantee that $u(t)$ is within S , the condition $t - t_0 \leq \rho$ is sufficient as it gives $M(t - t_0) \leq b$ and $t - t_0 \leq a$. In Fig. 1, this is illustrated in the case $D \subset \mathbb{R} \times \mathbb{R}$.

To prove the Cauchy–Peano existence theorem, a construction which goes back to Cauchy is used. This so-called Cauchy–Euler construction of approximate solutions uses joined line segments which are such that the tangent of each line segment is given by the vector field evaluated at the start of the line segment. To be specific, for any $N \in \mathbb{N}$, the interval $I_\rho(t_0) = [t_0, t_0 + \rho]$ is divided in N equal parts with intermediate points $t_k := t_0 + \frac{k}{N}\rho$, $k = 1, \dots, N$. The approximate solution is the function $u_N: I_\rho(t_0) \rightarrow \mathbb{R}$, with

$$\begin{aligned} u_N(t_0) &= u_0, \\ u_N(t) &= u_N(t_{k-1}) + f(t_{k-1}, u_N(t_{k-1}))(t - t_{k-1}), \quad (3) \\ t_{k-1} &< t \leq t_k, \quad k = 1, \dots, N, \end{aligned}$$

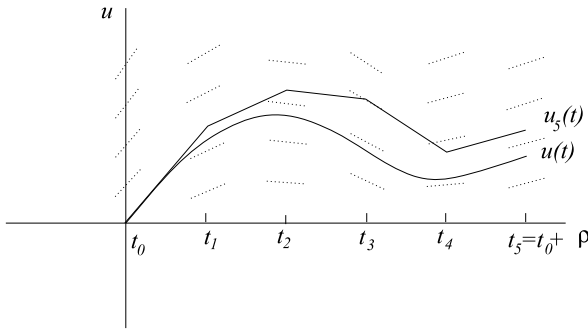
see also Fig. 2. One sees immediately that function u_N is continuous and piecewise differentiable and, using the bound on the vector field f ,

$$|u_N(t) - u_N(s)| \leq M|t - s|, \quad \text{for all } t, s \in I_\rho(t_0). \quad (4)$$



Existence and Uniqueness of Solutions of Initial Value Problems, Figure 1

The parameter $\rho = \min\left(\frac{b}{M}, a\right)$ guarantees that the solution is within the shaded area S , on the left in the case $a < \frac{b}{M}$ and on the right if $a > \frac{b}{M}$.



Existence and Uniqueness of Solutions of Initial Value Problems, Figure 2

The approximate solution u_N , with $N = 5$, and the solution $u(t)$. The dotted lines show the tangent given by the vector field $f(t, u)$ at the points (t_i, u) for various values of u and $i = 0, \dots, 5$

This estimate implies that the sequence of functions $\{u_N\}$ is uniformly bounded and that the functions u_N are equicontinuous. Indeed, the estimate above gives that for any $t \in I_\rho(t_0)$ and any $N \in \mathbb{N}$,

$$|u_N(t)| \leq |u_N(t_0)| + M|t - t_0| \leq |u_0| + M\rho.$$

And equicontinuity follows immediately from (4) as M does not depend on N .

Arzela–Ascoli’s lemma states that a sequence of functions which is uniformly bounded and equicontinuous on a compact set has a subsequence which is uniformly convergent on this compact set [5]. As the sequence $\{u_N\}$ is uniformly bounded and equicontinuous on the compact set $I_\rho(t_0)$, it follows immediately that it has a convergent subsequence. This convergent subsequence is denoted by u_{N-k} and its limit by u . To prove the Cauchy–Peano theorem, we will show that this limit function u satisfies the integral equation (2) and hence is a solution of the initial value problem (1).

Proof First it will be shown that if N is sufficiently large, then the functions u_N are close to solutions of the differential equation in the following sense: for all $\varepsilon > 0$, there is some N_0 such that for all $N > N_0$

$$|\dot{u}_N(t) - f(t, u_N(t))| < \varepsilon, \quad t \in I_\rho(t_0), \quad t \neq t_k, \quad k = 0, \dots, N. \quad (5)$$

Let $\varepsilon > 0$. As f is a continuous function on the compact set S , it follows that f is uniform continuous. Thus, there is some δ_ε such that for all $(t, u), (s, v) \in S$ with $|t - s| + |u - v| < \delta_\varepsilon$

$$|f(t, u) - f(s, v)| < \varepsilon. \quad (6)$$

Now define $N_0 = \left\lceil \frac{(M+1)\rho}{\delta_\varepsilon} \right\rceil$ and let $N > N_0$. Then for any $t \in I_\rho(t_0)$, $t \neq t_k$, $k = 0, \dots, N$, we have $\dot{u}_N(t) = f(t_l, u_N(t_l))$, where l is such that $t_l < t < t_{l+1}$. Hence, (6) gives

$$|\dot{u}_N(t) - f(t, u_N(t))| = |f(t_l, u_N(t_l)) - f(t, u_N(t))| < \varepsilon$$

as (4) shows that $|t_l - t| + |u_N(t_l) - u_N(t)| < (1 + M)|t_l - t| < \frac{(1+M)\rho}{N} \leq \frac{(1+M)\rho}{N_0} \leq \delta_\varepsilon$.

Next it will be shown that (5) implies that the functions u_N almost satisfy the integral equation (2) for N sufficiently large. From (5), it follows that $f(t, u_N(t)) - \varepsilon < \dot{u}_N(t) < f(t, u_N(t)) + \varepsilon$ for all $N > N_0$ and all $t \in I_\rho(t_0)$, except at the special points t_k , $k = 0, \dots, N$. Hence, for any $k = 1, \dots, N$ and all $t_{k-1} < t \leq t_k$, this gives

$$\begin{aligned} u_N(t) &< u_N(t_{k-1}) + \int_{t_{k-1}}^t (f(\tau, u_N(\tau)) + \varepsilon) d\tau \\ &\leq u_N(t_{k-1}) + \int_{t_{k-1}}^t f(\tau, u_N(\tau)) d\tau + \frac{\varepsilon\rho}{N}. \end{aligned}$$

Thus, also for any $k = 1, \dots, N$, $u_N(t_k) - u_N(t_{k-1}) < \int_{t_{k-1}}^{t_k} f(\tau, u_N(\tau)) d\tau + \frac{\varepsilon\rho}{N}$ and hence

$$\begin{aligned} u_N(t_k) - u_N(t_0) &= \sum_{j=1}^k [u_N(t_j) - u_N(t_{j-1})] \\ &< \int_{t_0}^{t_k} f(\tau, u_N(\tau)) d\tau + \varepsilon\rho. \end{aligned}$$

Combination of the last two results gives that for any $t \in I_\rho(t_0)$

$$u_N(t) < u_N(t_0) + \int_{t_0}^t (f(\tau, u_N(\tau)) + \varepsilon) d\tau + \varepsilon\rho.$$

In a similar way it can be shown that $u_N(t) > u_N(t_0) + \int_{t_0}^t f(\tau, u_N(\tau)) d\tau - \rho\varepsilon$ and hence for all $N > N_0$, the following holds:

$$\left| u_N(t) - u_N(t_0) - \int_{t_0}^t f(\tau, u_N(\tau)) d\tau \right| < \varepsilon\rho, \quad \text{for all } t \in I_\rho(t_0). \quad (7)$$

Thus, the function u_N satisfies the integral equation (2) up to an order ε error if N is sufficiently large.

As the subsequence u_{N_k} converges uniformly to u and f is continuous, this implies that $\int_{t_0}^t f(\tau, u_{N_k}(\tau)) d\tau \rightarrow \int_{t_0}^t f(\tau, u(\tau)) d\tau$. Thus, from (7), it can be concluded that u satisfies the integral equation (2) exactly. \square

Note that the full sequence $\{u_N\}$ does not necessarily converge. A counterexample can be found in exercise 12 in Chap. 1 of [5]. It is based on the initial value problem $\dot{u} = |u|^{\frac{1}{4}} \operatorname{sgn}(u) + t \sin(\pi/t)$ with $u(0) = 0$, and shows that on a small interval near $t_0 = 0$, the even approximations are bounded below by a strictly positive constant, while the odd approximations are bounded above by a strictly negative constant.

However, if it is known that the solution of the initial value problem is unique, then the full sequence of approximate solutions $\{u_N\}$ must converge to the solution. This can be seen by a contradiction argument. If there is a unique solution, but the sequence $\{u_N\}$ is not convergent, then there is a convergent subsequence and a remaining set of functions which does not converge to the unique solution. But Arzela–Ascoli’s lemma can be applied to the remaining set as well; thus, there exists a convergent subsequence in the remaining set, which must converge to a different solution. This is not possible as the solution is unique.

In case of the initial value problem as presented in Example 2, the sequence of Cauchy–Euler approximate solutions $\{u_N\}$ converges to the zero function; hence, no subsequence is required to obtain a solution. As there are many solutions of this initial value problem, this implies that not all solutions of the initial value problem in Example 2 solutions can be approximated by the Cauchy–Euler construction of approximate solutions as defined in (3). As indicated in exercise 2.2 in Chap. 2 of [13], it is possible to get any solution by using appropriate Lipschitz continuous approximations of the vector field f , instead of the original vector field itself, in the Cauchy–Euler construction. In the example below this is illustrated for the vector field $f(t, u) = \sqrt{|u|}$.

Example 3 Let $f: \mathbb{R} \times \mathbb{R}$ be defined as $f(t, u) = \sqrt{|u|}$. First we define the approximate functions $\hat{f}_n: [0, 2] \times [0, \infty) \rightarrow [0, \infty)$ as

$$\hat{f}_n(t, u) = \begin{cases} \sqrt{u}, & u > \frac{1}{n} \\ \frac{1}{\sqrt{n}} - \frac{1}{2} \sqrt{n} \left(\frac{1}{n} - u \right), & u \leq \frac{1}{n} \end{cases}$$

and the definition of the Euler–Cauchy approximate functions is modified to

$$u_N(t) = u_N(t_{k-1}) + \hat{f}_N(t_{k-1}, u_N(t_{k-1}))(t - t_{k-1}), \\ t_{k-1} < t \leq t_k, \quad k = 1, \dots, N.$$

This sequence converges to the solution $u(t) = 4t^2$ for $t \in [0, 2]$.

Next define the approximate functions $\tilde{f}_n: [0, 2] \times [0, \infty) \rightarrow [0, \infty)$ as

$$\tilde{f}_n(t, u) = \begin{cases} \sqrt{u}, & t \leq 1, \quad u > \frac{1}{n}, \\ \frac{1}{2n} - \left(\frac{1}{2n} - u \right) + a_2 \left(\frac{1}{2n} - u \right)^2 + a_3 \left(\frac{1}{n} - u \right)^3, & t \leq 1, \quad \frac{1}{2n} < u \leq \frac{1}{n}, \\ u, & t \leq 1, \quad u \leq \frac{1}{2n}, \end{cases}$$

with $a_2 = 4n(\sqrt{n}-1)$ and $a_3 = 12n^3(3\sqrt{n}-2)$, $\tilde{f}_n(t, u) = \hat{f}_n(t, u)$, for $t \geq 1 + \frac{1}{n}$ and a smooth connection between those two components of the approximation for $a < t < 1 + \frac{1}{n}$. Then the related approximate solutions will converge to the solution $u(t) = 4(t-1)^2$, $1 \leq t \leq 2$ and $u(t) = 0$, $0 \leq t < 1$. In a similar way all other solutions presented in Example 2 can be obtained.

Example 2 gives a connected and closed family of solutions to the initial value problem. The following theorem shows this is typical.

Theorem 2 ([15]) Assume that the assumptions of Theorem 1 are satisfied. Let $t_0 < c \leq t_0 + \rho$ and define A_c to be the set of points that can be reached at time $t = c$ by some solution of the initial value problem. Then A_c is closed, bounded and connected.

If the initial value problem is one-dimensional, i.e., $u \in \mathbb{R}$, then this implies that the set of points reached by possible solutions at time $t = c$ is the empty set, a point or a closed interval.

A proof of this theorem, based on [19], can be found in Theorem 4.1 in Chap. 2 of [13].

The existence theorems presented so far give existence of solutions near the initial value (t_0, u_0) . But this local result can be extended to a global one.

Theorem 3 (Extension Theorem) Let f be continuous on D . If $u(t)$ is a solution of the initial value problem (1) on some interval, then the solution can be extended to a maximal interval of existence (t_-, t_+) such that $(t, u(t))$ converges to the boundary of D as $t \uparrow t_+$ or $t \downarrow t_-$ (where $t_{\pm} = \pm\infty$ is possible).

Proof Let D_n be open subsets of D such that $\bigcup_{n \in \mathbb{N}} D_n = D$, the closures \bar{D}_n are compact and $\bar{D}_n \subset D_{n+1}$ (e.g., $D_n = \{(t, u) \in \mathbb{R} \mid |(t, u)| < n, \operatorname{dist}((t, u), \partial D) > 1/n\}$, see [13]). First it will be shown that for all $n \in \mathbb{N}$, there is some ρ_n such that for each $(t_0, u_0) \in D_n$, the initial value problem has a solution on the interval

$[t_0 - \rho_n, t_0 + \rho_n]$. Indeed, for each $n \in \mathbb{N}$, the function f is continuous on the compact set $\overline{D_n}$. Hence, there is some $M_n > 0$ such that $|f(t, u)| \leq M_n$ for all $(t, u) \in \overline{D_n}$. Furthermore, for each $n \in \mathbb{N}$, the distance $d_n := \text{dist}(\overline{D_n}, D_{n+1}) > 0$. Thus, the Cauchy–Peano theorem implies that for each $(t_0, u_0) \in \overline{D_n}$, the solution of the initial value problem exists on the interval $[t_0 - \rho_n, t_0 + \rho_n]$, with $\rho_n = 1/2 \min(d_n/M_{n+1}, d_n)$.

If the solution $u(t)$ is defined on an interval I , which is not a right maximal interval, then the argument above shows that the right endpoint of I can be included in the interval of existence. So it can be assumed that the interval I is of the form $[a_1, a_2]$. The continuity of the solution u gives that the set $\{(t, u(t)) \mid t \in [a_1, a_2]\}$ is a compact set in the open set D ; hence, there is some $n \in \mathbb{N}$ such that $\{(t, u(t)) \mid t \in [a_1, a_2]\} \subset D_n \subset \overline{D_n}$. But this implies that the solution can be extended to the interval $[a_1, a_2 + \rho_n]$. And if $\{(t, u(t)) \mid t \in [a_1, a_2 + \rho_n]\} \subset D_n$, then this can be repeated. Since $\overline{D_n}$ is compact, there is some $k \in \mathbb{N}$ such that $\{(t, u(t)) \mid t \in [a_1, a_2 + k\rho_n]\} \not\subset D_n$; hence, there exists some $(t_n, u(t_n))$ such that $(t_n, u(t_n)) \in D_{n+1} \setminus D_n$. This argument can be repeated for each $n \in \mathbb{N}$ to give sequence $(t_n, u(t_n)) \in D_{n+1} \setminus D_n$, with $t_n < t_{n+1}$. Thus, the sequence t_n is monotone increasing. Either it is unbounded, which implies that the solution exists on an interval $[a_1, \infty)$ and hence $t_+ = \infty$ or it is bounded and hence is convergent to some limit t_+ . Similarly, the sequence $(t_n, u(t_n))$ is either an unbounded sequence in D or has a cluster point on the boundary of D . In either case it is clear that the solution cannot be extended outside the interval $[a_1, t_+)$. A similar argument can be used for the left endpoint and it can be concluded that there exists a maximal interval of existence of the form (t_-, t_+) .

Finally it will be shown that the solution converges to the boundary of D by a contradiction argument. Assume that the solution does not converge to the boundary if $t \uparrow t_+$. Then there is some $\varepsilon_0 > 0$ and a sequence $\{\tau_n\}$ with $t_+ - \tau_n < \frac{1}{n}$ and $d(\tau_n, u(\tau_n), D) > \varepsilon_0$. This implies that there is some $N \in \mathbb{N}$ such that the sequence $\{(\tau_n, u(\tau_n))\} \subset D_N$. With the arguments above, this implies that the solution can be extended to an interval including t_+ , which contradicts t_+ being the maximal right point. \square

Example 1 gives an example of a solution for which $u(t) \rightarrow \infty$ if $t \rightarrow \frac{1}{u_0}$, the boundary of its interval of existence. Example 2 gives an example of a solution with an unbounded interval of existence. Example 4 shows that the endpoint can also be related to the failure of continuity of f .

Example 4 Consider the initial value problem with $D = \mathbb{R} \times (0, \infty)$ and

$$\dot{u} = -\frac{1}{u}, \quad \text{and} \quad u(0) = 2.$$

Then the solution is $u(t) = \sqrt{4 - 2t}$ for $t \in (-\infty, 2)$. If $t \rightarrow 2$, then $(t, u(t)) \rightarrow (2, 0)$, hence the boundary of D and the point where the continuity of f fails.

If the initial value problem has a nonunique solution, then the maximal interval of extension in the extension theorem will in general depend on the initial solution $u(t)$.

Uniqueness

Uniqueness follows if the vector field f is Lipschitz continuous with respect to its second variable u as shown first by Picard [21] and Lindelöf [18]. Their proof is based again on approximate solutions, but the approximate solutions are smoother than the Cauchy–Euler ones. The approximate solutions are obtained by successive iterations and are defined as

$$\begin{aligned} u_0(t) &:= u_0 \quad \text{and} \\ u_{n+1}(t) &:= u_0 + \int_{t_0}^t f(\tau, u_n(\tau)) d\tau, \quad n \in \mathbb{N}, \quad t \in I. \end{aligned} \tag{8}$$

Sometimes this iteration is called Picard iteration. Clearly, this iteration is based on the integral formula (2). The iteration process fails if $u_n(\tau) \notin D$ for some $\tau \in I$. If f is continuous, then there is some interval I for which the sequence is well-defined, as follows from the proof below.

Theorem 4 (Picard–Lindelöf Theorem) Let $a, b > 0$ be such that the cylinder $S := I_a(t_0) \times B_b(u_0) \subset D$. Let the vector field f be a uniformly Lipschitz continuous function on the cylinder S with respect to its second variable u and let M be the upper bound on f , i.e., $M = \max\{f(t, u) \mid (t, u) \in S\}$.

Define the parameter $\rho = \min\left(a, \frac{b}{M}\right)$. For $t \in [t_0, t_0 + \rho]$, there exists a unique solution $u(t)$ of the initial value problem (1).

Successive iterations play an important role in proving existence for more general differential equations. Thus, although the existence of solutions already follows from the Cauchy–Peano theorem, we will prove again existence by using the successive iterations instead of the Cauchy–Euler approximations as the ideas in the proof can be extended to more general differential equations.

Proof By using the bound on the vector field f , we will show that if $(t, u_n(t)) \in S$ for all $t \in [t_0, t_0 + \rho_0]$, then $(t, u_{n+1}(t)) \in S$ for all $t \in [t_0, t_0 + \rho_0]$. Indeed, for any $t \in [t_0, t_0 + \rho_0]$,

$$\begin{aligned} |u_{n+1}(t) - u_0| &\leq \int_{t_0}^t |f(\tau, u_n(\tau))| d\tau \\ &\leq \int_{t_0}^t M d\tau \leq M(t - t_0) \leq b; \end{aligned} \quad (9)$$

thus, $|u_{n+1}(t) - u_0| \leq M\rho \leq b$ and therefore $u_{n+1}(t) \in B_b(u_0)$.

The boundedness of f also ensures that the functions $\{u_n\}$ are equicontinuous as for any $n \in \mathbb{N}$ and any $t_0 \leq t_1 < t_2 \leq \rho + t_0$:

$$|u_n(t_1) - u_n(t_2)| \leq \int_{t_1}^{t_2} |f(\tau, u_n(\tau))| d\tau \leq M|t_2 - t_1|.$$

Up to this point, we have only used the boundedness of f (which follows from the continuity of f). But for the next part, in which it will be shown that for any $t \in [t_0, t_0 + \rho_0]$ the sequence $\{u_n(t)\}$ is a Cauchy sequence in \mathbb{R}^m , we will need the Lipschitz continuity. Let L be the Lipschitz constant of f on S , then for any $n \in \mathbb{N}$ and $t \in [t_0, t_0 + \rho_0]$, we have

$$\begin{aligned} |u_{n+1}(t) - u_n(t)| &\leq \int_{t_0}^t |f(\tau, u_n(\tau)) - f(\tau, u_{n-1}(\tau))| d\tau \\ &\leq \int_{t_0}^t L |u_n(\tau) - u_{n-1}(\tau)| d\tau \\ &\leq L \|u_n - u_{n-1}\|_\infty \int_{t_0}^t d\tau \\ &= L \|u_n - u_{n-1}\|_\infty (t - t_0), \end{aligned}$$

where $\|u_n - u_{n-1}\|_\infty = \sup\{|u_n(\tau) - u_{n-1}(\tau)| \mid t_0 \leq \tau \leq t_0 + \rho\}$. This implies that

$$\begin{aligned} |u_{n+2}(t) - u_{n+1}(t)| &\leq \int_{t_0}^t L |u_{n+1}(\tau) - u_n(\tau)| d\tau \\ &\leq L^2 \|u_n - u_{n-1}\|_\infty \int_{t_0}^t (\tau - t_0) d\tau \\ &= \frac{L^2}{2} \|u_n - u_{n-1}\|_\infty (t - t_0)^2. \end{aligned}$$

Repeating this process, it follows for any $k \in \mathbb{N}$ that

$$\begin{aligned} |u_{n+k+1}(t) - u_{n+k}(t)| &\leq \frac{L^k}{k!} \|u_n - u_{n-1}\|_\infty (t - t_0)^k \\ &\leq \frac{L^k \rho^k}{k!} \|u_n - u_{n-1}\|_\infty. \end{aligned}$$

By using the triangle inequality repeatedly this gives for any $n, k, \in \mathbb{N}$ and $t \in [t_0, t_0 + \rho_0]$,

$$\begin{aligned} |u_{n+k+1}(t) - u_n(t)| &\leq \sum_{i=0}^k |u_{n+i+1}(t) - u_{n+i}(t)| \\ &\leq \sum_{i=0}^k \frac{L^{n+i} \rho^{n+i}}{(n+i)!} \|u_1 - u_0\|_\infty \\ &\leq \frac{(L\rho)^n}{n!} \|u_1 - u_0\|_\infty \sum_{i=0}^k \frac{L^i \rho^i}{i!} \\ &\leq \frac{(L\rho)^n}{n!} b e^{L\rho}. \end{aligned}$$

Thus, for any $t \in [t_0, t_0 + \rho_0]$, the sequence $\{u_n(t)\}$ is a Cauchy sequence in \mathbb{R}^m ; hence, the sequence has a limit, which will be denoted by $u(t)$. In other words, the sequence of functions $\{u_n\}$ is pointwise convergent in $I_\rho(t_0)$. We have already seen that the functions u_n are equicontinuous; hence, the convergence is uniform and the limit function $u(t)$ is equicontinuous [23].

To see that u satisfies the integral equation (2), observe that for any $t \in [t_0, t_0 + \rho_0]$

$$\begin{aligned} \left| u(t) - u_0 - \int_{t_0}^t f(\tau, u(\tau)) d\tau \right| &\leq |u(t) - u_{n+1}(t)| \\ &\quad + \left| \int_{t_0}^t |f(\tau, u_n(\tau)) - f(\tau, u(\tau))| d\tau \right| \\ &\leq |u(t) - u_{n+1}(t)| + L \|u - u_n\|_\infty (t - t_0) \\ &\leq (1 + L\rho) (\|u - u_{n+1}\|_\infty + \|u - u_n\|_\infty) \end{aligned}$$

for any $n \in \mathbb{N}$. As the sequence $\{u_n\}$ is uniformly convergent, this implies that $u(t)$ satisfies the integral equation (2).

Finally the uniqueness will be proved using techniques similar to those in "Existence." Assume that there are two solutions u and v of the integral equation (2). Hence, for any $t_0 \leq t \leq t_0 + \rho$,

$$\begin{aligned} |u(t) - v(t)| &\leq \int_{t_0}^t |f(\tau, u(\tau)) - f(\tau, v(\tau))| d\tau \\ &\leq L \|u - v\|_\infty (t - t_0). \end{aligned}$$

Thus, for $k = 1$, it holds that

$$\begin{aligned} |u(t) - v(t)| &\leq \frac{L^k \|u - v\|_\infty (t - t_0)^k}{k!}, \\ &\text{for any } t_0 \leq t \leq t_0 + \rho. \end{aligned} \quad (10)$$

For the induction step, assume that (10) holds for some $k \in \mathbb{N}$. Then for any $t_0 \leq t \leq t_0 + \rho$,

$$\begin{aligned} |u(t) - v(t)| &\leq L \int_{t_0}^t \frac{L^k \|u - v\|_{\infty} (\tau - t_0)^k}{k!} d\tau \\ &= \frac{L^{k+1} \|u - v\|_{\infty} (t - t_0)^{k+1}}{(k+1)!}. \end{aligned}$$

By using the principle of induction, it follows that (10) holds for any $k \in \mathbb{N}$. This implies that for any $t_0 \leq t \leq t_0 + \rho$, the solutions satisfy $|u(t) - v(t)| \leq \frac{(L\rho)^{k+1} \|u - v\|_{\infty}}{(k+1)!}$ for any $k \in \mathbb{N}$. Since the expression on the right converges to 0 for $k \rightarrow \infty$, this implies that $u(t) = v(t)$ for all $t \in [t_0, t_0 + \rho]$ and hence the solution is unique. \square

Implicitly, the proof shows that the integral equation gives rise to a contraction on the space of continuous functions with the supremum norm and this could also have been used to show existence and uniqueness; see [11,25], where Schauder's fixed-point theorem is used to prove the Pi-

card-Lindelöf theorem. With Schauder's fixed-point theorem, existence and uniqueness can be shown for much more general differential equations than ordinary differential equations.

The iteration process in (8) does not necessarily have to start with the initial condition. It can start with any continuously differentiable function $u_0(t)$ and the iteration process will still converge. In specific problems, there are often obvious choices for the initial function u_0 which give a faster convergence or are more efficient, for instance, when proving unboundedness.

Techniques similar to those in the proof can be used to derive a bound on the difference between the successive approximations and the solution, showing that

$$|u_n(t) - u(t)| \leq \frac{ML^n(t - t_0)^n}{(n+1)!}, \quad t \in [t_0, t_0 + \rho];$$

see [13]. This illustrates that the convergence is fast for t near t_0 , but might be slow further away: see Fig. 3 for an example.

Continuous Dependence on Initial Conditions

In this section, the vector field f is uniformly Lipschitz continuous on D with respect to its second variable u and its Lipschitz constant is denoted by L . By combining the Picard-Lindelöf theorem (Theorem 4) and the extension theorem (Theorem 3), it follows that for every $(t_0, u_0) \in D$, there exists a unique solution of (1) passing through (t_0, u_0) with a maximal interval of existence $(t_-(t_0, u_0), t_+(t_0, u_0))$. The trajectory through (t_0, u_0) is the set of points $(t, u(t))$, where $u(t)$ solves (1) and $t_-(t_0, u_0) < t < t_+(t_0, u_0)$. Now define the set $E \subset \mathbb{R}^{n+2}$ as

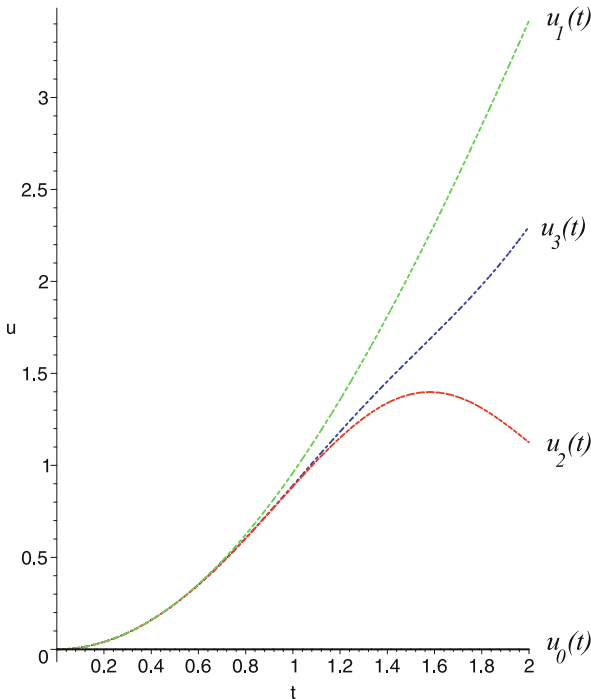
$$E := \{ (t, t_0, u_0) \mid t_-(t_0, u_0) < t < t_+(t_0, u_0), (t_0, u_0) \in D \}.$$

The flow map $\Phi: E \rightarrow \mathbb{R}^m$ is a mapping which describes the solution of the initial value problem with the initial condition varying through D , i.e.,

$$\begin{aligned} \frac{d}{dt} \Phi(t, t_0, u_0) &= f(t, \Phi(t, t_0, u_0)) \quad \text{and} \\ \Phi(t_0, t_0, u_0) &= u_0, \quad \text{for } (t, t_0, u_0) \in E. \end{aligned}$$

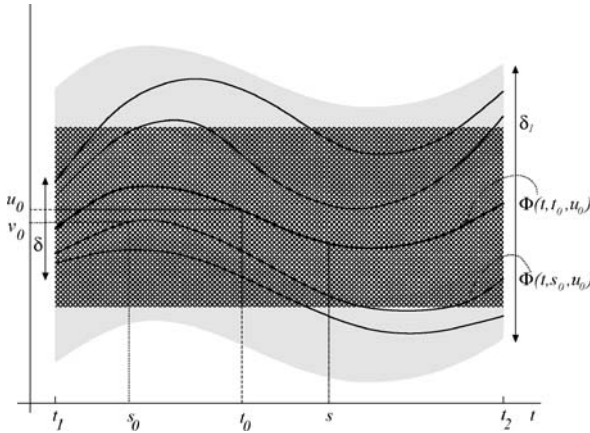
From the integral equation (2), it follows that the flow map satisfies

$$\begin{aligned} \Phi(t, t_0, u_0) &= u_0 + \int_{t_0}^t f(\tau, \Phi(\tau, t_0, u_0)) d\tau, \\ &\quad \text{for } (t, t_0, u_0) \in E. \end{aligned}$$



Existence and Uniqueness of Solutions of Initial Value Problems, Figure 3

The first four iterations for initial value problem with the vector field $f(t, u) = t \cos u + \sin t$ and the initial value $u(0) = 0$. Note that the convergence in the interval $[0, 1]$ is very fast (two steps seem sufficient for graphical purposes), but near $t = 2$, the convergence has not yet been reached in three steps



Existence and Uniqueness of Solutions of Initial Value Problems, Figure 4

The flow map $\Phi(t, t_0, u_0)$. Note the group property $\Phi(t, s, \Phi(s, t_0, u_0)) = \Phi(t, t_0, u_0)$. The lightly shaded area is the tube U_{δ_1} and the darker shaded area is the tube U_δ as used in the proof of Theorem 6. If a solution starts from within U_δ , then it will stay within the tube U_{δ_1} for $t_1 \leq t \leq t_2$, as the solution $\Phi(t, s_0, v_0)$ demonstrates

Furthermore, the uniqueness implies for the flow map that starting at the point (t_0, u_0) , flowing to $(t, \Phi(t, t_0, u_0))$ is the same as starting at the point (t_0, u_0) and flowing to the intermediate point $(s, \Phi(s, t_0, u_0))$ and continuing to the point $(t, \Phi(t, t_0, u_0))$. This implies that the flow map has a group structure: for any $(t_0, u_0) \in D$ and any $s, t \in (t_-(t_0, u_0), t_+(t_0, u_0))$, the flow map satisfies $\Phi(t, s, \Phi(s, t_0, u_0)) = \Phi(t, t_0, u_0)$; the identity map is given by $\Phi(t_0, t_0, u_0) = u_0$; and by combining the last two results, it follows that $\Phi(t_0, t, \Phi(t, t_0, u_0)) = \Phi(t_0, t_0, u_0) = u_0$, and hence there is an inverse. See also Fig. 4 for a sketch of the flow map for $D \subset \mathbb{R} \times \mathbb{R}$.

To show that the flow map is continuous in all its variables, Gronwall's lemma [9] will be very useful. There are many versions of Gronwall's lemma; the one presented here follows [1].

Lemma 5 (Gronwall's Lemma) Let $\psi: [a, b] \rightarrow \mathbb{R}$ and $\phi: [a, b] \rightarrow \mathbb{R}$ be nonnegative continuous functions on an interval $[a, b]$. Let $K \geq 0$ be some nonnegative constant such that

$$\phi(t) \leq K + \int_a^t \psi(\tau)\phi(\tau)d\tau, \quad \text{for all } a \leq t \leq b.$$

Then

$$\phi(t) \leq K \exp \left[\int_a^t \psi(\tau)d\tau \right], \quad \text{for all } a \leq t \leq b.$$

Proof Define $F(t) = K + \int_a^t \psi(\tau)\phi(\tau)d\tau$, for $a \leq t \leq b$. Then the assumption in the lemma gives $\phi(t) \leq F(t)$ for $a \leq t \leq b$. Furthermore, F is differentiable, with $\dot{F} = \phi\psi$; hence, for $a \leq t \leq b$

$$\begin{aligned} & \frac{d}{dt} \left(F(t) \exp \left[- \int_a^t \psi(\tau)d\tau \right] \right) \\ &= \psi(t)\phi(t) \exp \left[- \int_a^t \psi(\tau)d\tau \right] \\ & \quad - F(t)\psi(t) \exp \left[- \int_a^t \psi(\tau)d\tau \right] \leq 0 \end{aligned}$$

as $\phi(t) \leq F(t)$. Integrating the left-hand side gives

$$F(t) \exp \left[- \int_a^t \psi(\tau)d\tau \right] \leq F(a) = K,$$

which implies for ϕ that $\phi(t) \leq F(t) \leq K \exp \left[\int_a^t \psi(\tau)d\tau \right]$. \square

Gronwall's lemma will be used to show that small variations in the initial conditions give locally small variations in the solution, or

Theorem 6 The flow map $\Phi: E \rightarrow \mathbb{R}^{n+2}$ is continuous in E .

Proof Let $(t, t_0, u_0) \in E$. As $t, t_0 \in (t_-(t_0, u_0), t_+(t_0, u_0))$, there is some $t_1 < t_2$ such that $t, t_0 \in (t_1, t_2)$, $[t_1, t_2] \subset (t_-(t_0, u_0), t_+(t_0, u_0))$ and the solution $\Phi(s, t_0, u_0)$ exists for any $s \in [t_1, t_2]$. First we will show that there is some tube around the solution curve $\{(s, \Phi(s, t_0, u_0)) \mid t_1 \leq s \leq t_2\}$ such that for any (s_0, v_0) in this tube, the solution $\Phi(s, s_0, v_0)$ exists for $t_1 \leq s \leq t_2$.

As D is open, there is some $\delta_1 > 0$ such that the closed tube $U_{\delta_1} := \{(s, v) \mid |\Phi(s, t_0, u_0) - v| \leq \delta_1, t_1 \leq s \leq t_2\} \subset D$. As f is Lipschitz continuous, hence continuous, there is some $M > 0$ such that $|f(s, v)| \leq M$ for all $(s, v) \in U_{\delta_1}$. Recall that the Lipschitz constant of f on D is denoted by L . Now define $\delta := \delta_1 e^{-L(t_2-t_1)} < \delta_1$ and the open tube

$$U_\delta = \{(s, v) \mid |\Phi(s, t_0, u_0) - v| < \delta, t_1 < s < t_2\},$$

thus, $U_\delta \subset U_{\delta_1} \subset D$. Thus, for every $(s_0, v_0) \in U_\delta$, the solution $\Phi(s, s_0, v_0)$ exists for s in some closed interval in $[t_1, t_2]$ around s_0 and the solution in this interval satisfies

$$\Phi(s, s_0, v_0) = v_0 + \int_{s_0}^s f(\tau, \Phi(\tau, s_0, v_0))d\tau,$$

see also Fig. 4. Furthermore, for any $s \in [t_1, t_2]$, we have

$$\begin{aligned}\Phi(s, t_0, u_0) &= \Phi(s, s_0, \Phi(s_0, t_0, u_0)) \\ &= \Phi(s_0, t_0, u_0) + \int_{s_0}^s f(\tau, \Phi(\tau, s_0, \Phi(s_0, t_0, u_0))) d\tau \\ &= \Phi(s_0, t_0, u_0) + \int_{s_0}^s f(\tau, \Phi(\tau, t_0, u_0)) d\tau.\end{aligned}$$

Subtracting these expressions gives

$$\begin{aligned}|\Phi(s, s_0, v_0) - \Phi(s, t_0, u_0)| &\leq |v_0 - \Phi(s_0, t_0, u_0)| \\ &\quad + \int_{s_0}^s |f(\tau, \Phi(\tau, s_0, v_0)) - f(\tau, \Phi(\tau, t_0, u_0))| d\tau \\ &\leq \delta + \int_{s_0}^s L |\Phi(\tau, s_0, v_0) - \Phi(\tau, t_0, u_0)| d\tau,\end{aligned}$$

so Gronwall's lemma implies that $|\Phi(s, s_0, v_0) - \Phi(s, t_0, u_0)| \leq \delta e^{L(s-s_0)} \leq \delta_1$. Hence, for any s , $\Phi(s, s_0, v_0) \in U_{\delta_1} \subset D$, and hence this solution can be extended to its maximal interval of existence, which contains the interval $[t_1, t_2]$. So it can be concluded that for any $(s_0, v_0) \in U_\delta$, the solution $\Phi(s, s_0, v_0)$ exists for any $s \in [t_1, t_2]$.

Next we will show continuity of the flow map in its last two variables, i. e., in the initial value. For any $(s_0, v_0) \in U_\delta$ and $t_1 \leq t \leq t_2$, we have

$$\begin{aligned}|\Phi(t, s_0, v_0) - \Phi(t, t_0, u_0)| &\leq |v_0 - u_0| \\ &\quad + \left| \int_{s_0}^t f(\tau, \Phi(\tau, s_0, v_0)) d\tau - \int_{t_0}^t f(\tau, \Phi(\tau, t_0, u_0)) d\tau \right| \\ &\leq |v_0 - u_0| + \left| \int_{s_0}^{t_0} f(\tau, \Phi(\tau, s_0, v_0)) d\tau \right| \\ &\quad + \left| \int_{t_0}^t |f(\tau, \Phi(\tau, t_0, u_0)) - f(\tau, \Phi(\tau, s_0, v_0))| d\tau \right| \\ &\leq |v_0 - u_0| + M|t_0 - s_0| \\ &\quad + \left| \int_{t_0}^t L |\Phi(\tau, t_0, u_0) - \Phi(\tau, s_0, v_0)| d\tau \right|.\end{aligned}$$

Thus, Gronwall's lemma implies that

$$\begin{aligned}|\Phi(t, s_0, v_0) - \Phi(t, t_0, u_0)| &\leq (|v_0 - u_0| + M|t_0 - s_0|) e^{L|t-t_0|} \\ &\leq (|v_0 - u_0| + M|t_0 - s_0|) e^{L|t_2-t_1|},\end{aligned}$$

and it follows that Φ is continuous in its last two arguments. The continuity of Φ in its first argument follows

immediately from the fact that the solution of the initial value problem is continuous. \square

If the vector field f is smooth, then the flow map Φ is smooth as well.

Theorem 7 Let $f \in C^1(D, \mathbb{R}^m)$. Then the flow map $\Phi \in C^1(E, \mathbb{R}^m)$ and

$$\begin{aligned}\det(D_{u_0}\Phi(t, t_0, u_0)) &= \exp\left(\int_{t_0}^t \text{tr}(D_u f(\tau, \Phi(\tau, t_0, u_0))) d\tau\right)\end{aligned}$$

for any $(t, t_0, u_0) \in E$.

In this theorem, $\text{tr}(D_u f(\tau, \Phi(\tau, t_0, u_0)))$ stands for the trace of the matrix $D_u f(\tau, \Phi(\tau, t_0, u_0))$. For second-order linear systems, this identity is known as Abel's identity and $\det(D_{u_0}\Phi(t, t_0, u_0))$ is the Wronskian.

The proof of Theorem 7 uses the fact that $D_{u_0}\Phi(t, t_0, u_0)$ satisfies the linear differential equation

$$\frac{d}{dt} D_{u_0}\Phi(t, t_0, u_0) = D_u f(t, \Phi(t, t_0, u_0)) D_{u_0}\Phi(t, t_0, u_0),$$

with the initial condition $D_{u_0}\Phi(t_0, t_0, u_0) = I$. This last fact follows immediately from $\Phi(t_0, t_0, u_0) = u_0$. And the linear differential equation follows by differentiating the differential equation for $\Phi(t, t_0, u_0)$ with respect to u_0 . The full details of the proof can be found, for example, in [5,11].

Extended Concept of Differential Equation

Until now, we have looked for continuously differentiable functions $u(t)$, which satisfy the initial value problem (1). But the initial value problem can be phrased for less smooth functions as well. For example, one can define a solution as an absolute continuous function $u(t)$ which satisfies (1). A function $u: I \rightarrow \mathbb{R}^m$ is *absolutely continuous* on I if for every positive number ε , no matter how small, there is a positive number δ small enough so that whenever a sequence of pairwise disjoint subintervals $[s_k, t_k]$ of I , $k = 1, 2, \dots, N$ satisfies $\sum_{k=1}^N |t_k - s_k| < \delta$ then

$$\sum_{k=1}^n d(u(t_k), u(s_k)) < \varepsilon.$$

An absolute continuous function has a derivative almost everywhere and is uniformly continuous, thus continuous [24]. Thus, the initial value problem for an absolute continuous function can be stated almost everywhere.

For the existence of absolute continuous solutions, the continuity condition for the vector field in the Cauchy–

Peano theorem (Theorem 1) is replaced by the so-called Carathéodory conditions on the vector field. A function $f: D \rightarrow \mathbb{R}^m$ satisfies the Carathéodory conditions if [2]

- $f(t, u)$ is Lebesgue measurable in t for each fixed u ;
- $f(t, u)$ is continuous in u for each fixed t ;
- for each compact set $A \subset D$, there is a measurable function $m_A(t)$ such that $|f(t, u)| \leq m_A(t)$ for any $(t, u) \in A$.

Similar theorems to the ones in the previous sections about existence and uniqueness can be stated in this case.

Theorem 8 (Carathéodory) *If D is an open set in \mathbb{R}^{n+1} and $f: D \rightarrow \mathbb{R}^m$ satisfies the Carathéodory conditions on D , then for every $(t_0, u_0) \in D$ there is a solution of the initial value problem (1) through (t_0, u_0) .*

Theorem 9 *If D is an open set in \mathbb{R}^{n+1} , $f: D \rightarrow \mathbb{R}^m$ satisfies the Carathéodory conditions on D and $u(t)$ satisfies the initial value problem (1) on some interval, then there exists a maximal open interval of existence. Furthermore, if (t_-, t_+) denotes the maximal interval of existence, then the solution $u(t)$ tends to the boundary of D if $t \downarrow t_-$ or $t \uparrow t_+$.*

Theorem 10 *If D is an open set in \mathbb{R}^{n+1} , $f: D \rightarrow \mathbb{R}^m$ satisfies the Carathéodory conditions on D and for every compact set $U \subset D$ there is an integrable function $L_U(t)$ such that*

$$|f(t, u) - f(t, v)| \leq L_U |u - v|, \quad \text{for all } (t, u), (t, v) \in U,$$

then for every $(t_0, u_0) \in D$, there is a unique solution $\Phi(t, t_0, u_0)$ of the initial value problem (1). The domain E of the flow map Φ is open and Φ is continuous in E .

Proofs of those theorems can be found, for example, in Sect. I.5 in [11].

Many other generalizations are possible, for example, to vector fields which are discontinuous in u . Such vector fields play an important role in control theory. More details can be found in [6,7,17].

Further Directions

As follows from the previous sections, the theory of initial value problems for ordinary differential equations is quite well understood, at least for continuous vector fields. For more general differential equations, the situation is quite different. Consider for example a retarded differential equation, hence a differential equation with delay effects [12]. For retarded differential equations, there are local existence and uniqueness theorems which are quite similar to the ones for ordinary differential equations as

presented in the earlier sections and can be proved by using Schauder's fixed-point theorem. But not all solutions can be extended as in the extension theorem (Theorem 3); see Chap. 2 in [12].

If one considers differential equations with more variables, i. e., partial differential equations, then there are no straightforward general existence and uniqueness theorems. For partial differential equations, the existence and uniqueness depends very much on the details of the partial differential equation. One possible theorem is the Cauchy–Kowaleskaya theorem, which applies to partial differential equations of the form $\partial_t^m u = F(t, u, \partial_x u, \partial_t u, \partial_{tx}^2 u, \dots)$, where the total number of derivatives of u in F should be less than or equal to m [22]. But there are still many open questions as well. A famous open question is the existence and uniqueness of solutions of the three-dimensional Navier–Stokes equations, a system of equations which describes fluid motion. Existence and uniqueness is known for a fluid in a two-dimensional domain, but not in a three-dimensional domain. The question is one of the seven “Millennium Problems,” stated as prize problems at the beginning of the third millennium by the Clay Mathematics Institute [26].

Existence and uniqueness results are also used in dynamical systems. An area of dynamical systems is bifurcation theory and for bifurcation theory the smooth dependence on parameters is crucial. The following theorem gives sufficient conditions for the smooth dependence parameters; see Theorem 3.3 in Chap. 1 of [11].

Theorem 11 *If the vector field depends on parameters $\mu \in \mathbb{R}^k$, i. e., $f: D \times \mathbb{R}^k \rightarrow \mathbb{R}^m$ and $f \in C^1(D \times \mathbb{R}^k, \mathbb{R}^m)$, then the flow map is continuously differentiable with respect to its parameters μ . Furthermore, $D_\mu \Phi$ satisfies an inhomogeneous linear equation*

$$\begin{aligned} \frac{d}{dt} D_\mu \Phi(t, t_0, u_0, \mu) \\ = D_u f(t, \Phi(t, t_0, u_0, \mu), \mu) D_\mu \Phi(t, t_0, u_0, \mu) \\ + D_\mu f(t, \Phi(t, t_0, u_0, \mu), \mu) \end{aligned}$$

with initial condition $D_\mu \Phi(t_0, t_0, u_0, \mu) = 0$.

Results and overviews of bifurcation theory can be found, for example, in [4,8,10,14,16].

Another area of dynamical systems is stability theory. Roughly speaking, a solution is called stable if other solutions which start near this solution stay near it for all time. Note that the local continuity result in Theorem 6 is not a stability result as it only states that nearby solutions will stay nearby for a short time. For stability results, long-time existence of nearby solutions is a crucial property [10,14].

Bibliography

Primary Literature

- Bellman R (1943) The stability of solutions of linear differential equations. *Duke Math J* 10:643–647
- Carathéodory C (1918) *Vorlesungen über Reelle Funktionen*. Teubner, Leipzig (reprinted: (1948) Chelsea Publishing Company, New York)
- Cauchy AL (1888) *Oeuvres complètes* (1) 6. Gauthiers-Villars, Paris
- Chow S-N, Hale JK (1982) *Methods of bifurcation theory*. Springer, New York
- Coddington EA, Levinson N (1955) *Theory of Ordinary Differential Equations*. McGraw-Hill, New York
- Filippov AF (1988) *Differential equations with discontinuous righthand sides*. Kluwer, Dordrecht
- Flügge-Lotz I (1953) *Discontinuous automatic control*. Princeton University Press, Princeton
- Golubitsky M, Stewart I, Schaeffer DG (1985–1988) *Singularities and groups in bifurcation theory*, vol 1 and 2. Springer, New York
- Gronwall TH (1919) Note on the derivative with respect to a parameter of the solutions of a system of differential equations. *Ann Math* 20:292–296
- Guckenheimer J, Holmes P (1983) *Nonlinear oscillations, dynamical systems, and bifurcations of vector fields*. Springer, New York
- Hale JK (1969) *Ordinary Differential Equations*. Wiley, New York
- Hale JK, Verduyn Lunel SM (1993) *Introduction to Functional Differential Equations*. Springer, New York
- Hartman P (1973) *Ordinary Differential Equations*. Wiley, Baltimore
- Iooss G, Joseph DD (1980) *Elementary stability and bifurcation theory*. Springer, New York
- Kneser H (1923) Ueber die Lösungen eines Systems gewöhnlicher Differentialgleichungen das der Lipschitzschen Bedingung nicht genügt. *S-B Preuss Akad Wiss Phys-Math Kl* 171–174
- Kuznetsov YA (1995) *Elements of applied bifurcation analysis*. Springer, New York
- Lee B, Markus L (1967) *Optimal Control Theory*. Wiley, New York
- Lindelöf ME (1894) Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences* 114:454–457
- Müller M (1928) Beweis eines Satzes des Herrn H. Kneser über die Gesamtheit der Lösungen, die ein System gewöhnlicher Differentialgleichungen durch einen Punkt schickt. *Math Zeit* 28:349–355
- Peano G (1890) Démonstration de l'intégrabilité des équations différentielles ordinaires. *Math Ann* 37:182–228
- Picard É (1890) Mémoire sur la théorie de équations aux dérivées partielles et la méthode des approximations successives. *J Math*, ser 4, 6:145–210
- Rauch J (1991) *Partial Differential Equations*. Springer, New York
- Rudin W (1976) *Principles of Mathematical Analysis*, 3rd edn. McGraw-Hill, New York
- Rudin W (1987) *Real and Complex Analysis*, 3rd edn. McGraw-Hill, New York
- Zeidler E (1995) *Applied functional analysis*. Vol 1 Applications to mathematical physics. Springer, New York
- Clay Mathematics Institute (2000) The Millenium Problems: Navier Stokes equation. http://www.claymath.org/millennium/Navier-Stokes_Equations/

Books and Reviews

- Arnold VI (1992) *Ordinary Differential Equations*. Springer, Berlin
- Arrowsmith DK, Place CM (1990) *An Introduction to Dynamical Systems*. Cambridge University Press, Cambridge
- Braun M (1993) *Differential Equations and their Applications*. Springer, Berlin
- Brock WA, Malliaris AG (1989) *Differential Equations, stability and chaos in Dynamic Economics*. Elsevier, Amsterdam
- Grimshaw R (1993) *Nonlinear Ordinary Differential Equations*. CRC Press, Boca Raton
- Ince EL (1927) *Ordinary differential equations*. Longman, Green, New York
- Jordan DW, Smith P (1987) *Nonlinear Differential Equations*. Oxford University Press, Oxford
- Werner H, Arndt H (1986) *Gewöhnliche Differentialgleichungen: eine Einführung in Theorie und Praxis*. Springer, Berlin

Exobiology and Complexity

ERIC J. CHAISSON

Wright Center, Tufts University, Massachusetts, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Arrow of Time

Non-equilibrium Thermodynamics

Big-Bang Cosmology

Measuring Complexity

Complexity and Evolution, Broadly Considered

Conclusions and Future Directions

Bibliography

Glossary

Complexity A state of intricacy, complication, variety, or involvement, as in the interconnected parts of a system – a quality of having many interacting, different components.

Cosmic evolution A grand synthesis of the many varied changes in the assembly and composition of radiation, matter, and life throughout the history of the Universe.

Cosmology The study of the structure, evolution, and destiny of the Universe.

Energy The ability to do work or to cause change.

Energy rate density The amount of energy flowing through a system per unit time per unit mass.

Evolution Any process of growth and change with time, including an accumulation of historical information; in its broadest sense, both developmental and generational change.

Exobiology The study of the origin, evolution, and distribution of past and present life in the Universe; also known as astrobiology or bioastronomy.

Thermodynamics The study of the macroscopic changes in the energy of a system, for which temperature is a central property.

Definition of the Subject

Recent research, guided by theoretical searches for unification as much as by compilation of huge new databases, suggests that complex systems throughout Nature are localized, temporary islands of ordered structures within vastly larger, disordered environments beyond those systems. All such complex systems – including, for example, stars, life, and society – can be shown to obey quantitatively the principles of non-equilibrium thermodynamics, and all can be modeled in a common, integral manner by analyzing the energy passing through those systems. The concept of energy flow does seem to be as universal a process as anything yet found in Nature for the origin, maintenance, and evolution of ordered, complex systems. The optimization of such energy flows acts as an agent of evolution broadly considered, thereby affecting, and to some extent unifying, all of physical, biological, and cultural evolution.

More specifically, non-equilibrium thermodynamics, especially the energy flows resulting from contrasting temporal behaviors of matter and radiation energy densities, can generally explain the cosmic environments needed for the emergence of increasingly ordered structures over time. Furthermore, a necessary (though perhaps not sufficient) condition for the natural flow of energy, and hence for the growth of complexity, is the expansion of the Universe itself. Among all of Nature's diverse systems, energy – acquired, stored, and expressed – is a principal driver of the rising complexity among galaxies, stars, planets and life-forms throughout the cosmos. Neither new science nor appeals to non-science are required to appreciate the outstanding hierarchy of evolutionary change, from atoms to galaxies, from cells to society.

One way to approach the topic of exobiology and complexity – also known as astrobiology or bioastronomy – is to place it within the grand context of cosmic evolution.

This interdisciplinary subject seeks to combine all the natural sciences into a unified whole, thereby effectively creating a new scientific worldview for the 21st century. Evolution, broadly considered, has indeed become a powerful unifying concept in all of science. Life itself, including complex life, seems to be a natural, but not necessarily inevitable, result of the way things complexify in an expanding Universe.

Introduction

Cosmic evolution is the study of the sum total of the many varied developmental and generational changes in the assembly and composition of radiation, matter, and life throughout all space and across all time. These are the physical, biological, and cultural changes that have produced, in turn and among many other complex systems, our Galaxy, our Sun, our Earth, and ourselves. The result is an inclusive evolutionary synthesis bridging a wide variety of scientific specialties – physics, astronomy, geology, chemistry, biology, and anthropology – a genuine scientific narrative of epic proportions extending from the beginning of time to the present, from big bang to humankind.

The general idea of evolution – change writ large – extends well beyond the subject of biology, granting it a powerful unifying potential in all of science. Unquestionably, change is widespread throughout all of Nature, much as the Greek philosopher Heraclitus asserted 25 centuries ago: “All flows ... nothing stays”. Yet questions remain: How realistic is the quest for interdisciplinary unification? Can we reconcile the observed constructiveness of cosmic evolution with the inherent destructiveness of thermodynamics? Specifically, how have the amazing examples of order all around us arisen from chaos – and how does all this fit into complexity science?

We especially seek to understand the origins of the many diverse structures spanning the Universe today, notably those characterized by the term “complexity” – a state of intricacy, complication, variety, or involvement, as in the interconnected parts of a system. (In this article, no definitional distinctions are made among the words “order”, “form”, and “complexity”; we address only a general understanding of an entire spectrum of structures often described by the intuitive usage of the term complexity.) Particularly intriguing is the increase of complexity over the course of time, indeed dramatically so (with some exceptions) within the past half-billion years since the Cambrian period on Earth. Perhaps some underlying principle, a general law, or an ongoing process does create, organize, and maintain all complex structures in the

Universe, enabling us to study Nature's many changes on uniform, common ground – in the same quantitative way and with the same mental tools, in other words “on the same page”.

Both theory and experiment, as well as computer simulations, suggest affirmative answers to some of the above questions: Islands of ordered complexity – namely, open systems that are galaxies, stars, planets, and life-forms – are numerically more than balanced by great seas of increasing disorder elsewhere in the environments beyond those systems. All quantitatively agrees with the valued principles of thermodynamics, especially non-equilibrium thermodynamics. Furthermore, energy flows produced largely by the expanding cosmos do seem to be as universal a factor in the origin of structured systems as anything yet found in Nature. The optimization of such energy flows might well act as the motor of evolution broadly conceived, thereby affecting all of physical, biological, and cultural evolution, the combination of which constitutes cosmic evolution.

Therefore, a general outline for this article is:

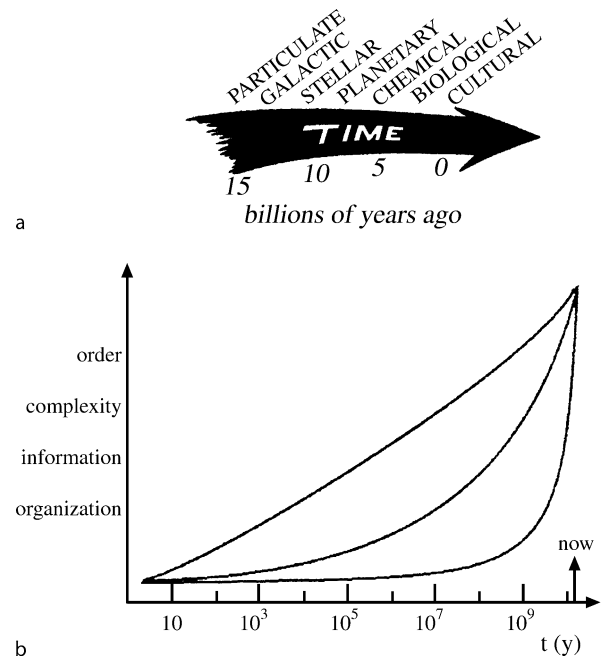
Cosmic Evolution

- Subsets in time: physical evolution → biological evolution → cultural evolution.

Other researchers have addressed life and complexity in a cosmic setting, originally Chambers [16], who anonymously wrote a pre-Darwinian study of wide interdisciplinary insight, and notably Shapley [62], who pioneered a “cosmography” that classified all known structures according to increasing dimensions. Among others, Spencer [64] championed the idea of growing complexity in biological and cultural evolution, Henderson [31] regarded the whole evolutionary process, both physical and organic, as one and the same, Whitehead [72] sought to broaden scientific thinking with his “organic philosophy”, von Bertalanffy [75] championed a systems theoretic approach to physical, biological, and social studies, and Sagan [57], Reeves [56], Jantsch [33] and Chaisson [7] widely advanced the concept of complex (intelligent) life within a cosmological framework.

Arrow of Time

Figure 1a sketches Nature's main historical epochs diagonally atop the so-called arrow of time [3]; these 7 epochs correspond to the major evolutionary phases comprising the whole of cosmic evolution [14]. Regardless of its shape or orientation, such an arrow symbolizes a *sequence* of events that have changed systems from simplicity to complexity, from inorganic to organic, from chaos in the early



Exobiology and Complexity, Figure 1

a An arrow of time symbolically chronicles the principal epochs of cosmic history, from the beginning of the Universe ~14 billion years ago (at left) to the present (at right). Labeled diagonally across the top are 7 major evolutionary phases (corresponding to the main historical epochs) that have produced, in turn, increasing amounts of order and complexity among all material systems: particulate, galactic, stellar, planetary, chemical, biological, and cultural. Cosmic evolution encompasses all of these phases, each of which represents a coarse temporal duration when the emergence of key systems flourished in Nature. Time is assumed to flow linearly and irreversibly, much as other basic tenets are presumed, including the fixed character of physical law and the idea that $2 + 2 = 4$ everywhere. **b** Sketched here qualitatively is the rise of order, form, and complexity typifying the evolution of *localized* material structures throughout the history of the Universe. This family of curves depicts a widespread, innate feeling, and not a rigid proof, that the complexity of ordered structures has *generally* increased over the course of time. It is unknown if this rise of complexity has been linear, exponential, or even faster (as drawn here for the 3 curves); current research aims to specify this curve and to describe it quantitatively

Universe to order more recently. That sequence, as determined by a large body of post-Renaissance data, accords well with the idea that a chain of knowledge – a loose continuity along an impressive hierarchy of complexity – links, in turn:

- The evolution of primal energy into elementary particles and then atoms
- The evolution of those atoms into galaxies and stars
- The evolution of stars into heavy elements

- The evolution of those elements into the molecular building blocks of life
- The evolution of those molecules into life itself
- The evolution of advanced life forms into intelligence
- The evolution of intelligent life into a cultured and technological civilization.

Despite the extreme specialization of modern science, evolution marks no disciplinary boundaries; complexity science is a truly interdisciplinary topic. A more specific outline for this article is then:

Cosmic Evolution

- Subsets: physical evolution → biological evolution → cultural evolution
- Phases: particulate → galactic → stellar → planetary → chemical → biological → cultural.

Accordingly, the most familiar kind of evolution – biological evolution, or neo-Darwinism – is just one, albeit important, subset of a broader evolutionary scenario including much more than life on Earth. In short, what Darwinian change does for plants and animals, cosmic evolution aspires to do for all things. And if Darwinism created a revolution in understanding by helping to free us from the anthropocentric belief that humans differ from other life-forms on our planet, then cosmic evolution extends that intellectual revolution by treating matter on Earth and in our bodies no differently from that in the stars and galaxies far beyond.

Note that time's arrow does not imply that primitive, "lower" life-forms have biologically changed directly into advanced, "higher" organisms, any more than galaxies have physically changed into stars, or stars into planets. Rather, with time – much time – the environmental conditions suitable for spawning simple life eventually changed into those favoring the biological origin and evolution of more complex species. Likewise, in the earlier Universe, the physical evolution of environments ripe for galactic formation eventually gave way more recently to conditions conducive to stellar and planetary formation. Now, at least on Earth, cultural evolution dominates, since our local planetary environment has once more changed to foster greater, societal complexity. Change in the surrounding environments usually precedes change in organized systems, and the resulting changes for those systems selected to endure have *generally* been toward greater amounts of diverse order and complexity.

Anthropocentrism is neither intended nor implied by the arrow of time; the arrow is not pointing at humankind. Anthropocentric principles notwithstanding, no logic supports

the idea that the Universe was conceived in order to produce specifically us. Humans are not the pinnacle or culmination of the cosmic-evolutionary scenario, nor are we likely to be the only technologically competent beings that have emerged in the organically rich Universe. The arrow merely provides a convenient symbol, artistically suggesting the building of increasingly complex structures, from spiral galaxies to rocky planets to thinking beings.

Figure 1b graphs the widespread impression that material systems have become more organized and complex, especially in relatively recent times. This family of curves graphs "islands" of complexity that comprise ordered systems per se – whether massive stars, colorful flowers, or busy urban centers – not their vastly, indeed increasingly disorganized surroundings. Modern science aims to explain this rise of complexity and to do so with accepted scientific principles and observational or experimental data.

Non-equilibrium Thermodynamics

Cosmic evolution, as understood today, is governed largely by the laws of physics, particularly those of thermodynamics. However, this does not mean classical reductionism, for here we seek to model change guided by a combination of randomness and determinism, of chance and necessity. Nor does the cosmic-evolutionary narrative employ mere equilibrium thermodynamics – the kind most often used to explain closed systems isolated from their surroundings and having maximum entropy states. All structures observed in Nature, among them most notably galaxies, stars, planets, and life-forms, are demonstrably open, non-equilibrium systems, with flows of energy in and out being an important feature. And it is this energy, often called available, or "free" energy that helps to build structures [27,28,38,54,60,74].

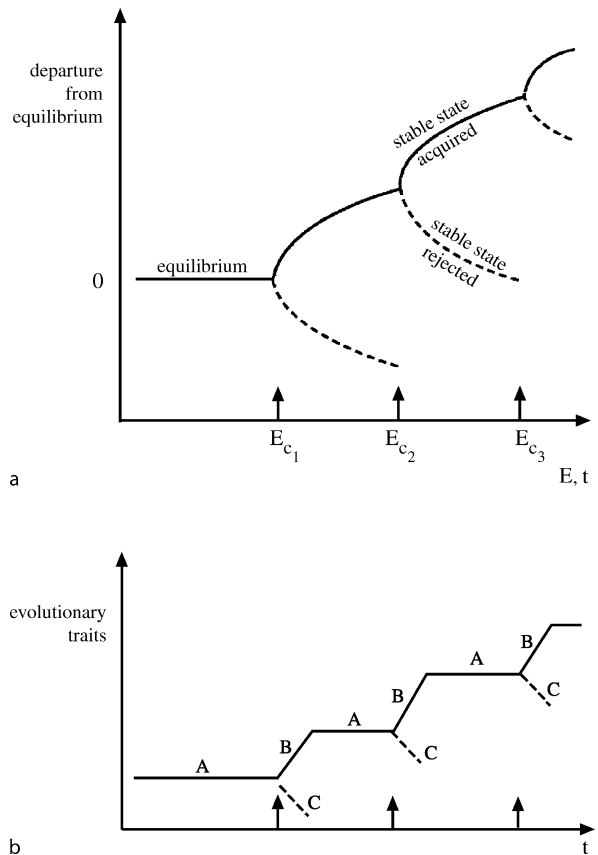
By utilizing energy, order can be achieved temporarily, or at least the environmental conditions made conducive for the potential rise of order within open systems ripe for growth. Energy flow plays a vital role in the creation, maintenance, and fate of complex systems – all quantitatively in accord with the second law of thermodynamics, which demands an overall rise in disorder. None of Nature's ordered structures, not even life itself, is a violation (or even a circumvention) of the second law. Considering both any ordered system as well as its disordered surroundings, non-equilibrium thermodynamics shows that the net entropy of the system and its environment always increases. (Quantitative details for many such systems can be found in [10].)

Energy is now recognized as a key ingredient, not only for biological systems such as plants and animals, but also

for physical systems such as stars and galaxies, indeed as well for social systems such as a city's inward flow of food and resources amidst its outward flow of products and wastes (Weber [69]; Dyke [21]). The analysis is much the same for all open systems, provided they are modeled in broad, interdisciplinary ways; energy flow seems indispensable for any system's origin and evolution.

Figure 2, adapted from the work of Prigogine et al. [55] and Salk [58], graphically diagrams the emergence of structure in the presence of energy flow. Physicists relate to the type of curves drawn in part (a); biologists are more familiar with those in part (b). By crossing certain energy thresholds that depend on a system's status, bifurcations can occur, fostering the origin of whole new structures that display surprising amounts of coherent behavior. Such "dissipative" structures can export some of their entropy (or expel some of their energy) into their external environments. Accordingly, order is created and sustained by routine consumption of substances rich in energy, followed by discharge of substances low in energy. This process, often misnamed, is not really *self-ordering*; it is ordering in the presence of energy. "Self-assembling" systems demonstrate an essential tension between energy inflow and dissipative outflow; such systems do no function magically by themselves.

The emergence of order from a condition where originally there was none (or less of it) is relatively straightforward [54,59]. Fluctuations – random deviations from some average, equilibrium value of, for example, density, temperature, or pressure – inevitably yet stochastically appear in any natural system having many degrees of freedom. Normally, as in equilibrium thermodynamics, such instabilities regress in time and disappear; they come and go by chance since the statistical fluctuations diffuse as quickly as they arise. Even in an isolated system, such internal fluctuations can generate local, microscopic reductions in entropy, but the second law ensures that they will always balance out. Microscopic temperature fluctuations, for instance, are said to be thermally relaxed, and entropy remains maximized in such systems. Nor can an open system *near* equilibrium change spontaneously to new and interesting structures. But should those fluctuations become too great for an open system to damp, that system can then depart far from equilibrium and have a chance to reorganize. Such restructuring generates a "dynamic steady state", provided the amplified fluctuations are continuously driven and stabilized by a flow of energy from the surroundings – namely, provided the energy flow rate exceeds the thermal relaxation rate. Systematic, coherent cycling is often the result, since under these conditions the spontaneous creation of macroscopic structures dissipates



Exobiology and Complexity, Figure 2

a The departure of an open system from equilibrium is drawn here as a function of both time, t , and energy, E . The time axis makes clear that this is an historical, evolutionary process, whereas the parallel energy axis tracks free energy flowing through the open system as a vital part of that process. At certain critical energies, labeled here E_c , a system can spontaneously change, or bifurcate, into new, non-equilibrium, dynamic steady states. Statistical fluctuations – that's chance – affect which fork the system selects – that's necessity – upon bifurcation (vertical arrows), namely which spatial structure is achieved. Not all new systems survive (solid curve); some are rejected (dashed curve). The process, as always, is a mixture of randomness and determinism, therefore the end result is inherently unpredictable, as with all of evolution. **b** Events in evolutionary biology mimic those of the diagram in **a**, although the results here are richer in structural detail, system function, and energy flow. In phases marked A, a species survives and thus persists until the environment changes (vertical arrows), after which further evolution occurs – along phase B toward renewed survival or phase C toward extinction. Both upwardly rising graphs (drawn by solid lines for both parts of this figure) imply neither progress nor inevitability, but they do suggest a *general trend* toward increasing complexity with time – a trend that cannot be denied among organized systems observed throughout Nature

energy more rapidly than the ensuing, and damaging, heat can damp the gradients and destroy those structures. Furthermore, since each successive reordering often causes more complexity than the preceding one, such systems become even more susceptible to fluctuations. Complexity itself consequently creates the conditions for greater instability, which in turn provides an opportunity for greater reordering. Nothing is guaranteed; thermodynamics specifies what can happen, not what actually does happen. The resulting phenomenon – termed “order through fluctuations” – is a distinctly evolutionary one, complete with feedback loops that help drive some systems further from equilibrium. And as the energy consumption and resulting complexity accelerate, so does the evolutionary process. This is the realm of true *thermodynamics*, the older, established subject of that name more properly labeled “thermostatics”.

Numerous examples abound throughout Nature, and not just among physical systems, but for biological and social systems as well. Naturally occurring phenomena such as convection cells, river eddies, atmospheric storms, and even artificially made devices such as kitchen refrigerators and coherent lasers among an array of many physical systems that experience coherent order when amply fed with sufficient energy, all display enhanced order when energy flows exceed some threshold. Biological systems also obey the rules of non-equilibrium thermodynamics, for we and our living relatives are demonstrable examples of dynamic steady states that emerge and function via energetically enhanced neo-Darwinism (though biologists often worry that such statements aim to reduce biology to physics – whereas in reality physics is broadened to include biology.) As are Lamarckian-type cultural systems of more recent times also dynamic steady states, for among the bricks and chips that civilization has built, energy has been a principal driver (although, again, sociologists and anthropologists often loathe their subjects being treated thermodynamically). The result is that life and its cultural inventions differ not in kind, but merely in degree – specifically, degree of complexity – among numerous ordered systems evident in Nature.

Big-Bang Cosmology

The origin of Nature’s many complex structures depends on the flow of free energy. And this, like the arrow of time itself, is a direct consequence of the expanding Universe – a much tested “standard cosmological model” based largely on three-fold observations of distant receding galaxies, microwave background radiation, and light-element abundances. Time marches on and free energies

flow because the cosmos dynamically evolves – building, maintaining, and often destroying systems. Indeed, it is cosmic expansion, and probably nothing more, that has caused the entire Universe to depart from its initial state of thermodynamic equilibrium. (Thus, the free energies are inevitable, not the resulting systems per se – which is why it’s called “available”, or potential, energy freely *capable* of doing work.) The stark contrast between localized hot stars and the vast, cold interstellar space surrounding them now guarantees a state of non-equilibrium, a cosmic condition that has pertained for billions of years [26,40].

Matter

Although modern cosmology stipulates that matter only later emerged from the primordial radiation of the early Universe, it is pedagogically useful to quantify first the role of matter and thereafter the primacy of radiation. In this way, perhaps the greatest change in the history of the Universe – the transformation from radiation to matter – can be mathematically justified.

Imagine an arbitrary shell of mass, m , and radius, r , expanding isotropically with the Universe at a velocity, v , from some central point. The sphere within the shell is not necessarily meant to represent the entire Universe, only an extremely large gas cloud within it – in fact, larger than the extent of a typical galaxy supercluster (~ 50 Mpc across), which comprises the highest rank in the known hierarchy of matter assemblages in the Universe. Invoking the principle of energy conservation, we find the Friedmann–Lemaître equation that describes a family of models for the Universe in bulk,

$$H^2 - \frac{8}{3}\pi G\rho_m = -kR^{-2}.$$

Here, H is Hubble’s constant (a measure of galaxy recession in the expanding Universe), G is the universal gravitational constant, ρ_m is the matter density, and k is a time-dependent curvature constant. R is a scale factor which relates the radius, r , at any time, t , in cosmic history to the current radius, r_0 , at the present time – namely, $r = Rr_0$. Solutions to the above equation specify three general models for the Universe:

- The Universe is “open” (i. e., k negative) and thus recedes forevermore
- The Universe is “closed” (i. e., k positive), meaning it eventually stops and thereafter contracts to a point much like that from which it began
- The Universe is precisely balanced between the open and closed models, in which case it eternally expands

and never contracts (because it can never reach infinity).

Even if the Universe is, as now suspected, accelerating in its outward expansion, the effect of “dark energy” (that supposedly causes the acceleration) on stars and galaxies is minimal, and that on smaller structures like planets and their organized living systems is inconsequential; cosmic acceleration likely affects only the dynamics of the Universe on the largest scales and not those of organized systems that are controlled by local energies within it.

The simplest case ($k = 0$, also known as the Einstein-deSitter solution) leads to the critical density for closure,

$$\rho_{m,c} = 3H^2/8\pi G,$$

which, when evaluated for G and H (~ 70 km/s/Mpc), equals $\sim 10^{-29}$ g/cm³. This is ~ 6 atoms in each cubic meter of space, or about a million times thinner than in the region between Earth and the Moon. Whether the actual current density, on average everywhere, is smaller or larger than this value, making the Universe open or closed, respectively, is currently unknown; there is too much uncertainty concerning “dark (non-baryonic) matter” that is implied, but not found yet needed, to gravitationally bind galaxies and their clusters. However, the above-noted acceleration of the Universe does imply that it is expanding ever faster, thereby giving it an open geometry that will recede toward infinity forevermore.

To follow the evolution of matter throughout cosmic history (up to the present), we appeal to the conservation of material particles in the huge sphere postulated above, $\rho_m = \rho_{m,0}R^{-3}$, substitute into the special ($k = 0$) case of the Friedmann–Lemaître equation, and manipulate,

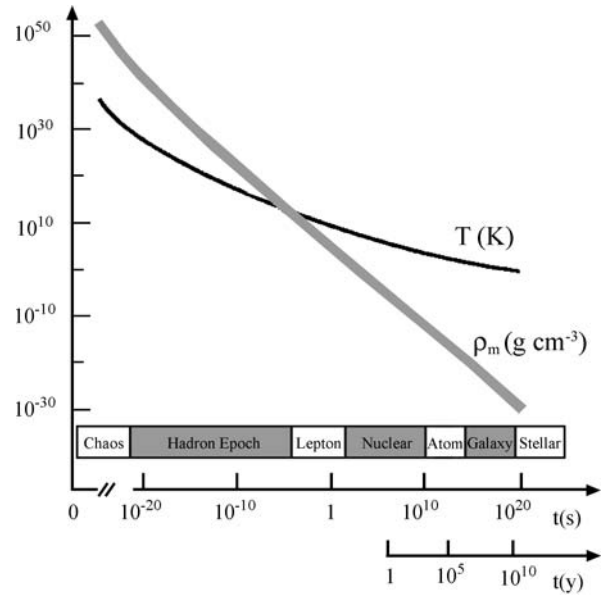
$$\int dt = \left(\frac{8}{3} \pi G \rho_{m,c} \right)^{-0.5} \int R^{0.5} dR.$$

The result suggests that the Universe is $\sim 14 \times 10^9$ y old ($\pm \sim 10\%$). This equation additionally stipulates how the average matter density thins with time,

$$\rho_m \approx 10^6 t^{-2},$$

where ρ_m is expressed in g/cm³ and t in seconds.

Figure 3 plots this evolution of matter throughout all of universal history. This run of density, ρ_m , in standard, big-bang cosmology demonstrates the essence of change on the largest scales – the broadest view of the biggest picture. Here displayed in this one plot is the thermodynamic history of the whole Universe, so the curve for ρ_m in this figure (as well as the curve for T discussed in the next section) pertain to nothing in particular, just everything in general.



Exobiology and Complexity, Figure 3

Log-log plot of the density (ρ_m) of matter on average and the temperature (T) of radiation on average, over the course of all time, to date. The thick width of the density curve displays the range of uncertainty in total mass density, whose true value depends on the amount of “dark matter” in the Universe. By contrast, the cold cosmic background temperature is very accurately measured today (2.7 K), and its thin curve here is equally accurately extrapolated back into the hot, early Universe. Recent findings that cosmic expansion is accelerating should not much affect these curves

Radiation

The same analysis regarding matter can be applied to radiation in order to follow the change of temperature with time. Again, for the simplest $k = 0$ case,

$$H^2 = \frac{8\pi G \rho_{r,c}}{3R^4},$$

where ρ_r is the equivalent mass density of radiation. Here the R^4 term derives from the fact that radiation scales not only as the volume ($\propto R^3$) but also by one additional factor of R because radiation (unlike matter) is also affected linearly by the Doppler shift. And noting that $\rho_r c^2 = aT^4$, where a is the universal radiation constant for any black-body emitter and T is the temperature of radiation, we find the temporal dependence of average temperature throughout all time (in seconds),

$$T \approx 10^{10} t^{-0.5}.$$

The universal radiation, having begun in a fiery expansion (popularly called the “big bang”), has now cooled to 2.7 K, the average value of the cosmic microwave back-

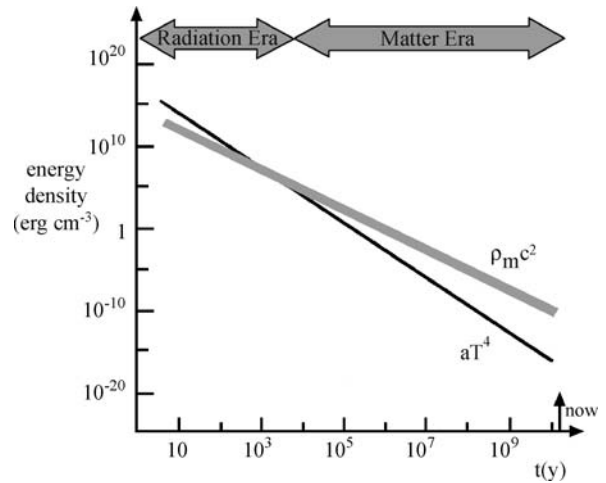
ground measured today by radio telescopes on the ground and satellites in orbit [2].

Figure 3 also plots this run of T versus t . Again, for emphasis, the two curves in this figure show the prime twin trends of big-bang cosmology: the cooling and thinning of radiation and matter, based largely on observations of the microwave background radiation and of the distant receding galaxies.

For the first few hundred millennia of the Universe, radiation reigned supreme over matter. Life was nonexistent and matter only a submicroscopic precipitate suspended in a glowing, chaotic fireball. All space was flooded with high-frequency photons, especially light, x-rays, and γ -rays, ensuring a non-structured, undifferentiated, (virtually) informationless, and highly uniform plasma. Matter and radiation were intimately coupled to each other – thermalized and equilibrated. Structure of any sort had yet to emerge; the energy density of radiation was too great. If single protons captured single electrons to form simple hydrogen atoms, the radiation was then so fierce as to destroy those atoms immediately. However, as the Universe expanded with time, the energy of radiation decreased faster than the energy equivalently contained in matter.

To confirm this statement, compare the energy densities of radiation and matter, and especially how these two quantities have change over time. First convert the matter density derived earlier to an equivalent energy density by invoking the Einsteinian mass (m)-energy (E) relation, $E = mc^2$ – that is, by multiplying the above equation for ρ_m by c^2 . Now, $\sim 14 \times 10^9$ y after the big bang, $\rho_{m,0}c^2 \approx 10^{-9}$ erg/cm³, whereas $aT_0^4 \approx 4 \times 10^{-13}$ erg/cm³; thus currently, $\rho_{m,0}c^2 > aT_0^4$ by several orders of magnitude, proving that matter is now in firm control (gravitationally) of cosmic changes, despite the Universe still being flooded today with long-wavelength radiation. However, given that $\rho_m c^2$ scales as R^{-3} and aT^4 scales as R^{-4} , there must have been a time in the past when $\rho_m c^2 = aT^4$, and an even earlier time when $\rho_m c^2 < aT^4$. Manipulation of the above equations shows that these two energy densities crossed at $t \approx 10^4$ y, well less than a million years after the big bang. Figure 4 is a graphical representation of this paragraph.

This crossover represents a preeminent change in all of cosmic history. The event, $\rho_m c^2 = aT^4$, separates the *Radiation Era* from the *Matter Era*, and designates the time ($\sim 10^4$ y) when the Universe gradually began to become transparent. Thermal equilibrium was destroyed and symmetry broken, causing the radiative fireball and disorganized matter to decouple; it was as though a fog had lifted. Photons, previously scattered aimlessly and destructively



Exobiology and Complexity, Figure 4

The temporal behavior of both matter energy density ($\rho_m c^2$) and radiation energy density (aT^4) illustrates perhaps the greatest change in all of natural history. Where the two curves intersect, neutral atoms began to form; by $t \approx 10^5$ y after the big bang the Radiation Era had changed into the Matter Era. A uniform, featureless state describing the early Universe was thus naturally transformed into one in which order and complexity were thereafter possible

by subatomic material particles (especially free electrons) in the expanding, hot, opaque plasma of the Radiation Era, were no longer so affected once the electrons were bound into atoms of the Matter Era. This crucial and dramatic change was over by $\sim 4 \times 10^5$ y, when the last remnants of the early ionized plasma state had finally transformed into neutral matter. The 2.7-K microwave radiation reaching Earth today is a relic of this critical phase transition, having streamed unimpeded (except for being greatly redshifted, $z \sim 10^3$) across space and time for most of the age of the Universe, granting a “view” of this grandest of all evolutionary events that occurred long ago.

With the onset of the Matter Era, matter literally began dominating radiation. Natural history became more interesting, for then structures could begin to form. The results of inevitable change, induced gradients, energy flows, and evolved systems, over billions of years and minus the details, are galaxies, stars, planets, and life-forms, one by-product of which is intelligence – at least on Earth. And this, in turn, has anthropogenically changed nearly everything on our planet.

Life

Now ~ 14 billion years after the beginning of space and time, the *Life Era* has begun, at least locally on Earth

(and possibly at many other places in the Universe). Here, the emergence of technologically intelligent life heralds a whole new era wherein life has gradually begun to dominate matter. This second of two great transformations was not caused by the origin of life per se several billion years ago; rather, it is technologically advanced life that differs significantly from primitive life and from other types of clustered inanimate matter scattered throughout the Universe. This is not an anthropocentric statement; we differ because we are the only species capable of knowing our past and worrying about our future, the only one able to control matter (albeit locally), much as matter evolved to control radiation long ago. Intelligent life on Earth is literally taking matter into its own hands – manipulating matter and energy, altering genes and terrestrial environments, indeed potentially changing evolution itself.

Some central questions before us are these: What caused the changes amid a wide spectrum of ordered structures throughout cosmic history and how has complexity increased with time? Have humans actually become the agents of change on Earth, able to tinker with both matter and energy, including now modifying genes and environments more than they affect us? How did the neural network within our human brains acquire the sophistication needed to fashion societies, weapons, cathedrals, philosophies, etc? In short, what caused us to become sentient enough to contemplate our complex selves?

Measuring Complexity

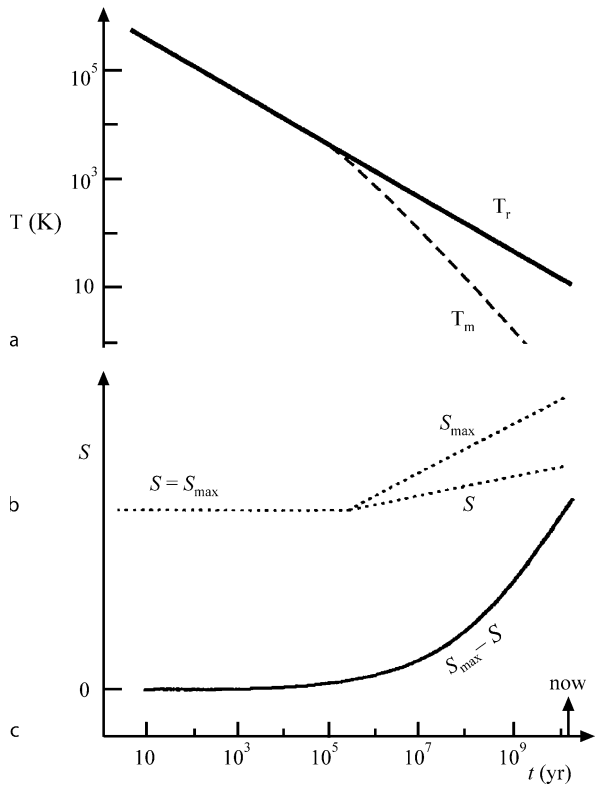
To appreciate the crux of the historical appearance of structured matter and life, we return to the greater cosmic environment and to some of the thermodynamic issues raised earlier. In brief, when the Universe broke its symmetry a few thousand centuries after the big bang, equilibrium was also destroyed. Temperature gradients were thereafter established naturally owing to the expansion of the cosmos. And that meant free energy began flowing, in fact increasingly so as the temperatures of matter and radiation diverged with time. These are the environmental conditions that are favorable for the potential growth of order, form, and complexity.

Cosmic Environment for the Growth of Complexity

When matter and radiation were still equilibrated in the Radiation Era, only a single temperature is needed to describe the thermal history of the Universe; the absence of any thermal gradients imply (virtually) zero information content, or zero macroscopic order, in the early Universe. However, once the Matter Era began, the gas-en-

ergy equilibrium was destroyed and a single temperature is insufficient to specify the bulk evolution of the cosmos. Since the random motions of the H and He atoms failed to keep pace with the rate of general expansion of the atoms away from one another [40], the matter cooled faster, $T_m \approx 6 \times 10^{16} t^{-1}$, than the radiation, $T_r \approx 10^{10} t^{-0.5}$. Figure 5a displays this thermal gradient, which has grown wider since $t \approx 10^5$ y.

Such a thermal gradient is the clear signature of a heat engine, and it is this ever-widening gradient that has enabled matter, in the main, to “build things” increasingly complex. Theoretically at least, the environmental conditions after 10^5 y naturally allowed a rise in “negentropy” [60] or “information content” [61] – both factors



Exobiology and Complexity, Figure 5

a In the expanding Universe, the temperatures of matter and radiation separated once these quantities became fully decoupled at $\sim 10^5$ y. Since that time, the Universe has been in a non-equilibrium state. **b** S increases less rapidly than S_{\max} , once the symmetry of equilibrium broke when matter and radiation decoupled at $\sim 10^5$ y. By contrast, in the early, equilibrated Universe, $S = S_{\max}$ for the prevailing conditions. **c** The potential for the growth of order, $S_{\max} - S$, has increased ever since the start of the Matter Era. This potential rise of order compares well with the subjectively drawn curves of Fig. 1b, thus providing a theoretical basis for the growth of system complexity

qualitatively synonymous with the term “complexity” [42]. But, as noted below, in practice both such terms are overly vague and subject to interpretation [5,73], so we resort here to a more conventional use of entropy, as agreed upon by most thermodynamicists. The important point – without getting lost in dubious semantics or contentious definitions – is that such non-equilibrium states are suitable, indeed apparently necessary, for the emergence of order, thus it can be reasoned that *cosmic expansion itself is the prime mover for the gradual construction of a hierarchy of structures throughout the Universe*.

Figure 5b plots the run of entropy, S , for a thermal gradient typical of a heat engine, but here graphed for the whole Universe. This is not a mechanical device running with idealized Newtonian precision, but a global engine capable of potentially doing work as locally emerging systems interact with their environments – especially those systems able to take advantage of increasing flows of free energy resulting from cosmic expansion and its naturally growing gradients. Although thermal and chemical (but not gravitational) entropy must have been maximized in the early Universe, hence complexity in the form of any structures then nonexistent, after decoupling the environmental conditions became favorable for the potential growth of order, taken here to mean a “lack of disorder.” At issue was timing: As ρ decreased, the equilibrium reaction rates ($\alpha\rho$) fell below the cosmic expansion rate ($\alpha\rho^{1/2}$) and non-equilibrium states froze in. Thus we have a paradoxical yet significant result that, in an expanding Universe, both the disorder (i.e. net entropy) and the order (maximum possible entropy minus actual entropy at any given time) can increase simultaneously – the former globally and the latter locally. All the more interesting when comparing the shape of this curve of potentially rising order, $S_{\max} - S$ in Fig. 5c, with our earlier intuited sketch of rising complexity in Fig. 1b [12,25,41].

Free Energy Rate Density

Theory aside, have the many diverse real structures known to exist in the Universe displayed this sort of progressive increase in order during the course of time? The answer is generally yes. At issue again is how to best characterize complexity numerically, given the varied connotations that this term presents for many researchers [46,47]. In biology alone, much as their inability to reach consensus on a definition of life, biologists cannot agree on a complexity metric. Some count non-junk genome size [66], others employ structural morphology or behavioral flexibility [4], while still others chart numbers of cell types in organisms [36] or appeal to cellular specialization [49]. All

these attributes of life have qualitative usefulness, yet all are hard to quantify in practical terms; nor do they apply to non-living things. If progress is to be made assessing a wide spectrum of complex systems in Nature, our analysis must extend beyond mere words, indeed beyond biology.

Putting aside as unhelpful (in the sense that it is too ambiguous and controversial) the above-noted concept of information content [32,34] as well as the concept of negative entropy (or negentropy, which Schroedinger [60] first adopted but then quickly abandoned), we return to the quantity with greatest appeal to physical intuition – energy. Given that energy – the ability to do work or to cause change – is the most universal currency known in the natural sciences, it might reasonably be expected to have a central role in any attempted unification of physical, biological, and cultural evolution.

Energy does act as an underlying, universal driver like no other in all of modern science. Whether living or non-living, dynamical systems need flows of energy to endure. If stars don’t convert gravitational potential into heat and light, they would collapse; if plants don’t photosynthesize sunlight, they would shrivel and decay; if humans don’t eat, we too would die. Likewise, society’s fuel is energy: Resources come in and wastes go out, all the while civilization goes about its daily business.

Not that energy has been ignored in previous studies of systems’ origin and assembly. Physicists (e.g., Morrison [52] and Dyson [22]), biologists (Lotka [44]; Morowitz [51]; Fox [24]), and ecologists (Odum [53], Ulano-witz [68], and Smil [63]), to cite only a few researchers, have noted energy’s organizational abilities. But the quantity of choice cannot be energy alone, for a star clearly has more energy than an amoeba, a galaxy much more than a single cell. Yet any living system is surely more complex than any inanimate object. Thus, absolute energies are not as suitable as normalized values, which depend on a system’s size, composition, and efficiency. Nor are maximum energy principles or minimum entropy states [43] likely relevant; rather, organizational complexity is more likely governed by the *optimum* use of energy – not too little as to starve a system, yet not too much as to destroy it.

To characterize complexity objectively – that is, to normalize all such ordered systems on the same, level page – a kind of energy density is useful, much like the competing energy densities of radiation and matter that dictated changing events in the earlier Universe (Fig. 4). In fact, for a proper treatment of the thermodynamics of non-equilibrium open systems, it is the *rate* at which free energy flows through such systems of given mass that is most practical. Hence, *free energy rate density*, symbolized by Φ_m , is an

operational term whose meaning, measurement, and units are clearly understood. In this way, neither new science nor appeals to nonscience are needed to justify the impressive hierarchy of the cosmic-evolutionary story, from stars to plants to society.

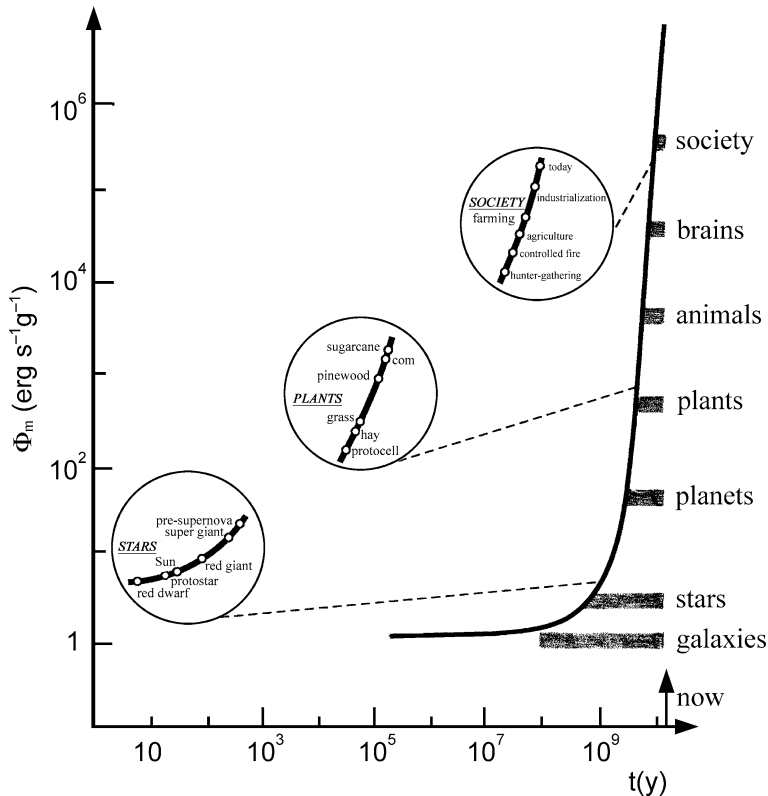
Note that this complexity metric is not an original term; free energy rate density is a mere revision of an old one. Moreover, for years the same term has been labeled differently by specialized researchers: Φ_m is familiar to astronomers as the luminosity-to-mass ratio, to physicists as the power density, to geologists as the specific radiant flux, to biologists as the specific metabolic rate, and to engineers as the power-to-mass ratio. Free energy rate density is central to many varied subjects; all the more reason to use it to build a true interdisciplinary subject and to use it in search of unity across the spectrum of all the natural sciences [8,9,10,11].

Exobiology and Complexity, Table 1

Free energy rate densities for several representative systems

System	Duration (10 ⁶ y)	Φ_m (erg/s/g)
Galaxy (Milky Way)	12,000	0.5
Star (main-sequence Sun)	10,000	2
Planet (Earth's climasphere)	5000	75
Plant (Earth's biosphere)	3000	900
Animal (hominid body)	10	20,000
Brain (human cranium)	1	150,000
Society (modern culture)	0	500,000

Table 1 lists values of Φ_m , in units of erg/sec/g, for seven representative systems (and their specific, computed cases in parentheses). Also listed is the duration, in millions of years, for each type of structure, dating back to their origins in the observational record. Clearly, Φ_m



Exobiology and Complexity, Figure 6

Increase in free energy rate density, Φ_m , plotted as *horizontal histograms* when various open structures prospered in Nature, has been especially rapid in the last few billion years, much as expected from subjective intuition (Fig. 1b) and objective thermodynamics (Fig. 5c). The *drawn curve* approximates the increase in normalized energy flows characterizing order, form, and structure for a range of systems throughout the history of the Universe. The *circled insets* show greater detail of further measurements or calculations of free energy rate density for three representative systems – stars, plants, and society – typifying physical, biological, and cultural evolution, respectively. The data in those circled insets are discussed in Sect. “Complexity and Evolution, Broadly Considered”. (Adapted from [9,10])

has increased as more intricately ordered systems have emerged throughout cosmic history, and dramatically so in relatively recent times.

The modeled flow of energy through a wide variety of open systems, alive or not, resembles the intuitive rise in complexity implied by Fig. 1b; it also mimics the potential rise of order in the above thermodynamic analysis of Fig. 5c. Complexity (at least as treated here energetically for localized structures) has indeed quantitatively increased over the course of natural history, and at a rate faster than exponential in more recent times ([9,11,13]; for details of energy computations and modeling, consult [10]).

Figure 6 plots the results listed in Table 1, where the Φ_m values are graphed as horizontal histograms for various systems' evolutionary durations to date. As expected:

- Stars and planets have small energy rate densities, $\Phi_m = 1\text{--}10^2$ erg/s/g
- Plants and animals have larger energy rate densities, $10^3\text{--}10^5$ erg/s/g
- Human societies have the largest known energy rate densities, $\sim 10^6$ erg/s/g.

Note that, although the total energy flowing through a star or planet is much greater than that through our individual bodies or brains, the *specific* rate (per unit mass) is larger for the latter – in fact, roughly a million times greater Φ_m for the human body than for the Sun.

This is not to say, by any means, that galaxies evolved into stars, or stars into planets, or planets into life. Rather, this analysis contends that galaxies gave rise to environments suited to the birth of stars, that some stars spawned environments conducive to the formation of planets, and that countless planets likely fostered environments ripe for the origin of life. Cosmic evolution, to repeat, incorporates both developmental and generational change.

Nor do these evolutionary phases, or historical durations, have well-determined start and stop times – or stop times necessarily at all. The horizontal histograms of Fig. 6 serve to stress that each of these phases once begun did not end; stars and galaxies, for example, first emerged in the earlier Universe, as also implied by the diagonal phases atop the arrow of time in Fig. 1a, but both such system types continue on presently originating, developing, and evolving, as do plants and animals that emerged much later. As depicted by those histograms yet unlike customary geological periods that do have set time intervals, currently all evolutionary phases noted in Figs. 1 and 6 operate simultaneously and indefinitely.

We thus arrive at a comprehensible reconciliation of the evident destructiveness of thermodynamics with

the observed constructiveness of cosmic evolution. The sources and sinks of such energy flows passing through complex yet disparate entities such as stars, plants, and civilization all relate back to the time of thermal decoupling in the early Universe, when the conditions naturally emerged for the origin and evolution of order and organization.

Complexity and Evolution, Broadly Considered

Evolution should not be the sole province of biology, nor should its utility be of value only to life scientists. Darwin [19] never used the word “evolution” as a noun, in fact only once as a verb in the very last sentence of his classic book, *On the Origin of Species*. Nor need the principle of natural selection be the only mechanism of evolutionary change, past or present.

Actually, the term “selection” is itself a misnomer, for no known agent in Nature deliberately selects. Selection is not an active force or promoter of change as much as a passive mechanism that weeds out the unfit. As such, selected systems are simply those that remain after all the poorly adapted or less fortunate ones have been removed from a population of such systems. A better term might be “non-random elimination” [48]. What we really aim to explain are the adverse circumstances responsible for the deletion of some members of a group. Accordingly, selection can be generally taken to mean favorable interaction of any system with its environment – a more liberal interpretation that also helps widen the concept of evolution.

Selection works alongside the flow of resources into and out of all open systems, not just life-forms. Ordered systems are selected partly for their ability to utilize energy; and this energy is the “force”, if there is any at all, in evolution. Broadly considered, selection occurs in the inanimate world as well as among animate objects, often providing a formative step in the production of order. It is energy flow and natural selection that together, working in tandem, underlie the “self”-assembly sketched in Fig. 2 – the former driving initial systems beyond equilibrium, the latter aiding the emergence of higher order in those systems that survive.

A handful of cases will suffice, among many others so documented [14], to illustrate the action of this energy-selection duo among a spectrum of increasingly ordered systems in successive phases of cosmic evolution:

- Red-giant stars are more complex than main-sequence stars
- Eukaryotes are more complex than prokaryotes
- Plants are more complex than protists
- Animals are more complex than plants
- Mammals are more complex than reptiles

- Brains are more complex than bodies
- Industrial society is more complex than hunter-gatherers.

Whether physical evolution of galaxies, stars and planets, or biological evolution of plants and animals on Earth, or cultural evolution of our technological civilization, a rather remarkable ranking order is apparent among all known organized structures. Stars, life, and society, all share a significant, common conclusion: Basic differences, both within and among these categories, are in degree, not in kind – namely, in degree of complexity arising from ongoing cosmic evolution. To justify this, consider below in greater detail each of the three major subsets of cosmic evolution noted in Sect. “[Introduction](#)”.

Physical Evolution

Stars are good examples of physical evolution. Growing complexity can serve as an indicator of stellar aging – a developmental process – allowing stars to be judged as their interiors undergo cycles of nuclear fusion that result in greater thermal and chemical gradients. More data are needed to describe the increasingly differentiated, onion-like layers of fused heavy elements within highly evolved stars; more energy also flows per unit mass. Stellar size, color, brightness, and composition all change while passing on up the hierarchy of complexity for all stars, each stage using more free energy rate density:

- From protostars at “birth” ($\Phi_m \approx 0.5 \text{ erg/s/g}$)
- To main-sequence stars at “mid-life” (~ 2)
- To red-giant stars in “old age” ($\sim 10^2$),
- To pre-supernovae near “death” ($\sim 5 \times 10^2$).

Those parenthetical values are the stars’ increased energy rate densities, plotted among other values in the lower circled inset of Fig. 6. At least as regards energy flow, material resources, and structural integrity while experiencing change, stars have much in common with life. This is not to say that stars are alive, which is an occasional misinterpretation of such a broad statement. Nor do stars evolve in the strict and limited biological sense; most researchers would agree that stars *develop*. Yet close parallels are apparent among stars, including selection, adaptation, and perhaps even a kind of stellar reproduction – a generational process – reminiscent of the following Malthusian-inspired scenario:

Galactic clouds spawn clusters of stars, only a few of which (the more massive ones unlike the Sun) enable other, subsequent populations of stars to emerge in turn, with each generation’s offspring showing slight variations, especially among the heavy elements contained

within. Waves of “sequential star formation” [23] propagate through many such clouds like slow-motion chain reactions over eons of time – shock waves from the death of old stars triggering the birth of new ones – neither any one kind of star displaying a dramatic increase in number nor the process of regeneration ever being perfect. Those massive stars selected by Nature to endure the fusion needed to produce heavy elements are in fact the very same stars that often produce shocks to create new populations of stars, thereby both episodically and gradually enriching the interstellar medium with greater elemental complexity on timescales measured in millions of millennia. As always, the necessary though perhaps not sufficient conditions for the growth of complexity depend on the environmental circumstances and on the availability of energy flows in such (here, galactic) environments. All of which is reminiscent of stellar “evolution”, minus any genes, inheritance, or overt function, for these are the value-added qualities of biological evolution that go well beyond physical evolution.

Continuing on and throughout the physical evolutionary subset of cosmic evolution, a *general* trend prevails, at least as pertains to Earth’s environment that set the stage for life:

- Young rocky planets have greater Φ_m ($\sim 10 \text{ erg/s/g}$) than normal stars and galaxies (~ 1)
- Hydrothermal vents on at least one of those planets have more (~ 50)
- Planetary climatespheres, such as Earth’s ocean-air interface, have even more (~ 100).

Note that some physical systems seem to be exceptions to the above findings, but upon closer inspection they are not exceptional at all. For example, that supernovae have very high values of Φ_m ($\gg 10^6 \text{ erg/s/g}$) does not violate our complexity metric. The reason is that supernovae are not organized systems, in fact just the opposite; as excellent examples of totally disorganized explosions of massive stars, they have too much energy flow that is well outside the optimal range for stars, and thus we should not expect to properly plot chaotic supernovae among other clearly ordered systems in Fig. 6. Pre-supernovae are noted there, representing an advanced stage of stellar evolution and growing complexity prior to explosion, but supernovae themselves are destructive events more typical of retreat from complexity toward simplicity. Likewise, bombs, flames, and many other damaging events do have large energy throughput yet do not belong on this curve, thus do not partake of a general trend toward rising complexity in Nature.

Biological Evolution

Plants are good examples of biological evolution. Here, we trace increasing complexity among plant life on Earth where neo-Darwinism is clearly at work, making use of free energy rate densities well higher than those for galaxies, stars, and planets. As shown in the middle circled inset of Fig. 6, energy-flow diagnostics display a definite increase in complexity among various plants that locally and temporarily decrease entropy. The most dominant process in Earth's biosphere – photosynthesis – well illustrates that complexity hierarchy [29]:

- From simple hay or grass ($\Phi_m \approx 5 \times 10^2$ erg/s/g)
- To inefficient pinewood ($\sim 3 \times 10^3$)
- To more efficient corn ($\sim 6 \times 10^3$)
- To well cultivated sugarcane ($\sim 10^4$).

System functionality and genetic inheritance are two factors, above and beyond mere system structure, which help to enhance order among animate systems that are clearly living compared to inanimate systems that are clearly not. Unsurprisingly, more complex life-forms require the acquisition of more energy per unit mass per unit time for their well being.

Energy flows through plants as captured solar energy during the act of photosynthesis converts H_2O and CO_2 into nourishing carbohydrates; the previous low-grade disordering sunlight becomes, in a relative sense, a higher-grade ordering form of energy compared to the even lower-grade (infrared) energy re-emitted by Earth. Likewise, as regards previously discussed physical evolution, energy flows through stars as gravitational potential energy during the act of star formation converts into radiation released by mature stars; high-grade energy produced by gravitational and nuclear events yield greater (thermal and elemental) organization, yet only at the expense of their environments into which stars emit low-grade light abundant in entropy. Either way, energy is a fuel for evolution, fostering some systems to utilize increased power densities while driving others to destruction and extinction.

Onward across the bush of life (or the arrow of time) – cells, tissues, organs, organisms – much the same metric holds for animals (all in units of erg/s/g):

- Cold-blooded reptiles have greater Φ_m ($\sim 10^4$) than globally averaged plants ($\sim 10^3$)
- Warm-blooded mammals typically have more ($\sim 5 \times 10^4$)
- Some birds, during complex flight, can achieve even more ($\sim 7 \times 10^4$).

Human life itself can also be examined on finer scale to show how energy usage continues upward (per unit mass) for more complex tasks [30,63]:

- Laboring humans have greater Φ_m ($\sim 6 \times 10^4$) than sedentary humans ($\sim 2 \times 10^4$)
- Vigorously bicycling and intricately sewing humans have more ($\sim 10^5$)
- Thinking human brains themselves have even more ($\sim 2 \times 10^5$).

Starting with life's precursor molecules (the realm of chemical evolution) and all the way to human brains exemplifying the most complex clump of animate matter known (neurological evolution), the same *general* trend characterizes the complexity of plants and animals as for stars and planets: The greater the perceived complexity of the system, the greater the flow of free energy density through that system – either to build it, or to maintain it, or both.

No strong distinctions are made here among Φ_m values for members of the animal kingdom, except to note that they are nearly all within a factor of ten of one another, confined between those for photosynthesizing plants on the one hand and central nervous systems on the other. The results are broadly consistent with measured specific metabolic rates scaling inversely with body mass, $M^{-1/4}$, among a wide variety of animal species [37,70]. Suffice it to say that animals in the main and in accord with Fig. 6 fit well within the complexity trends for the major evolutionary stages of life and for the intermediate phases of cosmic evolution.

Note, however, as for some non-living systems above, a minority of living systems seem exceptional, their values of Φ_m somewhat out of bounds among other equally advanced biological systems. Occasional life-forms also display retreat from complexity, such as some bats that move deeper into caves over generations and thus gradually lose their eyesight, or snakes and whales that eventually lost legs over time. Exceptions, real and apparent, to any rule will likely occur in a biosphere so rich in numbers and diversity as ours on Earth. For example, respiring bacteria are problematic at face value, having Φ_m values as much as 10^6 erg/s/g [45], thus comparable to higher forms of life. But microbes are so highly metabolic only when environmental resources warrant; none of them respire continuously. Measured rates are often quoted for peak periods of high reproductivity. By contrast, more than three-quarters of all soil bacteria are virtually dormant and thus have Φ_m values orders of magnitude less while eking out a living in nutrient-poor environments. When all microbial rates are time-weighted, microbes' average values range in the

thousands of erg/s/g, as expected for systems of intermediate complexity. Likewise and to note just one other seemingly exceptional animal, the Komodo Dragon can consume 80% of its body weight at one meal, yet not need another meal for a month – however, its time-averaged metabolic rate is much less than its maximum rate while eating.

Birds are another case in point, as they are well known to have high specific metabolic rates ($\sim 3 \times 10^4$ erg/s/g) during periods of peak activity, such as when earnestly foraging for food for their nestlings. But, once again, upon closer inspection, they are recognized not to be exceptions at all. That the smallest animals have the highest such rates is often taken [70] as an explanation of their frequent eating habits (hummingbirds ingest up to half of their body mass daily), extreme levels of activity (bumblebees flap their wings up to 160 times per second), and relatively short lifespans (few years typically, given the heavy toll on their metabolic functions); those are operational tasks, namely, function, not structure. Given that birds and bees normally function in a three-dimensional aerial environment while solving advanced tasks in spatial geometry, materials science, aeronautical engineering, molecular biochemistry, and social stratification, then perhaps they ought to have large values of Φ_m . That birds, while airborne, have higher values than for resting humans should not surprise us since we ourselves have not solved the art of flying, an admittedly complex task. By contrast, when bicycling vigorously or sewing intricately, our specific metabolic rates do exceed even those of birds in flight as noted above. Moreover, when humans do fly, aided by built aircraft, machine values of Φ_m are indeed higher ($\sim 10^7$ erg/s/g) than for even the most impressively ingesting hummingbirds, as discussed in the next section on cultural evolution.

Cultural Evolution

Society is a good example of cultural evolution. Here, the cosmic-evolutionary chronicle continues, yet with greater normalized energy flows to power our obviously complex civilization. As plotted in the upper circled inset of Fig. 6, social progress can be tracked, again in terms of energy consumption, for a variety of human-related cultural advances among our hominid ancestors. Quantitatively, that same energy rate density increases:

- From hunter-gatherers of a few million years ago ($\Phi_m \approx 10^4$ erg/s/g)
- To agriculturists of several thousand years ago ($\sim 10^5$)
- To industrialists of two hundred years ago ($\sim 5 \times 10^5$)
- To western society today, on average ($\sim 10^6$).

That a cluster of brainy organisms working collectively in a social group is more energy intensive per capita (and thus more complex) than each of its individual human members – at least as regards the present criterion for order of free energy rate usage per capita – is a good example of a “whole greater than the sum of its parts”, in this case for the open, non-equilibrated society that constitutes modern civilization [15].

The road to today’s technological society was unquestionably built with increased energy use, as has been earlier recognized by many cultural historians (e. g., White [71]; Cook [18]; Brown [6]; Jervis [35]; McNeill and McNeill [50]), who noted the importance of rising energy expenditure per capita, a factor also more recently emphasized by practitioners of “big history” (Christian [17]; Spier [65]; Aunger [1]), a newly emerging subject that treats conventional history more deeply, indeed parallels the scenario of cosmic evolution.

Machines, too, and not just computer chips, but also ordinary motors and engines that typified the fast-paced economy of the 20th century, can be cast in evolutionary terms – though here the mechanism is less Darwinian than Lamarckian [39], with the latter’s emphasis on accumulation of acquired traits. Either way, energy remains a driver, and with rapidly accelerating pace. Aircraft engines, for example, display clear evolutionary trends as engineering improvement and customer selection over generations of products have made engines more intricate, complex, and efficient, all the while utilizing enriched flows of energy density [63]:

- Gas-guzzling SUVs have greater Φ_m ($\sim 10^6$ erg/s/g) than model-T automobiles ($\sim 10^5$)
- Boeing-747 jumbo jets of the last few decades have more ($\sim 10^7$)
- Military F-117 stealth aircraft of the present have even more ($\sim 10^8$).

Finer-scale evolutionary analysis of many technological advancements display evident progress toward greater complexity, such as for the typical American passenger car over the past two decades that can be cast in terms of growing horsepower-to-weight ratios provided by the US Highway Traffic Safety Administration: $\Phi_m = 5.9 \times 10^5$ erg/s/g in 1978, 6.8×10^5 in 1988, and 8.3×10^5 in 1998. Not surprisingly, silicon chips – a cultural icon of today’s vibrant, digitized 21st-century economy – have immense flows of energy density, currently reaching values of $\sim 10^{10}$ erg/s/g mostly caused by chip miniaturization despite reduced power consumption.

Rare exceptions in cultural evolution’s apparent drive toward greater complexity sometimes cause regression to-

ward simpler systems, much as for minor aspects of the physical and biological subsets of cosmic evolution. Collapse of civilizations, either internally (because of societal conflict) or externally (owing to environmental change), that then resort to social chaos is are examples of infrequent retreat from society's overall drive toward greater complexity [20,67].

Occasional exceptions aside, increasingly sophisticated technological gadgets, under the Lamarckian pressure of dealer competition and customer selection, do in fact show increases in Φ_m values with product improvement over the years. Not only can the cultural evolution of machines be traced and their Φ_m values computed as noted above for engines, but similar advances can also be tracked for a whole array of silicon-based devices now undating our global economy. In keeping with the upper part of the curve in Fig. 6, many of these cultural devices do have complexity measures comparable to, and often greater than, biological systems, including brains. Technology clearly allows individual humans to accomplish things that cannot be done by us alone, and usually faster too, which partly explains why most of society continues to embrace technology, despite its pitfalls, to aid our senses and improve our increasingly complex lives.

Conclusions and Future Directions

This article has taken the liberty of extrapolating the word "evolution" in an intentionally broad way to analyze change on all spatial and temporal scales. Within the grand context of cosmic evolution, common threads have been identified linking a wide spectrum of ordered structures during an extremely long period of natural history, from big bang to humankind. More than any other single factor, energy flow seems to be a principal means whereby Nature's diverse systems naturally became increasingly complex in an expanding Universe, including not only galaxies, stars and planets, but also lives, brains and civilization.

The scenario of cosmic evolution accords well with observations demonstrating an entire hierarchy of structures to have emerged, in turn, throughout the history of the Universe: particles, galaxies, stars, planets, life, intelligence, and culture. As a general trend, an overall increase in complexity is apparent with the relentless march of time, without any progress, purpose or design implied. With cosmic evolution as our guide, we can begin to understand the environmental conditions needed for matter to have become increasingly ordered, organized, and complex. This rise in order, form, and structure violates no laws of physics, and certainly not those of modern thermodynamics. Nor is the idea of ubiquitous change novel to

modern worldviews. What is new and exciting is the way that frontier, non-equilibrium science now helps us unify a holistic cosmology wherein complex life plays an integral role – namely, to address the origin and evolution of all things by means of logic, rationality, and the methods of natural science.

When studying complexity in Nature, some researchers prefer the concept of information rather than energy, often becoming displeased when the former is put aside as done in this article. But information content has had a muddled history full of assorted interpretations – meaningful information, the value of information, Shannon information, algorithmic information, raw information. Furthermore, no one has yet shown quantitatively and unambiguously, that information content rises throughout the ages for physical, biological, and cultural systems. A useful future research direction would clarify the role of the information sciences in complexity studies and cosmic evolution, including the possibility that information is merely other forms of energy – energy acquired, energy stored, and energy expressed.

By contrast, it is encouraging that a single quantity such as free energy rate density, defined here clearly and with units well understood, affects all ordered systems, given that some systems are regulated by gravity and others practically not. Thermodynamics does pertain to all such systems universally, whether massive enough like stars subject to gravity or less so as for life-forms governed mostly by electromagnetism. Energy flow is a common feature of every open, non-equilibrium system, and it is insightful not only that one such quantity is uniformly applicable but also that it seems to map reasonably well the rise of complexity among many known systems. Gravitational force in physics, natural selection in biology, and technological innovation in culture are all examples of diversified actions that can give rise to accelerated rates of change at locales much smaller than the Universe per se – such as the islands of order that are stars, life, and civilization itself. Indeed, our use of energy wisely and optimally will likely guide our fate along the future arrow of time, for we humans are also part of the cosmic-evolutionary scenario.

Humankind is now moving toward a time, possibly as soon as within a few generations, when we shall no longer be able to expect Nature to easily provide for our own survival. Rather, civilization on Earth will either have to adapt to the natural environment with ever-accelerating speed, or to generate artificially controlled environments (either on or beyond Earth) needed for our ecological existence. From two of Nature's most advanced yet locally ordered systems – society and machines – will likely emerge

a symbiotic technoculture, the epitome (thus far as best we know) of complexity in the Universe – a new technology-based system that will likely require even greater values of energy rate density, as the curve in Fig. 6 continues racing upward. Can humanity endure despite its own increasing complexity? Or will our species transform into some other intricate entity as complexity continues to rise?

I thank *la Fondation Wright de Geneve* for support and encouragement of this research, which comprises the intellectual theme of the Wright Center for Science Education at Tufts University. The author is also associated with the Harvard College Observatory, where he teaches a course on the topic of cosmic evolution and complexity science, for which a multi-media web site has been built http://www.tufts.edu/as/wright_center/cosmic_evolution.

Bibliography

1. Aunger R (2007) Major transitions in 'big' history. *Tech Forecast Soc Chang* 68:27
2. Bennett CL et al (2003) Wilkinson Microwave Anisotropy Probe (WMAP) basic results. *Astrophys. J Suppl Ser* 148:1
3. Blum HF (1968) Time's arrow and evolution. Princeton University Press, Princeton
4. Bonner JT (1988) Evolution of complexity. Princeton University Press, Princeton
5. Brooks DR, Wiley EO (1988) Evolution as entropy. University of Chicago Press, Chicago
6. Brown H (1976) Energy in our future. *Ann Rev Energy* 1:1
7. Chaisson EJ (1981) Cosmic dawn: origins of matter and life. Atlantic Monthly Press, Boston
8. Chaisson EJ (1987) The life era (appendix). Atlantic Monthly Press, New York
9. Chaisson EJ (1998) The cosmic environment for the growth of complexity. *BioSystems* 46:13
10. Chaisson EJ (2001) Cosmic evolution: the rise of complexity in nature. Harvard University Press, Cambridge
11. Chaisson EJ (2003) A unifying concept for astrobiology. *Int J Astrobio* 2:91
12. Chaisson EJ (2004) Complexity; an energetics agenda. *Complex J Santa Fe Inst* 9:14
13. Chaisson EJ (2005) Non-equilibrium thermodynamics in an energy-rich universe. In: Kleidon A, Lorenz RD (eds) *Non-equilibrium Thermodynamics and the Production of Entropy*. Springer, Berlin
14. Chaisson EJ (2006) Epic of evolution: seven ages of the cosmos. Columbia University Press, New York
15. Chaisson EJ (2009) Cosmic evolution: state of the science. In: Dick S (ed) *Cosmos and Culture*. NASA, Washington
16. Chambers R (1844) *Vestiges of the natural history of creation*. Churchill, London
17. Christian D (2004) *Maps of time: introduction to big history*. University California Press, Berkeley
18. Cook E (1971) The flow of energy in an industrial society. *Sci Am* 224:135
19. Darwin C (1859) *On the origin of species*. J Murray, London
20. Diamond J (2005) *Collapse: how societies choose to fail or succeed*. Viking, New York
21. Dyke C (1988) Cities as dissipative structures. In: Weber BH et al (eds) *Entropy, Information, and Evolution*. MIT Press, Cambridge
22. Dyson F (1979) Time without end: physics and biology in an open universe. *Rev Mod Phys* 51:447
23. Elmegreen B, Lada C (1977) Sequential star formation of subgroups in OB associations. *Astrophys J* 214:725
24. Fox RF (1988) *Energy and the evolution of life*. Freeman, San Francisco
25. Frautschi S (1982) Entropy in an expanding universe. *Science* 217:593
26. Gold T (1962) The arrow of time. *Am J Phys* 30:403
27. Haken H (1978) *Synergetics*. Springer, Berlin
28. Haken H (1975) Cooperative phenomena in systems far from thermal equilibrium and in nonphysical systems. *Rev Mod Phys* 47:67
29. Halacy DS (1977) *Earth, water, wind and sun*. Harper & Row, New York
30. Hammond KA, Diamond J (1997) Maximal sustained energy budgets in humans and animals. *Nature* 386:457
31. Henderson L (1913) *Fitness of the environment*. Macmillan, New York
32. Hofkirchner W (ed) (1999) *The quest for a unified theory of information*. Gordon & Breach, Amsterdam
33. Jantsch E (1980) *Self-organizing universe*. Pergamon, Oxford
34. Jaynes ET (1957) Information theory and statistical mechanics. *Phys Rev* 108:171
35. Jervis R (1997) *System effects: complexity in political and social life*. Princeton University Press, Princeton
36. Kauffman S (1993) *The origins of order*. Oxford University Press, Press
37. Kleiber M (1961) *The fire of life*. Wiley, New York
38. Kleidon A, Lorenz RD (eds) (2005) *Non-equilibrium thermodynamics and the production of entropy*. Springer, Berlin
39. Lamarck J-B (1809) *Philosophie Zoologique*. Editions du Seuil, Paris
40. Layzer D (1976) The arrow of time. *Astrophys J* 206:559
41. Layzer D (1988) Growth of order in the universe. In: Weber BH et al (eds) *Entropy, Information, and Evolution*. MIT Press, Cambridge
42. Lewin R (1992) *Complexity*. Macmillan, New York
43. Lineweaver C (2005) Cosmological and biological reproducibility: limits on maximum entropy production principle. In: Kleidon A, Lorenz RD (eds) *Non-equilibrium Thermodynamics and the Production of Entropy*. Springer, Berlin
44. Lotka A (1922) Contribution to the energetics of evolution. *Proc Nat Acad Sci USA* 8:147
45. Margulis L, Sagan D (1986) *Microcosmos*. Simon & Schuster, New York
46. Marijuan PC et al (eds) (1996) *First conference on foundations of information sciences*. Biosystems 38:87
47. Matsuno K (1989) *Protobiology: Physical Basis of Biology*. CRC Press, Florida
48. Mayr E (1997) *This is biology*. Harvard University Press, Cambridge
49. McMahon T, Bonner JT (1983) *On size and life*. Freeman, San Francisco
50. McNeill JR, McNeill WH (2003) *The human web*. Norton, New York
51. Morowitz HJ (1968) *Energy flow in biology*. Academic Press, New York

52. Morrison P (1964) A thermodynamic characterization of self-reproduction. *Rev Mod Phys* 36:517
53. Odum HT (1971) *Environment, power, and society*. Wiley, New York
54. Prigogine I (1961) *Introduction to thermodynamics of irreversible processes*. Wiley, New York
55. Prigogine I, Nicolis G, Babloyantz A (1972) Thermodynamics of evolution. *Physics Today* 11:23
56. Reeves H (1981) *Patience dans l'azur: l'évolution cosmique*. Editions du Seuil, Paris
57. Sagan C (1980) *Cosmos*. Random House, New York
58. Salk J (1982) *An evolutionary approach to world problems*. UNESCO, Paris
59. Schneider ED, Kay JJ (1995) Order from disorder: thermodynamics of complexity in biology. In: Murphy M, O'Neill L (eds) *What is life*. Cambridge University Press, Cambridge
60. Schroedinger E (1944) *What is life?* Cambridge University Press, Cambridge
61. Shannon CE, Weaver W (1949) *Mathematical theory of communication*. University of Illinois Press, Champaign-Urbana
62. Shapley H (1930) *Flights from chaos*. McGraw Hill, New York
63. Smil V (1999) *Energies*. MIT Press, Cambridge
64. Spencer H (1896) *A system of synthetic philosophy*. Williams and Norgate, London
65. Spier F (2005) How big history works. *Soc Evol Hist* 4:25
66. Szathmari E, Maynard Smith J (1995) The major evolutionary transitions. *Nature* 374:227
67. Tainter JA (1988) *The collapse of complex societies*. Cambridge University Press, Cambridge
68. Ulanowicz RE (1986) *Growth and development*. Springer, Berlin
69. Weber BH, Depew DJ, Smith JD (eds) (1988) *Entropy, information, and evolution*. MIT Press, Cambridge
70. West GB, Brown JH, Enquist BJ (1999) Fourth dimension of life. *Science* 284:1677
71. White LA (1959) *The evolution of culture*. McGraw-Hill, New York
72. Whitehead AN (1925) *Science and the modern world*. Macmillan, New York
73. Wicken JS (1987) *Evolution, thermodynamics, and information*. Oxford University Press, Oxford
74. von Bertalanffy L (1932) *Theoretische biologie*. Borntraeger, Berlin
75. von Bertalanffy L (1968) *General system theory*. Braziller, New York

Exobiology (theoretical), Complexity in

AXEL BRANDENBURG
Nordita, AlbaNova University Center,
Stockholm, Sweden

Article Outline

[Glossary](#)
[Definition of the Subject](#)
[Introduction](#)

[Homochirality](#)
[Establishing Hereditary Information](#)
[Alteration of the Environment by Early Life](#)
[Conclusions](#)
[Bibliography](#)

Glossary

Chiral, achiral and racemic A molecule is chiral if its three-dimensional structure is different from its mirror image. Such molecules tend to be optically active and turn the polarization plane of linearly polarized light in the right- or left-handed sense. Correspondingly, they are referred to as D- and L-forms, which stand for dextrorotatory and levorotatory molecules. An achiral molecule is mirror-symmetric and does not have this property. A substance is racemic if it consists of equally many left- and right-handed molecules. A polymer is said to be isotactic if all its elements have the same chirality.

Enantiomers and enantiomeric excess Enantiomers are a pair of chiral molecules that have opposite handedness, but are otherwise identical. Enantiomeric excess, usually abbreviated as e.e., is a normalized measure of the degree by which one handedness dominates over the other one. It is defined as the ratio of the difference to the sum of the two concentrations, so e.e. always falls between -1 and $+1$.

Epimerization and racemization Epimerization is a spontaneous change of handedness of one sub-unit in a polymer. Racemization indicates the loss of a preferred handedness in a substance.

Catalysis and auto-catalysis Catalysts are agents that lower the reaction barrier. A molecule reacts with the catalyst, but at the end of the reaction, the catalyst emerges unchanged. This is called catalysis. In auto-catalysis the catalyst is a target molecule itself, so this process leads to exponential amplification of the concentration of this molecule by using some substrate. Biological catalysts are referred to as enzymes.

Nucleotides and nucleic acids Nucleotides are monomers of nucleic acids, e.g., of RNA (ribonucleic acid) or DNA (deoxyribonucleic acid). They contain one of four nucleobases (often just called bases) that can pair in a specific way. Nucleotides can form polymers, and their sequence carries genetic information. One speaks about a polycondensation reaction instead of polymerization because one water molecule is removed in this step. Other nucleotides of interest include peptide nucleic acid or PNA. Here the backbone is made of peptides instead of sugar phosphate.

Peptides and amino acids Amino acids are molecules of the general form $\text{NH}_3\text{-CHR-COOH}$, where R stands for the rest, which makes the difference between different amino acids. For glycine, the simplest amino acid, we have $\text{R}=\text{H}$, so two of the bonds on the central C atom are the same and the molecule is, therefore, chiral. For alanine, $\text{R}=\text{CH}_3$, so all four bonds on the central C atom are different, therefore this molecule is chiral. A peptide is a polymer generated through a polycondensation reaction of amino acids. Peptides are also referred to as proteins.

Solar constant and albedo The solar constant is the total flux of energy from the Sun above the Earth's atmosphere. Its current value is $S = 1.37 \text{ kW m}^{-2}$, but it was about 30% lower when the solar system was young (10^8 yr ago, say), so S is not a constant. The albedo A is the fraction of the Sun's energy that is reflected from the surface of the Earth, e. g., by clouds and snow and, to a lesser extent, by land masses and oceans.

Photosynthesis and carbon fixation Photosynthesis uses light to reduce CO_2 and to produce oxygen either as free molecular oxygen or in some other chemical form. This process removes CO_2 from the atmosphere and produces biomass, which is written in simplistic form as $(\text{CH}_2\text{O})_n$. This process is referred to as carbon fixation.

Life A preliminary definition of life involves replication and death, coupled to a metabolism that utilizes any sort of available energy. Life is characterized further by natural selection to adapt to environmental changes and to utilize available niches. A proper definition of life is difficult given that all life on Earth can be traced back to a single common ancestor. Any definition of life may need to be adjusted if extraterrestrial or artificial life is discovered.

Definition of the Subject

Astrobiology is concerned with questions regarding possible origins of life on Earth and elsewhere in the Universe. Although to date there has been no detection of extraterrestrial life, it is generally assumed that life could be widespread, provided certain conditions of habitability are met. A common implicit hypothesis in astrobiology is that life can emerge spontaneously once certain environmental conditions are met. This implies that there may well have been multiple geneses, separated only by global extinction events, such as major impacts by other celestial bodies [11].

Four important discoveries can be named that have provided impetus to the field of astrobiology.

1. More than 300 extrasolar planets have been discovered since 1995, providing explicit targets for detecting life outside the solar system.
2. Recent Mars missions have provided evidence for liquid water on the surface of Mars in the past and possibly even in the present time. This has fostered the search for techniques to detect microbial life on Mars.
3. On earth the carbon in very old sedimentary rocks dating back 3.8 Gyr ago shows a consistently lower ^{13}C to ^{12}C abundance ratio, which is normally indicative of life. This lends support to the notion that life may have been present as soon as the Earth's surface became hospitable.
4. The discovery of extremophiles on Earth has considerably extended the definition of habitability to include extreme temperatures, pressures and pH values, high salinity as well as high radiation levels. This has raised hopes of finding life elsewhere in our solar system.

Astrobiology thus comprises several scientific disciplines: astronomy, geology, chemistry, and biology. Therefore, much of the original literature tends to appear in journals in these various fields. We should also mention that there are technological attempts being made to produce artificial life [35]. While this approach is not aimed at reproducing the origin of life on Earth, it may still be useful for prompting our imaginations in understanding the transition from nonliving to living matter.

Introduction

Since the early days of nonlinear dynamics and non-equilibrium thermodynamics, it has been clear that one of the ultimate applications of this theory might be to facilitate an understanding of the transition from non-living to living matter. The main reason is obviously that living systems are very far from equilibrium – as indicated by the high degree of order, and hence the low entropy, of living systems relative to their environment.

As early as 1952, Turing [44] proposed the idea that chemical reaction-diffusion systems might provide a tool for studying biochemical pattern formation, which has increased our understanding of the laws of nature far from equilibrium, where life occurs. This idea was followed up in the late 1960s by Prigogine [33,34] who suggested that dissipative structures have great importance in establishing a physical description of living matter. A general theory of autocatalytic molecular evolution was developed in 1971 by Eigen [15], who argued that in a single micro-environment, only a single handedness can result from a single event. In particular, the famous chicken and egg problem that occurs in biology at different levels was identi-

fied as a Hopf bifurcation. A Hopf bifurcation describes the spontaneous emergence of an oscillating solution once some stability threshold has been crossed. The mathematics of this is familiar to any physicist, but it requires that the equations describing the relevant physics are known. In biology, it is not even clear that the various phenomena can be described by equations. A first detailed attempt in this direction was, indeed, that of Eigen. However, the equations governing the emergence of life are only phenomenological ones. Nevertheless, these approaches are invaluable in that they help give the origin of life question a mathematical basis.

One of the earliest anticipated forms of life that is still similar to present life is the RNA world [21], whereby simple RNA molecules with functional behavior self-reproduce using genetic information encoded either in the same or in other participating RNA molecules. Obviously, there are tremendous difficulties, given that RNA is too complicated a molecule to be synthesized abiologically. A significantly simpler molecule is peptide nucleic acid or PNA [30], in which the backbone consists of peptide instead of sugar phosphate. Nevertheless, the difficulty of producing RNA remains.

There is no firm idea where on Earth such molecular replication may have originally taken place. Frequently discussed scenarios include hydrothermal vent systems [37], and also beach scenarios that are subject to tides leading to cyclic changes in concentration [24] as well as to repeated wetting and drying [7].

An early experiment that contributed significantly to the research into the origins of life was the Urey–Miller experiment [26], which demonstrated the spontaneous production of amino acids in a reducing atmosphere consisting of H_2O vapor, CH_4 , NH_3 , and H_2 with an energy supply in the form of sparks. More recent experiments also allow for the presence of CO_2 , which now seems unavoidable on the early Earth, given that it is continuously replenished through outgassing by volcanoes.

In the following, we review some areas in which there has been considerable cross-fertilization between astrobiology and nonlinear dynamics. We begin by discussing a phenomenon that is believed to have taken place around the time of the origin of life, namely the establishment of a definitive handedness of biomolecules that is inherent in DNA and RNA (D-form) and in amino acids (L-form). Next, we discuss constraints on the evolution of hereditary information and, finally, we review some models that characterize the alteration of the terrestrial environment by early life. In addition to any of these physical effects, random fluctuations lead inevitably to local imbalances between the concentrations of molecules of D- and L-form.

In the following, we discuss mechanisms that can lead to an exponential amplification of enantiomeric excess. For a recent review of these ideas see [32].

Homochirality

Theories of a chemical origin of life involve polymerization of nucleotides that carry and utilize genetic information. Ribonucleotides possess chirality, i.e., they are different from their mirror images. All known life forms use ribonucleotides of the so-called D-form (right-handed), as opposed to the L-form (left-handed). These two molecules are referred to as opposite *enantiomers*. In most cases these different enantiomers are optically active, i.e., they turn the polarization plane of linearly polarized light in a right-handed or left-handed sense.

Any non-enzymatic synthesis of ribonucleotides would have produced a mixture of equally many right- and left-handed building blocks. Technically this is referred to as a *racemic* mixture of these molecules. However, it is known experimentally that, in a racemic mixture of mononucleotides, polymerization is quickly terminated after the first or second polymerization step [23]. This is generally referred to as *enantiomeric cross-inhibition*, which was long thought to be a serious obstacle to a chemical origin of life. It was therefore thought to be necessary that life evolved only in a homochiral environment. Moreover, it would then be necessary that the degree of enantiomeric purity must have been very high. This is important because it rules out a number of physical mechanisms based on the enantioselective effects of circularly polarized radiation, magnetic fields, and the parity-breaking property of the electroweak force.

The Frank-Mechanism

A general mechanism for producing complete homochirality was proposed in 1953 by Frank [19] based on the assumed effects of auto-catalysis and what he called mutual antagonism. In fact, the enantiomeric cross-inhibition mentioned above can be thought of as a possible example of mutual antagonism. Frank's model is characterized by the following set of three reactions:



where D and L denote monomers of the two enantiomers, S is a substrate from which monomers could be formed

via auto-catalysis, and DL are inactive dimers that are lost from the system. (At this level of simplification no distinction is made between DL and LD . This simplification will later be relaxed.) The parameters k_C and k_I characterize the reaction speeds. These reactions translate to the following set of equations for the concentrations of D , L , S , and DL ,

$$\frac{d}{dt}[D] = +k_C[S][D] - k_I[D][L], \quad (4)$$

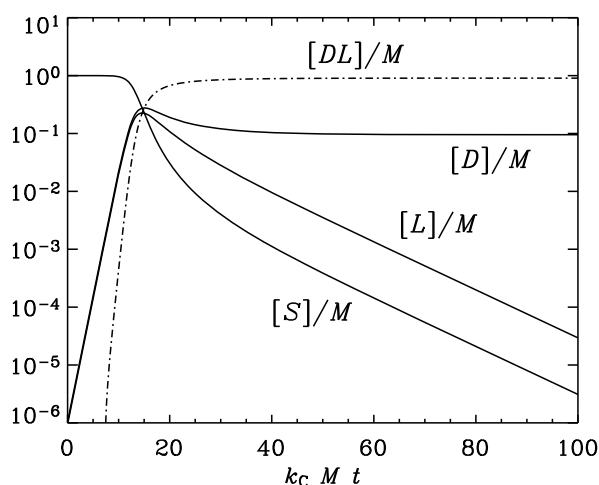
$$\frac{d}{dt}[L] = +k_C[S][L] - k_I[D][L], \quad (5)$$

$$\frac{d}{dt}[S] = -k_C[S]([D] + [L]), \quad (6)$$

$$\frac{d}{dt}[DL] = +2k_I[D][L]. \quad (7)$$

These equations imply that the total mass of all building blocks (including the substrate), is constant, i.e. $[D] + [L] + [S] + [DL] = \text{const} \equiv M$.

This system of equations describes the continued autocatalytic production of DL , D and L until the substrate S is exhausted, i.e., $[S] = 0$. However, as long as $[S]$ is still finite, the asymmetry, $\mathcal{A} = [D] - [L]$, grows quasi-exponentially, proportional to $\exp(\int [S] dt)$. A numerical example of this is shown in Fig. 1.



Exobiology (theoretical), Complexity in, Figure 1

Solution of Eqs. (4)–(6) for $k_I = k_C$. Both D and L grow exponentially until $[D] + [L]$ becomes comparable to the constantly declining substrate concentration $[S]$. At the same time the production of DL removes an equal amount of D and L , but this effect primarily affects those enantiomers that are already in the minority. In this calculation an initial asymmetry (here 10%) of $[D] - [L]$ grows until saturation. At the end, $[D]$ has reached 100% enantiomeric excess, but this happened at the expense of producing a large number of inactive heterochiral dimers DL .

In the numerical example above, we started with very small initial concentrations. Another possibility is to start with a perturbed racemic solution. The racemic solution is given by $[D] = [L] = \lambda/k_I$, where $\lambda = k_C[S]$ is the instantaneous growth rate due to auto-catalysis. Under the assumption that λ can be treated as a constant (i.e., when the system is still nearly racemic), a linear stability analysis shows that the enantiomeric excess,

$$\text{e.e.} = \frac{[D] - [L]}{[D] + [L]} \quad (8)$$

grows exponentially. This means that the racemic solution is unstable and that the mechanism for achieving homochirality is based on a linear instability.

Continued Polymerization

There is a priori no good reason to permit the production of heterochiral dimers DL , but not homochiral dimers DD and LL , i.e.,



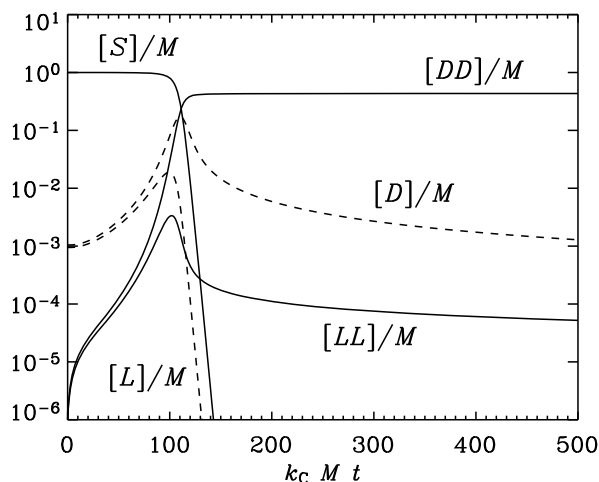
The importance of such reactions was stressed in a review by Blackmond [1], who also introduced an additional modification that consists in the assumption that, rather than monomers, the homochiral dimers DD and LL catalyze the production of monomers, i.e., reactions (1) and (2) are replaced by



This model is similar to the original Frank model provided there is a way of getting rid of those homochiral dimers that are in the minority. This requires enantiomeric cross-inhibition for dimers to form heterochiral trimers, i.e., we need the additional reactions



A solution to the corresponding reaction equations is given in Fig. 2. Reaction calorimetry supports the assumption that dimers and not monomers are the relevant catalysts [1]. This seems to apply, in particular, to the first autocatalytic reaction ever found that enhances enantiomeric excess [42]. In this reaction (sometimes referred to as the Soai reaction), the substrate is pyridine-3-carbaldehyde and the chiral molecule of either D- or L-form is 3-pyridyl

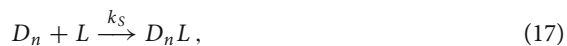


Exobiology (theoretical), Complexity in, Figure 2

Solution of Eqs. (4), (6) supplemented by kinetic equations corresponding to the reactions Eqs. (9), (14), for $k_S = k_I = k_C$. Again, an initial 10% asymmetry of $[D] - [L]$ grows until $[D] + [L]$ becomes comparable to the constantly declining substrate concentration $[S]$. The monomers polymerize into dimers DD and LL . Toward the end, $[DD]$ reaches 100% enantiomeric excess

alkanol, which thus acts as an asymmetric autocatalyst to produce more of itself. In this reaction, however, dialkylzinc acts as an additional achiral catalyst. While the Soai reaction is important as a first explicit example of an autocatalytic reaction that enhances the enantiomeric excess, it is not normally regarded as directly important for astrobiology.

The polymerization model was developed by Sanders [39], who included arbitrarily many polymerization steps of the form



The basic outcome of this and similar models is always the same as in the original Frank model, except that the polymerization model is also capable of displaying interesting wave-like dynamics in time-dependent histograms of different polymers [5].

Spatially Extended Models

In reality, there are limits as to the degree to which a system can be considered fully mixed. In general, $[D]$ and $[L]$

should be functions of time *and* space, i. e. $[D] = [D](t, \mathbf{x})$ and $[L] = [L](t, \mathbf{x})$. Assuming that there is only molecular diffusion, the relevant reaction equations are to be supplemented by additional diffusion terms,

$$\frac{d}{dt}[D_n] = R_n^{(D)} + \kappa \nabla^2 [D_n], \quad (19)$$

$$\frac{d}{dt}[L_n] = R_n^{(L)} + \kappa \nabla^2 [L_n], \quad (20)$$

where $R_n^{(D)}$ and $R_n^{(L)}$ are the right hand sides of the reaction equations.

If there were only one type of handedness, the resulting equation would be reminiscent of the Fisher equation [29],

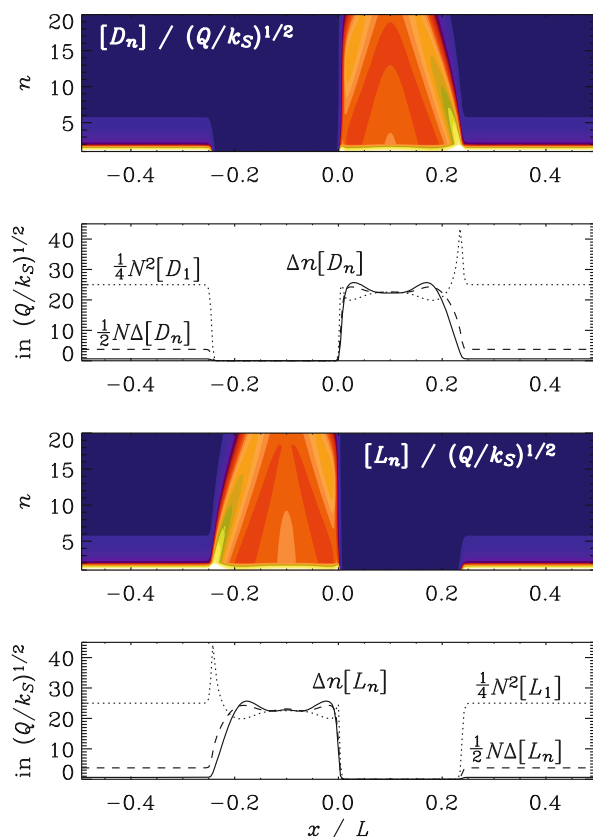
$$\frac{df}{dt} = \lambda(1-f)f + \kappa \nabla^2 f, \quad (21)$$

which admits propagating front solutions with front speed $v_{\text{front}} = 2\sqrt{\kappa\lambda}$. Here, f could represent the local concentration of some disease in models of the spread of epidemics, for example.

In the present case, there are two fields, one of each handedness. It is instructive to refer to these fields as populations, which is suggestive of their ability to replicate, migrate, become extinct, and to compete against a population of opposite handedness. Each population is able to expand into unpopulated space at a speed given approximately by v_{front} , but once two opposing handednesses come into contact, there is an impasse and the propagation comes to a halt. A snapshot of a one-dimensional model illustrating polymer length as a function of position is shown in Fig. 3 for populations of opposite handedness that have come into contact.

The overall dynamics of symmetry breaking are well characterized by a low order truncation, where the model is truncated at $n = 2$ and the evolution of the $n = 1$ modes is assumed to be enslaved by the evolution of the $n = 2$ modes [5]. An example of such a solution is shown in Fig. 4, which shows the evolution in a space-time diagram, where two populations of opposite handedness expand into unpopulated space until two opposite populations come into contact.

In two and three dimensions, a front between two opposing enantiomers is (in general) curved, in which case it can propagate diffusively in the direction of curvature. This is caused by the fact that the inner front between two populations is slightly shorter than the outer one. (Only the immediate proximity of a front matters; what lies behind it is irrelevant if it is of the same handedness.) Indeed, on a two-dimensional surface the inner front is shorter than the outer one by $2\pi d$ – independent of radius. Here, d is the front thickness, which is of the order of $d \approx (\kappa/\lambda)^{1/2}$.



Exobiology (theoretical), Complexity in, Figure 3

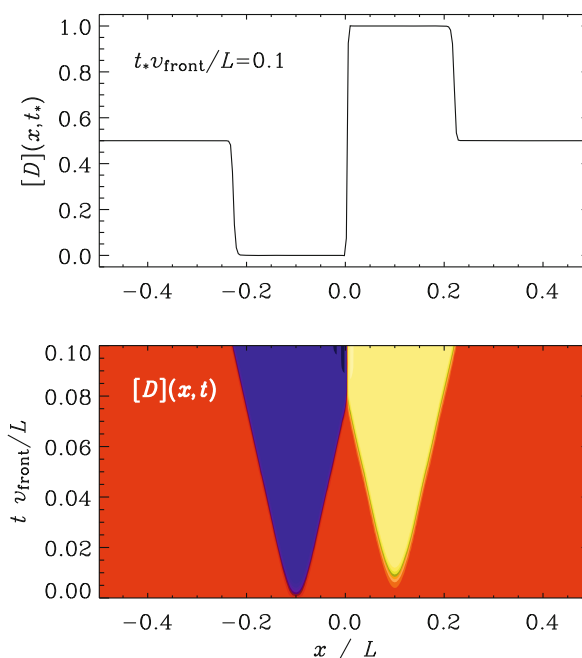
Color/gray scale plots of $[D_n]$ and $[L_n]$ for $t/\tau_{\text{diff}} = 0.8$ as a function of x and n , and the corresponding dependencies of $\sum_{n=1}^N n[D_n]$ and $\sum_{n=1}^N n[L_n]$ (solid line), compared with $\frac{1}{2}N \sum_{n=1}^N [D_n]$ and $\frac{1}{2}N \sum_{n=1}^N [L_n]$ (dashed line), and $\frac{1}{4}N^2[D_1]$ and $\frac{1}{4}N^2[L_1]$ (dotted line), all in units of $(Q/k_S)^{1/2}$. The normalized diffusivity is $\kappa/(L^2\lambda_0) = 10^{-2}$ and $N = 20$. Adapted from [4]

It turns out that in two dimensions the rate of change of the integrated asymmetry, $\mathcal{A} = \int ([D] - [L]) d^2x$, depends only on the number of topologically distinct rings or islands. Once an island is wiped out, the rate of change of \mathcal{A} changes abruptly and then stays constant until the next island gets wiped out. So the enantiomeric excess,

$$\text{e.e.} = \frac{\int ([D] - [L]) d^2x}{\int ([D] + [L]) d^2x}, \quad (22)$$

increases with time in a piecewise linear fashion.

Even if at each point homochirality could be reached rapidly (time scale λ^{-1}), global homochirality requires that one population wipes out the other one completely. Diffusion is usually too slow to lead to any significant mixing and hence to global homochirality. However, there could be circumstances where such mixing is sped up by

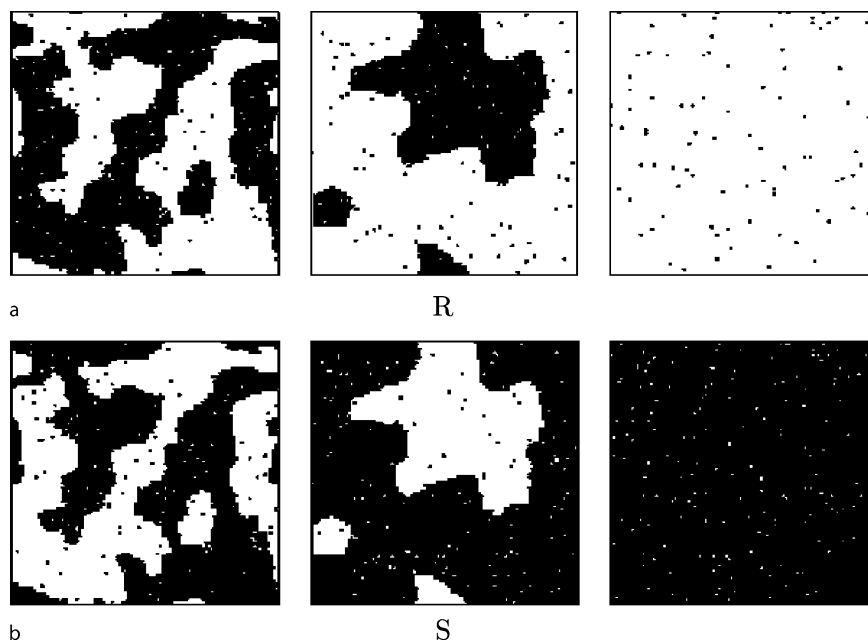


Exobiology (theoretical), Complexity in, Figure 4

Profile of $[D](x, t_*)$ and space-time diagram of $[D](x, t)$ for the one-dimensional problem without advection and an initial perturbation corresponding to a weak (amplitude 0.01) right-handed excess at $x/L = 0.1$ (marked in white or yellow) and a somewhat stronger (amplitude 0.3) left-handed excess at $x/L = -0.1$ (marked in dark or blue). Note the propagation of fronts with constant speed if the exterior is racemic (i.e. $[D] = [L] = 1/2$, shown in medium shades or red) and a non-propagating front when the chirality is opposite on the two sides of the front. The normalized diffusivity is $\kappa/(L^2\lambda_0) = 10^{-2}$, i.e. the same as in Fig. 3. Adapted from [4]

something like “turbulent” transport. In the case of the Earth, the slowest relevant transport is in the Earth’s mantle, part of which is now associated with what is called the deep biosphere. Assuming that multiple geneses of life is possible, this would raise the question of whether a simultaneous co-existence of different handednesses on different parts of the early Earth would have been possible. It is, however, unclear whether this possibility could have left any traces that would still be detectable today.

Another approach to solving the problem of spatially extended chemistry is by means of cellular automata. In this approach, points on a mesh can take different states corresponding to molecules of right or left handedness, achiral substrate molecules, or even empty states. An example of such a calculation by Shibata et al. [41] is shown in Fig. 5. Again, there are patches of populations of opposite handedness that grow and wipe each other out, such that in the end only one handedness survives.



Exobiology (theoretical), Complexity in, Figure 5

The evolution of molecules of D- and L-forms is shown in the upper and lower panels, respectively. Note the tendency toward complete homochiralization by gradually filling up isolated islands with the chirality of the surrounding molecules. The dimensionless times are 50, 250, and 1750 from left to right. Adapted from [41]

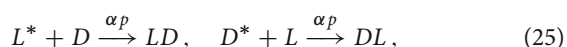
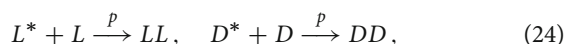
Epimerization

An interesting alternative to the Frank-type mechanism is a set of reactions based primarily on a phenomenon called epimerization, i. e., the spontaneous change of handedness in one part of the polymer. This mechanism is important in the chemistry of amino acids. Plasson et al. [31] identified four reactions: activation, polymerization, epimerization, and depolymerization as necessary ingredients that can, under certain conditions, lead to an instability of the racemic state with a bifurcation toward full homochirality. They called this the APED model, whose reactions can be summarized as follows:

A: activation:



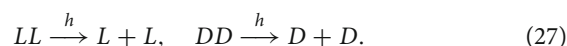
P: polymerization:



E: epimerization:



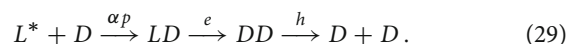
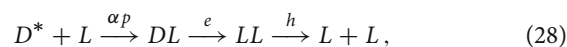
D: depolymerization:



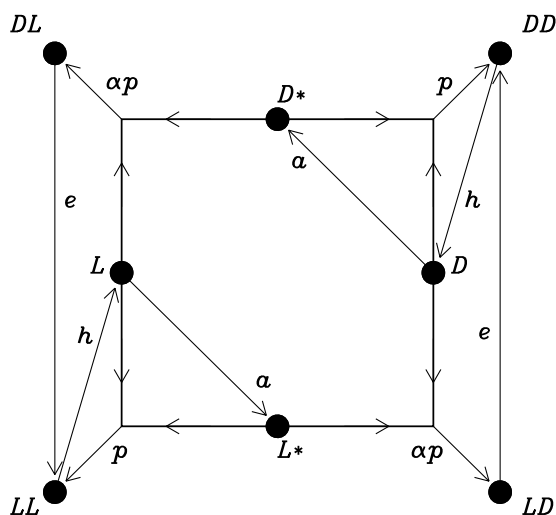
This minimal subset of reactions is shown in Fig. 6.

Compared with the Frank model, a major advantage of the APED model is that no hypothetical auto-catalysis is required. Indeed, all these reactions exist in principle, although it is as yet unclear what kind of manipulation of the environment is required to make all these reactions happen. Another advantage is that the system is closed, so no inflow or outflow of matter is required. The system is maintained away from equilibrium by energy input through the activation of amino acids.

Given that there is neither auto-catalysis nor enantiomeric cross-inhibition, one wonders whether the APED model still shares some similarities with Frank's original model. Some degree of similarity is immediately seen by writing the APED reactions in sequential form in one line, i. e.,



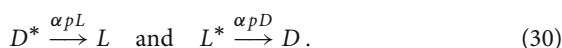
This shows that, as long as the reaction rates for epimerization and depolymerization are not limiting factors, we



Exobiology (theoretical), Complexity in, Figure 6

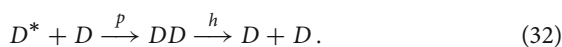
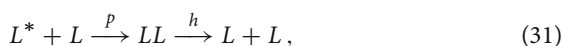
Representation of the minimal set of reactions necessary for allowing a spontaneous transition to homochirality. Adapted from [6]

have essentially the reactions



This way of writing these reactions emphasizes the roles of L and D in catalyzing the conversion of D^* into L and L^* into D , respectively. Just like the mechanism of mutual antagonism, these reactions disfavor a racemic state, but instead of producing unreactive waste, these reactions produce directly one of two possible homochiral states.

In addition, there are reactions of the form



These reactions simulate the autocatalytic conversion of L^* into L by L and of D^* into D by D . Again, linear analysis establishes that the racemic state is unstable provided α is in the range $0 < \alpha < 1$; see [6,31].

In conclusion we can say that the homochirality of life-bearing molecules might well have originated from the chemical reactions that led to their formation. Thus, the hypothetical RNA world may have been born into an environment surrounded by homochiral peptides (as described in Sect. “Epimerization”), or, alternatively, homochirality may have emerged as a consequence of enantiomeric cross-inhibition during the first stages of the RNA world (as discussed in Sect. “Continued Polymerization”).

In the next section we discuss some issues regarding possible strategies for establishing a primitive information-carrying system. This is also based on catalysis, but catalysis in the production of molecules other than itself.

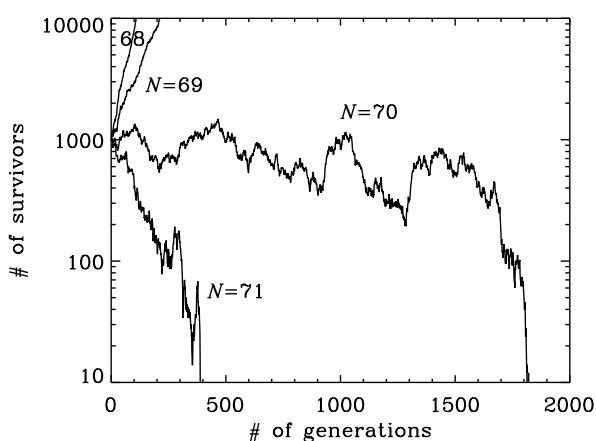
Establishing Hereditary Information

So far, we have ignored the fact that polymers can consist of different amino acid or nucleotide units, even though they would all have the same handedness. Therefore, such molecules could, in principle, carry information. Once such polymers can replicate, the question arises as to how to prevent them from becoming extinct due to errors in the copying process, and, instead, how to allow them to compete against parasites. It is generally believed that early self-replicating systems had a substantial error rate associated with each replication event. A certain error rate is obviously necessary for facilitating Darwinian evolution by natural selection, but it must be small enough to prevent extinction.

Assuming that with each generation a species produces σ offspring where the length of the genome is N bits, and that the probability for a copying error at any position in the genome is p , the necessary condition for long-term survival is given by [15,16]

$$pN < \ln \sigma. \quad (33)$$

The significance of this formula is illustrated in Fig. 7 with the help of a numerical example where the selective advan-



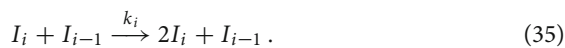
Exobiology (theoretical), Complexity in, Figure 7

Number of survivors as a function of the number of generations in a numerical example with an error rate of $p = 0.01$. This initial number of survivors is 1000. Note that for a genome length of $N = 70$ and 71 , the population dies out after 1800 and 400 generations, respectively. For $N = 69$ and less the number of survivors increases exponentially. This is compatible with the survival criterion $N \lesssim (\ln 2)/0.01 = 69.3$, derived from Eq. (33)

tage (i. e., the multiplication factor) is chosen to be $\sigma = 2$, the error rate is $p = 0.01$, and four different values of N between 68 and 71 are used. In this numerical experiment, σ new offspring are produced, but with a probability p that an error is introduced at each of the N positions. Following an error, only the intact copies can produce further offspring.

According to Eq. (33) the maximum genome length is, with the parameters of our example, $(\ln 2)/0.01 = 69.3$. This is compatible with Fig. 7, which shows that the dividing line between extinction and long-term survival is between $N = 69$ and 70. For contemporary genomes N is of the order of 10^8 , and p is of the order of 10^{-8} [14], or below, depending on the efficiency of error-correcting mechanisms available in contemporary organisms.

The first replicating systems are likely to have rather high error rates and no correction mechanism, making it virtually impossible to carry sufficient information for building more complex replicators. This difficulty can be removed by invoking the concept of hypercycles [17], whereby the full genetic information is carried collectively by several smaller systems (smaller N), each one small enough to obey Eq. (33). Mathematically, such a system can be described by the following set of reactions [2]:



Assuming, furthermore, that resources are limited, the total number of molecules, $M = \sum_i [I_i]$, is taken to be constant, i. e., I_i is assumed to be siphoned off from the system at a rate ϕ that is independent of i . Mathematically, such a system can be described by the following set of ordinary differential equations:

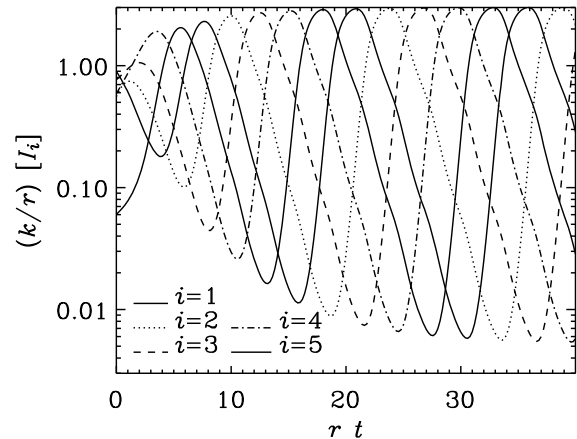
$$\frac{d}{dt}[I_i] = r_i[I_i] + k_i[I_i][I_{i-1}] - \phi[I_i], \quad (36)$$

where

$$\phi = \sum_i (r_i[I_i] + k_i[I_i][I_{i-1}]) / \sum_i [I_i] \quad (37)$$

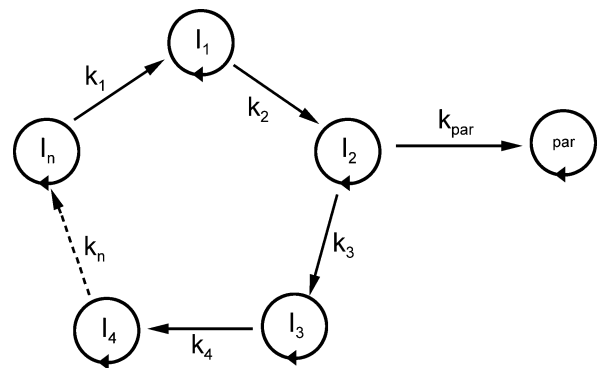
is the factor that keeps the total number of molecules constant. The kinetic coefficient r_i models the residual effects of birth and death, while k_i is the kinetic coefficient for the catalytic production of I_i , where I_{i-1} acts as a catalyst. The evolution of number densities in a model of five hypercycles is shown in Fig. 8 for a case in which all $k_i = k$ and $r_i = r$ are chosen to be the same for all values of i .

An interesting situation arises when the effects of parasites are included. Boerlijst & Hogeweg [2] considered an



Exobiology (theoretical), Complexity in, Figure 8

Evolution of the number densities of five hypercycles with equal parameters. Note that peaks of I_1 (solid line) are followed by peaks of I_2 (dotted line) and I_3 (dashed line), and so forth. Time is measured in units of r^{-1} and concentrations are measured in units of r/k

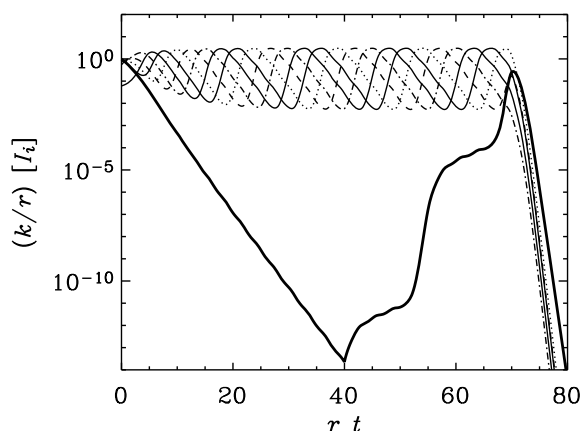


Exobiology (theoretical), Complexity in, Figure 9

Sketch showing the coupling of several hypercycles together with a parasite coupled to species I_2 . Adapted from [2]

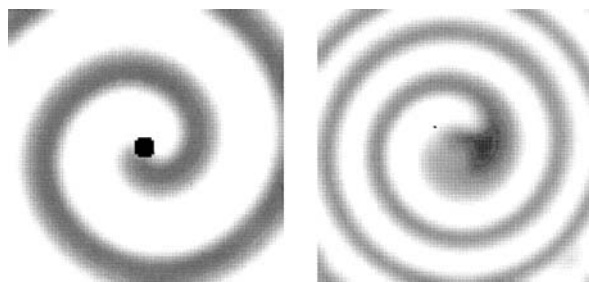
example in which a parasite was coupled to I_2 ; see Fig. 9. The effect on the above model is shown in Fig. 10 where $k_{\text{para}} = 2k$ and $r_{\text{para}} = r$ have been chosen. One sees that not much happens for a long time when the parasite is turned on. This is because the parasite has to grow to a level at which it can affect the entire system. When this point is reached, all components of the system decay exponentially – including the parasite itself. Unfortunately, the system can never recover from this disaster, so the hypercycle theory seems to have a problem.

Again, spatial extent can significantly alter the situation. Using a cellular automata approach, Boerlijst & Hogeweg [2] and Boerlijst [3] showed that the nonlinear spatial dynamics of spiral waves gives the system stabil-



Exobiology (theoretical), Complexity in, Figure 10

Evolution of the number densities of five hypercycles with equal parameters and a parasite where $k_{\text{para}} = 2k_i$ and $r_{\text{para}} = r$. Time is measured in units of r^{-1} and concentrations are measured in units of r/k



Exobiology (theoretical), Complexity in, Figure 11

Spatial patterns in the cellular automata approach of Boerlijst and Hogeweg [2]. Courtesy of Boerlijst MC

ity against otherwise deadly parasitic species. Interestingly enough, this approach tends to produce spiraling interfaces between different species; see Fig. 11

These equations are, by nature, similar to other chemical reaction-diffusion equations where several different substances catalyze each other's reactions. A particularly exciting example is the famous Belousov-Zhabotinsky reaction, where malonic acid, $\text{CH}_2(\text{COOH})_2$ is oxidized in the presence of bromate ions, BrO_3^- . To initiate the reaction, cerium is used as a catalyst to donate ions, although other metal ions may also be used. The color depends on the state of the cerium as it changes from Ce^{3+} to Ce^{4+} or, if iron is used, from Fe^{2+} to Fe^{3+} . The resulting reactions are of the form [28]



where $X = \text{HBrO}_2$, $Y = \text{Br}^-$, $Z = \text{Ce}^{4+}$, $A = B = \text{BrO}_3^-$, P and Q are reaction products that do not contribute further to the reactions, and k_1, \dots, k_5 are known rate constants. The reactions above lead to kinetic equations of the form

$$\frac{\partial[X]}{\partial t} = k_1[A][Y] - k_2[X][Y] + k_3[A][X] - k_4[X]^2, \quad (43)$$

$$\frac{\partial[Y]}{\partial t} = -k_1[A][Y] - k_2[X][Y] + k_5[Z], \quad (44)$$

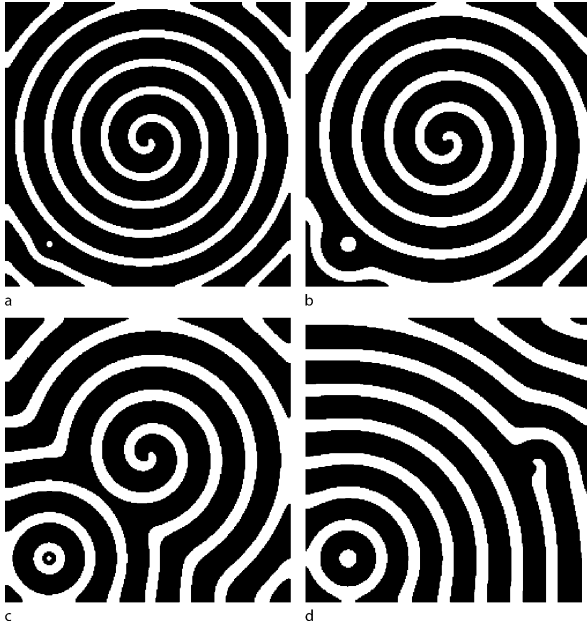
$$\frac{\partial[Z]}{\partial t} = 2k_3[A][X] - k_5[Z]. \quad (45)$$

This model of reaction equations is called the Oregonator, which refers to the affiliation of the authors at the time of publication [18].

If spatial extent is included via diffusion terms, this reaction exhibits, in certain cases, spiral patterns similar to those in the model of Boerlijst & Hogeweg [2]. In Fig. 12 we reproduce the pattern obtained by Zhang et al. [48] for a slightly modified model consisting of only two partial differential equations. Depending on the value of a certain control parameter in their model, spiral patterns of different size are produced. An extensive review of the physics of pattern formation in different settings is given by Cross and Hohenberg [10].

The connection between pattern formation and the origin of life may seem rather remote. However, the equations governing chemical pattern formation illustrate some of the critical steps that are thought to have played a role in the origins of life. In particular, the fact that different chemical compounds catalyze each other in a productive manner is an essential property behind the model proposed by Eigen. The Belousov-Zhabotinsky reaction also illustrates the phenomenon of auto-catalysis where, in the presence of A , the molecule X catalyzes the production of more X by using A as a substrate and producing Z as an additional side product.

The possibility of self-replication was demonstrated for simple RNA molecules by Spiegelman [27] back in the late 1960s. Now, there are examples of simple peptide chains that can catalyze the production of each other [38]. However, a serious shortcoming of any of the above examples is the fact that there is no possibility of natural selec-



Exobiology (theoretical), Complexity in, Figure 12

Spiral and ring-like patterns for the modified reaction equations by Zhang et al. [48]. On the boundaries a no-flux condition has been adopted, i. e., the normal components of all gradients vanish. Adapted from [48]

tion and hence Darwinian evolution. So, as far as the question of the origin of life is concerned, this pathway must be considered a dead end.

In summary, one can say that there are similarities in the mathematics of producing homochirality and in establishing hereditary information in the composition of the first replicating polymers. However, in the latter case, even less is known about the detailed nature of such polymers and their catalytic properties. A particularly important aspect is the possibility of spatial extent, which can substantially modify the behavior of any chemical system. In the present case, as shown in Ref. [2], the possibility of spatial extent is critical for stabilizing the system against destruction by parasites. The model also exhibits spiral pattern formation that has been at the heart of early work by Prigogine and others in connection with early ideas on biogenesis.

Alteration of the Environment by Early Life

In this last section, we discuss some physics problems within astrobiology that illustrate how life, once it has formed, might affect the environment of the early Earth and how it led to a planet so markedly different from a planet that does not harbour life.

Global Energy Balance of the Earth

The young Sun was about 30% fainter than today, and yet the young Earth was covered with liquid water and had temperatures higher than nowadays. This was caused by the presence of greenhouse gases such as water vapor, carbon dioxide, and probably methane. Life is responsible for reducing CO_2 to compounds of the form $(\text{CH}_2\text{O})_n$ and similar, and for oxidizing various minerals to produce O_2 . The resulting decrease of CO_2 weakens the greenhouse effect, so in this sense the emergence of life has an essentially cooling effect on the Earth's overall climate.

Without atmosphere, the planet would cool like a black body at a rate proportional to the local flux $\sigma_{\text{SB}} T^4$, where σ_{SB} is the Stefan–Boltzmann constant. Integrated over the entire surface of the planet, this corresponds to a loss of $4\pi R_{\text{E}}^2 \sigma_{\text{SB}} T^4$, which would need to be balanced against the rate of energy received by solar radiation. The solar “constant” is $S = 1.37 \text{ kW m}^{-2}$ and the total energy projected onto the disk of the Earth is $(1 - A)\pi R_{\text{E}}^2 S$, where A is the albedo, i. e., the fraction of energy reflected from the Earth. The resulting blackbody temperature would be

$$T_0 = \left[(1 - A) \frac{S}{4\sigma_{\text{SB}}} \right]^{1/4}. \quad (46)$$

Using $A = 0.3$ and $\sigma_{\text{SB}} = 5.67 \times 10^{-8} \text{ W m}^{-2} \text{ K}^{-1}$, the temperature of the Earth would be 255 K or about -18°C .

In the presence of an atmosphere, the rate of cooling is modified to $\sigma_{\text{SB}} T_{\text{eff}}^4$, where T_{eff} is the effective temperature equivalent to that of a black body. A positive greenhouse effect corresponds to $T_{\text{eff}} < T$, so the cooling is reduced and the atmosphere heats up according to the vertically integrated energy equation

$$C \frac{dT_0}{dt} = (1 - A) \frac{S}{4} - \sigma_{\text{SB}} T_{\text{eff}}^4, \quad (47)$$

where C is the vertically integrated specific heat.

The value of the effective temperature can be obtained from a radiative transfer calculation. A simplified model calculation¹ yields

$$T_{\text{eff}}^4 = \frac{\ell}{\ell_{\text{crit}}} T_0^4, \quad (48)$$

¹Under the assumption of local isotropy (Eddington approximation) radiative equilibrium implies that the flux is proportional to the negative gradient of the radiative energy density aT^4 , where a is the radiation-density constant, so

$$F = -\frac{1}{3} c \ell \nabla(aT^4).$$

Here, c is the speed of light and ℓ is the mean free path of photons. The latter can be expressed in terms of the opacity κ and the density ρ

where T_0 is the surface temperature, ℓ is an averaged mean free path of photons and ℓ_{crit} is the critical value above which there is a positive greenhouse effect. Again, a simplified calculation suggests $\ell_{\text{crit}} = 3H/16 \approx 0.19H$, where $H = \mathcal{R}T/(\mu g) \approx 8 \text{ km}$ is the pressure scale height of the atmosphere. So, an increase in opacity leads to a decrease of the cooling and hence to an increase in the surface temperature.

Another interpretation is that the greenhouse gases shift the radiating surface by a certain amount, ℓ_g , upward. The value of ℓ_g is related to ℓ . Ditlevsen [13] uses $\ell_g = 3 \text{ km}$. He also noted that a more accurate lapse rate of the temperature is $dT/dz = 10 \text{ K km}^{-1}$ instead of $T_0/H = 40 \text{ K km}^{-1}$, so that the temperature gain caused by greenhouse gases is $\ell_g \times dT/dz = 30 \text{ K}$. The reason for a shallower temperature gradient is the presence of convection that causes the specific entropy s to be nearly constant with height. In that case the temperature gradient is just the adiabatic one, $(dT/dz)_{\text{ad}} = g/c_p$, where c_p is the specific heat at fixed pressure, which in turn is related to the universal gas constant and the specific weight via $\mathcal{R}/\mu = c_p - c_v$, where c_v is the specific heat at fixed volume and $c_p/c_v = \gamma$ is the ratio of specific heats. With these formulae one does indeed get ²

$$\left(\frac{dT}{dz}\right)_{\text{ad}} = \left(1 - \frac{1}{\gamma}\right) \frac{\mu g}{\mathcal{R}} \approx 10 \text{ K km}^{-1}, \quad (49)$$

where we have used $\gamma = 7/5$ for air molecules with 5 degrees of freedom (3 for translation and 2 for rotation).

At certain times over the history of the Earth, other greenhouse gases such as methane may have played an important role in keeping the Earth above freezing temperatures. Indeed, the burial of oxides in the crust allowed methane to build up in the atmosphere, which may have led to concentrations of a few thousand times greater than modern levels. UV radiation in the upper atmosphere

breaks up methane into its components, letting H_2 escape into space, leading to a net gain of oxygen, that comes ultimately from H_2O .

According to a model of Catling et al. [8] methane (CH_4) may have been important 2.7...2.3 Gyr ago, just before the famous Snowball Earth deep freeze of the planet [22]. As discussed above, the associated loss of hydrogen may have led to a gradual accumulation of oxygen in the atmosphere, which then terminated the methane era and led to the Snowball Earth event. This event lasted until the continuous CO_2 production from volcanoes accumulated to large amounts so that the resulting greenhouse effect became sufficient to initiate partial melting of the ice cover.

Response to Changes in Greenhouse Gases

As we knew first from global climate models [20] and later from simplified models [9] using Eq. (47), with a relatively simple piecewise linear temperature dependence of $A(T)$, there can be three different equilibrium temperatures. This is illustrated in Fig. 13, where we compare the graph of $\sigma_{\text{SB}} T_{\text{eff}}^4$ versus surface temperature T_0 with the net radiation $(1 - A)S/4$. Here, $A(T)$ has been arranged such that $A = A_{\text{hot}}$ for $T \geq T_{\text{max}}$ (corresponding to no ice coverage) $A = A_{\text{cold}}$ for $T \leq T_{\text{min}}$ (corresponding to full ice coverage).

Ditlevsen [13] used Eq. (47) to study the response of the system to variable greenhouse gas concentrations. As the amount of CO_2 increases, the equilibrium temperature increases. Obviously, when the system is on the lower fixed point initially, there must be a critical CO_2 concentration above which the solution will jump discontinuously to the upper branch; see Fig. 14.

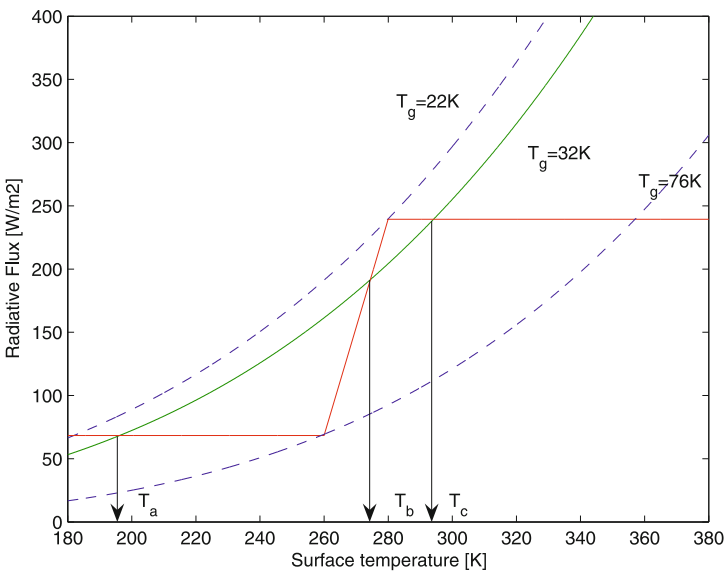
It is generally accepted that the rate of weathering increases with increasing temperature. This provides a stabilizing effect on the climate. As T increases, the rate of weathering increases, removing more CO_2 from the atmosphere, reducing the greenhouse effect, and thus leading to cooling. Ditlevsen [13] introduced the assumption that there is a continuous source of CO_2 through outgassing from volcanoes and a temperature-dependent sink of CO_2 from weathering when T exceeds a critical temperature T_w , but no weathering for $T < T_w$ [46]. This leads to a self-regulating effect for $T < T_w$, which Ditlevsen calls a greenhouse thermostat. Whenever $T < T_w$, since there is then no weathering and hence no sink of CO_2 , greenhouse gases will build up until the Earth's temperature has reached the value T_w ; see Fig. 15. This is the mechanism that is believed to have caused the early Earth to be above freezing through most of its history – with the

as $\ell = (\kappa\rho)^{-1}$. Hydrostatic equilibrium can be written in the form

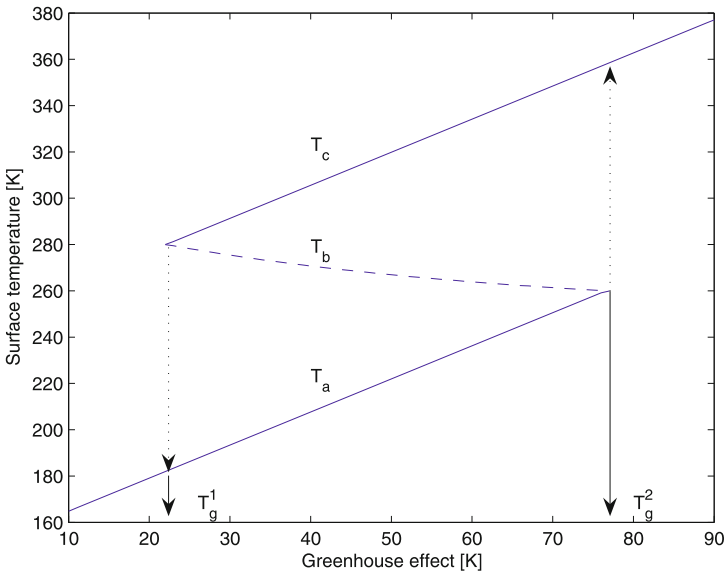
$$g = -\frac{\mathcal{R}}{\mu} \nabla T,$$

where \mathbf{g} is the gravitational acceleration, \mathcal{R} is the universal gas constant, and μ is the mean molecular weight. These equations can be solved by a polytrope, i. e., $T = T_0(1 - z/H)$ and $\rho = \rho_0(1 - z/H)^3$, where z is the distance from the surface and H is the vertical pressure scale height. This leads to a condition of the form Eq. (48) where $\ell_{\text{crit}} = 3H/16 \approx 0.19H$ is the critical mean free path of photons

²Hydrostatic equilibrium can be written as $-\rho^{-1}\nabla p - \nabla\phi = 0$, where p is the pressure and $\phi = gz + \text{const}$ is the gravitational potential. Using the thermodynamic relation $-\rho^{-1}\nabla p = -\nabla h + T\nabla s = 0$, where $h = c_p T$ is the specific enthalpy and $\nabla s = 0$ for adiabatic stratification, we have $d(c_p T)/dz = g$



Exobiology (theoretical), Complexity in, Figure 13
Plot of $\sigma_{\text{SB}} T_{\text{eff}}^4$ versus surface temperature T_0 for three different greenhouse temperature shifts, T_g , compared with the net radiation $\frac{1}{4}(1 - A)$ for a simple piecewise linear function $A(T)$. Courtesy of P. Ditlevsen [13]

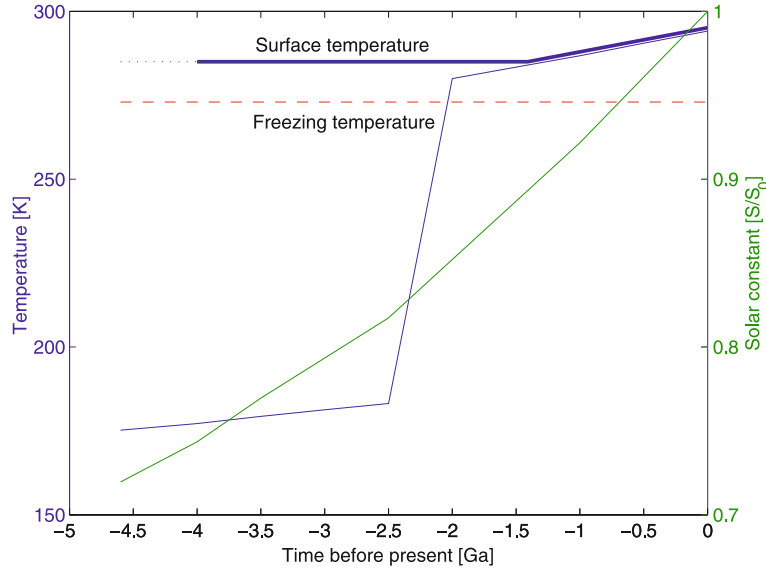


Exobiology (theoretical), Complexity in, Figure 14
Equilibrium temperature as a function of CO_2 concentration. Courtesy of P. Ditlevsen [13]

exception of intermediate Snowball Earth-like events that are caused by the emergence of other sinks of greenhouse gases, such as the onset of aerobic photosynthesis or the enhanced formation of mountain topography that leads to an increase in the erosion rate and, hence, weathering.

The Daisyworld Model

Life can also affect the planet's albedo, as has been demonstrated by Lovelock [47] in his Daisyworld model. For a tutorial on the Daisyworld model see [45]. This model also makes use of Eq. (47), but now the planet's albedo A is af-



Exobiology (theoretical), Complexity in, Figure 15

Dependence of surface temperature on time under the assumption of a continuous source of outgassing of CO₂ and the onset of a CO₂ sink for $T > T_w$. Courtesy of P. Ditlevsen [13]

ected by the plant population which is simplistically represented by black and white plants or flowers (daisies) with local albedos A_1 and A_2 , respectively. So the total albedo is a weighted average of the form

$$A = \sum_{i=1}^3 \alpha_i A_i, \quad (50)$$

where A_3 is the albedo of the unpopulated surface. The weights α_i depend on the surface coverage of the respective regions and obey evolution equations that are, in turn, governed by a temperature-dependent growth term, $\beta(T_i)$, and a fixed death rate, γ . The resulting equations for the rate of change of the albedo are

$$\frac{d\alpha_i}{dt} = [\beta(T_i) - \gamma]\alpha_i, \quad (51)$$

where $i = 1$ for black and $i = 2$ for white plant populations, $\beta(T)$ is assumed to be different from zero in the range $T_{\min} < T < T_{\max}$ with a maximum at $T_{\text{aver}} = \frac{1}{2}(T_{\min} + T_{\max})$. The weight for the unpopulated surface follows from the normalization $\sum \alpha_i = 1$, so $\alpha_3 = 1 - \alpha_1 - \alpha_2$.

The temperatures are higher in the regions of black plants and lower in regions of white plants according to the formula

$$T_i^4 = (A - A_i)q + T_0^4, \quad (52)$$

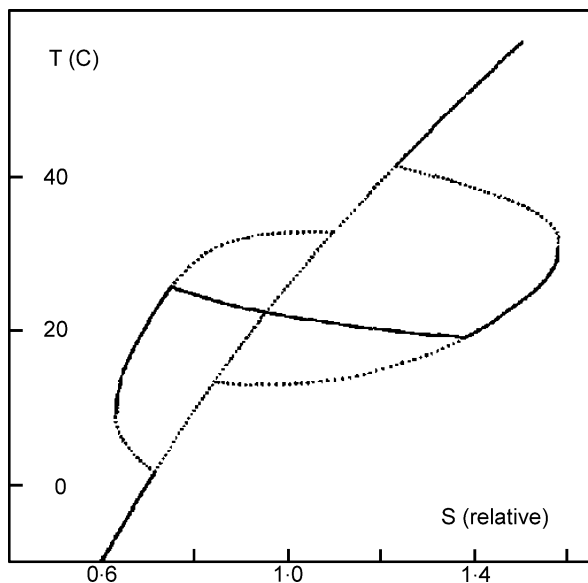
where q is a parameter that must be smaller than a critical value,

$$q < q_{\text{crit}} = S/(4\sigma_{\text{SB}}), \quad (53)$$

in order that heat flows against the temperature gradient [40]. For $q = q_{\text{crit}}$ the temperature is uniform for different values of A_i , while for $q < q_{\text{crit}}$ the regions of high albedo are cooler and those of low albedo warmer. Note also that Eq. (52) preserves heat balance, i. e., $\sum \alpha_i T_i^4 = T_0^4$.

The important point in the Daisyworld model is the fact that, for a certain range of S , the surface temperature of the planet, T_0 , is stabilized in a certain temperature range around the optimal value close to T_{aver} ; see Fig. 16.

Saunders makes another remarkable point. He showed that by changing the model to allow for Darwinian evolution such that each plant species works with an optimized temperature dependence, so $\beta(T_i) \rightarrow \beta_i(T_i)$ is modified to become dependent on i , the overall result changes only very little. More importantly, the range over which the model can stabilize the planet's temperature shrinks, making the planet as a whole more vulnerable. Although the amount of shrinkage is small, it emphasizes the dangers associated with adopting changes that lead only to short-term benefits. Saunders emphasizes in his work that the ability of life to regulate the surface temperature of a planet is not associated with natural selection as in the concept of Darwinian evolution. More generally he warns, therefore,



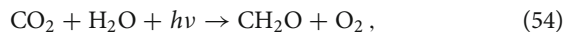
Exobiology (theoretical), Complexity in, Figure 16

Temperature (in Celsius) versus relative irradiation (normalized to the average temperature). Note that the temperature is stabilized around the value T_{aver} , provided the energy input S is within a certain range. Adapted from [40]

that not everything that is to an advantage needs to be the result of natural selection [40].

Oxidation of the Earth's Crust

It has recently been proposed that, in addition to the effects discussed above, life may also have profound effects on the Earth's crust. A possible scenario has recently been discussed by Rosing et al. [36]. The idea is that photosynthetic life may tap large amounts of solar energy that were used to reduce carbon from CO_2 to compounds of the form $(\text{CH}_2\text{O})_n$ and similar, via reactions of the form



where $h\nu$ denotes the energy taken from solar radiation. Furthermore, and even more surprisingly, the oxygen produced by photosynthesis may have been critical in oxidizing iron in the continental crust. Although other factors also played a substantial role, it is clear that biological processes can speed up the oxidation process substantially. Comparing the oceanic crust with the continental crust, a major difference is the enhanced fraction of SiO_2 (57% in the continental crust compared to 50% in the oceanic crust).

Granite, being one of the lightest rock types, was eventually able to escape subduction and to produce stable continents about 3.8 Gyr ago. This is also the time of the oldest

rock findings on Earth. Given that the rise of continents on the early Earth is associated with granite formation, the presence of granite on silicon-bearing rocky planets might thus be a possible biomarker for photosynthesis [36].

Although this idea is speculative, it may be supported quantitatively as follows. Firstly, the present day production rate of organically fixated carbon is estimated to be $9 \times 10^{15} \text{ mol C yr}^{-1}$ [12,25]. The amount of energy required for this can be calculated by using the fact that it costs 477 kJ to transfer one mole of carbon to hexose. The energy required for this is then 300 mW m^{-2} . Rosing et al. [36] argue that this amount could be supplied by only 0.1% of the effective solar energy flux, $S/4$. Assuming that the amount of carbon burial, relevant to estimating the usable fraction of oxygen for iron oxidation, is also about 0.1%, this corresponds to about $10^{13} \text{ mol C yr}^{-1}$. This would yield a comparable iron oxidation rate. Rosing et al. [36] argue further that annual basalt production contributes about $10^{14} \text{ mol Fe yr}^{-1}$, so a fraction of the magmatic iron flux could be used for building up the mantle reservoir of ferric iron.

In conclusion, the presence of life can lead to significant alterations of the planet in a number of different ways, as is quite clearly demonstrated by some of the differences between Earth and its neighboring planets Venus and Mars. Only the Earth has extensive reservoirs of oxygen and of granite. Within limits, the presence of life on a planet can also have a stabilizing effect on its climate. The relevant mathematical modeling of some of these processes resembles, in many ways, those encountered earlier in studies of homochirality and of the spread of hereditary information on the early Earth.

Conclusions

Astrobiology has developed into a rapidly growing research field involving expertise from a number of neighboring disciplines. Nonlinear dynamics and nonequilibrium thermodynamics find applications in all these subfields. Here, we have elaborated on a few such aspects. Closest to the onset of life is, perhaps, the emergence of homochirality of biomolecules. Given that RNA has been proven to form longer polymers only in a homochiral environment, one would expect that homochirality must be a prerequisite to the emergence of life at the level of a replicating RNA world. On the other hand, the very mechanism causing the polymerization to terminate, namely enantiomeric cross-inhibition, can also be the mechanism responsible for causing 100% homochirality by destroying RNA molecules whose chirality is already in the minority. This would, however, require the possibility of auto-

catalysis, which can be avoided in another scenario where a closed peptide system is kept away from equilibrium by continuous activation of amino acids.

Chemically speaking, the stabilization of a definite chirality is in some models similar to the subsequent establishment of hereditary information, in that catalysis plays a crucial role. Furthermore, in both cases, the possibility of chemistry in an extended system is crucial. On the one hand, spatial extent gives rise to the possibility of coexistence of life forms of opposite handedness on the early Earth. On the other hand, spatial extent can be critical in allowing the system to find unpopulated locations fast enough to avoid being overwhelmed by the effects of parasites that tap the same resources that are required for the maintenance and development of hereditary information.

Finally, life is invariably coupled to some kind of metabolism that is ultimately powered by solar energy. This clearly affects the environment by reducing carbon and oxidizing the crust of the Earth and, over the last two billion years, the atmosphere. How much these alterations of the environment are due to biological processes is less obvious. However, it is clear that biological factors greatly speed up weathering on the Earth. The extent of biologically induced alterations of the continental crust, for example, may therefore best be tested using quantitatively accurate model calculations. The outcome may ultimately hinge on energetic considerations and on the efficiency of photosynthesis as a solar energy collector.

With the scope of being able to explore, in the near future, not only the planets and other celestial bodies in the solar system in much more detail, but also planets of other planetary systems, research in astrobiology quickly develops into a field that will be driven more and more by new data, making this field less susceptible to speculation. Therefore, it is important to be prepared for upcoming discoveries in this field. Finally, it should be emphasized that astrobiology is efficient in communicating science to the general public, which may provide an additional boost to the field.

Bibliography

Primary Literature

- Blackmond DG (2004) Asymmetric auto-catalysis and its implications for the origin of homochirality. *Proc Natl Acad Sci* 101:5732–5736
- Boerlijst MC, Hogeweg P (1991) Spiral wave structure in prebiotic evolution – hypercycles stable against parasites. *Phys D* 48:17–28
- Boerlijst MC, (2000) Spiral ans spots: Novel evolution phenomena through spatial self-structing. In: Dieckmann U, Law R, Metz H (eds) *The geometry of ecological interactions: Simulating spatial complexity*. Cambridge University Press, Cambridge, pp 171–182
- Brandenburg A, Multamäki T (2004) How long can left and right handed life forms coexist? *Int J Astrobiol* 3:209–219
- Brandenburg A, Andersen AC, Höfner S, Nilsson M (2005) Homochiral growth through enantiomeric cross-inhibition. *Orig Life Biosph* 35:225–241
- Brandenburg A, Lehto HJ, Lehto KM (2007) Homochirality in an early peptide world. *Astrobiol* 7:725–732
- Bywater RP, Conde-Frieboesk K (2005) Did Life Begin on the Beach? *Astrobiol* 5:568–574
- Catling DD, Zahnle KJ, McKay CP (2001) Biogenic methane, hydrogen escape, and the irreversible oxidation of early Earth. *Science* 293:839–843
- Crafoord C, Källén E (1978) A note on the condition for existence of more than one steady-state solution in Budyko-Sellers type models. *J Atmos Sci* 35:1123–1125
- Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. *Rev Mod Phys* 65:851–1112
- Davies PCW, Lineweaver CH (2005) Finding a second sample of life on Earth. *Astrobiol* 5:154–163
- Des Marais DJ (2000) When did photosynthesis emerge on Earth. *Science* 289:1703–1705
- Ditlevsen PD (2005) A climatic thermostat making Earth habitable. *Int J Astrobiol* 4:3–7
- Dyson FJ (1999) *Origins of life*. Cambridge University Press, Cambridge
- Eigen M (1971) Selforganization of matter and evolution of biological macromolecules. *Naturwissenschaften* 58:465–523
- Eigen M (2002) Error catastrophe and antiviral strategy. *Proc Natl Acad Sci* 99:13374–13376
- Eigen M, Schuster P (1977) The hypercycle. *Naturwissenschaften* 64:541–565
- Field RJ, Körö E, Noyes RM (1972) Oscillations in chemical systems, part 2. Thorough analysis of temporal oscillations in the bromate-cerium-malonic acid system. *J Am Chem Soc* 94:8649–8664
- Frank FC (1953) On spontaneous asymmetric synthesis. *Biochim Biophys Acta* 11:459–464
- Ghil M (1976) Climate stability for a Sellers-type model. *J Theor Biol* 33:3–20
- Gilbert W (1986) Origin of life – the RNA world. *Nature* 319:618–618
- Hoffman PF, Kaufman AJ, Halverson GP, Schrag DP (1998) A Neoproterozoic Snowball Earth. *Science* 281:1342–1346
- Joyce GF, Visser GM, van Boeckel CAA, van Boom JH, Orgel LE, Westrenen J (1984) Chiral selection in poly(C)-directed synthesis of oligo(G). *Nature* 310:602–603
- Lathe R (2004) Fast tidal cycling and the origin of life. *Icarus* 168:18–22
- Martin JH, Knauer GA, Karl DM, Broenkow WW (1987) Vertex - carbon cycling in the Northeast Pacific. *Deep-Sea Res Part A* 34:267–285
- Miller SL (1953) A production of amino acids under possible primitive Earth conditions. *Science* 117:528–529
- Mills DR, Peterson RL, Spiegelman S (1967) An extracellular Darwinian experiment with a self-duplicating nucleic acid molecule. *Proc Natl Acad Sci* 58:217–224
- Murray JD (1974) On a model for the temporal oscillations in the Belousov-Zhabotinsky reaction. *J Chem Phys* 61:3610–3613

29. Murray JD (2002) *Mathematical Biology, An introduction*. Springer, New York
30. Nielsen PE (1993) Peptide nucleic acid (PNA): A model structure for the primordial genetic material. *Orig Life Evol Biosph* 23:323–327
31. Plasson R, Bersini H, Commeyras A (2004) Recycling Frank: spontaneous emergence of homochirality in noncatalytic systems. *Phys Rev* 101:16733–16738
32. Plasson R, Kondepudi DK, Bersini H, Commeyras A, Asakura K (2007) Emergence of homochirality in far-from-equilibrium systems: Mechanisms and role in prebiotic chemistry. *Chirality* 19:589–600
33. Prigogine I, Lefever R (1968) Symmetry breaking instabilities in dissipative systems II. *J Chem Phys* 48:1695–1700
34. Prigogine I, Nicolis G (1967) On symmetry-breaking instabilities in dissipative systems. *Chem J Phys* 46:3542–3550
35. Rasmussen S, Chen L, Nilsson M, Abe S (2003) Bridging nonliving and living matter. *Artif Life* 9:269–316
36. Rosing MT, Bird DK, Sleep NH, Glassley W, Albarede F (2006) The rise of continents—An essay on the geologic consequences of photosynthesis. *Palaeogeogr Palaeoclimatol Palaeoecol* 232:99–113
37. Russell M (2006) First life. *Am Sci* 94:32–39
38. Saghatelian A, Yokobayashi Y, Soltani K, Ghadiri MR (2001) A chiroselective peptide replicator. *Nature* 409:797–801
39. Sandars PGH (2003) A toy model for the generation of homochirality during polymerization. *Orig Life Biosph* 33:575–587
40. Saunders PT (1994) Evolution without natural selection: further implications of the Daisyworld parable. *J Theor Biol* 166:365–373
41. Shibata R, Saito Y, Hyuga H (2006) Diffusion accelerates and enhances chirality selection. *Phys Rev* 74:026117
42. Soai K, Shibata T, Morioka H, Choji K (1995) Asymmetric autocatalysis and amplification of enantiomeric excess of a chiral molecule. *Nature* 378:767–768
43. Taylor FW (1991) The greenhouse effect and climate change. *Rep Prog Phys* 54:881–918
44. Turing AM (1952) The chemical basis of morphogenesis. *Phil Trans Roy Soc B* 237:37–72
45. von Bloh W, Block A, Parade M, Schellnhuber HJ (1999) Tutorial Modelling of geosphere-biosphere interactions: the effect of percolation-type habitat fragmentation. *Rep Prog Phys* A 266:186–106
46. Walker JCG, Hays PB, Kasting JF (1981) A negative feedback mechanism for the long-term stabilization of the Earth's surface temperature. *J Geophys Res* 86:9776–9782
47. Watson AJ, Lovelock JE (1983) Biological homeostasis of the global environment: the parable of Daisyworld. *Tellus B* 35:284–289
48. Zhang XZ, Liao HM, Zhou LQ, Quyang Q (2004) Pattern selection in the Belousov–Zhabotinsky reaction with the addition of an activating reactant. *J Phys Chem B* 108:16990–16994
- Darwin C (1859) *The origin of species by means of natural selection*. Reprinted by: Penguin books, London, 1985
- Haken H (1983) *Synergetics – An Introduction*. Springer, Berlin
- Lovelock JE (1995) *Gaia. A new look at life on Earth*. Oxford University Press, Oxford
- Lunine J (2003) *Astrobiology: a multi-disciplinary approach*. Pearson Higher Education, Addison-Wesley, San Francisco
- Prigogine I (1980) *From being to becoming. Time and complexity in the physical sciences*. Freeman, New York
- Rauchfuß H (2005) *Chemische Evolution und der Ursprung des Lebens*. Springer, Berlin
- Ward PD, Brownlee D (2000) *Rare Earth: why complex life is uncommon in the Universe?* Copernicus, New York

Extreme Events in Socio-economic and Political Complex Systems, Predictability of

VLADIMIR KEILIS-BOROK^{1,2}, ALEXANDRE SOLOVIEV^{2,3}, ALLAN LICHTMAN⁴

¹ Institute of Geophysics and Planetary Physics and Department of Earth and Space Sciences, University of California, Los Angeles, USA

² International Institute of Earthquake Prediction Theory and Mathematical Geophysics, Russian Academy of Science, Moscow, Russia

³ Abdus Salam International Centre for Theoretical Physics, Trieste, Italy

⁴ American University, Washington D.C., USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Common Elements of Data Analyzes](#)

[Elections](#)

[US Economic Recessions](#)

[Unemployment](#)

[Homicide Surges](#)

[Summary: Findings and Emerging Possibilities](#)

[Bibliography](#)

Glossary

Complexity A definitive feature of nonlinear systems of interacting elements. It comprises high instability with respect to initial and boundary conditions, and complex but non-random behavior patterns (“order in chaos”).

Extreme events Rare events having a large impact. Such events are also known as critical phenomena, disas-

Books and Reviews

Barbieri M (2003) *The organic codes: an introduction to semantic biology*. CUP, Cambridge

Brack A (1998) *The molecular origins of life*. Cambridge University Press, Cambridge

ters, catastrophes, and crises. They persistently reoccur in hierarchical complex systems created, separately or jointly, by nature and society.

Fast acceleration of unemployment (FAU) The start of a strong and lasting increase of the unemployment rate.

Pattern recognition of rare events The methodology of artificial intelligence' kind aimed at studying distinctive features of complex phenomena, in particular – at formulating and testing hypotheses on these features.

Premonitory patterns Patterns of a complex system's behavior that emerge most frequently as an extreme event approaches.

Recession The American National Bureau of Economic Research defines recession as “a significant decline in economic activity spread across the economy, lasting more than a few months”. A recession may involve simultaneous decline in coincident measures of overall economic activity such as industrial production, employment, investment, and corporate profits.

Start of the homicide surge (SHS) The start of a strong and lasting increase in the smoothed homicide rate.

Definition of the Subject

At stake in the development of accurate and reliable methods of prediction for social systems is the capacity of scientific reason to improve the human condition. Today's civilization is highly vulnerable to crises arising from extreme events generated by complex and poorly understood systems. Examples include external and civil wars, terrorist attacks, crime waves, economic downturns, and famines, to name just a few. Yet more subtle effects threaten modern society, such as the inability of democratic systems to produce policies responsive to challenges like climate change, global poverty, and resource depletion.

Our capacity to predict the course of events in complex social systems is inherently limited. However, there is a new and promising approach to predicting and understanding complex systems that has emerged through the integration of studies in the social sciences and the mathematics of prediction. This entry describes and analyzes that approach and its real-world applications. These include algorithmic prediction of electoral fortunes of incumbent parties, economic recessions, surges of unemployment, and outbursts of crimes. This leads to important inferences for averting and responding to impending crises and for improving the functioning of modern democratic societies.

That approach was successfully applied also to natural disasters such as earthquakes. Ultimately, improved pre-

diction methods enhance our capacity for understanding the world and for protecting and sustaining our civilization.

Extreme events. Hierarchical complex systems persistently generate extreme events – the rare fast changes that have a strong impact on the system. Depending on connotation they are also known as critical phenomena, disasters, catastrophes, and crises. This article examines the development and application of the algorithmic prediction of extreme socio-economic and political events.

The prediction problem is formulated as follows:

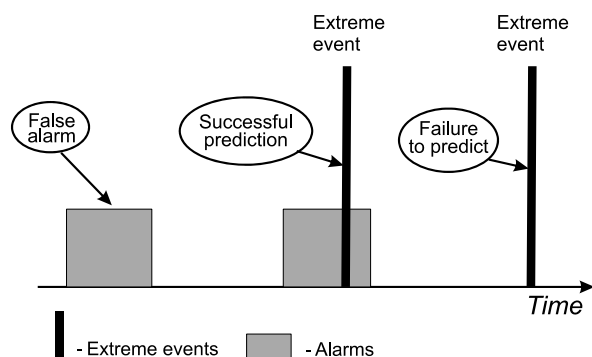
given are time series that describe dynamics of the system up to the current moment of time t and contain potential precursors of an extreme event;

to predict whether an extreme event will or will not occur during the subsequent time period $(t, t + \tau)$; if the answer is “yes”, this will be the “*period of alarm*”.

As the time goes by, predictions form a discrete sequence of alarms. The possible outcomes of such a prediction are shown in Fig. 1. The actual outcome is determined unambiguously, since the extreme events are identified independently of the prediction either by the actual happening (e.g. by an election result) or by a separate algorithm (e.g. homicide surge) after they occur.

Such “yes or no” prediction is aimed not at analyzing the whole dynamics of the system, but only at identifying the occurrence of rare extreme events. In a broad field of prediction studies this prediction is different from and complementary to the classical Kolmogoroff–Wiener prediction of continuous functions, and to traditional cause-and-effect analysis.

The problem includes estimating the predictions' accuracy: the rates of false alarms and failures to predict, and the total duration of alarms in relation to the total time considered. These characteristics represent the inevitable



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 1
Possible outcomes of prediction

probabilistic component of prediction; they provide for statistical validation of a prediction algorithm and for optimizing preparedness to predicted events (e. g. recessions or crime surges).

Twofold importance. The prediction problem is pivotal in two areas:

- *Fundamental understanding of complex systems.* Prediction algorithms quantitatively define phenomena that anticipate extreme events. Such quantitative definition is pivotal for fundamental understanding of a complex system where these events occur, including the intertwined mechanisms of system's development and its basic features, e. g. multiple scaling, correlation range, clustering, fragmentation etc. (see Sects. "Common Elements of Data Analyzes", "Elections", "US Economic Recessions", "Unemployment"). The understanding of complex systems remains a major unsolved problem of modern science, tantamount to transforming our understanding of the natural and human world.
- *Disaster preparedness.* On the practical side prediction is pivotal for coping with a variety of disasters, commonly recognized as major threats to the survival and sustainability of our civilization (e. g. [22]; see also materials of G8-UNESCO World Forum on "Education, Innovation and Research: New Partnership for Sustainable Development", <http://g8forum.ictp.it>). The reliable advance prediction of extreme events can save lives, contribute to social and economic stability, and to improving the governing of modern societies.

Introduction

Predictability vs. Complexity: The Need for Holistic Approach [7,12,13,15,17,27,32]

Natural science had for many centuries regarded the Universe as a completely predictable machine. As Pierre Simon de Laplace wrote in 1776, "... if we knew exactly the laws of nature and the situation of the universe at the initial moment, we could predict exactly the situation of the same universe at a succeeding moment." However, at the turn of the 20th century (1905) Jules Henry Poincare discovered, that "... this is not always so. It may happen that small differences in the initial conditions will produce very great ones in the final phenomena. Prediction becomes impossible".

This instability to initial conditions is indeed a definitive attribute of complex systems. Nonetheless, through the robust integral description of such systems, it is possible to discover regular behavior patterns that transcend the

inherent complexity. For that reason studying complexity requires the holistic approach that proceeds from the whole to details, as opposed to the reductionism approach that proceeds from details to the whole. It is in principle not possible "to understand a complex system by breaking it apart" [13].

Among the regular behavior patterns of complex systems are "premonitory" ones that emerge more frequently as an extreme event approaches. These premonitory patterns make complex systems predictable. The accuracy of predictions, however, is inevitably limited due to the systems' complexity and observational errors.

Premonitory patterns and extreme events are consecutive manifestations of a system's dynamics. These patterns may not trigger extreme events but merely signal the growth of instability, making the system ripe for the emergence of extreme events.

Methodology

The prediction algorithms described here are based on discovering premonitory patterns. The development of the algorithms requires the integration of complementary methods:

- Theoretical and numerical modeling of complex systems; this includes "universal" models considered in statistical physics and non-linear dynamics (e. g. [1,3,5,8,12,15,20,42]), and system-specific models, if available.
- Exploratory data analysis.
- Statistical analysis of limited samples, which is relevant since the prediction targets are by definition rare.
- Practical expertise, even if it is intuitive.
- Risk analysis and theory of optimal control for optimizing prediction strategy along with disaster preparedness.

Pattern Recognition of Rare Events This methodology provides an efficient framework for integrating diverse information into prediction algorithms [4,11,19]. This methodology has been developed by the artificial intelligence school of I. Gelfand for the study of rare phenomena of a highly complex origin. In terminology of pattern recognition, the "object of recognition" is the time moment t . The problem is to recognize whether it belongs to the period of alarm, i. e. to a time interval Δ preceding an extreme event. An alarm starts when certain combinations of premonitory patterns emerges.

Several features of that methodology are important for predicting extreme events in the absence of a complete

closed theory that would unambiguously define a prediction algorithm. First, this kind of pattern recognition relies on simple, robust parameters that overcome the bane of complexity analysis – incomplete knowledge of the system’s causal mechanisms and chronic imperfections in the available data. In its efficient robustness, pattern recognition of rare events is akin to exploratory data analysis as developed by J. Tukey [50]. Second, unlike other statistical methods, e.g. regression analysis, that methodology can be used for small samples such as presidential elections or economic recessions. Also, it integrates quantitative and judgmental parameters and thereby more fully captures the full dimensions of the prediction problem than procedures that rely strictly on quantitative variables.

Summing up, the methodology described here can help in prediction when there are (1) many causal variables, (2) qualitative knowledge about which variables are important, and (3) limited amounts of data [2].

Besides societal predictions, pattern recognition of rare events has been successfully applied in seismology and earthquake prediction (e.g. [11,19,20,44,46]), geological prospecting (e.g. [45]) and in many other fields. Review can be found in [21,47]. Tutorial materials are available at the web site of the Abdus Salam International Centre for Theoretical Physics (http://cdsagenda5.ictp.it/full_display.php?da=a06219).

Validation of Prediction Algorithms The algorithms include many adjustable elements, from selecting the data and defining the prediction targets, to specifying numerical parameters involved. In lieu of theory that would unambiguously determine these elements they have to be developed retrospectively, by “predicting” past extreme events. The application of the methodology to known events creates the danger of self-deceptive data-fitting: As J. von Neumann put it “*with four exponents I can fit an elephant*”. The proper validation of the prediction algorithms requires three consecutive tests.

- *Sensitivity analysis*: testing whether predictions are sensitive to variations of adjustable elements.
- *Out of sample analysis*: application of an algorithm to past data that has not been used in the algorithm’s development. The test is considered successful if algorithm retains its accuracy.
- *Predicting future events* – the only decisive test of a prediction algorithm (see for example Sect. “Elections” below).

A highly efficient tool for such tests is the error Diagram, showing major characteristics of prediction accuracy [33,

34,35,36,37,38,39]. Its example is given in Fig. 10. Exhaustive sets of these tests are described in [10,11,24,52].

Common Elements of Data Analyses

The methodology discussed above was used for predicting various kinds of extreme events, as illustrated in the next four Sections. Naturally, from case to case this methodology was used in different ways, according to specifics of phenomena considered. However in all cases data analysis has essential common elements described below.

Sequence of analysis comprises four stages: (i) Defining prediction targets. (ii) Choosing the data (time series), where premonitory patterns will be looked for and summing up a priori constrains on these patterns. (iii) Formulating hypothetical definition of these patterns and developing prediction algorithm; determining the error diagram. (iv) Validating and optimizing that algorithm.

Preliminary transformation of raw data. In predicting recessions (Sect. “US Economic Recessions”), fast acceleration of unemployment (Sect. “Unemployment”) and crime surges (Sect. “Homicide Surges”) raw data were time series of relevant monthly indicators, hypothetically containing premonitory patterns. Let $f(m)$ be such an indicator, with integer m showing time in months. Premonitory behavior of some indicators is better captured by their linear trends.

Let $W^f(l/q, p)$ be the local linear least-squares regression of a function $f(m)$ within the sliding time window (q, p) :

$$W^f(l/q, p) = K^f(q, p)l + B^f(q, p), \quad q \leq l \leq p, \quad (1)$$

where integers l, q , and p stand for time in months.

Premonitory behavior of most indicators was captured by the following two functions:

- The trend of $f(m)$ in the s months long window, $(m - s, m)$. For brevity we denote

$$K^f(m/s) = K^f(m - s, m) \quad (2)$$

- The deviation of $f(m)$ from extrapolation of its long-term regression (i.e. regression on a long time window $(q, m - 1)$):

$$R^f(m/q) = f(m) - W^f(m/q, m - 1). \quad (3)$$

Both functions can be used for prediction since their values do not depend on the information about the future (after the month m) which would be anathema in prediction.

Discretization. The prediction algorithms use one or several premonitory patterns. Each pattern is defined at

the lowest – binary – level of resolution, as 0 or 1, distinguishing only the presence of absence of a pattern at each moment of time. Then the objects of recognition are described by binary vectors of the same length. This ensures the robustness of the prediction algorithms.

Simple algorithm called Hamming distance is used for classification of binary vectors in applications considered here, [14,20,28]. Each vector is either premonitory or not. Analyzing the samples of vectors of each class (“the learning material”), the algorithm determines a reference binary vector (“kernel”) with components typical for the premonitory vector. Let D be the Hamming distance of a vector from the kernel (the number of non-coinciding binary components). The given vector is recognized as premonitory class, if D is below a certain threshold D^* . This criterion takes advantage of the clustering of precursors in time.

Summing up, these elements of the pattern recognition approach are common for its numerous applications, their diversity notwithstanding. Experience in the specific applications is described in Sects. “Elections”, “US Economic Recessions”, “Unemployment”, “Homicide Surges”. The conceptual summary of the accumulated experience is given in the final Sect. “Summary: Findings and Emerging Possibilities”.

Elections

This Section describes algorithms for predicting the outcome of the US Presidential and mid-term Senatorial elections [28,29,30,31]. Elections’ time is set by the law as follows.

- National elections are held every even-numbered year, on the first Tuesday after the first Monday in November (i. e., between November 2 and November 8, inclusively).
- Presidential elections are held once every 4 years, i. e. on every other election day. People in each of the 50 states and District of Columbia are voting separately for “electors” pledged to one or another of the Presidential candidates. These electors make up the “Electoral College” which directly elects the President. Since 1860, when the present two-party system was basically established, the Electoral College reversed the decision of the popular vote only three times, in 1888, 1912, and 2000. Algorithmic prediction of such reversals is not developed so far.
- A third of Senators are elected for a 6-year term every election day; “mid-term” elections held in the middle of a Presidential term are considered here.

Methodology

The prediction target is an electoral defeat of an “incumbent” party, i. e. the party holding the contested seat. Accordingly, the prediction problem is formulated as whether the incumbent party will retain this seat or lose it to the challenging party (*and not whether Republican or Democrat will win*). As is shown below, that formulation is crucial for predicting the outcomes of elections considered.

Data. The pre-election situation is described by robust common sense parameters defined at the lowest (binary) level of resolution, as the *yes* or *no* answers to the questionnaires given below (Tables 1, 2). The questions are formulated in such a way that the answer *no* favors the victory of the challenging party. According to the Hamming distance analysis (Sect. “Common Elements of Data Analyzes”) the victory of the challenging party is predicted when the number of answers *no* exceeds a threshold D^* .

Mid-term Senatorial Elections

The prediction algorithm was developed by a retrospective analysis of the data on three elections, 1974, 1978, and 1982. The questionnaire is shown in Table 1. Victory of the challenger is predicted if the number of answers *no* is 5 or more [28,29,30].

The meaning of these questions may be broader than their literal interpretation. For example, financial contributions (key 5 in Table 2) not only provide the resources required for an effective campaign, but may also constitute a poll in which the preferences are weighed by the money attached.

Predicting future elections. This algorithm (without any changes from year to year and from state to state) was applied in advance to the five subsequent elections, 1986–2002. Predictions are shown in Fig. 2. Altogether, 150 seats were put up for election. For each seat a separate prediction was made, 128 predictions were correct, and 22 – wrong.

Statistical significance of this score is 99.9%. In other words the probability to get such a score by chance is below 0.1% [28,29,30]. For some elections these predictions might be considered as trivial, since they coincide with prevailing expectation of experts. Such elections are identified by *Congressional Review*. Eliminating them from the score still results in 99% significance.

Presidential Elections

The prediction algorithm was developed by a retrospective analysis of the data on the past 31 elections, 1860–1980;

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 1
Questionnaire for mid-term Senatorial Elections [28]

1.	(Incumbency): The incumbent -party candidate is the sitting senator.
2.	(Stature): The incumbent -party candidate is a major national figure.
3.	(Contest): There was no serious contest for the incumbent -party nomination.
4.	(Party mandate): The incumbent party won the seat with 60% or more of the vote in the previous election.
5.	(Support): The incumbent -party candidate outspends the challenger by 10% or more.
6.	(Obscurity): The challenging -party candidate is not a major national figure or a past or present governor or member of Congress.
7.	(Opposition): The incumbent party is not the party of the President.
8.	(Contest): There is no serious contest for the challenging -party nomination (the nominee gains a majority of the votes cast in the first primary and beats the second-place finisher at least two to one).

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 2
Questionnaire for Presidential elections [29,30]

KEY 1	(Party Mandate): After the midterm elections, the incumbent party holds more seats in the US House of Representatives than it did after the previous midterm elections.
KEY 2	(Contest): There is no serious contest for the incumbent -party nomination.
KEY 3	(Incumbency): The incumbent -party candidate is the sitting president.
KEY 4	(Third party): There is significant third-party or independent campaign.
KEY 5	(Short-term economy): The economy is not in recession during the election campaign.
KEY 6	(Long-term economy): Real per -capita economic growth during the term equals or exceeds mean growth during the previous two terms.
KEY 7	(Policy change): The incumbent administration effects major changes in national policy.
KEY 8	(Social unrest): There is no sustained social unrest during the term.
KEY 9	(Scandal): The incumbent administration is unattained by a major scandal.
KEY 10	(Foreign/military failure): The incumbent administration suffers no major failure in foreign or military affairs.
KEY 11	(Foreign/military success): The incumbent administration achieves a major success in foreign or military affairs.
KEY 12	(Incumbent charisma): The incumbent -party candidate is charismatic or a national hero.
KEY 13	(Challenger charisma): The challenging -party candidate is not charismatic or a national hero.

that covers the period between victories of A. Lincoln and R. Reagan inclusively. The questionnaire is shown in Table 2. Victory for the challenger is predicted if the number of answers *no* is 6 or more [29,30].

Predicting of future elections. This algorithm (without any changes from year to year state) was applied in advance to the six subsequent elections, 1984–2004. Predictions are shown in Fig. 3. All of them happened to be correct. In 2000 the decision of popular majority was reversed by the Electoral College; such reversals are not targeted by this algorithm [29,30].

Understanding Elections

Collective behavior. The finding that aggregate-level parameters can reliably anticipate the outcome of both presidential and senatorial elections points to an electoral behavior highly integrated not only for the nation as a whole but also within the diverse American states.

- A presidential election is determined by the collective, integrated estimation of performance of incumbent administration during the previous four years.
- In case of senatorial elections the electorate has more diffused expectations of performance but puts more importance on political experience and status than in the case of presidential elections. Senate incumbents, unlike presidential ones, do not suffer from a bad economy or benefit from a good one. (This suggests that rather than punishing the party holding a Senate seat for hard times, the voters may instead regard the incumbent party as a safe port in a storm).

Similarity. For each election year in all states the outcomes of elections follow the same pattern that transcends the diversities of the situations in each of the individual elections.

The same pattern of the choice of the US President prevails since 1860, i. e. since election of A Lincoln, despite

0	1	2	3	4	5	6	7
			OK98				
			CO98				
			FL98				
			GA98				
			HA98	TN02			
			ID98	SC02			
			MA98	NC02			
			ND98	NE02			
			PN98	KY02			
			SD98	IA02			
			UT98	CO02			
			FL94	AL02			
			HA94	AK98			
			IN94	CA98			
			MT94	CT98			
			NB94	NE98			
			NJ94	OR98			
			TX94	SC98			
			WA94	VT98			
		AS98	WV94	WA98			
		KA98	WI94	CT94			
		LA98	AK90	MD94			
		MI98	IN90	NV94			
		NH98	KN90	WY94			
		MS94	ME90	CO90			
	AL98	NM94	MA90	HA90			
	AZ98	ND94	MT90	KY90			
	IO98	RI94	NB90	MI90			
	DL94	VT94	NC90	AZ86			
	MA94	AS90	TX90	CO86			
	NY94	IO90	WY90	ID86			
	AL90	MS90	AR86	LA86			
	DE90	NM90	CA86	NY86			
	IL90	OR90	IL86	OK86	WI98	MN94	
	LA90	RI90	IN86	WI86	CA94	MO94	
	OK90	SD90	IA86	NC86	ID90	VA94	
	SC90	VA90	NH86	WA86	PA86	NH90	
	TN90	WV90	OR86	MN90	IL98	IN98	
	HI86	AK86	VT86	OK94	ME94	OH98	
	OH86	CT86	TN94	PA94	AL86	MI94	
UT94	SC86	KS86	TX02	TN294	FL86	MD86	KY98
GA90	UT86	KY86	OK02	NC98	GA86	NV86	AZ94
NJ90	NH02	ND86	NJ02	NY98	MO86	SD86	OH94
0	1	2	3	4	5	6	7

OK98 – incumbent won, KY98 – challenger won, errors are highlighted.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 2

Made-in-advance predictions of the mid-term senatorial elections (1986–2002). Each election is represented by the two-letter state abbreviation with the election year shown by two last digits. Each column shows elections with certain number D of answers “no” to the questionnaire given in Table 1 (such answers are favorable to challenging party). Value of D , indicated at the top, is the Hamming distance from the kernel

D (number of answers NO)	0	1	2	3	4	5	6	7	8	9
Predictions published months in advance			1984	1988	2004	2000* 1996	1992			
Learning				1964 1928 1916 1908 1944 1956	1900 1872 1924	1972 1948			1980 1976 1968 1952 1932 1920	1960
	1904	1936	1868	1864	1880	1888*	1892	188 4 186 0	1896	1876*

1904 years when incumbent won popular vote

1892 years when challenger won popular vote

* years when popular vote was reversed by electoral vote

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 3

Division of presidential elections (1860–2004) by the number D of answers “no” to the questionnaire given in Table 2 (such answers are favorable to challenging party). D is the Hamming distance from the kernel

all the overwhelming changes in the electorate, the economy, the social order and the technology of politics during these 130 years. (For example, the electorate of 1860 did not include the groups, which constitute 3/4 of present electorate, such as women, African Americans, or most of the citizens of the Latin American, South European, Eastern European, and Jewish descent [30].

An alternative (and more traditional) concept of American elections focuses on the division of voters into interest and attitudinal groups. By this concept the goal of the contestants is to attract the maximum number of voting blocks with minimal antagonism from other blocks. Electoral choice depends strongly on the factors irrelevant to the essence of the electoral dilemma (e. g. on the campaign tactics). The drawbacks of this concept are discussed in [18,30]. In sum, the work on presidential and senatorial elections described above suggests the following new ways of understanding American politics and perhaps the politics of other societies as well.

1. Fundamental shifts in the composition of the electorate, the technology of campaigning, the prevailing economic and social conditions, and the key issues of campaigns do not necessarily change the pragmatic basis on which voters choose their leaders.
2. It is governing not campaigning that counts in the outcomes of presidential elections.
3. Different factors may decide the outcome of executive as compared to legislative elections.
4. Conventional campaigning will not improve the prospects for candidates faced with an unfavorable

combination of fundamental historical factors. Disadvantaged candidates have an incentive to adopt innovative campaigns that break the pattern of conventional politics.

5. All candidates would benefit from using campaigns to build a foundation for governing in the future.

US Economic Recessions

US National Bureau of Economic Research (NBER) has identified the seven recessions that occurred in the US since 1960 (Table 3). The starting points of a recession and of the recovery from it follow the months marked by a peak and a trough of economic activity, respectively.

A peak indicates the last month before a recession, and a trough – the last month of a recession.

Prediction targets considered are the first month after the peak and after the trough (“the turns to the worst and to the best”, respectively). The start of the first recession, in 1960, is not among the targets, since the data do not cover a sufficient period of time preceding the recession.

The data used for prediction comprise the following six monthly leading economic indicators obtained from the CITIBASE data base, Jan. 1960–June 2000 (abbreviations are the same, as in [49]).

G10FF = FYGT10 – FEDFUN Difference between the annual interest rate on 10 year US Treasury bonds, and federal fund annual interest rate.

IP Industrial Production, total: index of real (constant dollars, dimensionless) output in the entire economy.

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 3

American Economic Recessions since 1960

#	Peaks	Troughs
1	1960:04	1961:02
2	1969:12	1970:11
3	1973:11	1975:03
4	1980:01	1980:07
5	1981:07	1982:11
6	1990:07	1991:03
7	2001:03	2001:11

This represents mainly the manufacturing industry, because of the difficulties in measuring the quantity of the output in services (such as travel agents, banking, etc.).

LHELL Index of “help wanted” advertising. This is put together by a private publishing company that measures the amount of job advertising (column-inches) in a number of major newspapers.

LUINC Average weekly number of people claiming unemployment insurance.

INVTQ Total inventories in manufacturing and trade, in real dollars. Includes intermediate inventories (for example held by manufacturers, ready to be sent to retailers) and final goods inventories (goods on the shelves in stores).

FYGM3 Interest rate on 90 day US treasury bills at an annual rate (in percent).

These indicators were already known [48,49], as those that correlate with a recession’s approach.

Prediction of a Recession Start

Single indicators exhibit the following premonitory patterns:

G10FF: small value

IP and INVTQ: small deviation from the long-term trend R^f (3)

FYGM3: large deviation from the long-term trend R^f

LHELL: small trend K^f (2)

LUINC: large trend K^f

The prediction algorithm triggers an alarm after a month when most of the patterns emerge simultaneously. It lasts Δ months and can be extended by the same rule, if premonitory patterns keep emerging. Formal quantitative definition of the algorithm can be found in [23] along with its validation by sensitivity and out-of-sample analyzes.

Alarms and recessions are juxtaposed in Fig. 4. We see that five recessions occurring between 1961 and 2000 were predicted by an alarm. The sixth recession started in April 2001, one month before the corresponding alarm. (Recession of 1960 was not considered for prediction, since data analyzed start just before it.)

Only the first six recessions listed in Table 1 were considered in the developing of the algorithm [23]. Duration of each alarm was between 1 and 14 months. Total duration of all alarms was 38 months, or 13.6% of the time interval considered. There were no false alarms. No alarms were yielded so far by subsequent prediction in advance and no recession was identified during that time.

Prediction of a Recession End

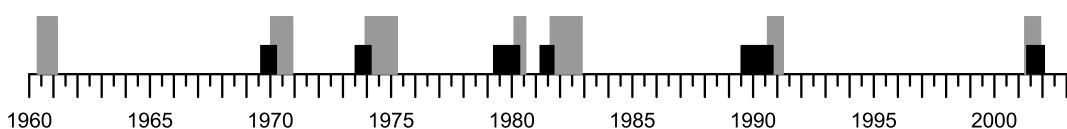
Prediction targets are the starting points of recovery from recessions; these points are indicated in the last column of Table 3.

The data comprise the same six indicators that indicate the approach of a recession (see Subsect. “**Prediction of a Recession Start**”); they are analyzed only within the recessions’ periods.

Data analysis shows intriguing regularity illustrated in Fig. 5:

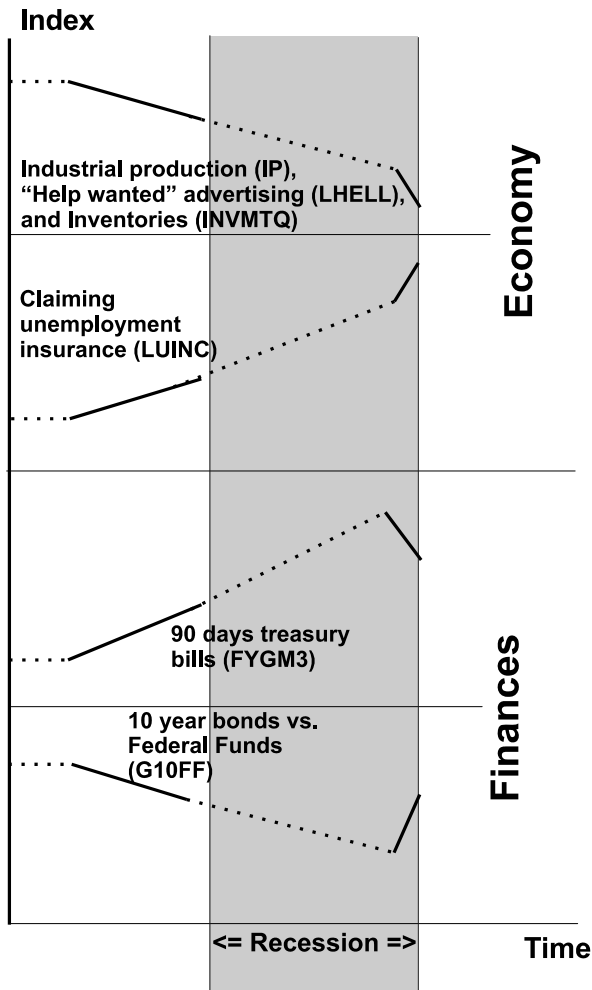
- Financial indicators change in opposite directions before the recession and before the recovery.
- Economic indicators change in the same direction before the recession and the recovery; but the change is stronger before the recovery, i. e., the economic situation worsens.

Prediction algorithm is formulated in the same terms as in the previous case but an alarm is triggered after *three* consecutive months when most of the patterns emerge simultaneously. The alarms predict when the recovery will start. Alarms and prediction targets are juxtaposed in Fig. 6. Du-



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 4

Alarms (black bars) and recessions (gray bars)



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 5

Premonitory changes of indicators before the start of a recession and before its end. See explanations in the text

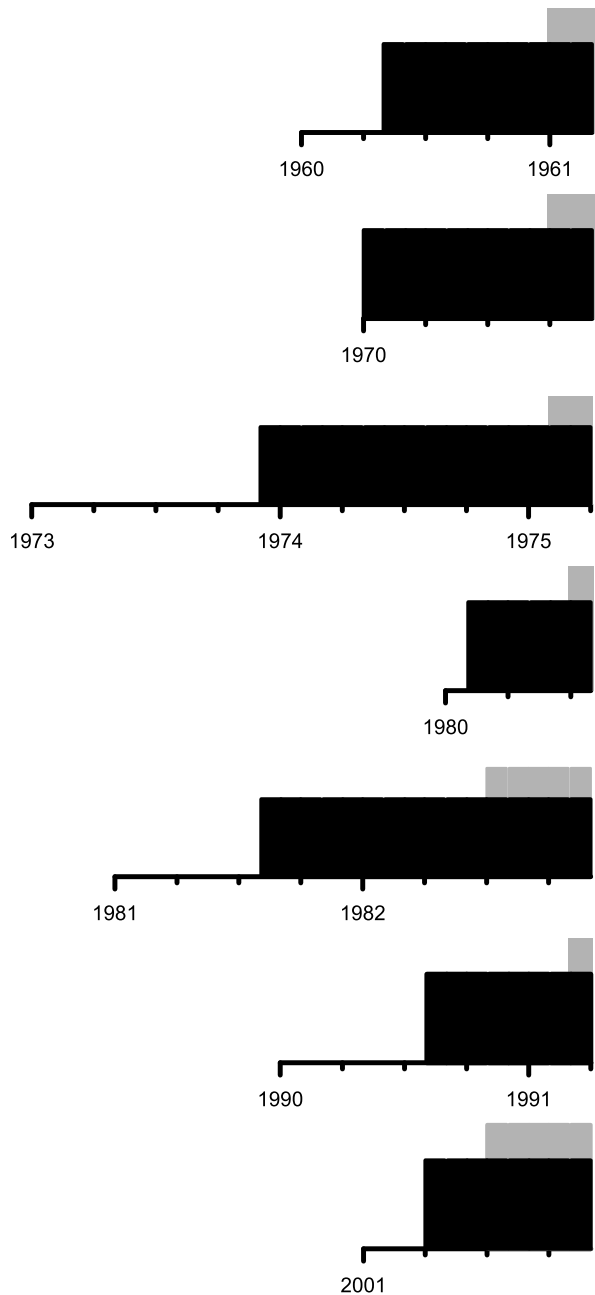
ration of a single alarm is one to five months. Total duration of alarms is 16 months, which is 22% of time covered by all recessions. There are neither false alarms nor failures to predict.

Unemployment

Here we describe uniform prediction of the sharp and lasting unemployment surge in France, Germany, Italy, and the USA [25].

Prediction Target

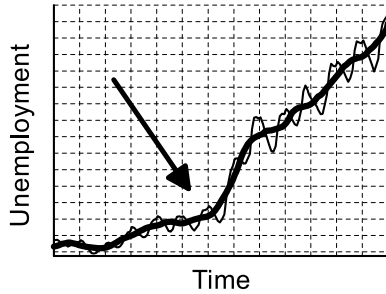
A prediction target is schematically illustrated in Fig. 7. Thin curve shows monthly unemployment with seasonal



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 6

Prediction of recovery from a recession. Black bars – periods of recessions. Gray bars – alarms preceding the end of a recession

variations. On the thick curve seasonal variations are smoothed away. The arrow indicates a sharp upward bend of the smoothed curve. The moment of that bend is the prediction target. It is called by the acronym FAU, for “Fast Acceleration of Unemployment”.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 7

Fast acceleration of unemployment (FAU): schematic definition. Thin line – monthly unemployment; with seasonal variations. Thick line – monthly unemployment, with seasonal variations smoothed away. The arrow indicates a FAU – the sharp bend of the smoothed curve. The moment of a FAU is the target of prediction

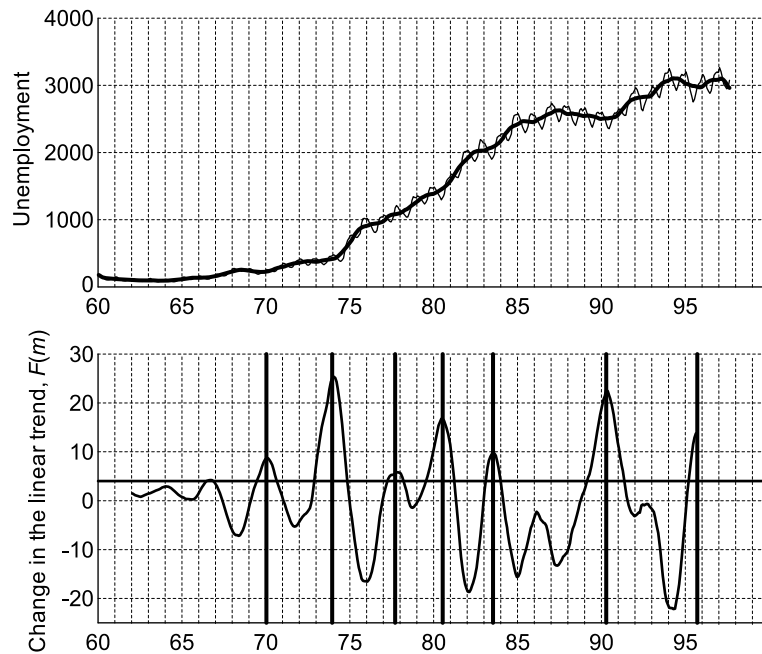
Smoothing was done as follows: Let $u(m)$ be number of unemployed in a month $m = 1, 2, \dots$. After smoothing out the seasonal variation we obtain time series $U(m) = W^u(m/m - 6, m + 6)$; this is the linear regression over the year-long time interval $(m - 6, m + 6)$. A natural robust measure of unemployment acceleration at the time m is the bend of the linear trend of U ; in notations used

in (1) this is the function $F(m/s) = K^U(m + s, m) - K^U(m, m - s)$. The FAUs are identified by the local maxima of $F(m)$ exceeding a certain threshold F . The time m^* and the height F^* of such a maximum are, respectively, the time and the magnitude of a FAU. Subsequent local minimum of $F(m)$ identifies the month m_e when acceleration ends. Figure 8 shows thus defined FAUs for France.

The Data

The analysis has been initially made for France and three groups of data have been analyzed.

- *Composite macroeconomic indicators of national economy*
 1. **IP:** Industrial production indicator, composed of weighted production levels in numerous sectors of the economy, in % relative to the index for 1990.
 2. **L:** Long-term interest rate on 10-year government bonds, in %.
 3. **S:** Short-term interest rate on 3-month bills, in %.
- *Characteristics of more narrow areas of French economy*
 4. **NC:** The number of new passenger car registrations, in thousands of units.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 8

Unemployment in France. Top: Monthly unemployment, thousands of people. Thin line: $u(m)$, data from the OECD database; note the seasonal variations. Thick line: $U(m)$, data smoothed over one year. Bottom: Determination of FAUs. $F(m)$ shows the change in the linear trend of unemployment $U(m)$. FAUs are attributed to the local maxima of $F(m)$ exceeding threshold $F = 4.0$ shown by horizontal line. The thick vertical lines show moments of the FAUs

5. **EI**: Expected prospects for the national industrial sector.
 6. **EP**: Expected prospects for manufacturers.
 7. **EO**: Estimated volume of current orders.
- Indicators 5–7 distinguish only “good” and “bad” expectations determined polling 2,500 manufacturers, whose answers are by the size of their businesses.
- *Indicators related to US economy.*
 - 8. **FF/\$**: Value of US dollar in French francs.
 - 9. **AR**: The state of the American economy: is it close to a recession or not? This indicator shows the presence or absence of a current pre-recession alarm (see Subsect. “[Prediction of a Recession Start](#)”).

The data bases with above indicators for Europe are issued by the Organization for Economic Cooperation and Development [43] and the International Monetary Fund [16].

American analogues of indicators **IP**, **L**, and **S** are provided by CITIBASE; they are described in Sect. “[US Economic Recessions](#)” under abbreviations **IP**, **FYGM3** and **FIGT10** respectively.

Prediction

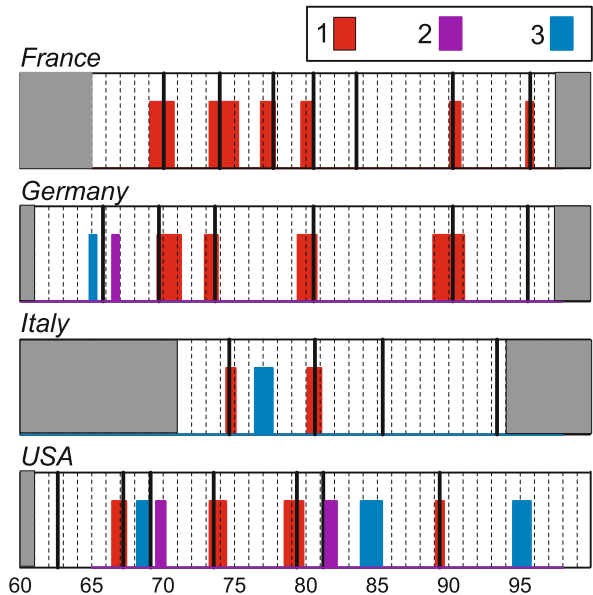
Single indicators exhibit the following premonitory behavior.

- Steep upward trends of composite indicators (#1–#3). This behavior reflects “overheating” of the economy and may sound counterintuitive for industrial production (#1), since the rise of production is supposed to create more jobs. However, a particularly steep rise may create oversupply.
- Steep downward trends of economic expectations by general public (#4) and business community (#5–#8).
- Proximity of an American recession (#9). Before analysis was made such and opposite precursors might be expected for equally plausible reasons, so that this finding, if further confirmed, does provide a constraint on understanding unemployment’s dynamics.

Among different combinations of indicators the macroeconomic ones (#1–#3) jointly give relatively better predictions, with smallest rates of errors and highest stability in sensitivity tests.

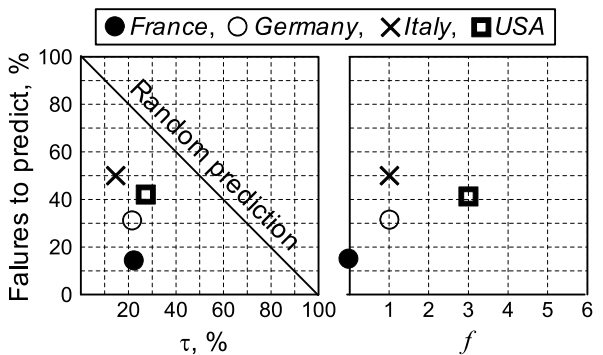
Retrospective prediction. Macroeconomic indicators were used jointly in the Hamming distance prediction algorithm (Sect. “[Common Elements of Data Analyzes](#)”). Being robust and self-adjusting to regional conditions, this algorithm was applied without any changes to the four countries considered here.

Alarms and *FAUs* are juxtaposed in Fig. 9. Error diagram in Fig. 10 shows quality of prediction for different



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 9

Retrospective predictions for four countries: *FAUs* and alarms obtained by the prediction algorithm. The *thick vertical lines* show the moments of *FAUs* in a country. *Bars* – the alarms with different outcome: 1 – alarms that predict *FAUs*, 2 – alarms starting shortly after *FAUs* within the periods of unemployment surge, 3 – false alarms. *Shaded areas* on both sides indicate the times, for which data on economic indicators were unavailable

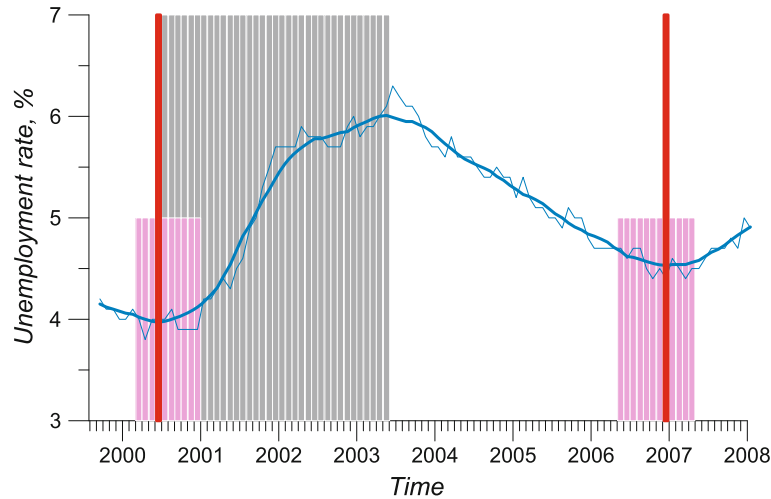


Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 10

Error diagram for prediction of *FAUs* in different countries; τ is total duration of alarms in % to the time interval considered, f – total number of false alarms

countries. For US the quality is lower than for European countries, though still higher than in random predictions.

Prediction of the future FAUs was launched for USA. The results are shown in Fig. 11. It shows that by January 2008 two correct predictions have been made, without other false alarms or failures to predict. In November



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 11

Experiment in predicting future FAUs, September (1999)–January (2008). Thin blue curve shows monthly unemployment rate in USA, according to data of Bureau of Labor Statistics, US Department of Labor (<http://www.data.bls.gov>). Thick curve shows this rate with seasonal variation smoothed away. Vertical red lines show prediction targets – the moments of FAU, gray bar – the period of unemployment's growth; pink bars – periods of alarms

2006 the second prediction was filed on the web site of the Anderson School of Management, University of California, Los Angeles (<http://www.uclaforecast.com/>). This started the documented experiment in testing the algorithm by predicting future FAUs on that website.

Homicide Surges

This section analyzes the prediction of homicide rates in an American megacity – Los Angeles, CA [24].

Prediction Target

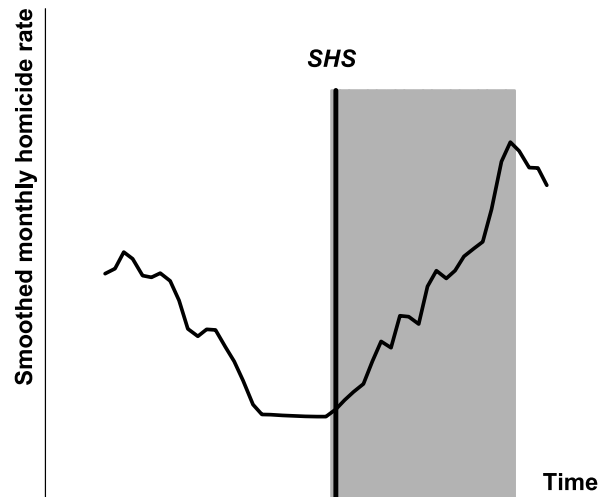
A prediction target is the start of a sharp and lasting acceleration of the homicide rate; it is called by the acronym SHS, for "Start of the Homicide Surge." It is formally determined by the analysis of monthly homicides rates, with seasonal variations smoothed out, as described in Subsect. "Prediction Target". Prediction targets thus identified are shown by vertical lines in Figs. 12 and 14 below.

The Data

The analyzed data include monthly rates of the homicides and 11 types of lesser crimes, listed in Table 2. Definitions of these crimes are given in [6].

The data are taken from two sources:

- The National Archive of Criminal Justice Data, placed on the web site (NACJD), 1975–1993.

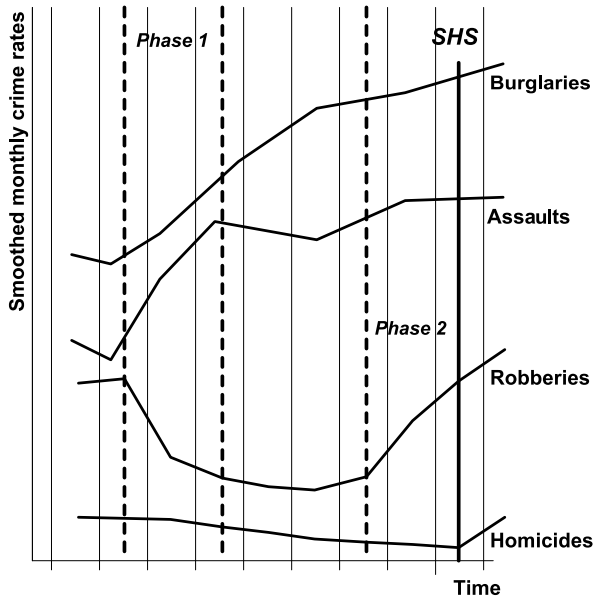


Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 12

Target of prediction – the Start of the Homicide Surge ("SHS"); schematic definition. Gray bar marks the period of homicide surge

- Data bank of the Los Angeles Police Department (LAPD) Information Technology Division), 1990–2003.

The algorithm does not use socio-economic determinants of crime, or other data that might be also useful. The objective was to develop a simple, efficient prediction model; development of comprehensive causal model would be a complementary objective.



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 13

Scheme of premonitory changes in crime statistics

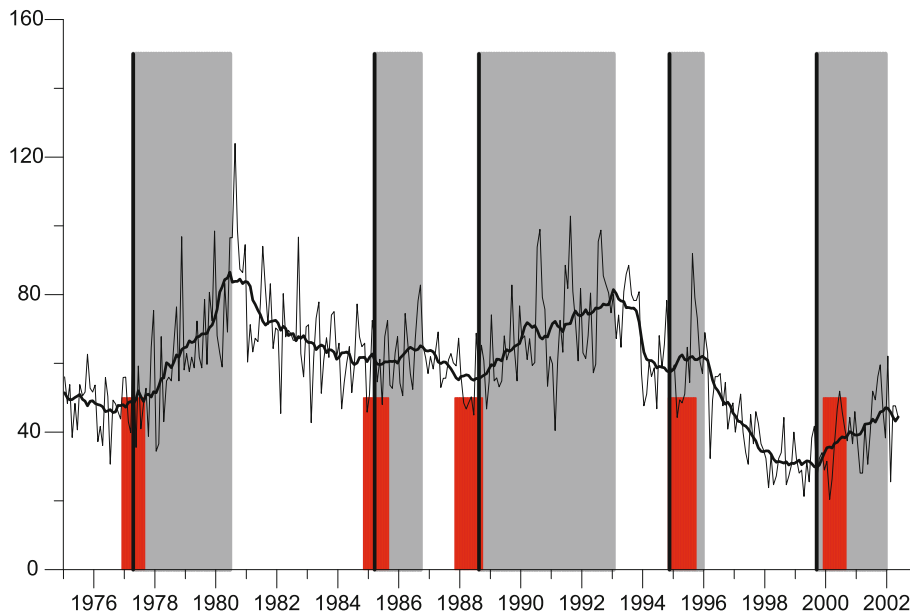
Prediction

Premonitory behavior of indicators is illustrated in Fig. 13. The first phase is characterized by an escalation of burglaries and assaults, but not of robberies. Later on, closer to a homicide surge, robberies also increase.

The Prediction algorithm based on Hamming distance (see Sect. “Common Elements of Data Analyzes”) uses seven indicators listed in Table 4. Other five indicators marked by * are used in sensitivity tests; and the homicide rate is used for identification of targets SHS.

Alarms and homicide surges are juxtaposed in Fig. 14. The SHS episode in November 1994 has occurred simultaneously with the corresponding alarm. It is captured by an alarm, which starts in the month of SHS without a lead time. Prediction missed the October 1999 episode: it occurred two months before the start of the corresponding alarm. Such delays should be taken into account for validating the algorithm. Note, however, that the last prediction did remain informative.

Altogether alarms occupy 15% of the time considered. During phase 2 (as defined in Fig. 13) this rate might be reduced [24].



Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Figure 14

Performance of prediction algorithm through 1975–2002. Thin curve – original time series, total monthly number of homicides in Los Angeles city, per 3,000,000 inhabitants. Data from NACJD [6] have been used for 1975–1993 and from the Data Bank of the Los Angeles Police Department (LAPD Information Technology Division) for subsequent 9 years. Thick curve – smoothed series, with seasonal variations eliminated. Vertical lines show the targets of prediction – episodes of SHS (Subjective Homicide Surge). Gray bars show the periods of homicide surge. Red bars show the alarms declared by the prediction algorithm [24]

Extreme Events in Socio-economic and Political Complex Systems, Predictability of, Table 4

Types of crimes considered (after [6])

Homicide	Robberies	Assaults	Burglaries
• All	• All	• All*	• Unlawful not forcible entry
	• With firearms	• With firearms	• Attempted forcible entry*
	• With knife or cutting instrument	• With knife or cutting instrument	
	• With other dangerous weapon	• With other dangerous weapon*	
	• Strong-arm robberies*	• Aggravated injury assaults*	

*Analyzed in sensitivity tests only

Summary: Findings and Emerging Possibilities

The findings described above enhance predictive understanding of extreme events and indicate yet untapped possibilities for further R&D in that field.

Pattern Recognition Approach

Information extracted from the already available data is indeed increased by this approach. To each problem considered here one may apply the following conclusion of J. Stock, a leading expert in the field: “Prediction/of recessions/requires fitting non-linear, high-dimensional models to a handful of observations generated by a possibly non-stationary economic environment... The evidence presented here suggests that these simple binary transformations of economic indicators have significant predictive content for recessions. It is striking that these models, in which the information in the data is reduced to binary indicators, have predictive contents comparable to or, in many cases, better than that of more conventional models.” Importantly, this is achieved by using not more detailed data and models, but more robust aggregation (Subsect. “Predictability vs. Complexity: The Need for Holistic Approach”).

Partial “universality” of premonitory patterns is established by broad research in modeling and data analysis. This includes the common definition of the patterns, their self-adjustment, scaling, and similarity [9,10,20,26,42]; see also references in Sects. “Elections”, “US Economic Recessions”, “Unemployment”, “Homicide Surges”).

Relation to “cause and effect” analysis (perpetrators or witnesses?). Premonitory patterns might be either “perpetrators” contributing to causing extreme events, or the “witnesses” – parallel manifestations of the system’s development. The cause that triggered a specific extreme event is usually identified, at least in retrospect. It may be, for example, a certain governmental decision, a change in the international situation, a natural disaster, the depletion of natural resources etc. However an actual extreme event

might materialize only if the system is destabilized and “ripe” for it. Patterns of each kind signal such a ripe situation.

What premonitory patterns to use for prediction? Existing theories and experience reduce the number of such patterns, but too many of them remain hypothetically promising and have to be chosen by a trial and error procedure. Inevitably a prediction algorithm begins with a limited number of promising patterns. They should be sufficient for prediction, but other patterns may be equally or more useful and should be considered in further development of the algorithm. Most relevant “perpetrators” might not be included in the most useful patterns (e.g. due to their sensitivity to too many factors).

Relation to policy-making: prediction and disaster preparedness. Reliable predictions of future extreme events in complex societal systems would allow policy-makers to take remedial action before rather than after the onset of such afflictions as economic disasters, crime surges, etc. As in case of military intelligence predictions would be useful if their accuracy is known, albeit not necessarily high. Analysis of error diagrams allows to regulate the tradeoff between the rates of failures to predict and false alarms according to the needs of a decision-maker.

Relation to governing and campaigning. The findings presented here for the USA elections show that top elected officials would have better chances for reelection, if they focus on effective governing, and not on rhetoric, packaging and image-making. Candidates will benefit themselves and their parties if they run substantive campaigns that build a foundation for governing during the next term.

Further Possibilities

A wealth of yet untapped data and models is readily available for the continuation of the kinds of studies described and analyzed in this article. Following are some immediate possibilities; specific examples can be found in the given references.

- *Continuing experiments in advance prediction*, for which the above findings set up a base (Sect. “[Elections](#)”). Successes and errors are equally important [37,38].
- *Incorporating other available data into the analysis* (Sects. “[US Economic Recessions](#)”, “[Unemployment](#)”)
- *Predicting the same kind of extreme events in different contexts* (Sect. “[Unemployment](#)”)
- *Predicting the end of a crisis* (Sect. “[US Economic Recessions](#)”).
- *Multistage prediction with several lead times* (Sect. “[Homicide Surges](#)”)
Less imminent, but within reach are:
- “*Universal*” scenarios of extreme development and low-parametric definition of an ensemble of premonitory patterns [9,51,52].
- *Validation of an algorithm and joint optimization of prediction and preparedness strategy* [38].
- *Developing prediction algorithms for other types of extreme events.*

The authors would be glad to provide specific information upon request.

Generalizations

The problems considered here have the following common features:

- *The absence of a closed theory* that would unambiguously determine prediction methodology. This leads to the need for intense intertwining of mathematics, statistical physics and non-linear dynamics, a range of societal sciences, and practical experience (Subsect. “[Methodology](#)”). In reality this requires long-term collaboration of respective experts. As can be seen from the references to Sects. “[Elections](#)”, “[US Economic Recessions](#)”, “[Unemployment](#)”, “[Homicide Surges](#)” previous applications inevitably involved the teams of such experts.
- *Predictions in advance* is the only final validation of the results obtained.
- *The need for holistic analysis* driven to extreme robustness.
- *Considerable, albeit limited, universality* of the premonitory phenomena.

Two classical quotations shed the light on these features:

A. N. Kolmogoroff. “It became clear for me that it is unrealistic to have a hope for the creation of a pure theory [of the turbulent flows of fluids and gases] closed in itself. Due to the absence of such a theory we have to rely upon

the hypotheses obtained by processing of the experimental data.”

M. Gell-Mann: “... if the parts of a complex system or the various aspects of a complex situation, all defined in advance, are studied carefully by experts on those parts or aspects, and the results of their work are pooled, an adequate description of the whole system or situation does not usually emerge. ... The reason, of course, is that these parts or aspects are typically entangled with one another. ... We have to supplement the partial studies with a transdisciplinary crude look at the whole.”

In the general scheme of things the problem considered belongs to a much wider field – the quest for a universal theory of complex systems extended to predicting extreme events – the Holy Grail of complexity studies. This quest encompasses the natural and human-made complex systems that comprise what some analysts have called “the global village”. It requires entirely new applications of modern science, such as algebraic geometry, combinatorics, and thermodynamics. As a means for anticipating, preventing and responding to natural and manmade disasters and for improving the outcomes of economic and political systems, the methods described here may hold one key for the survival and sustainability of our civilization.

Bibliography

Primary Literature

1. Allègre CJ, Le Mouél J-L, Ha Duyen C, Narteau C (1995) Scaling organization of fracture tectonics (SOFT) and earthquake mechanism. *Phys Earth Planet Inter* 92:215–233
2. Armstrong JS, Cuzan AG (2005) Index methods for forecasting: An application to american presidential elections. *Foresight Int J Appl Forecast* 3:10–13
3. Blanter EM, Shnirman MG, Le Mouél JL, Allègre CJ (1997) Scaling laws in blocks dynamics and dynamic self-organized criticality. *Phys Earth Planet Inter* 99:295–307
4. Bongard MM, Vaintsveig MI, Guberman SA, Izvekova ML, Smirnov MS (1966) The use of self-learning prog in the detection of oil containing layers. *Geol Geofiz* 6:96–105
5. Burridge R, Knopoff L (1967) Model and theoretical seismicity. *Bull Seismol Soc Am* 57:341–360
6. Carlson SM (1998) Uniform crime reports: Monthly weapon-specific crime and arrest time series 1975–1993 (National, State, 12-City Data), ICPSR 6792 Inter-university Consortium for Political and Social Research. Ann Arbor
7. Farmer JD, Sidorowich J (1987) Predicting chaotic time series. *Phys Rev Lett* 59:845
8. Gabriellov A, Keilis-Borok V, Zaliapin I, Newman WI (2000) Critical transitions in colliding cascades. *Phys Rev E* 62:237–249
9. Gabriellov A, Keilis-Borok V, Zaliapin I (2007) Predictability of extreme events in a branching diffusion model. arXiv:0708.1542

10. Gabrielov AM, Zaliapin IV, Newman WI, Keilis-Borok VI (2000) Colliding cascade model for earthquake prediction. *Geophys J Int* 143(2):427–437
11. Gelfand IM, Guberman SA, Keilis-Borok VI, Knopoff L, Press F, Ranzman IV, Rotwain IM, Sadovsky AM (1976) Pattern recognition applied to earthquake epicenters in California. *Phys Earth Planet Inter* 11:227–283
12. Gell-Mann M (1994) *The quark and the jaguar: Adventures in the simple and the complex*. Freeman, New York
13. Crutchfield JP, Farmer JD, Packard NH, Shaw RS (1986) *Chaos* *Sci Am* 255:46–57
14. Gvishiani AD, Kosobokov VG (1981) On found of the pattern recognition results applied to earthquake-prone areas. *Izvestiya Acad Sci USSR. Phys Earth* 2:21–36
15. Holland JH (1995) *Hidden order: How adaptation builds complexity*. Addison, Reading
16. IMF (1997) International monetary fund, international financial statistics. CD-ROM
17. Kadanoff LP (1976) Scaling, universality and operator algebras. In: Domb C, Green MS (eds) *Phase transitions and critical phenomena*, vol 5a. Academic, London, pp 1–34
18. Keilis-Borok VI, Lichtman AJ (1993) The self-organization of American society in presidential and senatorial elections. In: Kravtsov YA (ed) *Limits of predictability*. Springer, Berlin, pp 223–237
19. Keilis-Borok VI, Press F (1980) On seismological applications of pattern recognition. In: Allègre CJ (ed) *Source mechanism and earthquake prediction applications*. Editions du centre national du la recherché scientifique, Paris, pp 51–60
20. Keilis-Borok VI, Soloviev AA (eds) (2003) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Berlin
21. Keilis-Borok V, Soloviev A (2007) Pattern recognition methods and algorithms. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-11
22. Keilis-Borok VI, Sorondo MS (2000) (eds) *Science for survival and sustainable development. The proceedings of the study-week of the Pontifical Academy of Sciences*, 12–16 March 1999. Pontificiae Academiae Scientiarum Scripta Varia, Vatican City
23. Keilis-Borok V, Stock JH, Soloviev A, Mikhalev P (2000) Pre-recession pattern of six economic indicators in the USA. *J Forecast* 19:65–80
24. Keilis-Borok VI, Gascon DJ, Soloviev AA, Intriligator MD, Pichardo R, Winberg FE (2003) On predictability of homicide surges in megacities. In: Beer T, Ismail-Zadeh A (eds) *Risk science and sustainability*. Kluwer, Dordrecht (NATO Sci Ser II Math, Phys Chem 112), pp 91–110
25. Keilis-Borok VI, Soloviev AA, Allègre CB, Sobolevskii AN, Intriligator MD (2005) Patterns of macroeconomic indicators preceding the unemployment rise in Western Europe and the USA. *Pattern Recogn* 38(3):423–435
26. Keilis-Borok V, Soloviev A, Gabrielov A, Zaliapin I (2007) Change of scaling before extreme events in complex systems. In: *Proceedings of the plenary session on “predictability in science: Accuracy and limitations”*, Pontificiae Academiae Scientiarum Scripta Varia, Vatican City
27. Kravtsov YA (ed) (1993) *Limits of predictability*. Springer, Berlin
28. Lichtman AJ, Keilis-Borok VI (1989) Aggregate-level analysis and prediction of midterm senatorial elections in the United States, 1974–1986. *Proc Natl Acad Sci USA* 86(24):10176–10180
29. Lichtman AJ (1996) *The keys to the White House*. Madison Books, Lanham
31. Lichtman AJ (2005) The keys to the White House: Forecast for 2008. *Foresight Int J Appl Forecast* 3:5–9
30. Lichtman AJ (2008) *The keys to the White House*, 2008 edn. Rowman/Littlefield, Lanham
32. Ma Z, Fu Z, Zhang Y, Wang C, Zhang G, Liu D (1990) *Earthquake prediction: Nine major earthquakes in china*. Springer, New York
33. Mason IB (2003) Binary events. In: Jolliffe IT, Stephenson DB (eds) *Forecast verification. A practitioner's guide in atmospheric science*. Wiley, Chichester, pp 37–76
34. Molchan GM (1990) Strategies in strong earthquake prediction. *Phys Earth Planet Inter* 61:84–98
35. Molchan GM (1991) Structure of optimal strategies of earthquake prediction. *Tectonophysics* 193:267–276
36. Molchan GM (1994) Models for optimization of earthquake prediction. In: Chowdhury DK (ed) *Computational seismology and geodynamics*, vol 1. Am Geophys Un, Washington, pp 1–10
37. Molchan GM (1997) Earthquake prediction as a decision-making problem. *Pure Appl Geophys* 149:233–237
38. Molchan GM (2003) Earthquake prediction strategies: A theoretical analysis. In: Keilis-Borok VI, Soloviev AA (eds) *Nonlinear dynamics of the lithosphere and earthquake prediction*. Springer, Berlin, pp 209–237
39. Molchan G, Keilis-Borok V (2008) Earthquake prediction: Probabilistic aspect. *Geophys J Int* 173(3):1012–1017
40. NACJD: <http://www.icpsr.umich.edu/NACJD/index.html>
41. NBER: <http://www.nber.org/cycles/cyclesmain.html>
42. Newman W, Gabrielov A, Turcotte DL (eds) (1994) *Nonlinear dynamics and predictability of geophysical phenomena*. Am Geophys Un, Int Un Geodesy Geophys, Washington
43. OECD (1997) *Main economic indicators: Historical statistics 1960–1996*. Paris, CD-ROM
44. Press F, Briggs P (1975) Chandler wobble, earthquakes, rotation and geomagnetic changes. *Nature* 256:270–273, London
45. Press F, Briggs P (1977) Pattern recognition applied to uranium prospecting. *Nature* 268:125–127
46. Press F, Allen C (1995) Patterns of seismic release in the southern California region. *J Geophys Res* 100(B4):6421–6430
47. Soloviev A (2007) Application of the pattern recognition techniques to earthquake-prone areas determination. Ninth workshop on non-linear dynamics and earthquake prediction, Trieste ICTP 1864-9
48. Stock JH, Watson MW (1989) New indexes of leading and coincident economic indicators. *NBER Macroecon Ann* 4:351–394
49. Stock JH, Watson MW (1993) A procedure for predicting recessions with leading indicators. In: Stock JH, Watson MW (eds) *Business cycles, indicators, and forecasting (NBER Studies in Business Cycles, vol 28)*, pp 95–156
50. Tukey JW (1977) *Exploratory data analysis*. Addison-wesley series in behavioral science: Quantitative methods. Addison, Reading
51. Turcotte DL, Newman WI, Gabrielov A (2000) A statistical physics approach to earthquakes. In: *Geocomplexity and the physics of earthquakes*. Am Geophys Un, Washington
52. Zaliapin I, Keilis-Borok V, Ghil M (2003) A Boolean delay model of colliding cascades, II: Prediction of critical transitions. *J Stat Phys* 111(3–4):839–861

Books and Reviews

- Bongard MM (1967) The problem of recognition. Nauka, Moscow
- Brito DL, Intriligator MD, Worth ER (1998) In: Eliasson G, Green C (eds) Microfoundations of economic growth: A Schumpeterian perspective. University of Michigan Press, Ann Arbor
- Bui Trong L (2003) Risk of collective youth violence in french suburbs. A clinical scale of evaluation, an alert system. In: Beer T, Ismail-Zadeh A (eds) Risk science and sustainability. Kluwer, Dordrecht (NATO Sci Ser II Math Phys Chem 112)
- Engle RF, McFadden DL (1994) (eds) Handbook of econometrics, vol 4. North-Holland, Amsterdam
- Klein PA, Niemira MP (1994) Forecasting financial and economic cycles. Wiley, New York
- Messner SF (1983) Regional differences in the economic correlates of the urban homicide rate. *Criminology* 21:477–488
- Mitchell WC (1951) What happens during business cycles: A progress report. NBER, New York
- Mitchell WC, Burns AF (1946) Measuring business cycles. NBER, New York
- Moore GH (ed) (1961) Business cycle indicators. NBER, New York
- Mostaghimi M, Rezayat F (1996) Probability forecast of a downturn in US economy using classical statistical theory. *Empir Econ* 21:255–279
- Watson MW (1994) In: Engle RF, McFadden DL (eds) Handbook of econometrics, vol IV. North-Holland, Amsterdam

Extreme Value Statistics

MARIO NICODEMI

Department of Physics, University of Warwick,
Coventry, UK

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Extreme Value Distributions](#)

[The Generalized Extreme Value Distribution](#)

[Domains of Attraction and Examples](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Random variable When a coin is tossed two random outcomes are permitted: head or tail. These outcomes can be mapped to numbers in a process which defines a ‘random variable’: for instance, ‘head’ and ‘tail’ could be mapped respectively to +1 and −1. More generally, any function mapping the outcomes of a random process to real numbers is defined a random variable [15]. More technically, a random variable is any function from a probability space to some measurable

space, i. e., the space of admitted values of the variable, e. g., real numbers with the Borel σ -algebra. The amount of rainfall in a day or the daily price variation of a stock are two more examples. It’s worth to stress that, formally, the outcome of a given random experiment is not a random variable: the random variable is the function describing all the possible outcomes as numbers. Finally, two random variables are said independent when the outcome of either of them has no influence on the other.

Probability distribution The probability of either outcomes, ‘head’ and ‘tail’, in tossing a coin is 50%. Similarly, a discrete random variable, X , with values $\{x_1, x_2, \dots\}$ has an associate discrete probability distribution of occurrence $\{p_1, p_2, \dots\}$. More generally, for a random variable on real numbers, X , the corresponding probability distribution [15] is the function returning the probability to find a value of X within a given interval $[x_1, x_2]$ (where x_1 and x_2 are real numbers): $\Pr[x_1 \leq X \leq x_2]$. In particular, the random variable, X , is fully characterized by its cumulative distribution function, $F(x)$, which is: $F(x) = \Pr[X < x]$ for any x in \mathcal{R} . The probability distribution density, $f(x)$, can be often defined as the derivative of $F(x)$: $f(x) = dF(x)/dx$.

The probability distribution of two independent random variables, X and Y , is the product of the distributions, F_X and F_Y , of X and Y : $F(x, y) \equiv \Pr[X < x; Y < y] = \Pr[X < x] \cdot \Pr[Y < y] = F_X(x) \cdot F_Y(y)$.

Expected value The expected value [15] of a random variable is its average outcome over many independent experiments. Consider, for instance, a discrete random variable, X , with values in the set $\{x_1, x_2, \dots\}$ and the corresponding probability for each of these values $\{p_1, p_2, \dots\}$. In probability theory, the expected, or average, value of X (denoted $E(X)$) is just the sum: $E(X) = \sum x_i p_i$. For instance, if you have an asset which can give two returns $\{x_1, x_2\}$ with probability $\{p_1, p_2\}$, its expected return is $x_1 p_1 + x_2 p_2$.

In case we have a random variable defined on real numbers and $F(x)$ is its probability distribution function, the expected value of X is: $E(X) = \int X dF$. As for some $F(x)$ the above integral may not exist, the ‘expected value’ of a random variable is not always defined.

Variance and moments The variance [15] of a probability distribution is a measure of the average deviations from the mean of the related random variable. In probability theory, the variance is usually defined as the mean squared deviation, $E((X - E(X))^2)$, i. e., the expected value of $(X - E(X))^2$. The square root of the

variance is named the standard deviation and is a more sensible measure of fluctuations of X around $E(X)$. Alike $E(X)$, for some distributions the variance may not exist. In general, the expected value of the k th power of X , $E(X^k)$, is called the k th moment of the distribution.

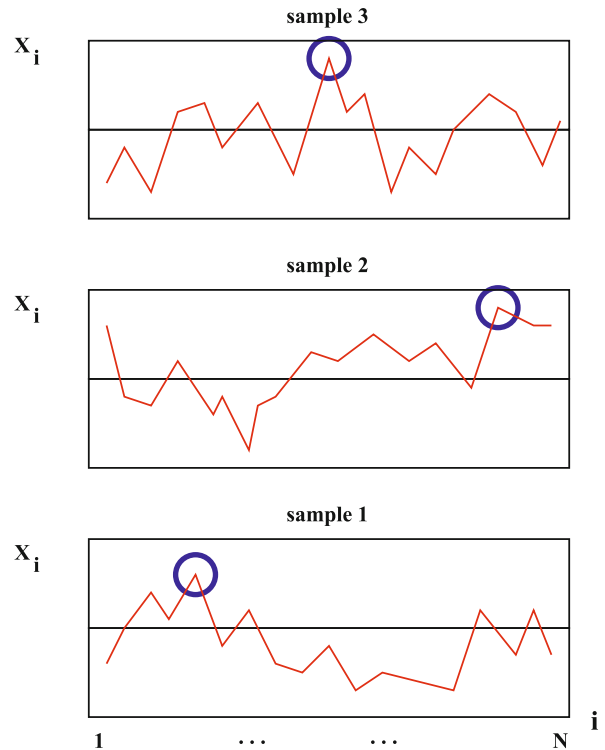
The Central limit theorem The Central Limit Theorem [15] is a very important result in probability theory stating that the sum of N independent identically-distributed random variables, with finite average and variance, has a Gaussian probability distribution in the limit $N \rightarrow \infty$, irrespective of the underlying distributions of the random variables. The domain of attraction of the Gaussian as a limit distribution is, thus, very large and can explain why the Gaussian is so frequently encountered. The theorem is, in practice, very useful since many real random processes have a finite average and variance and are approximately independent.

Definition of the Subject

Extreme value theory is concerned with the statistical properties of the extreme events related to a random variable (see Fig. 1), and the understanding and applications of their probability distributions. The methods and the practical use of such a theory have been developed in the last 60 years, though, many complex real-life problems have only recently been tackled. Many disciplines use the tools of extreme value theory including meteorology, hydrology, ocean wave modeling, and finance to name just a few.

For example, in economics, extreme value theory is currently used by actuaries to evaluate and price insurance against the probability of rare but financially catastrophic events. An other application is for the estimation of Value at Risk. In hydrology, the theory is applied by environmental risk agencies to calculate, for example, the height of sea-walls to prevent flooding. Similarly, extreme value theory is also used to set strength boundaries in engineering materials, as well as for material fatigue and reliability in buildings (e. g., bridges, oil rigs), and estimating pollution levels.

This paper aims to give a simple, self contained, introduction to the motivations and basic ideas behind the development of extreme value theory, and briefly covers a few more technical topics such as extreme r order statistics and the generalization of extreme value distribution theory. We refer to textbooks on probability theory, such as [15] (or, for simplicity, to the Glossary), for the definition of the basic notions of probability used here.



Extreme Value Statistics, Figure 1

We show three samples, 1 to 3, each with N realizations of a random variable $X_i (i \in \{1, \dots, N\})$. Extreme value theory is concerned with the statistical properties of occurrence of the extreme values in those samples, such as the maxima (circled points)

The extensive research on extreme value statistics is reviewed in excellent books published over the past years, e. g., [6,11,12,13,14,17,18,21,24]; we provide here only an overview of the basic concepts and tools. To make this paper as self contained as possible, the Glossary gives a beginner introduction to all the elementary notions of probability theory encountered in the following sections.

Introduction

Extreme events, exceeding the typical expected value of a random variable, can have substantial relevance to problems raising in disciplines as diverse as sciences, engineering and economics.

Extreme value theory is a sub-field of applied statistics, early developed [13,14,17,18] by mathematician such as Fisher, Tippet, Gnedenko, and, in particular, Emil Julius Gumbel, dealing precisely with the problems related to extreme events (see Fig. 1). One of its key point is the so called ‘three types theorem’, relating the properties of

the distribution of probability of the underlying stochastic variable to its extreme value distributions, i. e., the limiting distributions for the extreme (minimum or maximum) value of a large collection of random observations. Interestingly, for a comparatively large class of random variables, the theory points out that only a few species of limit extreme value distributions are found.

In some respect, the ‘three types theorem’ can be considered the analogous of the well known central limit theorem applying to ordinary sums, or averages, of random variables. From a practical point of view it is as important, since it opens a way to estimate the asymptotic distribution of extreme values without any a priori knowledge, or guess assumption, on the parent distribution. In this way, we have a solid ground to estimate the parameters of the limit distributions along with their confidence intervals, an issue crucial, for instance, to proper risk assessment.

In Finance, for example, market regulators and financial institutions face the important and complex task to estimate and manage risk. Assessing the probability of rare and extreme events is, of course, a crucial issue and reliable measures of risk are needed to minimize undesirable effects on portfolios from large fluctuations in market conditions, e. g., exchange rates or prices of assets. Similar issues about risk and reliability are faced in insurance and banking, which are deeply concerned with unusually large fluctuations. Extreme value theory provides the solid theoretical foundation needed for the statistical modeling of such events and proper computation of risk and related confidence intervals.

The study of natural and environmental hazards is also strongly interested to extreme events; for instance, reported hydrology and meteorology applications of extreme event theory concern flood frequency analysis, estimation of precipitation probabilities, or extreme tide levels. Predictions of events such as strong heat waves, rainfall, occurrence of huge sea waves are deeply grounded on such a theory as well. Analogous problems are found in telecommunications and transport systems, such as traffic data analysis, Internet traffic, queuing modeling; problems from material science, pollutants hazards on health add examples from a very long list of related phenomena.

It is impossible to summarize here the huge, often technical, literature on all these topics and we refer to the general books cited in the bibliography. Actually, to give an idea of the variety of applications of the theory, we only mention a few more example from still an other class of disciplines, Physical Sciences. In Physics, for instance, the equilibrium low-temperature properties of disordered systems are characterized by the statistics of ex-

tremely low-energy states. Several problems in this class, including the Random Energy Model and models for decaying Burgers turbulence, have been connected to extreme value distributions [3]. In GaAs films, extreme values in Gaussian $1/f$ correlations of voltage fluctuations were shown to follow one of the limit distributions of extreme value theory, the Gumbel asymptote [1]. Hierarchically correlated random variables representing the energies of directed polymers [9] and maximal heights of growing self-affine surfaces [20] exhibits extreme value statistics as well. The Fisher–Tippett–Gumbel asymptote is involved in distribution of extreme height fluctuations for Edwards–Wilkinson relaxation of fluctuating interfaces on small-world-coupled interacting systems [16]. A connection was also established between the energy level density of a gas of non-interacting bosons and the distribution laws of extreme value statistics [7].

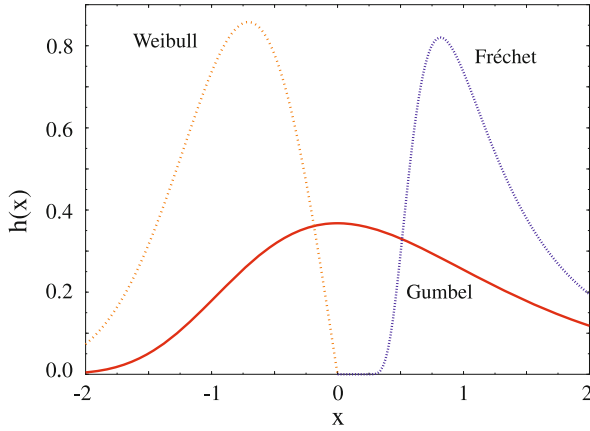
In case of systems with correlated variables, the application of the extreme value theory is far from trivial. A theorem states that the statistics of maxima of stationary Gaussian sequences, with suitable correlations, asymptotically converges to a Gumbel distribution [2]. Evidences supporting similar results were derived from numerical simulations and analysis of long-term correlated exponentially distributed signals [10]. In general, however, the scenario is non trivial. For instance, in Physics a variant of Gumbel distribution was observed in turbulence [4] and derived in the two-dimensional XY model [5], which are systems where correlations play an important role. Similarly, correlated extreme value statistics were discovered in the Sneppen depinning model [8]. In models for fluctuating correlated interfaces, such as Edwards–Wilkinson and Kardar–Parisi–Zhang equations, an exact solution for the distribution of maximal heights was recently derived and it turns out to be an Airy function [19].

After the above brief picture of the field, we illustrate next the general properties of extreme value distributions of independent random variables.

The Extreme Value Distributions

Extreme value distributions are the limit distributions of extremes (either maxima or minima) of a set of random variables (see Fig. 1). For definiteness, we will deal here with maxima, as minima can be seen as ‘maxima’ of a set of variables with opposite signs.

Consider a set $\{X_1, X_2, \dots, X_N\}$ of N independent identically distributed random variables, X_i , with a cumulative distribution function, $F(x) \equiv \Pr\{X_i \leq x\}$. The maximum in the set, $Y_N = \text{Max}\{X_1, X_2, \dots, X_N\}$, has a distribution function, $H_N(x)$, which is simply related to



Extreme Value Statistics, Figure 2

As an example, we plot in this figure the Gumbel density distribution, $h(x) = dH(x)/dx$, from Eq. (3), and the Fréchet and Weibull density distributions from Eqs. (4) and (5), for $\alpha = 2$

F , since by definition of Y_N , we have:

$$\begin{aligned} H_N(x) &\equiv \Pr\{Y_N \leq x\} \\ &= \Pr\{X_1 \leq x, X_2 \leq x, \dots, X_N \leq x\} \\ &= \Pr\{X_1 \leq x\} \cdot \Pr\{X_2 \leq x\} \cdot \dots \cdot \Pr\{X_N \leq x\} \\ &= F^N(x). \end{aligned} \quad (1)$$

In the limit of large samples, $N \rightarrow \infty$, it is possible to show that, under some general hypotheses on F described below, we can find a suitable sequence of scaling constants a_N and b_N , such that the scaled variable $y_N = (Y_N - b_N)/a_N$ has a non degenerate probability distribution function $H(y)$. Specifically, as $N \rightarrow \infty$, the distribution $\Pr\{y_N \leq y\}$ has a non trivial well defined limit $H(y)$:

$$\begin{aligned} \Pr\{y_N \leq y\} &= \Pr\{(Y_N - b_N)/a_N \leq y\} \\ &= \Pr\{Y_N \leq a_N y + b_N\} \\ &= F^N(a_N y + b_N) \rightarrow H(y) \text{ for } N \rightarrow \infty. \end{aligned} \quad (2)$$

For a given underlying distribution, F , the individuation of the precise sequence of scaling constants a_N and b_N required in Eq. (2) is a non trivial technical problem in the mathematics of extreme values [18], which we briefly discuss in the next sections. Such an issue is overshadowed by the simplicity of the result of the ‘three types theorem’, which states that there are only three types (apart from a scaling transformation of the variable) of limiting distribution $H(y)$ (see Fig. 2):

I) Gumbel type:

$$H(y) = \exp[-\exp(-y)] \quad (3)$$

with $-\infty < y < \infty$;

II) Fréchet type:

$$H(y) = \exp[-y^{-\alpha}] \quad (4)$$

where α is a fixed exponent and $0 < y < \infty$ (with $H(y) = 0$ for $y < 0$);

III) Weibull type:

$$H(y) = \exp[-(-y)^\alpha] \quad (5)$$

where α is a fixed exponent and $-\infty < y < 0$ (with $H(y) = 1$ for $y > 0$).

The Generalized Extreme Value Distribution

In the extreme value statistics literature, the three types of limiting distributions, Gumbel, Fréchet and Weibull, are often represented as a single family including all of them, the so called generalized extreme value distribution:

$$H(y; \mu, \sigma, \xi) = \exp \left[- \left(1 + \xi \frac{y - \mu}{\sigma} \right)^{-1/\xi} \right] \quad (6)$$

with support in the interval where $1 + \xi(y - \mu)/\sigma > 0$, as otherwise H is either zero or one. Out of the three parameters μ, σ, ξ of Eq. (6), ξ is called the ‘shape’ parameter and it is especially important as it selects the specific type of asymptote:

I) The case $\xi = 0$ corresponds to the Gumbel asymptote, since it is easy to show that

$$\lim_{\xi \rightarrow 0} H(y; \mu, \sigma, \xi) = \exp \left[- \exp \left(\frac{-y - \mu}{\sigma} \right) \right]; \quad (7)$$

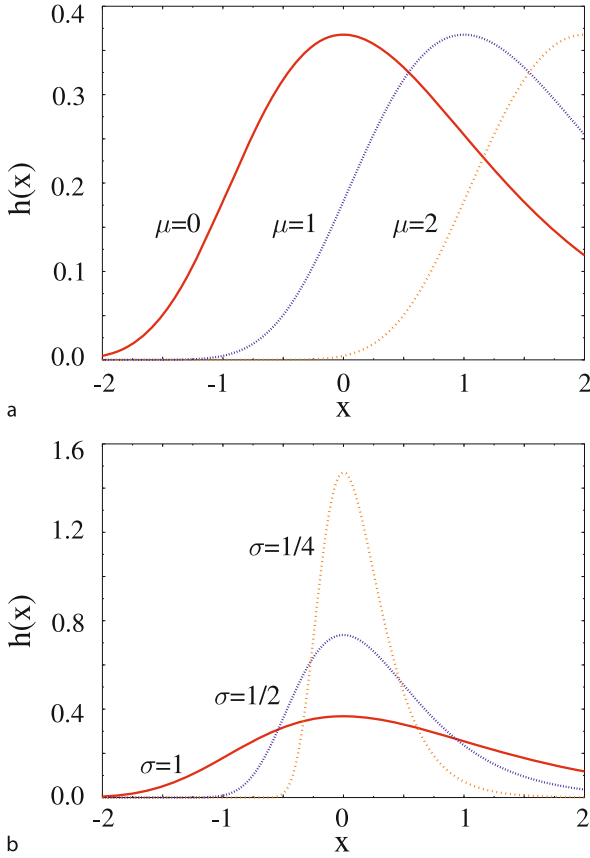
II) Similarly, the case $\xi > 0$ corresponds to the Fréchet asymptote of Eq. (4), where the exponent is $\alpha = 1/\xi$;

III) And, finally, the case $\xi < 0$ corresponds to the Weibull asymptote of Eq. (5), where the exponent is $\alpha = -1/\xi$.

The parameters μ and σ of Eq. (6) are called the ‘location’ and the ‘scale’ parameters, since they are related to the moments of the generalized extreme value distribution of Eq. (6). Figure 3 plots the effects of changes in μ and σ on the form of $H(y)$ from Eq. (6) in the Gumbel case, $\xi = 0$. It is possible to show [18] that the k th moment is finite only if $\xi < 1/k$. The mean, which exists only if $\xi < 1$, can be expressed in the following general form

$$E(y) = \mu + \frac{\sigma}{\xi} [\Gamma(1 - \xi) - 1], \quad (8)$$

where $\Gamma(x)$ is the Gamma function. In the Gumbel limit, $\xi \rightarrow 0$, the above result is simplified to: $E(y) = \mu + \sigma\gamma$, where $\gamma = 0.577 \dots$ is the Euler γ constant.



Extreme Value Statistics, Figure 3

In this figure we show the effects of the μ and σ parameters on the appearance of the generalized extreme value density distribution, $h(x) = dH(x)/dx$, from Eq. (6), in the case $\xi = 0$, i.e., the Gumbel type. In the upper panel, we plot $h(x)$ for $\sigma = 1$ and $\mu = 0, 1, 2$. In the lower panel, we plot $h(x)$ for $\mu = 0$ and $\sigma = 1/4, 1/2, 1$.

Analogously, the variance, existing for $\xi < 1/2$, can be written as:

$$E([y - E(y)]^2) = \left(\frac{\sigma}{\xi}\right)^2 [\Gamma(1-2\xi) - \Gamma^2(1-\xi)], \quad (9)$$

which in the $\xi \rightarrow 0$ limit becomes $E([y - E(y)]^2) = \sigma^2 \pi^2/6$.

The r Largest Order Statistics

The results on distributions of the maximum (or minimum) discussed above can be extended to the set of the r th largest value of an ensemble. Consider a set $\{X_1, X_2, \dots, X_N\}$ of N identically distributed random variables which, for simplicity of notation, are arranged in order of magnitude: $X_1 < X_2 < \dots < X_N$. As before,

$F(x) \equiv \Pr\{X_i \leq x\}$ is their common cumulative distribution function. The statistics of X_N and X_1 are the distribution of, respectively, the maximum and the minimum seen before. Similarly, X_r (with $r \in \{1, N\}$) is called the r (largest) order statistic.

The r order statistic has a distribution function, $H_r(x)$, simply related to F

$$\begin{aligned} H_r(x) &\equiv \Pr\{X_r \leq x\} \\ &= \Pr\{X_1 \leq x, X_2 \leq x, \dots, X_r \leq x\} \\ &\quad \cdot \Pr\{X_{r+1} > x, \dots, X_N > x\} \\ &= \sum_{i=r}^N \frac{N!}{(N-i)!i!} F^i(x) [1 - F(x)]^{N-i}. \end{aligned} \quad (10)$$

The theory of the generalized extreme value distribution can be extended to the r order statistic. Actually, in the limit $N \rightarrow \infty$, if a suitable sequence of scaling constants, a_N and b_N , can be found such that the scaled maximum variable $y_N = (X_N - b_N)/a_N$ has a limit distribution function $H(y)$ given in Eq. (6), then the r order statistic has a limit distribution which can be easily expressed in terms of $H(y)$ [18].

In order to describe a broader panorama of the available results in extreme value theory, we give here some details on the more general case of the limit probability distribution density of the vector of the first r largest values $(y_1, y_2, \dots, y_r) = ((X_N - b_N)/a_N, (X_{N-1} - b_N)/a_N, \dots, (X_{N-r+1} - b_N)/a_N)$. Such a limit distribution density can be shown to be [18]

$$\begin{aligned} h(y_1, \dots, y_r) &= \frac{1}{\sigma^r} \exp \left[- \left(1 + \xi \frac{y_r - \mu}{\sigma} \right)^{-\frac{1}{\xi}} \right. \\ &\quad \left. - \left(1 + \frac{1}{\xi} \right) \sum_{i=1}^r \ln \left(1 + \xi \frac{y_i - \mu}{\sigma} \right) \right]. \end{aligned} \quad (11)$$

Most of the other results of the previous sections can be generalized to the r order statistics, as shown for instance in [18].

Domains of Attraction and Examples

The problem of finding the domains of attraction of the classes of limiting distribution is a complex, partially still open, topic in extreme value theory [18]. Even in the case of independent identically distributed random variables, understanding which asymptote a given distribution, F , converges to and which is the sequence of scaling constants, a_N and b_N , can be a non trivial task. As the extreme events of a random variable are characterized by the tail of

F , a simple approximate approach to guess the domain of attraction F falls into is to consider its behavior for large x . We summarize below a few well known examples having a broad validity which can guide practical applications of the theory to the case where the random variables are independent and identically distributed.

- E I) Many common distributions, $F(x)$, have exponential tails in x , very important examples being the Gaussian and the exponential distributions. In this case, their extreme value statistic is the Gumbel asymptote.

A more formal condition for F to belong to the domain of attraction of the Gumbel limiting distribution was established by von Mises. Take a function F and denote x_{\max} the largest value in its support, i. e., where $F(x_{\max}) = 1$ (the point x_{\max} can be also infinite). Consider the derivative $f(x) = dF(x)/dx$ and the rate of $F(x)$ in approaching 1 as $x \rightarrow x_{\max}$: when $x \rightarrow x_{\max}$, if

$$\frac{d}{dx} \left[\frac{1 - F(x)}{f(x)} \right] \rightarrow 0 \quad (12)$$

then $\Pr\{y_N \leq y\}$ tends to the Gumbel asymptote given in Eq. (3).

The above criterion can be rephrased in a more colloquial way: the Gumbel type is the limiting distribution when $1 - F(x)$ decays faster than a polynomial for $x \rightarrow x_{\max}$. Beyond the Gaussian and exponential distributions, the lognormal, the Gamma, the Weibull, the Benktander-type I and II, and many more common distributions, with x_{\max} either finite or infinite, belong to this class.

- E II) Distributions such as the Pareto, Cauchy, Student, Burr have the Fréchet asymptote. More generally, when x_{\max} is infinite and F has a power law tail for $x \rightarrow \infty$

$$1 - F(x) \simeq x^{-\alpha} \quad (13)$$

with an exponent $\alpha > 0$, then the domain of attraction of the extreme value statistics is the Fréchet type given in Eq. (4), with precisely the same exponent α of Eq. (13).

- E III) Finally, when x_{\max} is finite and F has a power law behavior for $x \rightarrow x_{\max}$

$$1 - F(x) \simeq (x_{\max} - x)^{-\alpha} \quad (14)$$

with an exponent $\alpha > 0$, then the domain of attraction of the extreme value statistics is the Weibull type of Eq. (5), with the same exponent α of

Eq. (14). The Uniform and Beta distributions have, for instance, the Weibull asymptote.

Extremes of Correlated Random Variables

When the underlying random variables are not independent, as in many cases of practical relevance ranging from Meteorology to Finance, the problem to individuate the form, or even the existence, of the limiting distribution is, in general, open. The existing broad technical literature on the topic [18] shows that the three types, Gumbel, Fréchet and Weibull, summarized in the generalized extreme value distribution of Eq. (6), often arise as well. For instance, a recent theorem has shown that in the case of stationary Gaussian sequences with suitable correlations the distribution of maxima asymptotically follows the Gumbel type [2]. Analysis of numerical simulations of long-term correlated exponentially distributed signals has given evidences supporting similar conclusions [10]. Sometimes, when considering ‘time’ series of N correlated variables, the approximate rule of thumb that N must be much bigger than the ‘correlation length’ of the sequence is used as a guide to decide whether Eq. (6) is likely to be the right asymptote.

Some of the example mentioned in the Introduction can help, however, in delineating the strong limits of the above approximate criteria and the lack of a general picture. For instance, in the XY model for magnetic systems used in Statistical Physics, in the Kosterlitz–Thouless low temperature phase the magnetization has a distribution which is a generalized Gumbel [5], but not the one in Eq. (3), a result expected to hold in a broad class of systems. In models for fluctuating interfaces, developing correlations, described by Edwards–Wilkinson and Kardar–Parisi–Zhang like equations, it has been derived that the exact distribution of maximal heights is an Airy function [19]. These examples show the variety of situations which can arise in practical cases and indicate that the theorems derived for independent identically distributed variables must be applied with caution.

Future Directions

In the sections above, we reviewed at an introductory level the mathematics of extreme value theory, with a special focus on the ‘three types theorem’ on the limiting distributions. We also discussed their domains of attraction and many examples on random extreme events. We have not covered, instead, other important, though, more technical and still evolving topics such as the theoretical approach to the problem of ‘exceedances over thresholds’, and the methodology for estimating from real sample data the

parameters of extreme distributions, such as maximum likelihood and Bayesian methods. These are covered, for instance, in the general references listed in the bibliography. Actually, there is a number of excellent textbooks on these topics, ranging from the original book by E.J. Gumbel [17], to more recent volumes illustrating in details extreme value theory in the formal framework of the theory of probability [11,13,14,18]. Volumes more focused on applications to Finance and Insurances are, e. g., [6,11,12,21], as applications to climate, hydrology and meteorology research are found in [6,12,21,24]. Finally, there is a number of more technical review papers on the topic, including [10,22,23].

Bibliography

1. Antal T, Droz M, Gyrgyi G, Racz Z (2001) *Phys Rev Lett* 87:240601
2. Berman SM (1964) *Ann Math Stat* 35:502
3. Bouchaud J-P, Mézard M (1997) *J Phys A* 30:7997
4. Bramwell ST, Holdsworth PCW, Pinton J-F (1998) *Nature (London)* 396:552
5. Bramwell ST, Christensen K, Fortin J-Y, Holdsworth PCW, Jensen HJ, Lise S, Lopez JM, Nicodemi M, Pinton J-F, Sellitto M (2000) *Phys Rev Lett* 84:3744
6. Bunde A, Kropp J, Schellnhuber H-J (eds) (2002) *The science of disasters-climate disruptions, heart attacks, and market crashes*. Springer, Berlin
7. Comtet A, Leboeuf P, Majumdar SN (2007) *Phys Rev Lett* 98:070404
8. Dahlstedt K, Jensen HJ (2001) *J Phys A* 34:11193; [Inspec] [ISI]
9. Dean DS, Majumdar SN (2001) *Phys Rev E* 64:046121
10. Eichner JF, Kantelhardt JW, Bunde A, Havlin S (2006) *Phys Rev E* 73:016130
11. Embrechts P, Klüppelberg C, Mikosch T, Karatzas I, Yor M (eds) (1997) *Modelling extremal events*. Springer, Berlin
12. Finkenstadt B, Rootzen H (2004) *Extreme values in finance, telecommunications, and the environment*. Chapman and Hall/CRC Press, London
13. Galambos J (1978) *The asymptotic theory of extreme order statistics*. Wiley, New York
14. Galambos J, Lechner J, Simin E (eds) (1994) *Extreme value theory and applications*. Kluwer, Dordrecht
15. Gnedenko BV (1998) *Theory of probability*. CRC, Boca Raton, FL
16. Guclu H, Korniss G (2004) *Phys Rev E* 69:065104(R)
17. Gumbel EJ (1958) *Statistics of extremes*. Columbia University Press, New York
18. Leadbetter MR, Lindgren G, Rootzen H (1983) *Extremes and related properties of random sequences and processes*. Springer, New York
19. Majumdar SN, Comtet A (2004) *Phys Rev Lett* 92:225501; *J Stat Phys* 119, 777 (2005)
20. Raychaudhuri S, Cranston M, Przybyla C, Shapir Y (2001) *Phys Rev Lett* 87:136101
21. Reiss RD, Thomas M (2001) *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields*. Birkhäuser, Basel
22. Smith RL (2003) *Statistics of extremes, with applications in environment, insurance and finance*, chap 1. In: *Statistical analysis of extreme values: with applications to insurance, finance, hydrology, and other fields*. Birkhäuser, Basel
23. Smith RL, Tawn JA, Yuen HK (1990) *Statistics of multivariate extremes*. *Int Stat Rev* 58:47
24. v. Storch H, Zwiers FW (2001) *Statistical analysis in climate research*. Cambridge University Press, Cambridge