

Identification of Cellular Automata

ANDREW ADAMATZKY

Computing, Engineering and Mathematical Sciences,
University of the West of England, Bristol, UK

Article Outline

Glossary
Definition of the Subject
Introduction
Background and Basics of Identification
Identification Using Machine Learning
Identification Using Polynomial Representation
Binary Tree Representations
and Genetic Programming
Identification Using Decision Trees
Identification by Immunocomputing
Application of Identification in Automatic Design
of Cellular-Automata Processors
Future Directions
Bibliography

Glossary

Cellular automaton is an array of finite automata connected locally, which update their states in discrete time and at the same moments; every automaton updates its next state depending on the states of its closest neighbors.

Decision tree is a mapping from a classified set of observations about an event to the conclusion about its outcome.

Deterministic automaton has only one next state for each pair of internal and input states.

Finite automaton is an abstract machine which takes a finite number of states and transitions between the states; the machine changes its states depending on the input states.

Immunocomputing replicates principles of information processing by immune networks to perform computation.

Learning automaton modifies its transition rules depending on its past experience.

Learning classifier system is a rule-based system, a population of rules, which are processed, selected and updated using reinforcement learning techniques.

Machine learning is a subfield of artificial intelligence concerned with the design and development of algorithms and techniques that allow computers to learn – to improve automatically through experience.

Orthogonalization is subdividing a system into its distinct components.

Polynomial representation of cell-state transition rules interpret local transition rules of a cellular automaton as a Boolean or arithmetic polynomial.

Definition of the Subject

Identification of the cellular automaton is the reconstruction of cell-state transition rules from a given series of global transformations, i. e., the approximation of the minimal cellular automaton that implements given global transformations.

Introduction

The functional synthesis of finite automata problem was first formulated in the 1960s, and is as follows: given an operator on “superwords”, we wish to construct a finite automaton that realizes this operator. In other words, given some language, we wish to build an automaton that represents this language [37,38]. A meta-language of regular expressions as well as algorithms for synthesis was discussed by Kleene [23]. As to the synthesis of cellular-automata-like networks, only the results obtained by [26] on the homogeneous structures with external inputs and outputs are known. The possibility of finding cellular-automata rules and initial conditions that generate a specified time series

was indicated as yet another task in cellular automaton synthesis by Voorhees [40].

Well-known methods of automata synthesis with determined structure and behavior include Markov chains and fuzzy systems for indeterministic automata, and genetic programming for deterministic automata. One of the most efficient ways is to use genetic programming in the design of automaton parts [24,25]. Thus, Andre et al. [14] used genetic programming with automatically defined functions to evolve a rule for the majority classification task for one-dimensional cellular automata. They obtained a rule with an accuracy of 82 percent.

Another way is to involve results founded on the identification of finite automata, i.e. the reconstruction of the automaton structure from the given snapshots of automaton behavior. Two pioneering works on this subject are based on the theory of experiments with finite automata [31,37]. A detailed historical overview of automata experiments can be found in a textbook on finite automata by Brauer [15]. Identification is generally very similar to the classic algorithms proposed by Trakhtenbrot and Bardzin [38], and their modifications [32].

Adamatzky [1,2,3,4,5,6,7,8,9,10,11,12,13] developed algorithms for the identification of various classes of cellular automata, i.e., approximating minimal cellular automata from a finite series of the snapshots, or configurations, recorded in the global evolution of the automaton. A basic scheme for the identification of a deterministic cellular automaton is implemented in the following steps. The minimal radius for the cell's neighborhood is chosen, and for all configurations all observable transitions in the table of cell-state transitions are collected, in the format "cell's neighborhood state at time t " \rightarrow "cell's state at time $t + 1$ ". If there are only two transitions with identical left sides but different right sides – this is a sign of indeterminism – then it is assumed the wrong neighborhood has been chosen (when identified automaton is known to be deterministic). So the radius of the neighborhood is increased and the table of cell-state transitions is collected again. Identification is complete if the state transitions for all possible states of the neighborhood are found.

Background and Basics of Identification

Cellular automaton is an array of uniform finite automata connected locally. Each finite automaton, called a *cell*, of the array takes a finite number of states. All cells update their states simultaneously, in parallel, by the same cell-state transition rule. A cell calculates its next state depending on states of its closest neighbors, known as a "cell neighborhood".

Cellular automaton can be defined by the tuple $\langle \mathbf{L}, \mathbf{Q}, u, f \rangle$, where \mathbf{L} is a lattice, or an array, of cells; each cell $x \in \mathbf{L}$ takes a state from the finite set \mathbf{Q} ; u is a cell neighborhood $u(x) = \{y \in \mathbf{L} : |x - y| \leq r\}$ (r is a radius, and $k = |u(x)|$ is the size of the neighborhood); $f: \mathbf{Q}^k \rightarrow \mathbf{Q}$ is cell-state transition function. Let x^t be the state of cell x at time step t and $u(x)^t = (y_1^t, \dots, y_k^t)$, the next state x^{t+1} of cell x is calculated as $x^{t+1} = f(u(x)^t) = f(y_1^t, \dots, y_k^t)$. The configuration, or global configuration, of cellular automaton is an array of the cells' states.

The basic idea of identifying *deterministic* cellular automaton is as follows. Given a sequence of cellular automaton configurations $C_1 \rightarrow \dots C_l \rightarrow C_{l+1} \rightarrow C_m$, we want to reconstruct a minimal and correct description of the cell-state transition function f . At a low level, the function f can be represented as a set, or table, of local transitions $w \rightarrow a$, where $w = \langle w_1, \dots, w_k \rangle$ is the string of states of cell x 's neighbors and $a = f(w)$. Therefore to reconstruct function f one has to select some minimal radius for the cell neighborhood, scan the given sequence $C_1 \rightarrow \dots C_l \rightarrow C_{l+1} \rightarrow C_m$ of cellular automaton configurations, and collect all observable transitions in the form $w \rightarrow a$. Then the consistency of the set of local transitions is checked. If there are two local transitions $w \rightarrow a$ and $w \rightarrow b$, where $w \in \mathbf{Q}^k$ and $a, b \in \mathbf{Q}$, with the same left part w and different right parts, $a \neq b$, then we consider that the calculated candidate for f violates the determinism of the identified cellular automaton. This may mean that we have chosen too small of a neighborhood which does not include all neighbors influencing the cell's state transition. So, the radius of the neighborhood is increased, the sequence of the given global configurations is re-scanned, and the set of local transitions is rebuilt. The procedure is carried out until a complete set of non-contradicting transitions is calculated.

Example 1 (Identification of a one-dimensional cellular automaton) Given two configurations of a one-dimensional deterministic cellular automaton of ten cells with periodic boundary conditions, $c^t = 0011101000$ and $c^{t+1} = 0110101100$, we want to reconstruct the cell-state transition function. We start identification with a minimal neighborhood $u(x_i) = (x_{i-1}, x_{i+1})$, where $i = 1, \dots, 10$ and collect all observable transitions $(x_{i-1}^t, x_{i+1}^t) \rightarrow x_i^{t+1}$:

00 \rightarrow 0
 01 \rightarrow 1
 11 \rightarrow 0
 10 \rightarrow 1
 00 \rightarrow 1.

Transitions $00 \rightarrow 0$ and $00 \rightarrow 1$ contradict each another. This means that the chosen neighborhood is insufficient in describing the conditions of local transitions. Let us include the central cell in the neighborhood: $u(x_i) = (x_{i-1}, x_i, x_{i+1})$ and collect local transitions in the form $(x_{i-1}^t, x_i^t, x_{i+1}^t) \rightarrow x_i^{t+1}$:

000 \rightarrow 0
 001 \rightarrow 1
 010 \rightarrow 1
 011 \rightarrow 1
 100 \rightarrow 1
 101 \rightarrow 0
 110 \rightarrow 1
 111 \rightarrow 0.

Now we do not have contradicting transitions, so we assume that we reconstructed the minimal cellular automaton which implements the global transformation $c' \rightarrow c''$. The automaton is identified completely. Having the cell-state transition rules, the behavior of the automaton can be investigated further, e. g. global behavior can be studied in terms of basin attraction fields (see Fig. 1).

Identification is assumed to be complete if state transitions for all possible states of the neighborhood are found. The identification is incomplete, or situational, otherwise. More algorithms and examples of identification of deterministic cellular automata, automata with memory, structurally dynamic cellular automata, and asynchronous automata are provided in [10].

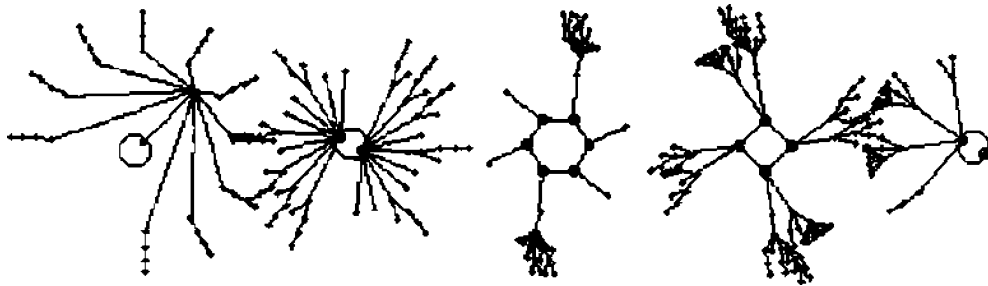
Classical identification of d -dimensional cellular automata implies explicit construction of a cell-state transition, or look-up, table. Let each cell of an identified cellular automaton have q states, and each cell updates its states in discrete time depending on the states of its im-

mediate neighbors in a neighborhood of radius r . There are q^k possible local states, or configurations of cell neighborhood, where $k = (2r + 1)^d$; so, the complete look-up table has q^k lines, each of length $k + 1$. To find a minimal possible model we start by identifying the automaton with a neighborhood radius of $r=1$ and gradually increase it during identification. This usually leads to an enormous computational effort for the identification, e. g., time complexity $O(\sqrt{k} q^k)$, which becomes an unrealistic number even for small numbers of cell-states and quite a modest size of the cell neighborhood. In the rest of the article we consider several techniques aimed to reduce complexity of identification.

Identification Using Machine Learning

Tools of evolutionary computing and genetic programming have already been used extensively to evolve cellular automata aimed to perform certain range of tasks, see e. g. [19,20,22,25,29,30]. However, in these works the authors used some global measure, e. g. the density of cells at some specified state (density classification task), to select the most successful rules. In the present section we will discuss how cell-state transition rules can be determined automatically from a given sequence of cellular automaton configurations using machine learning techniques.

We consider the machine learning approach for identification of binary cellular automata. It employs accuracy-based learning classifier system designed by Bull [16]. This is a memory-less system in which the rulebase consists of N action rules, and the condition is a string of characters from the ternary alphabet $\{0,1,\#\}$, where 0 and 1 are cell states, and # is a value of undefined cell-state transition. The action is also represented by a binary string. With each rule we associate a predicted payoff value p , a scalar error ε , and an estimate σ of the average size of action sets in which that rule participates. When the input message is re-



Identification of Cellular Automata, Figure 1

Global transition graphs of a one-dimensional cellular automaton, identified from the snapshot $0011101000 \rightarrow 0110101100$. Nodes correspond to the configurations, edges to the global transitions (produced by DDLab, www.ddlab.com)

ceived by the learning system, its rulebase is scanned, and the rules with conditions matching the input message at each position are marked as members of the current match set M . The action is selected from actions proposed by elements M . The rules proposing the selected action form the action set A .

Bull [16] uses immediate reward and reinforcement via updating the error, the niche size estimate, and the payoff estimate of each element of the current action set A using the Widrow-Hoff delta rule with learning rate β . Offsprings are produced via mutation and single-point crossover. Existing members of the rulebase are replaced randomly based on the estimated size of action set.

Let us consider an example of the identification of the one-dimensional cellular automaton with a three-cell neighborhood and two cell states, see details in [18]. Cells of the automaton update their states by the following rule: $\langle \text{State of the left neighbor, state of the central cell, state of the right neighbor} \rangle \rightarrow \text{new state of the central cell}$.

000 \rightarrow 1
 001 \rightarrow 0
 010 \rightarrow 1
 011 \rightarrow 1
 100 \rightarrow 0
 101 \rightarrow 1
 110 \rightarrow 1
 111 \rightarrow 0

Examples of space-time snapshots of the automaton developing from random initial configurations are shown in Fig. 2.

At the beginning of identification, the learning classifier system has no information about the size of the cell neighborhood or the cell-state transition rules. The system represents automaton as a random automaton, where every cell takes its next state with probability 0.5, irrelevant to the state of the cell's neighborhood, see an example of such development in Fig. 3a.

After 2000 trials the learning classifier system extracted the rules (we also show prediction p , error ε and action set size σ estimates):

	p	ε	σ
110 \rightarrow 1	: 1000.00000	: 0.00000	: 53.00009
#00 \rightarrow 1	: 555.93876	: 527.68224	: 66.33652
01# \rightarrow 1	: 000.00000	: 0.00000	: 76.05551
##1 \rightarrow 0	: 32.45343	: 456.82484	: 50.10989

Quite an accurate generalization of $01\# \rightarrow 1$ has already been identified, i. e., the fact that the right neighbor state is redundant for states 010 and 011 has been learned ('#' indicates a wildcard). The accurate rule 110 has also been found. The two other general rules mean the learning classifier system put the wrong states for two of the remaining five possible states. Thus the partially incorrect rule table is as shown below (contradicting transitions are underlined):

000 \rightarrow 1 (rule#00 \rightarrow 1)
 001 \rightarrow 0 (rule##1 \rightarrow 0)
 010 \rightarrow 1 (rule01# \rightarrow 1)
 011 \rightarrow 1 (rule01# \rightarrow 1)
100 \rightarrow 1 (rule#00 \rightarrow 1)
101 \rightarrow 0 (rule##1 \rightarrow 0)
 110 \rightarrow 1 (rule110 \rightarrow 1)
 111 \rightarrow 0 (rule##1 \rightarrow 0)

Development of this partially identified cellular automaton is shown in Fig. 3b.

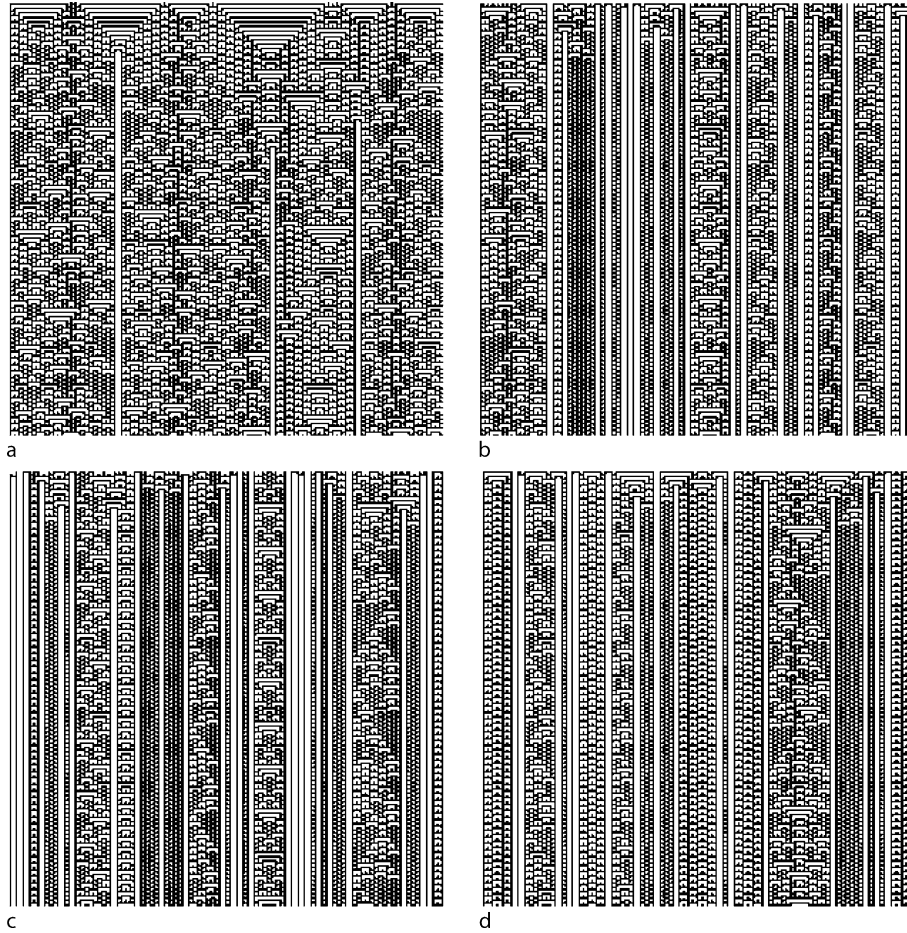
After ten times more steps of learning and identification, on the 20 000th trial, more accurate cell-state transition rules are discovered:

001 \rightarrow 0	: 1000.00000	: 0.00000	: 19.73689
100 \rightarrow 0	: 1000.00000	: 0.00000	: 23.87451
110 \rightarrow 1	: 1000.00000	: 0.00000	: 25.97547
0#0 \rightarrow 1	: 1000.00000	: 0.00000	: 22.78546
01# \rightarrow 1	: 1000.00000	: 0.00000	: 21.89581
##1 \rightarrow 0	: 526.06540	: 495.97291	: 36.40984

The table includes one more possible generalization: $0\#0 \rightarrow 1$. The learning classifier system got the wrong rule $101 \rightarrow 1$ because of the persistence of the rule $\#1 \rightarrow 1$. At this stage of identification, the cell-state transition rule table looks as follows:

000 \rightarrow 1 (rule0#0 \rightarrow 1)
 001 \rightarrow 0 (rule##1 \rightarrow 0)
 010 \rightarrow 1 (rule01# \rightarrow 1)
 011 \rightarrow 1 (rule01# \rightarrow 1)
 100 \rightarrow 0 (rule100 \rightarrow 0)
101 \rightarrow 0 (rule##1 \rightarrow 0)
 110 \rightarrow 1 (rule110 \rightarrow 1)
 111 \rightarrow 0 (rule##1 \rightarrow 0)

The development of a cellular automaton governed by such transition table is shown in Fig. 3c.



Identification of Cellular Automata, Figure 2

Example space-time configurations of a one-dimensional cellular automaton with 200 cells. Initially each cell is assigned the state 1 with probability a 0.01, b 0.3, c 0.6, and d 0.9, and state 0 with probabilities, 0.99, 0.7, 0.4 and 0.1, respectively. Time goes down. A black pixel represents state 1, and a blank pixel represents state 0 [18]

The minimum radius and complete table of cell-state transitions are computed by the learning classifier system in 100,000 steps as follows:

001 → 0	: 1000.00000	: 0.00000	: 18.78942
100 → 0	: 1000.00000	: 0.00000	: 17.33453
101 → 1	: 1000.00000	: 0.00000	: 18.21793
110 → 1	: 1000.00000	: 0.00000	: 17.65666
111 → 0	: 1000.00000	: 0.00000	: 17.89178
0#0 → 1	: 1000.00000	: 0.00000	: 20.78546
01# → 1	: 1000.00000	: 0.00000	: 21.89581

The appropriate generalization is evolved:

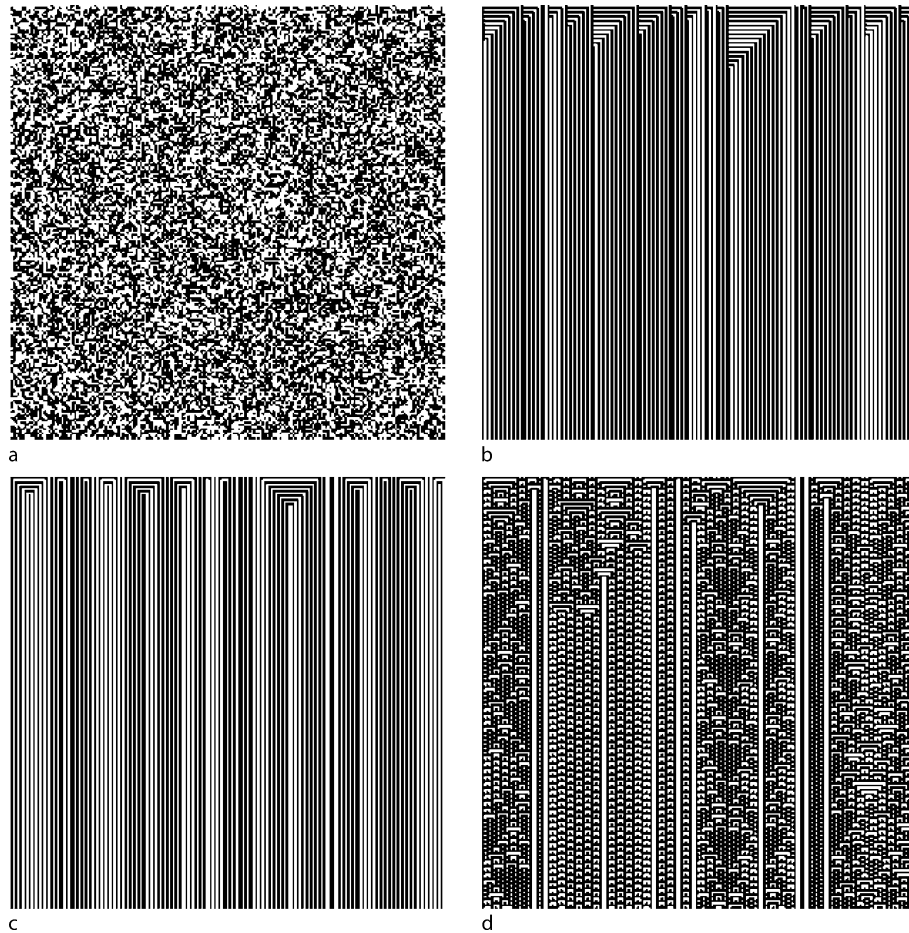
000 → 1	(rule0#0 → 1)
001 → 0	(rule001 → 0)
010 → 1	(rule01# → 1)

011 → 1	(rule01# → 1)
100 → 0	(rule100 → 0)
101 → 1	(rule101 → 1)
110 → 1	(rule110 → 1)
111 → 0	(rule111 → 0)

These rules give an identical space-time evolution from the same random initial configuration as the underlying cellular automaton shown in Figure 3d. More details regarding the performance and behavior of Bull's learning classified system used in the identification of cellular automata can be found in [18].

Identification Using Polynomial Representation

In developing their approaches to identify cellular automata, Billings and colleagues [43,44,45,46,47] used poly-



Identification of Cellular Automata, Figure 3

Examples of space-time configurations of cellular automaton (CA). **a** cells which update their states by the rules extracted after the 1st trial, **b** after the 2000th trial, **c** after the 20 000th trial, and **d** after the 100 000th trial. Initially, each cell is assigned the state 1 with probability 0.5, and is assigned the state 0 otherwise. For all rules, the CA started its development in the same random configuration. Time goes down. A black pixel represents state 1, and a blank pixel represents state 0 [18]

nomial representations of local transition rules. It is a well know and explored fact (see [40,41] for overview and further details) that cell-state transition rules of binary cellular automata can be represented as an expression from Boolean algebra, with conjunction and exclusive disjunction operations. The elements of the cell neighborhood are variables in such a logical expression. The logical expression in turn can be represented by an arithmetical polynomial with operations of multiplication and binary addition.

Yang and Billing [44] are applying orthogonalisation to derive an algorithm for detecting significant terms and estimate coefficients of the polynomial representation of local transition rules. Significant terms indicate which elements of a cell neighborhood should be necessarily taken

into consideration when extracting local transitions. They are using error reduction ratios to evaluate how much each element of the cell neighborhood contributes to updating the state of the cell. The methods of selecting the correct neighborhood can also be based on statistics associated with neighborhood elements and mutual information [46].

The polynomial-representation-based-identification seems to also work for stochastic cellular automata. Thus, Yang and Billing [43] demonstrate that binary-state stochastic automata are well modeled by an integer-parametrized polynomial disturbed by noise. To detect the minimal and sufficient neighborhood one should select correct terms of the model.

Binary Tree Representations and Genetic Programming

El Yacoubi and Jacewicz [20,22] use genetic algorithms to discover and design cell-state transition functions. They are fulfilling goal-based identification tasks, where cellular automata are not extracted from a set of given configuration but rather designed from scratch to satisfy certain global conditions in their space-time evolution. We will discuss the El Yacoubi-Jacewicz approach briefly because similar ideas are explored in more depth by Maeda and Sakama [27,28], as shown in the next section.

When binary state cellular automata are concerned, a cell-state transition function is represented as a binary tree (Fig. 4), where each node corresponds to some binary logic operator, and the leaves represent cells, or elements, of the neighborhood. Genetic algorithms can then be used to evolve 'optimal' trees, which in turn give us best version of cell-state transition function, where the fitness criterion is a measure of how good the desired global dynamics can be represented by the cell-state transition function [20,22].

The local transition functions are developed in the following manner. We select the initial size of the tree, the crossover and mutation probabilities, the number of iterations, the population size, and the fitness values. Then a Boolean function is selected, and a terminal set of accepted neighborhoods, or neighborhood sizes. The random population of cell-state transition functions creates a set of configurations. For each configuration a fitness value is calculated (as some global characteristic, e.g. number of cells in certain state). Based on the fitness values, the local transition functions are selected and subject to crossover and mutation operations.

Using their approach, El Yacoubi and Jacewicz [20,22] discovered the cell-state transition rule for the uniformity

problem, the synchronization problem, and the density classification problem.

Identification Using Decision Trees

Maeda and Sakama [27,28] employ decision trees and genetic programming to identify cellular automata. Their approach is basically very similar to the identification algorithms designed by Adamatzky [10], with some improvements in computational complexity. Given a sequence of automaton configurations, local transitions are collected as evidences, which are then classified as a decision tree. A cell-state transition table is derived from the decision tree using genetic programming.

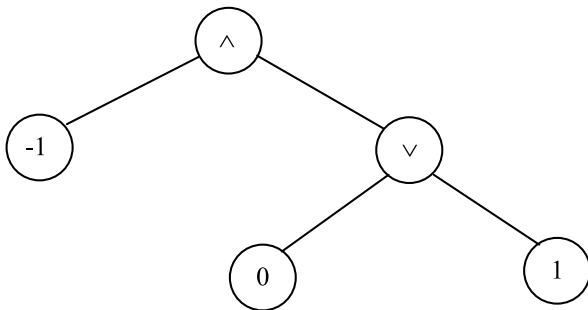
An evidence is a pair: the state of the neighborhood, and the next state of the neighborhood's central cell. The first element of each evidence is selected from a cellular automaton configuration at a time step t and, the second element at time step $t + 1$. Evidences are collected in a set. Identification starts with some minimal neighborhood. The radius of the neighborhood is increased if there are contradicting evidences, i.e. two evidences with similar first elements and different second elements. Redundant cells are removed from the final neighborhood by the standard procedure suggested in [10].

What are the classification conditions in Maeda-Sakama's identification techniques? Evidences are classified by their neighborhood patterns [27,28]. The explicit, spatial, state of the cell neighborhood is converted to if-then rules involving the enumeration of neighbors and the direct indication of their states; e.g., "If the South-West and North-East neighbors have state 1, then the central cell will take state 1". These are the classification conditions. Classification conditions are represented by condition trees, as in El Yacoubi-Jacewicz [20,22].

Genetic algorithms are applied to condition trees to find classification conditions which correctly classify all evidences. Several condition trees can be connected in one, so a condition tree is extended until it represents a classification condition which classifies all evidences.

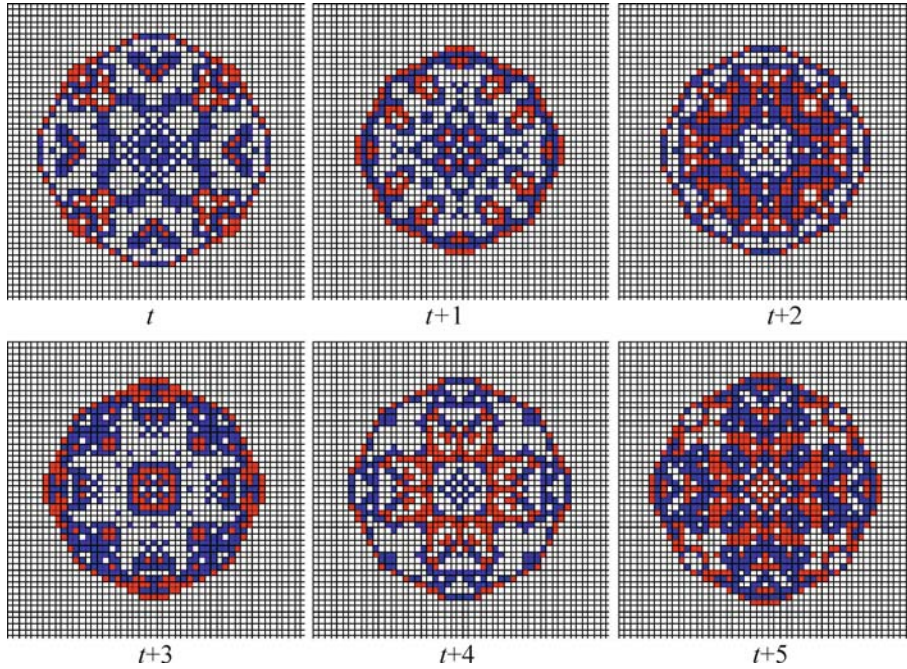
Each node of the decision tree has a certain condition value which is computed from the condition of a classification condition of the previous node and the next state of the cell based on the classification condition. Given a neighborhood state of a cell, a decision tree returns next state of the cell.

Let us consider the following example as it is discussed by Maeda and Sakama [28]. Given a sequence of configurations (Fig. 5) for a two-dimensional cellular automaton with three cell-states and a five-cell von Neumann neighborhood, we want to extract the rules of cell-state



Identification of Cellular Automata, Figure 4

Binary tree representation of the cell-state transition function:
 $x_i^{t+1} = x_{i-1}^t \wedge (x_i^t \vee x_{i-1}^t)$



Identification of Cellular Automata, Figure 5

Given snapshots of two-dimensional cellular automaton to identify

transitions. Let us mark cells in the neighborhood according to their geographical position relative to the neighborhood's central cell: South (S), North (N), West (W), East (E) and Central (C) cells. The condition and decision trees obtained from the sequence of configurations in Fig. 5 are primitive. The condition tree T has one root, an addition operation, and six leaves: S, N, W, E, C, C (the central cell is used twice). The decision tree has one root and eight leaves, expressed by the following if-then rules (where R is the outcome of a condition in a decision tree):

If	$T = 0$	then	$R = 0$
If	$T = 1$	then	$R = 0$
If	$T = 2$	then	$R = 1$
If	$T = 3$	then	$R = 1$
If	$T = 4$	then	$R = 2$
If	$T = 5$	then	$R = 2$
If	$T = 6$	then	$R = 2$
If	$T = k$	then	$R = 0$,

where $7 \leq k \leq 12$.

This can be minimized to the following set of rules. $P(x, t)$ is the arithmetical sum of the cell-states in a neighborhood of the cell x at the time step t , and x^{t+1} is the state

of the cell x at the time step $t + 1$. Then

if	$P(x, t) \leq 1$	then	$x^{t+1} = 0$
else if	$P(x, t) \in \{2, 3\}$	then	$x^{t+1} = 1$
else if	$P(x, t) \in \{4, 5, 6\}$	then	$x^{t+1} = 2$
if	$P(x, t) \geq 7$	then	$x^{t+1} = 0$.

Identification by Immunocomputing

Tarakanov and Prokaev [34] represent cell-state transitions of cellular automaton by an artificial immune network. The number of possible entries in the state transition table is reduced by using apoptosis (programmed cell death, or cell suicide) and immunization [35].

A set of parameters, or state vector, is defined for each cell of the identified automaton: $X = [x_1 \dots x_n]'$. The vector is comprised of a sequence of cell states (extracted from series of given snapshots) $c^t, c^{t-1}, \dots, c^{t-p}$ and states of the cell's neighbors, along an identified period: $u^t, u^{t-1}, \dots, u^{t-p}$. Vector X does not have to be defined completely.

To identify a cellular automata, Tarakanov and colleagues [34,35] employ their techniques for pattern recognition by immune-computing. This is simulated by molecular recognition, as a computation of binding energy between an antigen, n -dimensional input vector, and anti-

bodies, singular vectors of the single value decomposition of a training matrix [35]. The procedure is built of two parts: two stages of training, and recognition.

At the first stage of training, data are mapped to a formal immune network: training patterns are acquired, a training matrix is formed, and a single vector decomposition (binding energies) of the training matrix is calculated. Single vectors obtained are analogs of antibody-probes. Then the obtained data are compressed by apoptosis – if there are two neighboring cells with the same value just one cell is kept alive, and immunization – if the ‘killed’ cell has no neighbors with the same value as the cell has then the cell is restored. Immunization is used to correct mistakes accumulated during apoptosis.

At the second part, recognition is implemented. The cellular automaton pattern (analog of antigen) is mapped into the formal immune network. Then the nearest cell, corresponding to the pattern of the network, is detected, and the class, or local transition function of identified cellular automaton, of the nearest cell is assigned to the pattern [34,35].

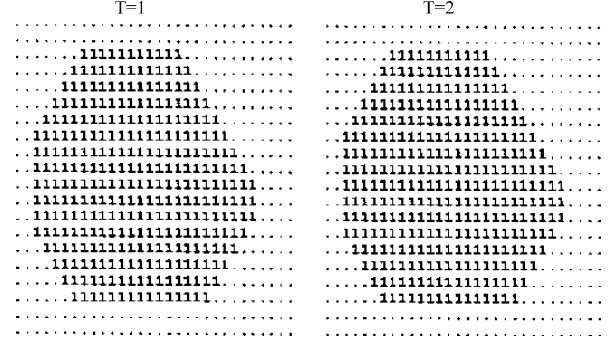
Application of Identification in Automatic Design of Cellular-Automata Processors

Identification strategies can be used in developing automatic programming techniques for massively parallel cellular automaton processors [11]. Let us consider two examples of the programming of two-dimensional semi-totalistic cellular automata, with a binary set of cell states and an eight-cell neighborhood.

Example 1 (Computation of a discrete convex hull)
A 45-convex hull of the subset S of nodes of an integer lattice is an intersection of all discrete half-planes containing S [21,33]. A discrete 45-halfplane is the set of all such nodes (i, j) of an integer lattice that satisfy the inequality $a \cdot i + b \cdot j \leq c$ for $|a|, |b| \leq 1$ and integer c .

Given a connected set of black pixels (a cellular automaton configuration), we want to design a cellular-automaton processor which computes the discrete convex hull of this set. The given set must be mapped on to the lattice in such a manner that cells having the same indices as the nodes of the given set take state 1, whereas others remain in rest state 0. The domain grows until it reaches the shape of the convex hull. The state 1 is an absorbing state because every cell initially “excited” with state 1 remains in the state 1 forever.

In discovering the cell-state transition function, we will employ the fact that convex hull must be a stationary configuration of a cellular automaton. Therefore we take two identical configurations as shown in Fig. 6. From the tran-



Identification of Cellular Automata, Figure 6

Source configuration ($T = 1$) and target configuration ($T = 2$) are the input for the tool of automatic programming of a cellular automaton that builds a discrete convex hull

sition ‘given global transition source configuration’ \rightarrow target configuration we extracted conditions of the cell-state transition $0 \rightarrow 0$. It was found that a cell in state 0 remains in state 0 if it has 0, 1, 2, or 3. This means that a cell in state 0 takes state 1 if more than three of its neighbors are in state 1.

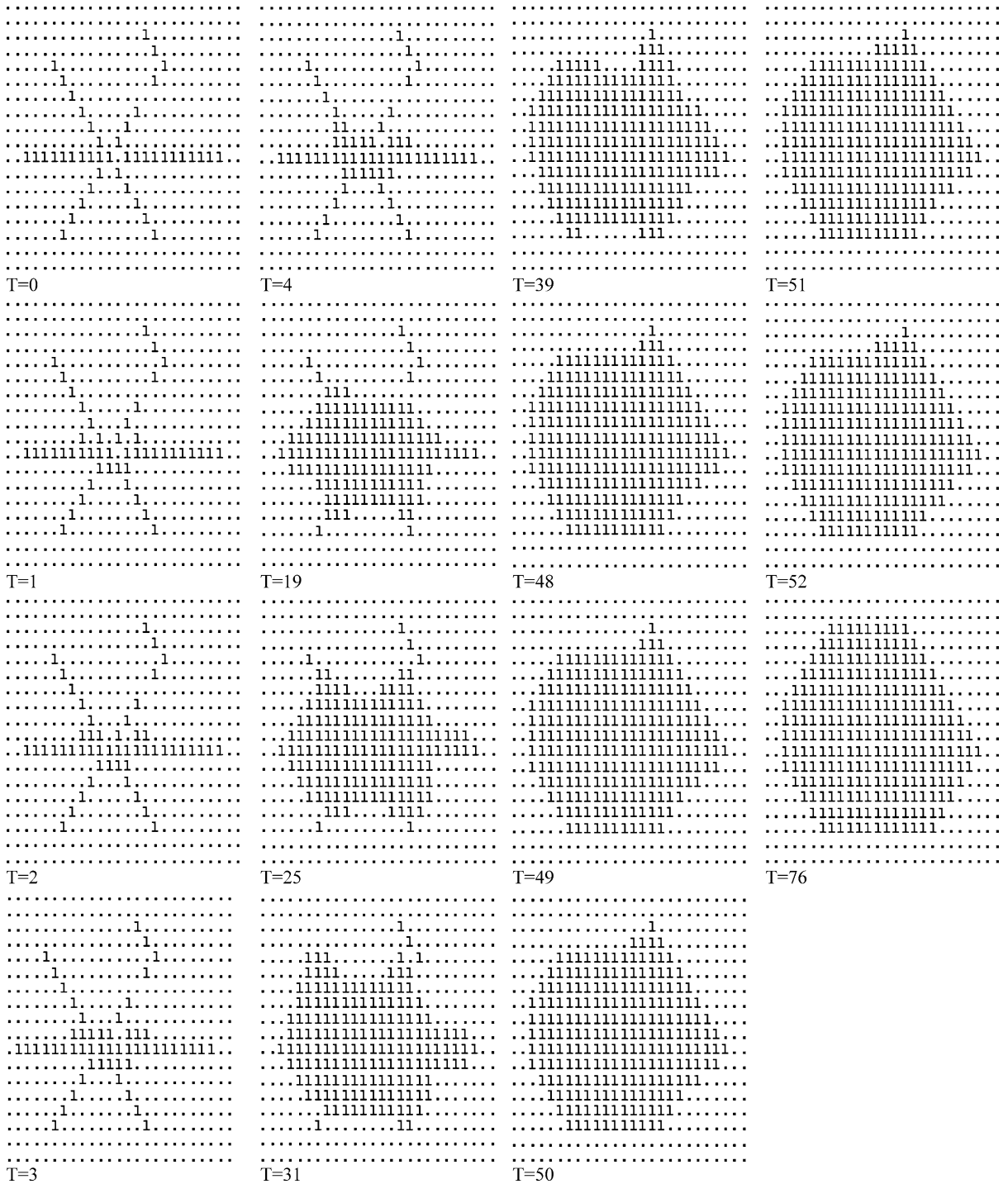
The resultant local transition rule is as follows:

$$x^{t+1} = \begin{cases} 1, & (x^t = 1) \text{ or} \\ & ((x^t = 0) \text{ and } (\sum_{y \in u(x)} y^t > 3)) \\ 0, & \text{otherwise,} \end{cases}$$

where x^t is the state of cell x in time step t , and $u(x)$ is a neighborhood of the cell x . The evolution of a designed cellular automaton is presented in Fig. 7.

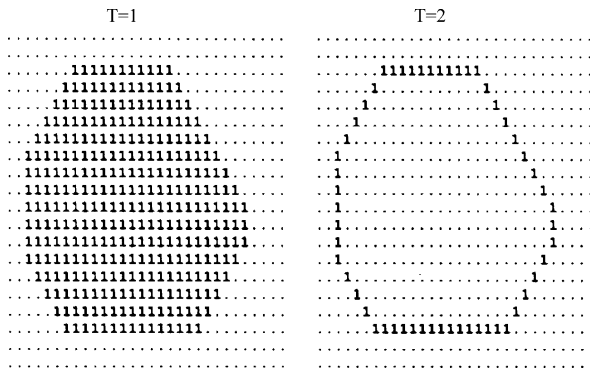
Example 2 (Detecting the contour of a convex image)
Given a convex solid image filled entirely with black pixels, we want to design a cellular automaton which computes the contour of the image. We represent a given image and its contour on the integer lattice as two separate configurations of a cellular automaton. The image is a source configuration, and the contour is a target configuration (Fig. 8).

Taking for granted a minimal eight-cell neighborhood, all observable states of the neighborhood for the transitions $1 \rightarrow 1$, $1 \rightarrow 0$ and $0 \rightarrow 0$ are collected. The transition $0 \rightarrow 0$ is unconditional. It takes place independently of the current state of the neighborhood. The two other transitions are conditional. The black pixel (cell state 1) remains black if four or five of its neighbors are black pixels; otherwise it becomes white (cell state 0). A two-dimensional, semi-totalistic cellular automaton was designed, which has an eight-cell neighborhood, two cell states, and



Identification of Cellular Automata, Figure 7

The computation of a convex hull in a cellular automaton identified from the two given configurations at Fig. 6. The discrete data set is represented in the initial configuration of the automaton ($T = 0$), and the completed convex hull in the configuration at time step $T = 76$



Identification of Cellular Automata, Figure 8

The source configuration ($T = 1$) and the target configuration ($T = 2$) are the input for the automatic programming of a cellular automaton that computes counter of convex image

the following function of the local transitions:

$$x^{t+1} = \begin{cases} 1, & (x^t = 1) \text{ and } (4 \leq \sum_{y \in u(x)} y^t \leq 5) \\ 0, & (x^t = 0) \text{ or } (x^t = 1) \text{ and} \\ & (\sum_{y \in u(x)} y^t \in [0..3] \cup [6..8]), \end{cases}$$

where x^t is the state of cell x in time step t , and $u(x)$ is a neighborhood of cell x .

Future Directions

Theoretical studies are only powerful when they can be applied in real life. Identification of real-world spatially extended systems proved to be a non-trivial task. In computational models, the sequence of global transformations presented for identification are always discrete, noise-free, and can be easily handled by identification software systems. Nevertheless, some progress has been made.

In 1996 Adamatzky and Tolmachev [13,36] managed to identify local transitions rules of a cellular-automaton that constructs the Voronoi diagram, the tessellation of a plane from given planar set. The cell-state transition rules discovered helped to locate a proper set of chemical species, suitable for the task, and fabricate a chemical laboratory processor which approximates the Voronoi diagram [13,36].

In 2003, Bull et al. [17] proposed to undertake an experimental identification of real-world reaction-diffusion chemical systems, including the Belousov-Zhabotinsky excitable chemical medium, the gel mediated chlorite-iodide-malonic acid reaction, the inorganic gel mediated pattern forming reactions based on the reaction of copper chloride and potassium ferrocyanide, and many other systems. The ideas were independently developed by Zhao et al. [45,46,47], who approximated the local transition

function of a minimal cellular automaton from snapshots of an experimental Belousov-Zhabotinsky medium [46].

Tarakanov and Prokaev [34] demonstrated how to employ identification techniques to build a cellular-automaton model of a temperature field generation using real field data from Barents Sea.

We expect that identification techniques will be applied in reconstructing local transition rules of natural spatially extended systems: interacting populations, morphogenesis, reaction-diffusion systems, homogeneous neural networks, as well in design of more efficient massively parallel processors with cellular automaton architecture.

Bibliography

Primary Literature

- Adamatzky A (1990) Identification of probabilistic cellular automata. *Izv AN SSSR Ser Tekhnicheskaya Kibernet* 3:95–100 (in Russian). Translated in Soviet (1990) *J Comput Syst Sci* 30:118–123
- Adamatzky A (1991) Identification of fuzzy cellular automata. *Autom Comput* 6:75–80
- Adamatzky A (1992) Complexity of cellular automata identification. *Avtom Telemekh* 9:72–86 (in Russian). Translated (1992) Complexity of identification of cellular automata. *Autom Remote Control* 53 9/2:1449–1458
- Adamatzky A (1992) Identification of nonstationary cellular automata. *J Comp Sci Tech* 7:379–382
- Adamatzky A (1992) On complexity of identification of nonstationary cellular automata. *Izv AN SSSR Ser Tekhnicheskaya Kibernet* 3:74–79 (in Russian)
- Adamatzky A (1993) Implantation of cellular automata. *Appl Math Comput* 55:49–71
- Adamatzky A (1993) Identification of distributed intelligence. *Izv AN SSSR Ser Tekhnicheskaya Kibernet* 6:359–369 (in Russian). Translated (1995) Recognition of distributed intelligence. *J Comp Syst Sci Int* 33:160–169
- Adamatzky A (1994) Hierarchy of fuzzy cellular automata. *J Fuzzy Sets Syst* 62:167–174
- Adamatzky A (1994) On complexity of serial simulation of cellular-automata mappings. *Avtom Telemekh* 3 (in Russian). Translated (1994) Complexity of sequential realisation of cellular-automata maps. *Autom Remote Control* 55 2/2:271–280
- Adamatzky A (1994) Identification of cellular automata. Taylor and Francis, London
- Adamatzky A (1997) Automatic programming of cellular automata: Identification approach. *Kybernetes* 26:126–135
- Adamatzky A, Bronnikov V (1989) Identification of additive cellular automata. *Izv AN SSSR Ser Tekhnicheskaya Kibernet* 3:200–205 (in Russian). Translated in Soviet (1990) *J Comput Syst Sci* 28:47–51
- Adamatzky A, Tolmachev D (1997) Chemical processor for computation of skeleton of planar shape. *Adv Mater Opt Electron* 7:535–539
- Andre D, Bennet FH III, Koza JR (1996) Discovery by genetic programming of a cellular automata rule that is better than any known rule for the majority classification problem. In: Koza

- JR (ed) Genetic programming 1996. Proceedings of the 1st Annual Conference. MIT Press, Cambridge
15. Brauer W (1984) Automaton-theorie. Teubner BG, Stuttgart
 16. Bull L (2005) Two simple learning classifier systems. In: Bull L, Kovacs T (eds) Foundations of learning classifier systems. Springer, New York, pp 63–90
 17. Bull L, Adamatzky A, De Lacy Costello B (2003) The automatic identification of spatially extended reaction-diffusion systems. EPSRC Proposal GR/S68798/01, University of the West of England
 18. Bull L, Adamatzky A (2007) A learning classifier system approach to the identification of cellular automata. *J Cell Autom* (in press)
 19. Das R, Crutchfield J, Mitchell M, Hanson J (1995) Evolving globally synchronized cellular automata. In: Eshelman L (ed) Proc 6th Int Conf on Genetic Algorithms. Morgan Kaufmann, New York, pp 336–343
 20. El Yacoubi S, Jacewicz P (2007) A genetic programming approach to structural identification of cellular automata. *J Cell Autom* (in press)
 21. Heijmans HJAM (1990) Iteration of morphological transformations. *CWI Q* 9:19–36
 22. Jacewicz P, El Yacoubi S (1999) A genetic programming approach to structural identification of cellular automata. In: Proc of 3th International Conference on Parallel Processing and Applied Mathematics, Kazimierz Dolny, Poland, 1999, pp 148–157
 23. Kleene SC (1956) Representation of events in nerve nets. In: Shannon CE, McCarthy J (eds) Automata Studies. Princeton University Press, Princeton pp 3–41
 24. Koza JR (1994) Genetic programming II: Automatic discovery of reusable programs. MIT Press, Cambridge
 25. Koza J, Bennett III F, Andre D, Keane M (1999) Genetic programming III: Darwinian invention and problem solving. Morgan Kaufmann, New York
 26. Kudrjatzev VB, Podkolzin AS, Bolotov AA (1990) The Foundations of the theory of homogeneous structures. Nauka Publishers, Moscow
 27. Maeda K, Sakama C (2003) Discovery of cellular automata rules using cases. In: Proceedings of the 6th International Conference on Discovery Science. Lecture Notes in Artificial Intelligence, vol 2843. Springer, Heidelberg, p 357–364
 28. Maeda K, Sakama C (2007) Identifying cellular automata rules. *J Cell Autom* (in press)
 29. Mitchell M, Crutchfield J, Hraber P (1994) Evolving cellular automata to perform computations: Mechanisms and impediments. *Phys D* 75:361–391
 30. Mitchell M, Hraber PT, Crutchfield JP (1993) Revisiting the edge of chaos: Evolving cellular automata to perform computations. *Complex Syst* 7:89–130
 31. Moore EF (1956) Gedanken-experiments on sequential machines. In: Shannon CE, McCarthy J (eds) Automata studies. Princeton University Press, Princeton, pp 129–153
 32. Murphy KP (1996) Passively learning finite automata. Technical Report 96-04-017, Santa-Fe Institute, Santa Fe, CA
 33. Ronse C (1985) Definition of convexity and convex hulls in digital images. *Bull Soc Math Belg Ser B* 37:71–85
 34. Tarakanov A, Prokaev A (2007) Identification of cellular automata by immunocomputing. *J Cell Autom* (in print)
 35. Tarakanov A, Skormin V, Sokolova S (2003) Immunocomputing: Principles and applications. Springer, New York
 36. Tolmachiev D, Adamatzky A (1996) Chemical processor for computation of Voronoi diagram. *Adv Mater Opt Electron* 6:191–196
 37. Trakhtenbrot BA (1957) On operators, realizable in logical nets. DAN SSSR, Proceedings of the USSR Academy of Sciences 112:1005–1007 (in Russian)
 38. Trakhtenbrot BA, Bardzin YAM (1970) Finite automata (Behaviour and synthesis). Nauka Publishers, Moscow
 39. Von Neuman J (1990) Theory of self-reproducing automata. University of Illinois, Urbana
 40. Voorhees B (1996) Computational analysis of one-dimensional cellular automata. World Scientific, Singapore
 41. Wolfram S (1994) Cellular automata and complexity: Collected papers. Addison-Wesley, Reading
 42. Wuensche A, Lesser M (1992) The global dynamics of cellular automata. Addison Wesley, Reading
 43. Yang YX, Billings SA (2003) Identification of probabilistic cellular automata. *IEEE Trans Syst Man and Cybern Part B* 33:225–236
 44. Yang YX, Billings SA (2003) Identification of the neighbourhood and CA rules from spatio-temporal CA patterns. *IEEE Trans Syst Man Cybern Part B* 30:332–339
 45. Zhao Y, Billings SA (2006) Neighborhood detection using mutual information for the identification of cellular automata. *IEEE Trans Syst Man Cybern Part B* 36:473–479
 46. Zhao Y, Billings SA (2007) The identification of cellular automata. *J Cell Autom* (in press)
 47. Zhao Y, Billings SA, Routh A (2005) Identification of excitable media using cellular automata. *Int J Bifur Chaos* 17:153–168

Additional Reading

- Adamatzky A (2001) Computation in nonlinear media and automata collectives. IoP Publishing, Bristol
- Brauer W (1984) Automaton-theorie. Teubner, London
- Chopard B, Droz M (1998) Cellular automata modeling of physical systems. Cambridge University Press, Cambridge
- Crutchfield JP, Hanson JE (1999) Computational mechanics of cellular processes. Princeton University Press, Princeton
- Narendra K, Thathachar MAL (1989) Learning automata. Prentice-Hall, New Jersey
- Prokaev A, Sokolova L, Tarakanov A (2007) Using immunocomputing to forecast hydrophysical fields in the ocean. In: Int Workshop on Information Fusion and GIS (IF GIS 07), St. Petersburg, Russia (accepted for publication in LNCS)
- Sipper M (1997) Evolution of parallel cellular machines. In: The cellular programming approach, lecture notes in computer science, vol 1194. Springer, Berlin
- Tarakanov A, Adamatzky A (2002) Virtual clothing in hybrid cellular automata. *Kybernetes (Int J Syst Cybernetics)* 31:394–405
- Tarakanov A, Goncharova L, Tarakanov O (2005) A cytokine formal immune network. *Lecture Notes in Artificial Intelligence* 3630:510–519
- Tarakanov A, Prokaev A, Varnaskikh E (2007) Immunocomputing of hydroacoustic fields. *Int J Unconv Comput* (accepted for publication)
- Tarakanov A, Skormin V, Sokolova S (2003) Immunocomputing: Principles and applications. Springer, New York
- Toffoli T, Margolus N (1987) Cellular automata machines. A new environment for modeling. MIT Press, MIT Press Series in Scientific Computation, Cambridge
- Weimar J (1998) Simulation with cellular automata. Logos, Berlin

Image Based State Estimation

NICHOLAS GANS, GUOQIANG HU, WARREN E. DIXON
University of Florida, Gainesville, USA

Article Outline

Glossary
Definition of the Subject
Introduction
Pinhole Projection Model
Linear Estimation of Position and Orientation
from Two Images
Extensions to the Linear Pose Estimation Algorithms
Estimation of Linear and Angular Velocity
Image Based Kalman Filter Estimation
Range Identification
Future Directions
Bibliography

Glossary

Pose The position and orientation of an object is referred to as the pose. The pose has a translation component that is an element of \mathbb{R}^3 (i. e., Euclidean-space), and the rotation component is an element of the special orthogonal group $SO(3) \subset \mathbb{R}^{3 \times 3}$, though local mappings of rotation exist in \mathbb{R}^3 (e. g., Euler angles, angle/axis) or \mathbb{R}^4 (e. g., unit quaternions). Pose of an object has specific meaning when describing the relative position and orientation of one object to another or the position and orientation of an object at different time instances. Pose is typically used to describe the position and orientation of one reference frame with respect to another frame.

Camera-in-hand configuration The camera-in-hand configuration refers to the case when the camera is attached to a moving robotic system (e. g., held by the robot end-effector). In this case, camera pose is typically the state being estimated.

Camera-to-hand configuration The camera-to-hand configuration refers to the case when the camera is stationary and observing moving targets (e. g., a fixed camera observing a moving robot end-effector). In this case, pose of a viewed object is typically the state being estimated.

Feature point Different computer vision algorithms have been developed to search images for distinguishing features in an image (e. g., lines, points, corners, textures). The three-dimensional Euclidean-space representation of a point in the two-dimensional image

plane is called a feature point. Feature points are selected based on contrast with surrounding pixels and the ability for the point to be tracked from frame to frame. Sharp corners (e. g., the windows of a building), or center of a homogenous region in the image (e. g., the center of car headlights) are good examples of feature points.

Calibration A camera has a set of parameters collectively known as calibration parameters that describe the interaction of optics and imaging surface to produce an image. The camera calibration parameters are composed of the intrinsic parameters (i. e., image center, camera scale factors, and camera magnification factor) and extrinsic parameters (i. e., relative camera pose with respect to a fixed reference frame). The process of learning the parameters for a particular camera is known as calibration. Using the calibration parameters to recover geometric data from raw image data is also referred to as calibration.

Epipolar constraint Given two images of a three-dimensional scene taken at two different poses, the Epipolar constraint restricts the positions of two-dimensional image points in both images. If the relative pose between the cameras is known and a point is identified in one image, the Epipolar constraint restricts the locations of the feature points in the second image. Conversely, if points are matched in two images, the Epipolar constraint restricts the possible relative camera poses that could yield the two images and the relative three-dimensional coordinates of the points in the scene. It is possible to codify the Epipolar constraint as a 3×3 matrix, known as the Essential Matrix. Given a sufficient number of points in two images, the Essential Matrix can be determined and decomposed to solve for the translation and rotation between the two camera poses. The Epipolar constraint exists even when image points have not been calibrated. In which it is codified in the Fundamental Matrix. It is possible to codify the epipolar constraint as a 3×3 matrix, known as the Essential Matrix. Given a sufficient number of points in two images, the Essential Matrix can be determined and decomposed to solve for the translation and rotation between the two camera poses. The Epipolar constraint exists even when image points have not been calibrated, in which it is codified in the Fundamental Matrix.

Homography The geometric concept of homography is a one-to-one and on-to transformation or mapping between two sets of points. In computer vision, homography generally refers to the mapping between points in two Euclidean-planes (Euclidean homogra-

phy), or to the mapping between points in two images (projective homography).

Definition of the Subject

Image-based state estimation (IBSE) is the use of image data to approximate a set of variables that determine the current conditions of a system. Image data commonly refers to visual images captured by an optical camera (such as the development in this chapter); however, images can also be synthesized from other sensors, such as SOund Navigation And Ranging (SONAR), RAdio Detection And Ranging (RADAR), LAser Detection And Ranging (LADAR) or cross-section information from Computed Tomography (CT) or Magnetic Resonance Imaging (MRI). For robotic systems that employ state-estimation, the states that are observed typically include the kinematics of a physical object such as position, orientation, and linear and angular velocity in the Euclidean-space.

The task of determining the pose and/or velocity of a moving camera observing a static scene is described as the “camera-in-hand” problem since early IBSE for robotics was achieved by a manipulator gripping the camera. The opposite task where the camera is stationary observing a moving target is called the “camera-to-hand” to denote the scenario where the camera is observing the motion of the robotic system from some fixed location.

Vision is arguably the primary environmental sensor used by human beings and many other animals. As members of a visual society, people can easily relate to information provided from a camera. Identification and interpretation of data is simple to convey in both directions of the man-machine interaction. From an engineering perspective, cameras are passive sensors (i. e., do not emit energy for sensing), and thus undetectable unlike active sensors modalities such as; SONAR, RADAR, and lasers. Passive sensing is important for surveillance and security tasks and environments where energy emitting sensors are hazardous. IBSE methods are important because they provide a means to determine information from images such as the six-dimensional pose and velocity of some identified and tracked object. Relative pose information, proximity to a target, and velocity of a target/camera are examples of common feedback information that robots require for autonomous operation.

Introduction

The use of images to determine the pose of a camera with respect to a viewed object can be traced back to at least 1913 when Kruppa [1] proved that two camera views of five Euclidean points could be used to estimate

the translation and rotation separating the two camera views. However, the lack of sufficient computational resources has limited the growth of such IBSE until modern times. In recent literature, the eight-point algorithm was a landmark result introduced in [2] and [3] which uses the epipolar constraint (see Subsect. “[The Essential Matrix and the Eight-Point Algorithm](#)”) to solve for camera motion and/or scene structure from eight or more feature points. Algebraically, the algorithm is an implicit function of two images of a Euclidean point and the rotation and translation between the camera views. Geometrically, the epipolar constraint restricts the set of Euclidean coordinates of a feature point between two camera views separated by a finite rotation and translation. The development of the epipolar constraint gave rise to a large body of work in reconstruction of motion and structure from multiple camera views using linear algorithms. Specifically, advances in computer vision research later reduced the number of requisite non-coplanar feature points to as few as five feature points [4,5]; however, the five-points algorithm requires solving the roots of a tenth degree polynomial and results in additional valid solutions. The planar homography (see Subsect. “[The Planar Homography Algorithm](#)”) algorithm [6], which solves for motion and structure from multiple views of four or more feature points that lie in the same 3D plane, is also based on outcomes of the epipolar constraint. The linear algorithms in works such as [2,3,6] provide the geometric backbone of the IBSE methods presented in this chapter. These linear methods are attractive due to their relative simplicity and mathematical tractability. These methods are easily implemented as computer programs with limited resources such as embedded systems, and can operate much faster than thirty Hertz (considered “real time” for video) with modern desktop or laptop resources. The linear algorithms that relate multiple images also require no a priori information about the scene.

The previously described linear geometric relationships are widely used in literature and have been successfully demonstrated in a number of applications. Yet, these methods have some limitations. Like any linear technique, the approximation can give erroneous results in the presence of nonlinearities such as optical effects, normalization, and noise. Thus, in the 1990s some methods began to develop as a means to compensate for these effects using nonlinear methods. These results use linear techniques for an initial estimate of the relative geometry followed by an optimization step through gradient descent techniques (e. g., [7,8,9]), typically referred to as bundle adjustment. Accommodating for such disturbances in the geometric transformation is an important area of research, and the

reader should be aware of these methods. However, this chapter will not discuss them in detail.

Even under the assumption that the linear geometric relationships are sufficient, a significant issue that impacts IBSE is that to maintain an estimate of the motion of an object (or of the camera), the feature points must remain in the field-of-view (FOV). Motivated by this issue, researchers have investigated various methods to expand the FOV. Techniques to indefinitely relate multiple pose estimates from multiple, overlapping views, have been explored in [10,11]. Both of these papers are motivated by IBSE and control of unmanned aerial vehicles. Subsect. “[Long Term Pose Estimation Through Chained Homography Decomposition](#)” provides an example technique that can be used for camera pose estimation over large scenes where features leave the FOV. Cameras with curved lenses or curved mirrors can deliver a 360° (albeit distorted) view. Several researches have developed nonlinear observers for state estimation with such omnidirectional cameras (e. g., [12,13]).

Another challenge for IBSE is that the image-plane geometry only provides a scaled pose of a camera or object in the current image relative to a pose at reference goal image. An additional set of linear equations can deliver the depths of feature points up to an unknown scale factor. However, the unknown scaling can be a significant obstacle for autonomous robotic tasks. Subsect. “[Solving for Relative Object Poses Through Knowledge of a Single Length](#)” presents an example method that exploits some additional knowledge, such as a single known length on an object, to attach a reference frame to the object and deliver relative pose of the object with respect to the camera.

In the late 1980s, researchers began applying optimal filter techniques, namely the Kalman Filter (KF) and Extended Kalman Filter (EKF) for IBSE. Some approaches revolve around the Essential Matrix and use the EKF to refine the estimate of the Essential Matrix to determine the pose and/or velocity [14]. Other methods eschew external pose estimation schemes and use image features as inputs to the KF/EKF with pose and/or velocity as an output [15,16,17,18]. KF approaches to IBSE are discussed in Sect. “[Image Based Kalman Filter Estimation](#)”.

In the 1990s and 2000s nonlinear estimators/observers were investigated as IBSE methods. Similar to the KF developments, some methods use linear pose estimation techniques to estimate velocity or pose of an object [19], while others develop estimation approaches without using the linear methods [20,21,22,23,24]. In Subsect. “[Velocity Estimation Through a Nonlinear Observer](#)” a nonlinear estimator based on the planar Homography Matrix is provided as an example that estimates velocity of a mov-

ing object given geometric knowledge of the shape of the object. In Subsect. “[Nonlinear Estimator for Range Identification](#)”, an example of a nonlinear method to estimate the range (i. e., depth) to feature points is provided that exploits the additional information that the camera/object velocity is known. Given knowledge of the camera/target motion, the example Subsect. “[Nonlinear Estimator for Range Identification](#)” is developed without linear pose estimation methods.

This chapter serves as an introduction and brief survey to the field of IBSE. Several books have been written which can be recommended to provide in-depth treatment of these subjects. The bedrock methods based in epipolar geometry are covered in great detail in [25,26,27,28]. Further discussion on Kalman Filtering for IBSE is provided in [26].

Pinhole Projection Model

There are several approaches to model the projection of an object’s image through a lens and imaging surface. A simple, widely used method for accurately modeling a well focused imaging system is the pinhole camera model (i. e., a perspective projection). For the pinhole camera model, a Euclidean reference frame is attached to the camera at the center of projection. The imaging surface is a plane perpendicular to the z -axis of the camera frame, and located a distance f from the center of projection. The distance f is the focal length, and is a physical characteristic of the camera and lens. The pinhole camera model is illustrated in Fig. 1.

The coordinates of a feature point in the Euclidean space with respect to the camera frame are denoted as

$$\tilde{m}(t) = [x(t), y(t), z(t)]^T.$$

The point $\tilde{m}(t)$ projects to a point $m(t)$ on the image plane, with normalized coordinates

$$m(t) = \left[\frac{x(t)}{z(t)}, \frac{y(t)}{z(t)}, 1 \right]^T$$

in the camera frame. This projection can be described as a nonlinear projection function $\Pi(\cdot)$ as

$$m = \Pi(\tilde{m}).$$

In a digital camera, the normalized coordinates of points are measured in pixel coordinates as

$$p(t) = [u(t), v(t), 1]^T,$$

where $u(t)$ and $v(t)$ are integers that are typically measured from a corner of the image. The relationship between the

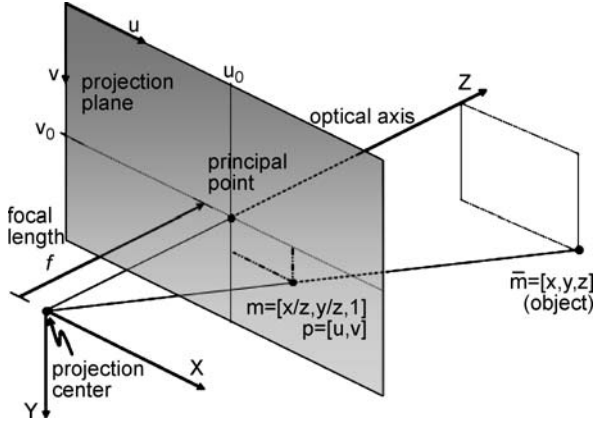


Image Based State Estimation, Figure 1
Pinhole projection model

j th image-based feature point and the corresponding Euclidean point is given as

$$p_j(t) = A m_j(t), \quad (1)$$

where $A \in \mathbb{R}^{3 \times 3}$ is a constant, invertible, upper-triangular camera calibration matrix given by [28]

$$A = \begin{bmatrix} f\sigma_x & f\sigma_x \cot \alpha & u_0 \\ 0 & f\sigma_y \sec \alpha & v_0 \\ 0 & 0 & 1 \end{bmatrix}. \quad (2)$$

In (2), f is the focal length, σ_x , σ_y are the pixel size in the x and y directions of the camera frame (also the u and v directions in the image plane), α is the skew angle, which defines how rectilinear the pixels are, and u_0 and v_0 define the center of projection in pixels. Together, f , σ_x , σ_y , α , u_0 , v_0 are the intrinsic calibration parameters, and can be determined through camera calibration procedures [29,30,31,32,33]. The methods presented in this chapter are based on the assumption that the camera is calibrated. Given knowledge of A , it is possible to recover the normalized coordinates of a feature point from the pixel coordinates. More complicated geometric primitives such as lines, curves and surfaces are composed of infinite sets of points $\bar{M} = \{\bar{m}_j\}$, where $\bar{m}_j(t)$ is every point in the primitive. Likewise a line, curve, or surface is described in the image by a set $M = \{m_j\}$, where $m_j = \Omega(\bar{m}_j)$ is every point in the two-dimensional primitive.

The development in this chapter will also focus on the use of points as image features for estimation. Algorithms exist for lines, curves, surfaces, etc., but points are simple to extract from an image, track between images, and to mathematically represent. Furthermore, many methods that use other primitives are simply alternative ways to



Image Based State Estimation, Figure 2
Feature points from corner detection



Image Based State Estimation, Figure 3
Feature points from the critical points of a contour

solve for the structures such as the Essential or Homography Matrices, which can also be determined by just using points. Extraction of feature points in images is outside the scope of this chapter, but there are many resources available. Points in an image are often found as contrasting areas such as corners (e.g., [34,35]), as an averaged area such as the centroid of a region, as local extrema [36], as critical points of a contour, or at the intersection of lines. Figures 2–4 show examples of corner extraction from corner detection, critical points of contours and intersection of lines. There are many methods of contour extraction (cf. [37,38]) and line/curve extraction (cf. [34,39,40]).

Linear Estimation of Position and Orientation from Two Images

While techniques for estimating camera pose and scene structure from a series of images have been developed

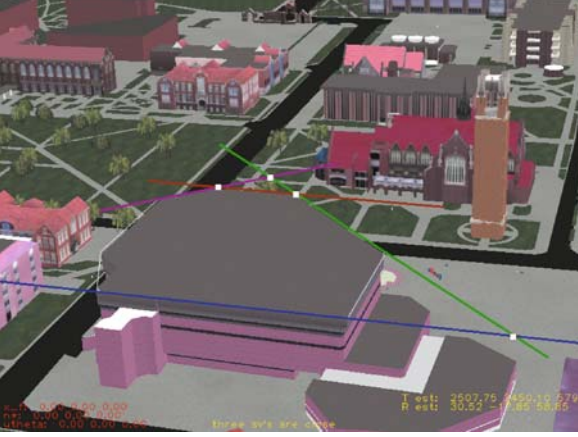


Image Based State Estimation, Figure 4
Feature points from the intersection of lines

since the early 20th century, computational resources limited the advancement and usage of these methods until the early 1980s. The first techniques developed were linear methods, in the sense that they rely solely on linear algebra. The first of these methods involved the epipolar constraint, and is commonly referred to as the Essential Matrix or the eight-point algorithm, due to the fact that it requires at least eight points in the scene (e.g., [2,3]). A similar method was developed for planar scenes [6], and is often referred to as the Homography Matrix.

The Essential Matrix and the Eight-Point Algorithm

Consider a camera that is viewing a collection of N feature points with Euclidean coordinates denoted by

$$\bar{m}_j^* = [x_j^*, y_j^*, z_j^*]^T, \quad \forall j \in \{1 \dots N\} \quad (3)$$

expressed in the camera reference frame, denoted by \mathcal{F}_c^* . An image of the points is captured, resulting in a projection to a set of points in the image-plane denoted by the normalized coordinates

$$m_j^* = \left[\frac{x_j^*}{z_j^*}, \frac{y_j^*}{z_j^*}, 1 \right]^T, \quad \forall j \in \{1 \dots N\}. \quad (4)$$

By translating the camera by $x(t)$ and rotating the camera by $R(t)$, the camera will obtain a new pose denoted by the reference frame \mathcal{F}_c , as illustrated in Fig. 5. The Euclidean and normalized coordinates of the collection of feature points expressed in \mathcal{F}_c are defined as

$$\bar{m}_j(t) = [x_j(t), y_j(t), z_j(t)]^T, \quad \forall j \in \{1 \dots N\} \quad (5)$$

$$m_j(t) = \left[\frac{x_j(t)}{z_j(t)}, \frac{y_j(t)}{z_j(t)}, 1 \right]^T, \quad \forall j \in \{1 \dots N\}. \quad (6)$$

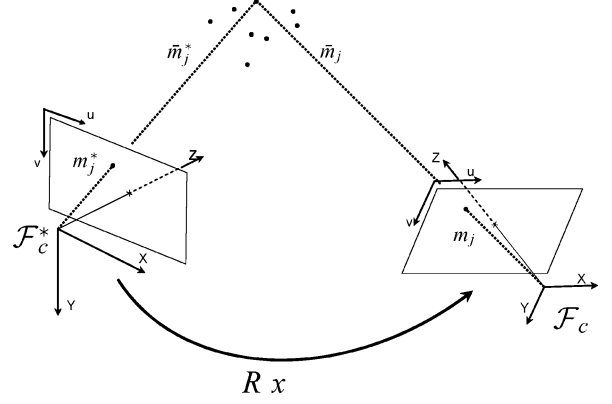


Image Based State Estimation, Figure 5

Projection of a point onto two image planes, and the resulting epipole

The Euclidean coordinates of the feature points expressed in \mathcal{F}_c^* can be related to the coordinates expressed in \mathcal{F}_c through the following algebraic relationship:

$$z_j(t)m_j(t) = z_j^*R(t)m_j^* + x(t), \quad \forall j \in \{1 \dots N\}, \quad (7)$$

where the relationships in (3)–(7) have been used. Multiplying both sides of (7) by the skew symmetric form of $x(t)$, denoted by $[x(t)]_\times \in \mathbb{R}^{3 \times 3}$, yields

$$z_j(t)[x(t)]_\times m_j(t) = z_j^*[x(t)]_\times R(t)m_j^*, \quad \forall j \in \{1 \dots N\}, \quad (8)$$

after using the property that $[x(t)]_\times x(t)$ is equal to a vector of zeros. After multiplying both sides of (8) by $m_j^T(t)$ and exploiting the property that

$$z_j(t)m_j^T(t)[x(t)]_\times m_j(t) = 0$$

the essential constraint or epipolar constraint is given as

$$m_j^T(t)E(t)m_j^*(t) = 0 \quad (9)$$

where $E(t) \in \mathbb{R}^{3 \times 3}$ is known as the Essential Matrix, and is defined as

$$E(t) = [x(t)]_\times R(t). \quad (10)$$

Eight or more noncoplanar feature points are required to create a set of linear equations to solve for $E(t)$. The Essential Matrix can be decomposed to recover $R(t)$ and $\lambda x(t)$, where λ is an unknown scale factor [2,3]. This decomposition only involves the use of linear algebra techniques including singular value decomposition.

The inability to recover the scale of $x(t)$ is a fundamental problem in IBSE. The inherent scale factor uncer-

tainty is due to the loss of depth information when three-dimensional Euclidean coordinates are normalized (i. e., projected to the two-dimensional image plane). Mathematically, the scale factor uncertainty is evident from the fact that there is no unique solution to (9). Furthermore, there are two possible rotations and two possible translations that satisfy (10) for a particular $E(t)$, giving four possible solutions to the decomposition, of which only one solution is physically valid. Thus, given some physical knowledge of the system, the correct solution can be selected. For further detail, see [2,3].

The epipolar constraint in (9) relates two planes of normalized Euclidean coordinates. If the calibration matrix A is unknown, the epipolar constraint can be applied to points expressed in pixel coordinates as

$$p_j(t) = F(t)p_j^* . \quad (11)$$

In (11) $F(t) \in \mathbb{R}^{3 \times 3}$ is known as the Fundamental Matrix, and has important functions in fields of computer vision. However, the Essential Matrix must be recovered for meaningful pose reconstruction.

The Planar Homography Algorithm

Consider a camera with reference frame \mathcal{F}_c^* . The camera views a collection of $N \geq 4$ or more feature points lying in a plane π in front of the camera, as illustrated in Fig. 6. These points have Euclidean coordinates and normalized Euclidean coordinates defined as in (3) and (4). The plane π has a normal vector $-n^*$ with respect to \mathcal{F}_c^* . By translating the camera by $x(t)$ and rotating the camera by $R(t)$, the camera will obtain a new pose denoted by the reference frame \mathcal{F}_c . The Euclidean and normalized coordinates of the collection of feature points expressed in \mathcal{F}_c are defined

as in (5) and (6). The Euclidean coordinates of the points in two views are related by the relationship

$$\bar{m}_j(t) = R(t)\bar{m}_j^* + x(t) , \quad \forall j \in \{1 \dots N\} . \quad (12)$$

The projective relationship

$$d^* = n^{*T} \bar{m}_j^* , \quad \forall j \in \{1 \dots N\} , \quad (13)$$

where $d^* \in \mathbb{R}$ is the distance from the origin of \mathcal{F}_c^* to the plane, can be used to rewrite (12) in terms of the normalized Euclidean coordinates as [6,41]

$$m_j(t) = \alpha_j(t)H(t)m_j^* . \quad (14)$$

In (14), $\alpha_j(t)$ is a time-varying ratio of the depth coordinates defined as

$$\alpha_j(t) = \frac{z_j^*}{z_j(t)} , \quad \forall j \in \{1 \dots N\} , \quad (15)$$

and $H(t) \in \mathbb{R}^{3 \times 3}$ denotes the Euclidean homography defined as

$$H(t) = R(t) + \left(\frac{x(t)}{d^*} \right) n^{*T} . \quad (16)$$

The matrix $H(t)$ in (16) is known by many names such as the Planar Essential Matrix, the Planar Homography Matrix, and the Euclidean Homography Matrix. In the case of digital imaging, the calibration matrix (1) is used with the Euclidean relationship in (14) to give

$$\begin{aligned} p_j &= \alpha_j A H A^{-1} p_j^* \\ &= \alpha_j G p_j^* . \end{aligned} \quad (17)$$

Given a set of $N \geq 4$ image points (i. e., p_j , $j \in \{1 \dots N\}$) matched in two images and knowledge of A , a set of linear equations can be solved to acquire $G(t)$ and recover $H(t)$. Linear decomposition methods (e. g., [6,41]) can then be applied to recover, $R(t)$, $x(t)/d^*$, n^* and $\alpha_j(t)$.

In the general case, linear decomposition methods will return four mathematically valid solutions. Two solutions are physically invalid, as $n^{*T} \bar{m}_j^* = d^* < 0$, which would place the points behind the camera; hence, there are only two physically valid solutions for the displacement. Scene knowledge, such as the expected value of n^* could be used to choose the correct solution. If such knowledge is not available, a small, arbitrary translation can be performed with respect to the frame \mathcal{F}_c^* , and a second Homography Matrix can be computed and decomposed. Since both homographies give a solution with respect to the constant

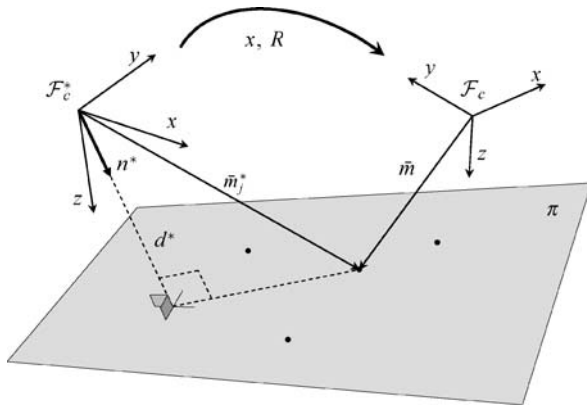


Image Based State Estimation, Figure 6
Projection of coplanar points onto two image planes

frame \mathcal{F}_c^* , one solution from each must have the same n^* . The solution with agreeing n^* is the correct solution. This method of comparing the normals is referred to as temporal coherence of the homography solution.

The translation $x(t)$ is recovered from the homography decomposition only up to the scale factor d^* , which is typically not known and may not be recovered from images alone. If $d^*(t)$ can be measured with a separate sensor, the translation $x(t)$ can be recovered. Depending on the task, additional information can be used to recover d^* and $x(t)$. For example, the camera is translated a known distance x' from \mathcal{F}_c to \mathcal{F}_c' , and the planar homography algorithm is used to solve for the rotation R'' and scaled translation x''/d^* from \mathcal{F}_c^* to \mathcal{F}_c' . Since \mathcal{F}_c and \mathcal{F}_c' are separated by pure translation, $R'' = R$, and the known scaled translations x''/d^* and x/d^* and the known translation x' are related by

$$d^* \frac{x''}{d^*} = d^* \frac{x}{d^*} + x'. \quad (18)$$

The constant d^* can then be determined as

$$d^* = \frac{\left(\frac{x''}{d^*} - \frac{x}{d^*} \right)^T x'}{\left\| \left(\frac{x''}{d^*} - \frac{x}{d^*} \right) \right\|^2}. \quad (19)$$

The planar homography algorithm is sufficient to give orientation and (scaled) translation from the current camera frame \mathcal{F}_c to a reference frame \mathcal{F}_c^* . The alternate case where the camera is fixed, and the planar object moves from a reference pose \mathcal{F}_s^* to a new pose $\mathcal{F}_s(t)$, can also be considered in a similar framework. By solving and decomposing the Homography Matrix $H(t)$, the rotation $R_s(t)$ and scaled translation $x_s(t)/d^*$ of the object can be recovered from the perceived camera rotation $R(t)$ and translation $x(t)/d^*$ as

$$R_s(t) = R^T(t) \quad (20)$$

$$\frac{x_s(t)}{d^*} = -R^T(t) \frac{x(t)}{d^*}. \quad (21)$$

If the camera can be moved precisely while the viewed object is known to be stationary (or vice versa), then d^* can be recovered as in (18). The distance d^* can not be recovered if both the object and camera are moving with an unknown motion.

Characteristics of the Essential Matrix Algorithm and Planar Homography Matrix

The pose estimation methods that are discussed in this chapter are examples of different approaches, and each approach has different advantages and disadvantages. The

characteristics of the Essential Matrix algorithm and planar homography algorithm are compared in this section to distinguish the strengths and weaknesses. Some characteristics are shared because the Essential and Planar Homography Matrices are formally related. Comparing (10) and (16) indicates that

$$E(t) = [x(t)]_{\times} H(t). \quad (22)$$

In the general case that $x(t)$ is known only to a scale factor λ , $E(t)$ is proportional to $[x(t)]_{\times} H(t)$. Given $H(t)$ from four coplanar points and two additional points outside of the plane, it is possible to recover $E(t)$ [26]. Similarly, given $E(t)$ from eight non-coplanar points, it is possible to solve for the $H(t)$ matrix corresponding to any three of the eight points and their corresponding plane (see [26,42]).

A notable difference between the Essential Matrix and the Homography Matrix is that at least eight points are required to solve for the Essential Matrix using linear methods, while at least four points are necessary to solve for the Homography Matrix. The target/landmark/scene involved in the state estimation will determine how many points are available to track, and will determine which methods may be used. The Essential Matrix can not be solved if the Euclidean points $\tilde{m}_j(t)$ lie in the same plane. Specifically, there is a numerical dependence on the system of equations used to solve $E(t)$, and the Essential Matrix can not be used in the case that the eight points are coplanar.

Similarly, the points used to solve for the Homography Matrix in (16) must be coplanar. Besides using linear algebra techniques to solve multiple equations, there are alternative methods to solve the Homography Matrix that may be used in cases that feature points are not coplanar. It is possible to use the Essential Matrix to recover the Homography for a subset of three points. A similar method, coined the Virtual Parallax [43], skips the initial computation of the Essential Matrix. Given eight points, three points define a plane π_s , and the remaining five or more points are projected onto π_s , delivering five virtual points. These points can be used to deliver the Homography Matrix where n^* and d^* are defined with respect to π_s . Solving $H(t)$ from $E(t)$, or using the Virtual Parallax extends the ability to use the Homography Matrix for general scenes at the cost of requiring more points, slightly increased computation cost, and possibly increased noise sensitivity.

From (10), if $x(t) = 0$, then $E(t) = 0$, even for non-zero rotation. Thus, if there is no translation, rotation can not be recovered from the Essential Matrix. Generally, as the amount of translation becomes closer to zero, the estimate of $R(t)$ will incur greater error. For some applications where the camera or target is always translating, the Essential Matrix may be useful, but for many autonomous

robotic applications, the Essential Matrix will be insufficient and the Homography Matrix may be preferable.

A third difference is the additional information delivered from the Homography Matrix. While both methods can estimate translation $x(t)$ and rotation $R(t)$, the Homography gives the normal vector n^* which is used in Subsect. “Estimating the Pose of a Planar Object Relative to the Camera” to estimate the pose of an object. Homography decomposition also gives the ratios $\alpha_j(t)$, $\forall j \in \{1 \dots N\}$. This ratio can be used as an error signal for control (e.g., [44,45]) or IBSE, as shown in Subsect. “Velocity Estimation Through a Nonlinear Observer”.

Extensions to the Linear Pose Estimation Algorithms

The methods discussed in Subsects. “The Essential Matrix and the Eight-Point Algorithm” and “The Planar Homography Algorithm” give a foundation for pose estimation. However, as presented the methods have limitations. These methods can only estimate translation up to a scale factor, are limited to estimating camera or object pose relative to an initial pose corresponding to a fixed camera view, and are constrained to limited motions such that the viewed object remains in the FOV. This section provides examples that can be used to extend the linear pose estimation methods to overcome these deficiencies. Often these additional methods will require some additional information or sensor, such as a known length on the viewed object, a known distance to the object, or accurate odometry for the camera frame.

Long Term Pose Estimation Through Chained Homography Decomposition

The methods discussed in Subsects. “The Essential Matrix and the Eight-Point Algorithm” and “The Planar Homography Algorithm” allow for the estimation of camera motion provided the feature points used in the Essential or Homography Matrices remain in the FOV. Many tasks will require state estimation continue as objects leave and enter the field of view. Consider an unmanned air vehicle (UAV) equipped with a GPS and a camera capable of viewing a landscape. The Essential or Homography Matrix can be used to recover the motion of the plane, but the limited camera FOV and motion of the vehicle can cause observed feature points to leave the image. Motivated by the FOV issue, recent research (e.g., [10,11]) has been developed to allow a transition from estimating the camera pose from previously viewed feature points to estimating the camera pose from incoming feature points. For example, the work in [10] assumes coplanarity of all points and is focused on building and maintaining a mosaic of images. The pose es-

timization efforts in [11] do not assume coplanarity of all sets of feature points and uses a homography relationship to “daisy-chain” relative pose estimates together. The approach in [11] provides a mechanism to work around the FOV obstacle by indefinitely chaining together a series of pose estimates.

To further illustrate the daisy-chaining method provided in [11], consider an inertial world frame \mathcal{F}_w , and a frame attached to some vehicle, \mathcal{F}_c . Since the methods discussed in Sect. “Linear Estimation of Position and Orientation from Two Images” provide means to determine camera motion relative to a fixed camera frame, the following example is based on the assumption that the vehicle pose is known through some inertial sensor (e.g., a global positioning system (GPS)) and other sensors at some initial time t_0 . That is, the translation and rotation, $x_0(t_0)$ and $R_0(t_0)$, between \mathcal{F}_w and $\mathcal{F}_c(t_0)$ is known. Figure 7 illustrates this scenario with a UAV. At time t_0 , the aircraft enters a GPS denied environment and relies solely on on-board sensors including a camera. Without loss of generality, the GPS unit is assumed to be fixed to the origin of the aerial vehicle’s coordinate frame, and the constant pose of the camera frame is known with respect to the pose of the UAV coordinate frame. The subsequent development further assumes that the GPS is capable of delivering altitude, perhaps in conjunction with an altimeter, so that the altitude above the ground $a(t_0)$ is known, i.e., the scalar distance to the ground in the direction of gravity (see Fig. 8).

As illustrated in Fig. 7, the initial set of tracked coplanar and noncoplanar feature points, denoted $p_a(t)$, are projections from the set of Euclidean points, $m_a(t)$, contained in the plane π_a . These feature points have Euclidean coordinates $\bar{m}_{a_j}(t_0) \in \mathbb{R}^3 \forall j \in \{1 \dots N\}$ in $\mathcal{F}_c(t_0)$. The plane π_a is perpendicular to the unit vector $n_a(t_0)$ in the camera frame, and lies at a distance $d_a(t_0)$ from the camera frame origin. At time t_1 , the vehicle has some rotation $R_{01}(t_1)$ and translation $x_{01}(t_1)$ that can be determined from the images by decomposing the relationships given

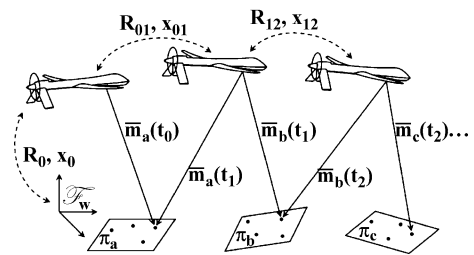


Image Based State Estimation, Figure 7
Illustration of pose estimation chaining

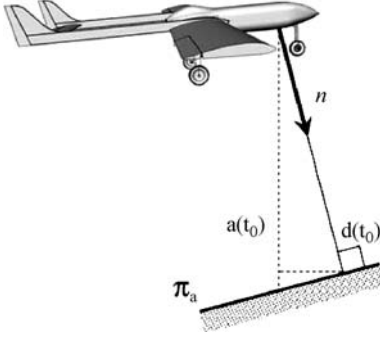


Image Based State Estimation, Figure 8
Depth estimation from altitude

in (17). For notational simplicity, the subscript j is omitted in subsequent development.

As described in Subsect. “The Planar Homography Algorithm”, $R_{01}(t_1)$ and $x_{01}(t_1)/d_a(t_0)$ can be determined from two corresponding images of the feature points $p_a(t_0)$ and $p_a(t_1)$. A measurement or estimate for $d_a(t_0)$ is required to recover $x_{01}(t_1)$. This estimation is possible with distance sensors or with a priori knowledge of the geometric distances between the points in π_a . However, with an additional assumption that the plane is above the plane π_a , it is possible to estimate $d_a(t_0)$ geometrically using altitude information from the last GPS reading, range finder and/or an altimeter. From the illustration in Fig. 8, if $a(t_0)$ is the height above π_a (e. g., the slope of the ground is constant between the feature points and projection of the plane’s location to the ground), then the distance $d_a(t_0)$ can be determined as

$$d_a(t_0) = n_a(t_0) \cdot a(t_0). \quad (23)$$

where $n_a(t_0)$ is known from the homography decomposition.

Once $R_{01}(t_1)$, $d_a(t_0)$, and $x_{01}(t_1)$ have been determined, the rotation $R_1(t_1)$ and translation $x_1(t_1)$ can be determined with respect to \mathcal{F}_w as

$$\begin{aligned} R_1 &= R_0 R_{01} \\ x_1 &= R_{01} x_{01} + x_0. \end{aligned}$$

As illustrated in Fig. 7, a new collection of feature points $p_b(t)$ can be obtained that correspond to a collection of points on a planar patch denoted by π_b . At time t_2 , the two images of the sets of feature points on π_b taken at times t_1 and t_2 can be used to determine $R_{12}(t_2)$ and $x_{12}(t_2)/d_b(t_1)$, which provides the rotation and scaled translation of \mathcal{F}_c with respect to \mathcal{F}_w . If π_b and π_a are the same plane, then $d_b(t_1)$ can be determined as

$$d_b(t_1) = d_a(t_1) = d_a(t_0) + x_{01}(t_1) \cdot n(t_0). \quad (24)$$

When π_b and π_a are the same plane $x_{12}(t_2)$ can be correctly scaled, and $R_2(t_2)$ and $x_2(t_2)$ can be computed in a similar manner as described for $R_1(t_1)$ and $x_1(t_1)$. The pose estimates can be propagated by chaining them together at each time instance without further use of GPS.

In the general case, sets of feature points on π_b and π_a are not coplanar, and (24) cannot be used to determine $d_b(t_1)$. If the sets of feature points on π_b and π_a are both visible for two or more frames, it is still possible to calculate $d_b(t)$ through geometric means. Let t_{1-} denote as some time before the daisy chain operation is performed, when sets of feature points on π_b and π_a are visible in the image. At time t_{1-} , an additional set of homography equations can be determined for the points on π_b and π_a at times t_1 and t_{1-}

$$m_{ai}(t_1) = \alpha_a \left(R + \frac{x n_a(t_{1-})^T}{d_a(t_{1-})} \right) m_{ai}(t_{1-}) \quad (25)$$

$$m_{bi}(t_1) = \alpha_b \left(R + \frac{x n_b(t_{1-})^T}{d_b(t_{1-})} \right) m_{bi}(t_{1-}) \quad (26)$$

where

$$\alpha_a = \frac{z_{ai}(t_{1-})}{z_{ai}(t_1)} \quad \text{and} \quad \alpha_b = \frac{z_{bi}(t_{1-})}{z_{bi}(t_1)}.$$

Note that R and x have the same values in Eqs. (25) and (26), but the distance and normal to the plane are different for the two sets of points, therefore the decomposition methods will give different scaled translations

$$x_a(t_1) = \frac{x(t_1)}{d_a(t_{1-})} \quad \text{and} \quad x_b(t_1) = \frac{x(t_1)}{d_b(t_{1-})}.$$

The distance $d_a(t_{1-})$ can be found using (24). The translation $x(t_1)$ is solved as

$$x(t_1) = d_a(t_{1-}) x_a(t_1).$$

The distance $d_b(t_{1-})$ can then be determined as

$$d_b(t_{1-}) = \frac{x_b^T x}{\|x_b\|^2}.$$

The distance $d_b(t_1)$ can then be found by using (24) with $d_b(t_{1-})$ in place of $d_a(t_0)$. Additional sensors, such as an altimeter, can provide an additional estimate in the change in altitude. These estimates can be used in conjunction with (24) to update depth estimates.

Simulations To provide a practical explanation of the daisy-chaining method this section provides simulation results for an autonomous UAV flying into a zone with

denied GPS availability. For the simulation, five patches of four feature points are evenly placed 100 m apart in a straight line. For simplicity, all planar patches of feature points lie in the same plane. The objective is to perform IBSE during a maneuver of a 10 m lateral shift to the right and a 10 m longitudinal increase in altitude. This particular maneuver results in the vehicle simultaneously pitching, rolling, and yawing, while translating. For simulation purposes, the camera is considered to be mounted underneath the fuselage looking downwards. The camera model

for this exercise is intended to be representative of a typical 640×480 lines of resolution CCD camera equipped with a 10 mm lens. To more accurately capture true system performance, pixel coordinates were rounded to the nearest integer to model errors due to camera pixilation effects (i. e., quantization noise). Furthermore, a five percent error was added to the estimated vehicle altitude to examine the robustness.

The first simulation was designed to test the accuracy of the vision-based estimation. Vision was not used in the

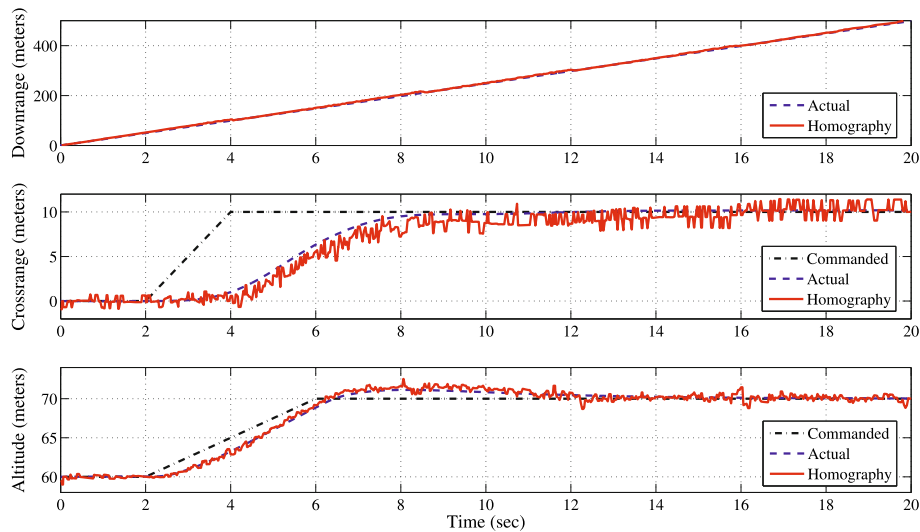


Image Based State Estimation, Figure 9

Simulation results of actual position versus estimated position when the UAV is using GPS for control

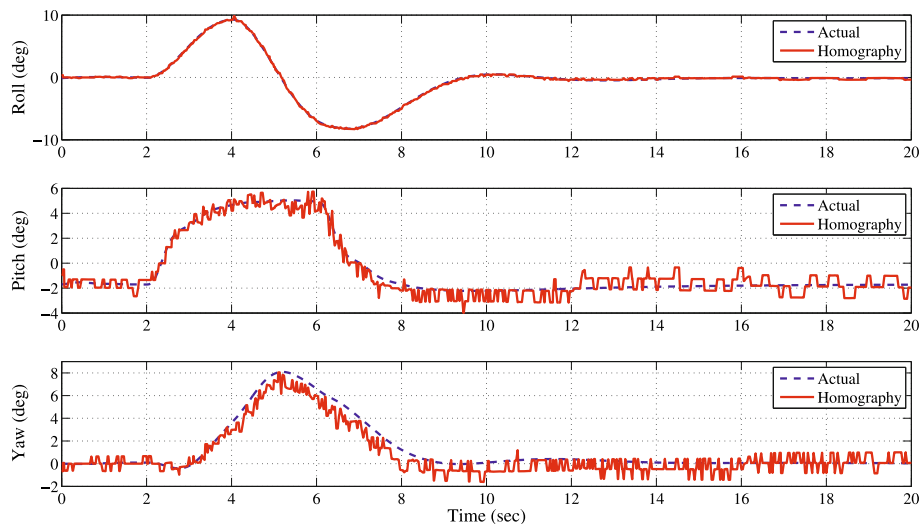


Image Based State Estimation, Figure 10

Simulation results of actual attitude versus estimated attitude when the UAV is using GPS for control

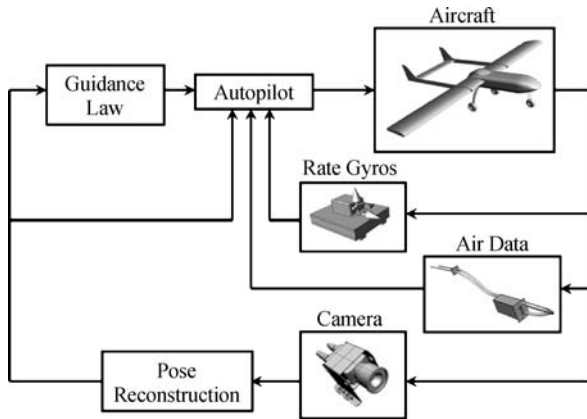


Image Based State Estimation, Figure 11

Block diagram control system architecture

feedback in this maneuver, and the estimated pose is compared to the true pose. The results of this preliminary analysis are given in Figs. 9 and 10. The effects of noise are visible but the estimated pose is accurate.

The second simulation illustrates the effects of using the IBSE as a sensor in closed-loop control. This simulation involved replacing the perfect pose measurements used in the guidance system and autopilot with the results from the IBSE. The resulting control architecture and sensor suite for this UAV is given in Fig. 11. The noise content of the estimated pose required filtering prior to being used by the autopilot to prevent the high frequency noise from being passed to the aircraft actuators. As expected, the noise occurs at 30 Hz and corresponds to the frame

rate of the camera. First-order, low-pass filters (cutoff frequency as low as 4 rad/s) were used to filter the noise. The noise also prevented effective differentiation of the position and attitude and necessitated the use of rate gyros for yaw and roll damping, as depicted in Fig. 11. The air data system is also included, as shown in Fig. 11, for the initial altitude measurement, since it is more accurate for altitude than current GPS solutions. The results of the camera-in-the-loop system performing the same autonomous maneuver are given in Figs. 12 and 13.

The simulation results indicate that a camera supplemented with minimal sensors such as rate gyros and barometric altitude can be used for completely autonomous flight of a fixed wing vehicle; however, some residual oscillation effects due to noise is present in the vehicle attitude response. A majority of the noise source can directly be attributed to camera pixilation effects and the corresponding phase lag introduced by the first order filtering.

Experiments Based on the results of the simulation, a flight test experiment was conducted to establish the feasibility of the proposed IBSE method. Artificial features were placed along a stretch of a runway. A radio controlled aircraft with an onboard camera was flown over the runway. The video was overlaid with GPS data. An example of a single frame of this video is given in Fig. 14. A second GPS unit was also onboard to test inter-GPS accuracy. The use of two GPS units provides a comparison for the IBSE method, which is intended to compute GPS-like information. Video data was captured using a DV tape recorder

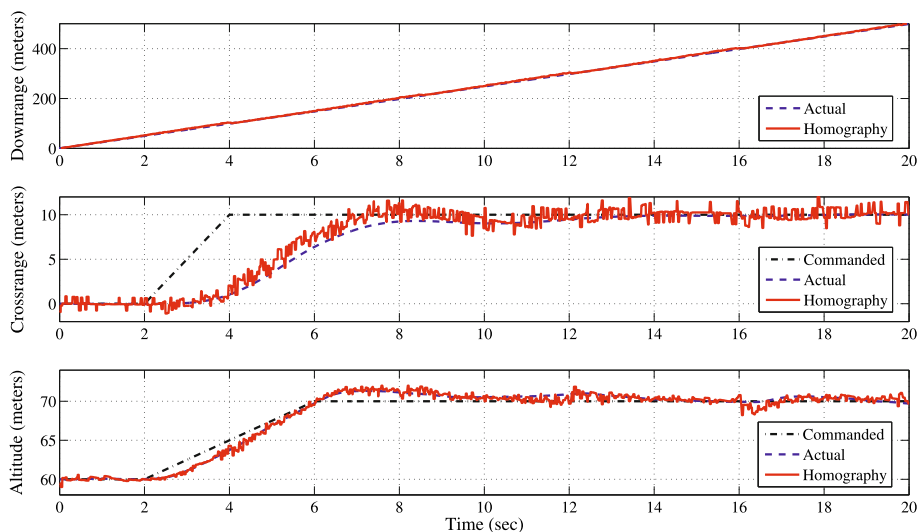


Image Based State Estimation, Figure 12

Simulation results of actual position versus estimated position when the UAV control-loop uses estimated pose information

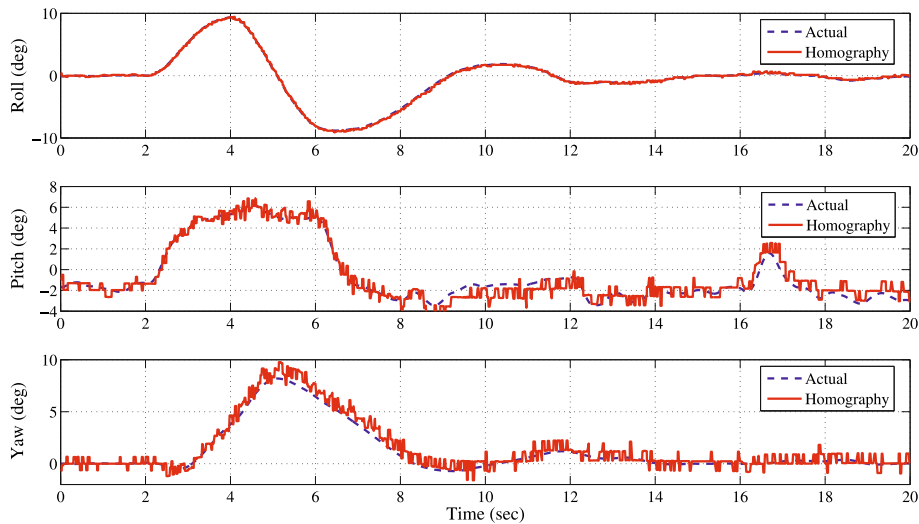


Image Based State Estimation, Figure 13
Simulation results of actual attitude versus estimated attitude when the UAV control-loop uses estimated pose information



Image Based State Estimation, Figure 14
Single video frame with GPS overlay illustrating landmarks placed along inside edge of the runway

and analyzed offline. Due to poor image quality, including focus, motion blur and interlacing of the DV video, it became necessary to extract features by hand from individual frames for this particular example. Features were extracted every sixth frame, resulting in a 5 Hz input signal.

Results of the experiment are given in Fig. 15. In the legend for Fig. 15, GPS2 represents the overlaid GPS data, and GPS1 represents the onboard data logger GPS values. The symbol ‘*’ indicates the times when daisy-chaining was performed and pose reconstruction is performed using a new set of feature points. Significant mismatch exists between the two GPS measurements, and the IBSE results remain proportionate to the two GPS measurements. Furthermore, the estimates agree closely with GPS2 for down-

range and crossrange translation, and with GPS1 for altitude translation. There is no discernible discontinuity or increased error at the daisy-chain handoff times. Note that the 5 Hz refresh rate of the vision-based estimation is also higher than the 1 Hz rate of both GPS units. The pose estimation code can be executed in real time (> 30 Hz) on a typical laptop.

**Solving for Relative Object Poses
Through Knowledge of a Single Length**

It is often not enough to recover camera pose relative to a reference pose. This section expands the techniques in Subject. “[The Planar Homography Algorithm](#)” to recover camera pose relative to a planar object. Recovering relative pose between the camera and multiple objects, as well as the relative poses between objects, is also possible. To achieve these results, some additional information is required to recover the scale factor. It is sufficient to know the distance to a plane, or to recover it from precise motion, as in (19). The following example is based on work in [46,47], where relative pose with respect to an object is estimated given a single known geometric length between two feature points in the Euclidean-space.

Estimating the Pose of a Planar Object Relative to the Camera The planar homography algorithm is sufficient to give rotation and scaled translation of the camera or object with respect to some reference pose. However, it is not sufficient to solve for the complete pose of the camera with respect to a viewed object or the pose of one viewed object with respect to another. The following development pro-

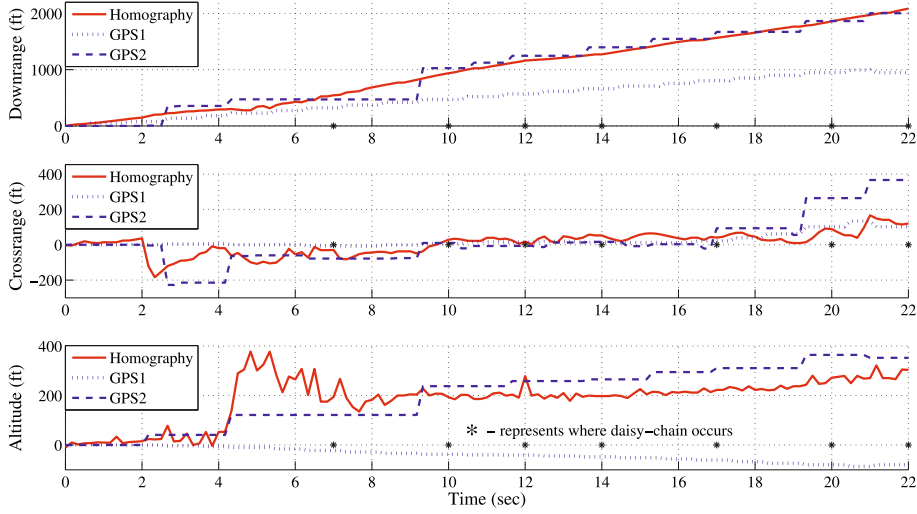


Image Based State Estimation, Figure 15

Experimental flight test results that compare the IBSE results to two GPS signals

vides an example where the planar homography algorithm can be used along with some additional information to attach a reference frame \mathcal{F}_s to a planar object, with rotation and translation given with respect to the camera frame \mathcal{F}_c . That is, the development in this section can be used to determine the rotation and translation between \mathcal{F}_s and \mathcal{F}_c .

Consider a camera viewing a planar object with four or more distinguishable feature points, and denote the feature point plane as π . A reference frame \mathcal{F}_c is attached to the camera. If the camera and/or object move over time, a set of linear equations can be solved for the Homography Matrix $H(t)$ as in (16). The matrix $H(t)$ can be decomposed into $R(t)$, $x(t)/d^*$, and n^* as in (16), where the reference frame \mathcal{F}_c^* can be taken as $\mathcal{F}_c(t_0)$, i. e., the initial frame. The normal vector of the planar object in the current frame \mathcal{F}_c is given as

$$n(t) = R(t)n^*.$$

The goal is to attach a reference frame $\mathcal{F}_s(t)$ to the planar object. The orthonormal vectors $i_x, i_y, i_z \in \mathbb{R}^3$ that define $\mathcal{F}_s(t)$ form a rotation matrix $R_s(t)$ in \mathcal{F}_c as

$$R_s = [i_x, i_y, i_z]. \quad (27)$$

Without loss of generality, the origin of $\mathcal{F}_s(t)$ is assigned to the feature point \tilde{m}_1 . The columns of $R_s(t)$ in (27) are defined as

$$i_z = -n \quad (28)$$

$$i_x = \frac{\tilde{m}_2 - \tilde{m}_1}{\bar{s}_1} \quad (29)$$

$$i_y = -n \times \frac{\tilde{m}_2 - \tilde{m}_1}{\bar{s}_1}, \quad (30)$$

where the constant distance between the two feature points $\bar{s}_1 = \|\tilde{m}_1 - \tilde{m}_2\|$ is assumed to be the extra information that can be exploited to complete the pose estimation. If $\tilde{m}_1(t)$ and $\tilde{m}_2(t)$ were known, then i_x and i_y can be determined from (29) and (30) since \bar{s}_1 is assumed to be known. To solve for $\tilde{m}_1(t)$ and $\tilde{m}_2(t)$, a new plane π'_s is defined with normal $-n(t)$ (so π'_s is parallel to π_s) and containing the normalized image point $m_1(t)$. A line l is defined from the origin of \mathcal{F}_c through $m_2(t)$ and $\tilde{m}_2(t)$. The plane π'_s intersects l at a point m'_2 . The unknown distance between $m_1(t)$ and $m'_2(t)$ is s_1 , as illustrated in Fig. 16.

The primitives l and π'_s are defined by the sets of points $q \in \mathbb{R}^3$ that satisfy the implicit functions

$$l = \{q \mid q - um_2 = 0, \forall u \in \mathbb{R}\} \quad (31)$$

$$\pi'_s = \{q \mid n \cdot (q - m_1) = 0, q, n, m_1 \in \mathbb{R}^3\}. \quad (32)$$

The intersection of π'_s and l occurs when

$$u = \frac{n \cdot m_1}{n \cdot m_2}. \quad (33)$$

The expressions in (31) and (33) can be combined to solve for the point $q = m'_2$ as

$$m'_2 = \frac{n \cdot m_1}{n \cdot m_2} m_2.$$

The solution for $m'_2(t)$ is used to solve for s_1 as

$$s_1 = \|m'_2 - m_1\|, \quad (34)$$

and the properties of similar triangles can be used to calculate the following:

$$\frac{s_1}{\bar{s}_1} = \frac{\|m_1\|}{\|\tilde{m}_1\|} = \frac{\|m'_2\|}{\|\tilde{m}_2\|}. \quad (35)$$

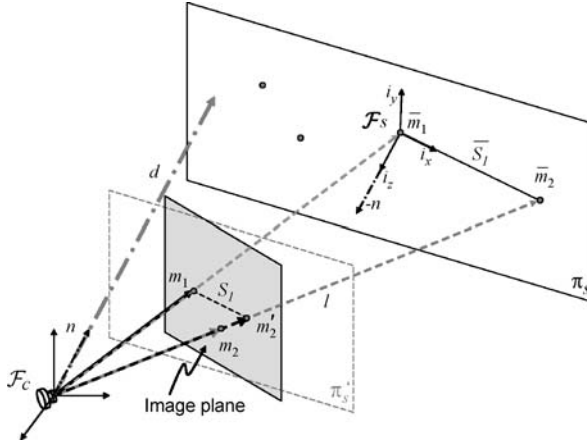


Image Based State Estimation, Figure 16

Geometric elements used to determine frame \mathcal{F}_s with respect to \mathcal{F}_c

Since $s_1, \bar{s}_1, \|m_1(t)\|$, and $\|m'_2(t)\|$ are now known or measurable, (35) can be used to solve for $\|\bar{m}_1(t)\|$ and $\|\bar{m}_2(t)\|$, which can be used to recover $\bar{m}_1(t)$ and $\bar{m}_2(t)$ as

$$\bar{m}_1 = \frac{\|\bar{m}_1\|}{\|m_1\|} m_1, \quad \bar{m}_2 = \frac{\|\bar{m}_2\|}{\|m_2\|} m_2.$$

Solutions for i_x, i_y , and $R_s(t)$ can now be determined from (29), (30), and (27), respectively. Since $\mathcal{F}_s(t)$ is attached to $\bar{m}_1(t)$, the translation is simply given by $x_s(t) = \bar{m}_1(t)$. Furthermore, since $\bar{m}_1(t)$ can be determined, it is now possible to solve for the distance $d(t)$ as

$$d = n \cdot \bar{m}_1. \quad (36)$$

If the constant length \bar{s}_1 is not known, but $d(t)$ is known or solved using methods such as (19), then the fact that $d(t) = z_j(t)n^T(t)m_j(t)$ can be used to solve for the Euclidean coordinates of $\bar{m}_j(t)$ as

$$\bar{m}_j = \frac{dm_j}{n \cdot m_j}, \quad \forall j \in \{1 \dots N\}. \quad (37)$$

Solving for the Euclidean coordinates of any two points $\bar{m}_1(t)$ and $\bar{m}_2(t)$ will allow the estimation of \bar{s}_1 which can be used to solve for the frame $\mathcal{F}_s(t)$.

Estimating the Pose of Multiple Planar Objects Relative to the Camera The previous development is now extended to the case of multiple planar patches and piecewise planar objects, given knowledge of only a single geometric length on a single static object in the scene. Consider a large sample of points \mathcal{P} visible to the camera. These points have been grouped into k sets of coplanar points $\mathcal{P}_h \subset \mathcal{P}, \forall h \in \{1 \dots k\}$, where all points in \mathcal{P}_h lie

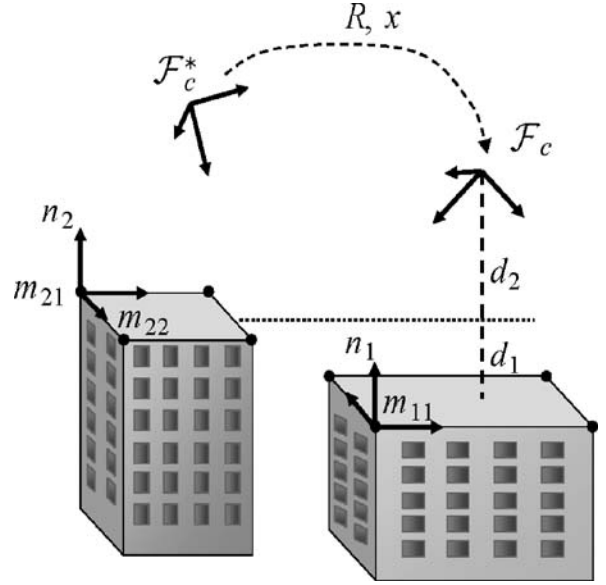


Image Based State Estimation, Figure 17

Example of multiple planar surfaces with respect to a camera

in a plane π_h . The sets \mathcal{P}_h may overlap, i.e., a point can be in more than one set.

Segmenting a set of points into coplanar sets is not a trivial task. The sets can be distinguished through human interaction, scene knowledge (e.g., multiple light objects on a dark background) or various automatic methods [48,49]. This example assumes that each set \mathcal{P}_h is well conditioned in the sense that it contains no more than three collinear points, as illustrated in Fig. 17.

In the following development, the camera is assumed to undergo a rotation $R(t)$ and translation $x(t)$ from a reference frame \mathcal{F}_c^* to a frame \mathcal{F}_c . The points in each set \mathcal{P}_h have coordinates

$$\begin{aligned} \bar{m}_{hj}^* &= [x_{hj}^*, y_{hj}^*, z_{hj}^*]^T, \\ &\forall j \in \{1 \dots N_h\}, \quad \forall h \in \{1 \dots m\} \\ \bar{m}_{hj}(t) &= [x_{hj}, y_{hj}, z_{hj}]^T, \\ &\forall j \in \{1 \dots N_h\}, \quad \forall h \in \{1 \dots m\} \end{aligned}$$

in the frames \mathcal{F}_c^* and \mathcal{F}_c , respectively. These points project to image points with normalized coordinates m_{hj}^* and $m_{hj}(t)$ as described in Subsect. “The Planar Homography Algorithm”. Each set of points in the two images are related by a homography $H_h(t)$ defined by

$$m_{hj} = \frac{z_{hj}^*}{z_{hj}} H_h m_{hj}^* \quad (38)$$

$$m_{hj} = \frac{z_{hj}^*}{z_{hj}} \left(R(t) + \frac{x(t)}{d_h^*} n_h^{*T} \right) m_{hj}^*. \quad (39)$$

Note that $R(t)$ and $x(t)$ are the same for all point sets \mathcal{P}_h , since all coordinate changes are due to the motion of the single camera. However, each plane π_h is different; therefore, each set of points will have different d_h^* , n_h^* and $H_h(t)$.

From m_{hj}^* and m_{hj} , it is possible to recover $H_h(t)$, n_h^* , $R(t)$, and $x_h(t) = x(t)/d_h^*$ for all $h \in \{1 \dots m\}$. The subsequent development is based on the assumption that a single geometric length between two points in a single set is known. Without loss of generality, this length is assumed to be known in set \mathcal{P}_1 . Given this geometric length, a reference frame \mathcal{F}_{s1} can be attached to plane π_1 and the development in Subsect. “Estimating the Pose of a Planar Object Relative to the Camera” can be used to solve for $R_s(t)$, $x_s(t)$, $d_1(t)$, d_1^* and all $\tilde{m}_{1j} \forall j \in \{1 \dots N_1\}$. The translation $x(t)$ can then be recovered from d_1^* as

$$x(t) = d_1^* x_1(t).$$

Given $x(t)$, each d_h^* , $\forall h \in \{2 \dots m\}$ can be recovered from the scaled translations $x_h(t)$ as

$$d_h^* = \frac{x_h^T x}{\|x_h\|}.$$

Once each d_h^* has been determined, all $\tilde{m}_{hj}^*, \forall j \in \{1 \dots N_h\}, \forall h \in \{2 \dots m\}$ can be recovered as in (37). From knowledge of \tilde{m}_{hj}^* , a constant length \bar{s}_h between two feature points can be estimated for each plane, and the frames \mathcal{F}_{sh} can be attached to the plane π_h . Thus, (27)–(30) can be used to solve for R_s and $x_s \forall h \in \{2 \dots m\}$. Given the rotation and translation from each plane π_h to the reference camera frame \mathcal{F}_c , the rotation and translation between each planar patch can be recovered.

For a stationary camera viewing multiple moving planar objects, the analysis cannot be performed because there is not a common $R(t)$ or $x(t)$. If a geometric length is known on each object, then the analysis in Subsect. “Estimating the Pose of a Planar Object Relative to the Camera” can be performed for each moving plane.

Experiments of Pose Estimation of a Single Object An experiment using the moving vehicle in Fig. 18 is performed to demonstrate the method presented in Subsect. “Estimating the Pose of a Planar Object Relative to the Camera”. Four bright LED arrays were fixed to the back of a truck to facilitate simple image segmentation, where the centroid of each detected array provides four feature points used to construct the Homography Matrix. Each of the four centroids is indicated in Fig. 18 by a cross and a number. The truck was equipped with a differential GPS



Image Based State Estimation, Figure 18

A processed video frame from the pose estimation experiment

unit to provide a reference to compare the pose estimation. The road was marked at approximately 20 f (6.1 m) intervals and the car was driven forward and stopped approximately every 20 f.

The results of the experiment are seen in Fig. 19. The expected periodically increasing step function along the camera frame z direction is evident. Furthermore, the change in pose estimate agrees closely to the GPS Northing measurement. There is also a small periodic step increase estimated in the camera x direction. The estimate becomes noticeably degraded as the distance to the vehicle increases. This is primarily due to sensor noise. As the car moves farther from the camera, the perceived lights become dimmer and it is harder to extract the centroids. This increases the effects of pixilation (i. e., quantization noise).

Experiments of Pose Estimation of Multiple Objects

Experimental results are also provided to demonstrate the efficacy of the algorithm presented in Subsect. “Estimating the Pose of Multiple Planar Objects Relative to the Camera”. Two images of a cluttered desktop were taken with a camera with a resolution of 1280×1024 . One geometric length, the width of the photograph frame, is known. In each image, the four corners of the picture frame, four corner points on a computer monitor and four points on a speaker give feature points. From two of the images, reference frames are attached to all three objects as seen in Fig. 20 and the location of all corner points are reconstructed in the camera frame as seen in Fig. 21.

The true pose of these objects with respect to the camera is impossible to ascertain, so two heuristic tests are performed. First the estimated dimensions of the objects are compared to hand-measured dimensions in Table 1. The maximum error is in the height of the monitor screen, with an estimate error of 5.7%. The average estimate error is 1.8%. For the second test, a third image was then taken from another viewpoint, as seen in Fig. 22. The corners of the picture frame were identified, but no other points on any other object are selected. The current location of

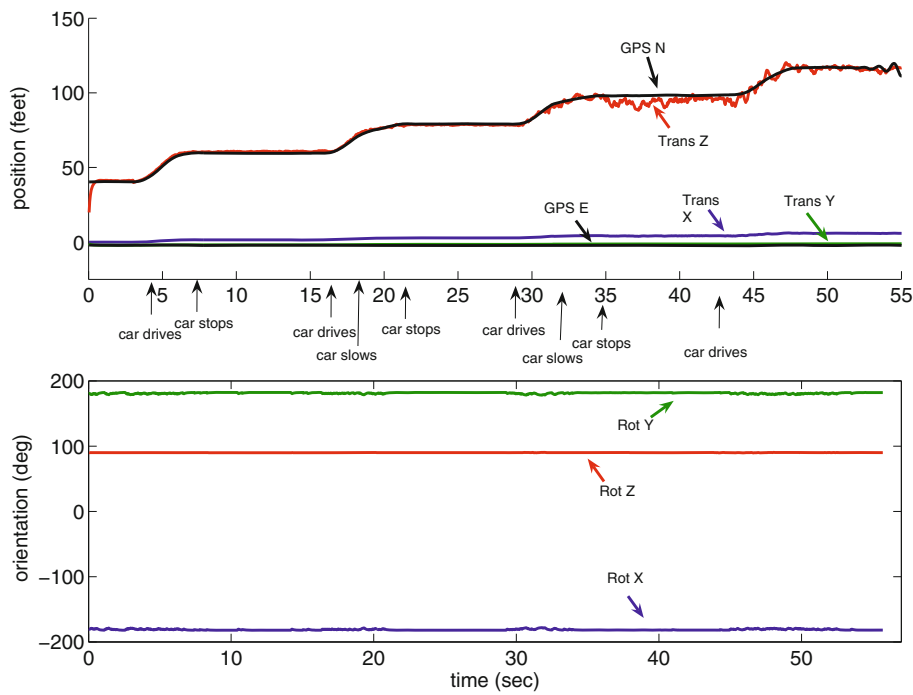


Image Based State Estimation, Figure 19
Results of experiment to estimate the pose of a moving vehicle

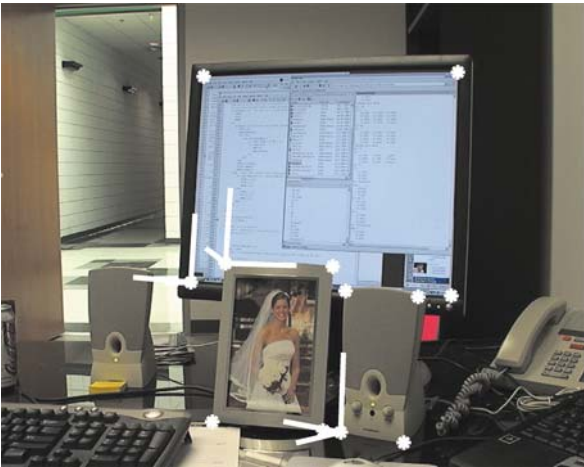


Image Based State Estimation, Figure 20
Image with reference frames attached to a photograph, monitor and speaker

Image Based State Estimation, Table 1
Measured vs Estimated Lengths

Object	Width	Height	Est. width	Est. height
Photo	6"	8"	Known	8.03"
Screen	16"	12"	16.92"	12.04"
Speaker	4.125"	6.875"	4.19"	6.96"

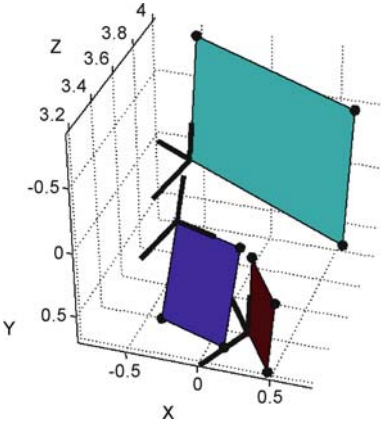


Image Based State Estimation, Figure 21
Geometric reconstruction of the photograph, monitor and speaker in the camera frame \mathcal{F}_c . Units are in feet

the corners of the monitor and speaker in the image are estimated through the estimated relative positions solved in the previous reconstruction. From the location of the frames and points known in the camera frame, the pose of each reference frame can be determined with respect to the photograph frame, and the location of each of the object's points can be determined within its own reference frame. With the picture frame localized in the new camera frame,

Attached reference frames



Image Based State Estimation, Figure 22

Reprojected feature points from their estimated Euclidean positions

the locations of the other frames and points are estimated and projected to their estimated locations in the third image.

Estimation of Linear and Angular Velocity

Image based estimation of velocity has a shorter history than estimation of pose. One method of velocity estimation is a continuous version of the eight-point algorithm that uses optical flow (i. e., velocity of image points). Early efforts were developed by Zhuang and Haralick [50]. Ma et al. [51] also introduced a continuous version of the eight-point algorithm that very closely resembled the classic eight-point algorithm. An optical flow based version of the planar homography algorithm is also provided in [26], which is discussed in Subsect. “Continuous Planar Homography Algorithm”.

Another method of velocity estimation is to use standard pose estimation methods to generate an initial estimate that is input to functions or filters that estimate velocity. Soatto et al. used a logarithmic map of the results of the eight-point algorithm to estimate velocity [14]. Chitrakaran et al. [19] use the results of the planar homography as input to a nonlinear observer, which will be discussed in Subsect. “Velocity Estimation Through a Nonlinear Observer”.

Continuous Planar Homography Algorithm

Consider a camera with reference frame \mathcal{F}_c moving with linear velocity $v(t)$ and angular velocity $\omega(t)$. The cam-

era views a collection of $N \geq 4$ or more feature points lying in a plane π in front of the camera. The plane π has normal $n(t)$ with respect to \mathcal{F}_c , and lies a distance $d(t)$ from the origin of \mathcal{F}_c along $n(t)$. These feature points have Euclidean coordinates $\tilde{m}_j(t) \in \mathbb{R}^3$ and velocities given as a function of the linear and angular velocity of the camera, v and ω , respectively,

$$\dot{\tilde{m}}_j = \omega \times \tilde{m}_j + v, \quad \forall j \in \{1 \dots N\}. \quad (40)$$

The Euclidean coordinates of the coplanar feature points obey the projection relationship

$$d = n^T \tilde{m}_j, \quad \forall j \in \{1 \dots N\}. \quad (41)$$

Substituting (41) into (40) gives

$$\dot{\tilde{m}}_j = \left([\omega]_{\times} + \frac{v}{d} n^T \right) \tilde{m}_j, \quad \forall j \in \{1 \dots N\}, \quad (42)$$

where $[\omega(t)]_{\times} \in \mathbb{R}^{3 \times 3}$ is the skew-symmetric matrix form of the vector $\omega(t)$ such that $[\omega(t)]_{\times} \tilde{m}_j(t) = \omega(t) \times \tilde{m}_j(t)$. From the normalized image points $m_j(t)$ and the corresponding optical flow vectors $\dot{m}_j(t)$, the optical flow version of the Homography Matrix can be obtained as

$$\dot{m}_j = \frac{z_j}{\dot{z}_j} \left([\omega]_{\times} + \frac{v}{d} n^T \right) m_j \quad (43)$$

$$= \alpha_c H_c m_j \quad (44)$$

where $\alpha_c = z_j/\dot{z}_j$ is a ratio of depth to change in depth, and $H_c = ([\omega]_{\times} + (v/d)n^T)$ is the Continuous Homography Matrix. Similar to the pose estimation method of Subsect. “The Planar Homography Algorithm”, it is possible to solve for α_c , $H_c(t)$, $n(t)$, $\omega(t)$ and $v(t)/d$. Similar to translation estimates, the velocity can only be solved up to the scale factor d . Also like the pose estimation case, there exist multiple mathematical solutions for $n(t)$, $\omega(t)$ and $v(t)/d$, where two solutions can be eliminated based on the physics of the problem see [26] for further details.

Velocity Estimation Through a Nonlinear Observer

Given an image sequence (e. g., video), the pose estimation methods discussed in Sect. “Linear Estimation of Position and Orientation from Two Images” can be used to estimate velocity using established observer design methods. For example, Chitrakaran et al. [19], used homography-based techniques to design a nonlinear observer to identify an object’s unknown velocity. The method in [19] requires knowledge of the object’s initial pose with respect to the camera, and knowledge of a single length on the object. As demonstrated in Subsect. “Estimating the Pose of

a Planar Object Relative to the Camera” the knowledge of the single length is sufficient to recover the initial pose.

Consider an inertial frame \mathcal{I} attached to a fixed camera that is viewing a planar object with an attached reference frame $\mathcal{F}_s(t)$, moving with unknown linear and angular velocities $v_s(t), \omega_s(t) \in \mathbb{R}^3$ expressed in frame \mathcal{I} . Four or more feature points are visible on the object at all time, and the coordinates of one point are known with respect to $\mathcal{F}_s(t)$.

The feature points have pixel coordinates in the image plane as given in (1). The extended image coordinates [52] of the image point $p_1(t)$, denoted by $p_e(t)$, seen when the camera is at a pose $\mathcal{F}_s(t)$ with respect to the viewed object are defined as

$$p_e(t) = [u_1(t), v_1(t), \ln(z_1(t))]^T, \quad (45)$$

where $\ln(\cdot)$ denotes the natural logarithm. The constant vector $p_e^* \in \mathbb{R}^3$ denotes the extended image coordinates of the corresponding image point seen in a reference image taken when the object is at a known reference pose \mathcal{F}_s^* , and is defined as

$$p_e^* = [u_1^*, v_1^*, \ln(z_1^*)]^T. \quad (46)$$

If \mathcal{F}_s^* is not known, the methods of Subsect. “Estimating the Pose of a Planar Object Relative to the Camera” can be used to estimate \mathcal{F}_s^* given a known geometric length.

The Homography Matrix $H(t)$ can be solved from the images of the object taken at \mathcal{F}_s^* and $\mathcal{F}_s(t)$, and can be decomposed to give $R(t)$ and $\alpha_1(t)$. The translation of the object, denoted by $e_v(t) \in \mathbb{R}^3$, is defined as

$$e_v = p_e - p_e^*. \quad (47)$$

The first two elements of $e_v(t)$ are directly measured from the images, and

$$\ln(z_1(t)) - \ln(z_1^*) = -\ln(\alpha_1)$$

where $\alpha_1(t)$ is reconstructed from the homography. The time derivative of (47), is given by

$$\dot{e}_v = \dot{p}_e = \frac{\alpha_1}{z_1^*} A_e L_v [v_s - R[s_1]_{\times} R^T \omega_s] \quad (48)$$

where s_1 denotes the known, constant coordinates of feature point 1 in $\mathcal{F}_s(t)$. In (48), $A_e \in \mathbb{R}^{3 \times 3}$ is defined as

$$A_e = A - \begin{bmatrix} 0 & 0 & u_0 \\ 0 & 0 & v_0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (49)$$

the Jacobian-like matrix $L_v(t) \in \mathbb{R}^{3 \times 3}$ is defined as

$$L_v = \begin{bmatrix} 1 & 0 & -\frac{x_1}{z_1} \\ 0 & 1 & -\frac{y_1}{z_1} \\ 0 & 0 & 1 \end{bmatrix}, \quad (50)$$

and $R(t)$ is reconstructed from the Homography Matrix.

The rotation of the object at $\mathcal{F}(t)$, with respect to the fixed coordinate system \mathcal{F}^* , is defined using the angle-axis representation of the rotation $R(t)$ as

$$e_\omega \triangleq u(t)\theta(t). \quad (51)$$

In (51), $u(t) \in \mathbb{R}^3$ is a unit vector giving the axis of rotation, and $\theta(t) \in \mathbb{R}$ denotes the rotation angle about $u(t)$, where $\theta(t) \in (-\pi, \pi]$. The time derivative of (51) is given by

$$\dot{e}_\omega = L_\omega \omega_s, \quad (52)$$

where the Jacobian-like matrix $L_\omega(t) \in \mathbb{R}^{3 \times 3}$ is defined as

$$L_\omega = I_3 - \frac{\theta}{2} [u]_{\times} + \left(1 - \frac{\text{sinc}(\theta)}{\text{sinc}^2(\theta/2)}\right) [u]_{\times}^2. \quad (53)$$

The IBSE objective is to develop an estimator that can be used to identify the velocity of an object given by

$$v(t) = [v_s^T(t), \omega_s^T(t)]^T \in \mathbb{R}^6.$$

Combining (48) and (52), the combined linear and angular velocity error system can be determined as a function of $v(t)$ as

$$\dot{e} = [\dot{e}_v^T, \dot{e}_\omega^T]^T = Jv, \quad (54)$$

where the nonsingular Jacobian-like matrix $J(t) \in \mathbb{R}^{6 \times 6}$ is defined as

$$J = \begin{bmatrix} \frac{\alpha_1}{z_1^*} A_e L_v & -\frac{\alpha_1}{z_1^*} A_e L_v R[s_1]_{\times} R^T \\ 0 & L_\omega \end{bmatrix}. \quad (55)$$

The open-loop system in (54) can be inverted in as

$$v = J^{-1} \dot{e} \quad (56)$$

where

$$\begin{aligned} J^{-1} &= \begin{bmatrix} \frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1} & \left(\frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1}\right) \frac{\alpha_1}{z_1^*} A_e L_v R[s_1]_{\times} R^T L_\omega^{-1} \\ 0 & L_\omega^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \frac{z_1^*}{\alpha_1} L_v^{-1} A_e^{-1} & R[s_1]_{\times} R^T L_\omega^{-1} \\ 0 & L_\omega^{-1} \end{bmatrix}, \end{aligned}$$

and

$$L_{\omega}^{-1} = I_3 + \frac{\theta}{2} \text{sinc}^2\left(\frac{\theta}{2}\right) [u]_{\times} + (1 - \text{sinc}(\theta)) [u]_{\times}^2.$$

Based on (56), and the fact that $J^{-1}(t)$ exists, the velocity estimation goal can be expressed as the desire to estimate $\dot{e}(t)$ and map it to $v(t)$. The estimated pose is denoted as $\hat{e}(t)$, and the estimator error is defined as

$$\tilde{e} = e - \hat{e}. \quad (57)$$

Based on Lyapunov analysis and design methods detailed in [19] and [53], an estimate update law $\dot{\hat{e}}(t)$ is generated from the known error signal $\tilde{e}(t)$ as

$$\dot{\hat{e}} = \int_0^t (K + I_6) \tilde{e}(\tau) d\tau + \int_0^t \rho \text{sgn}(\tilde{e}) d\tau + (K + I_6) \tilde{e}(t) \quad (58)$$

where K and $\rho \in \mathbb{R}^{6 \times 6}$ are positive definite constant diagonal gain matrices, $I_6 \in \mathbb{R}^{6 \times 6}$ denotes the 6×6 identity matrix, and the notation $\text{sgn}(\tilde{e})$ denotes the standard signum function applied to each element of the vector $\tilde{e}(t)$. The estimator in (58) ensures that

$$\|\tilde{e}(t)\|, \|\dot{\tilde{e}}(t)\| \rightarrow 0 \quad \text{as } t \rightarrow \infty. \quad (59)$$

Based on (59), the relationship in (56) can be used to recover the velocity $v(t)$.

Experiments of Velocity Estimate

A simulation is presented to illustrate the performance of this type of the image-based velocity estimate. Four coplanar points are simulated undergoing a periodic motion involving translation and rotation along all directions. Figs. 23 and 24 show the true velocity, the estimated velocity and resulting estimation error.

An experiment was also performed to estimate the velocity of the truck in Subsect. “Experiments of Pose Estimation of a Single Object”. A low-pass filter was applied to the output of the nonlinear estimator to mitigate the effects of signal noise. Results of the experiment are seen in Fig. 25. The GPS data was backwards differenced to provide a comparison velocity. The truck periodically accelerated and stopped without changing direction. The camera optical axis (z-axis) and the truck were bearing approximately due North. It is expected that image-based estimate of the truck velocity in direction of the camera z-axis should be very close to the GPS-based estimate of velocity North. Indeed the velocity estimation is in close agreement with the GPS estimate, though the effects of signal noise are prevalent as the vehicle moves farther away.

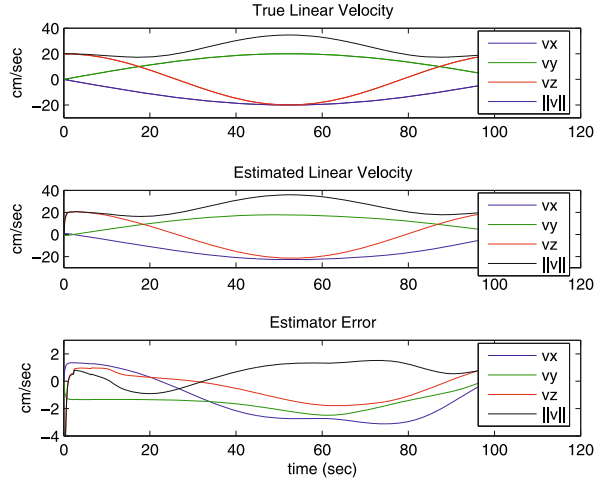


Image Based State Estimation, Figure 23

Linear velocity estimation results for simulation of nonlinear velocity observer

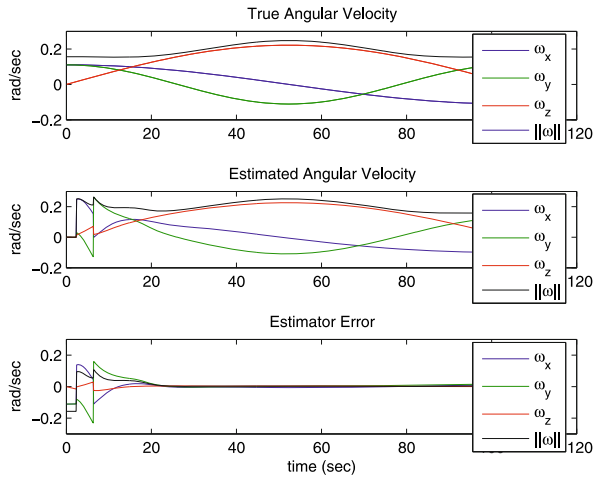


Image Based State Estimation, Figure 24

Angular velocity estimation results for simulation of nonlinear velocity observer

Image Based Kalman Filter Estimation

The KF and its variations enjoy widespread use as state estimators in many fields of engineering and science. Described simply, the KF is a linear, recursive estimator. For linear systems, the KF is an optimal estimator that gives the estimate with the least error covariance. For nonlinear systems (such as IBSE), the EKF provides an effective estimate by linearizing the state equations. A brief review of the KF/EKF is provided to introduce the notation. References for further insight into KF/EKF are provided in [54,55].

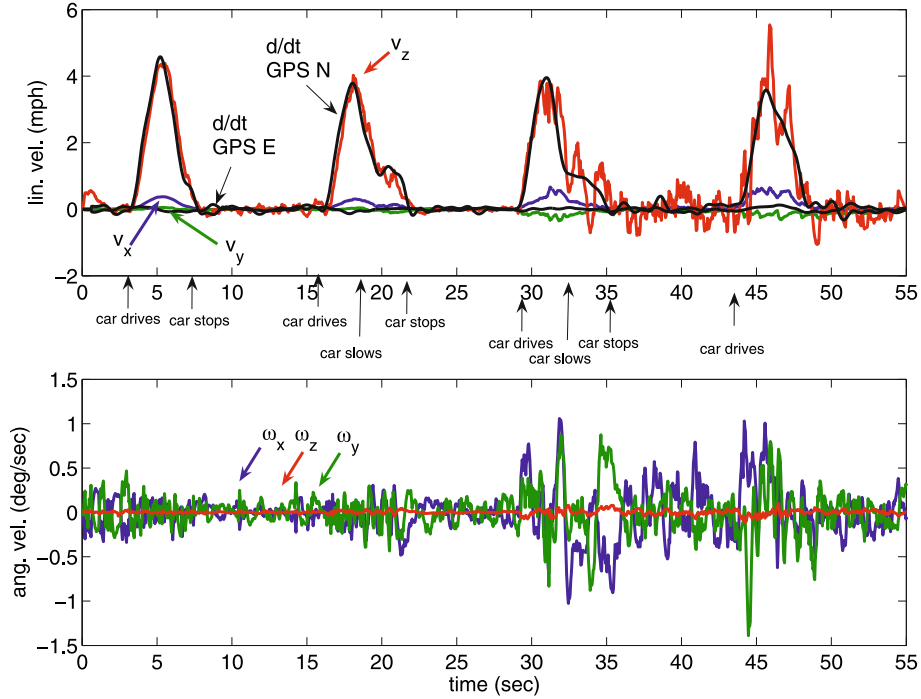


Image Based State Estimation, Figure 25
Estimated velocity of a moving vehicle

Given its relative simplicity and prominence, it is no surprise that the KF/EKF has been widely used in IBSE. KF/EKF approaches also explicitly account for signal/measurement noise, in contrast to the methods presented in Sect. “Linear Estimation of Position and Orientation from Two Images”. Some IBSE approaches use pose estimation methods as in Sect. “Linear Estimation of Position and Orientation from Two Images” as an initial pose measurement that is supplied to the KF/EKF for further pose and/or velocity estimation. The advantage of this solving for the pose estimate as an input to the KF/EKF is that the system can be expressed in the linear form

$$\dot{x} = Ax,$$

thus the estimate is expected to be optimal. For example, Soatto et al. [14] give two Kalman filter approaches based on the Essential Matrix. The first approach solves for motion from the Essential Matrix and passes the estimation through a Kalman filter. The second approach solves for the Essential Matrix and passes the result as a measurement into a Kalman filter to be refined. However, the space of Essential Matrices is not a vector space (e.g., the sum of two Essential Matrices is not necessarily an Essential Matrix), so the output of the Kalman filter is projected to space of Essential Matrices before being decomposed to find the rotation and translation. Other methods

eschew external pose estimation schemes and use image features as inputs to the KF/EKF with pose and/or velocity as output [15,16,18]. The advantage of this approach is lower computational cost and avoiding the restrictions and limitations of the pose reconstruction methods detailed in Subsect. “Characteristics of the Essential Matrix Algorithm and Planar Homography Matrix”.

The state to be estimated is denoted by the vector $x(t) \in \mathbb{R}^n$ and consists of pose, orientation, and velocity signals. The system output (i.e., the signals available for measurement) consists of image information, such as feature points, and is denoted by $y(t) \in \mathbb{R}^l$. The system is expressed as a discrete-time nonlinear system as [15]

$$x[k+1] = f(x[k], w[k], k) \quad (60)$$

$$y[k] = h(x[k], v[k], k) \quad (61)$$

where $w[k]$ and $v[k]$ are uncorrelated, white, zero-mean noise signals with known covariance matrices $Q[k]$ and $R[k]$, respectively. There is varying notation and terminology associated with the KF/EKF. This chapter denotes the estimated state by $\hat{x}[k]$, the time update or a priori estimate as $\hat{x}^-[k]$, and measurement update or a posteriori estimate as $\hat{x}^+[k]$. Likewise, the a priori and a posteriori error covariance matrices are denoted by $P^-[k]$ and $P^+[k]$, respectively.

The a priori state estimates at time k are given by

$$\hat{x}^-[k] = f_{k-1}(\hat{x}[k], 0, k) \quad (62)$$

$$P^-[k] = F[k]P^+[k-1]F^T[k] + L[k]Q[k-1]L^T[k], \quad (63)$$

where

$$F[k] \in \mathbb{R}^{n \times n} = \frac{\partial f}{\partial x} \Big|_{x=\hat{x}^+[k-1]} \quad \text{and}$$

$$L[k] \in \mathbb{R}^{n \times n} = \frac{\partial f}{\partial w} \Big|_{x=\hat{x}^+[k-1]}.$$

The a posteriori state estimates at time k are given by

$$K[k] = P^-[k]H^T[k] \cdot (K[k]P^-[k]H^T[k] + M[k]R[k]M^T[k])^{-1} \quad (64)$$

$$= \hat{x}^-[k] + K[k] (y[k] - h(\hat{x}^-[k], 0, k)) \quad (65)$$

$$P^+[k] = (I - K[k]H[k])P^-[k](I - K[k]H[k])^T + K[k]L[k]R[k]L^T[k]K[k] \quad (66)$$

where $K[k]$ is called the *Kalman Gain*, and

$$H[k] \in \mathbb{R}^{l \times n} = \frac{\partial h}{\partial x} \Big|_{x=\hat{x}^-[k]} \quad \text{and}$$

$$L[k] \in \mathbb{R}^{l \times l} = \frac{\partial h}{\partial v} \Big|_{x=\hat{x}^-[k]}.$$

Assuming the camera is stationary and viewing a rigid body undergoing general rigid body motion, a reference frame \mathcal{F}_s is attached to the object with an origin at a point $\tilde{m}_0(t) = [x_0(t), y_0(t), z_0(t)] \in \mathbb{R}^3$, which is not necessarily visible (e. g., it can be in the interior of the rigid body). Translation and linear velocity $v(t)$ is measured as the displacement of $\tilde{m}_0(t)$, and orientation is measured as the orientation $R \in \text{SO}(3)$ between the inertial camera frame and body frame. Broida et al. [15] choose to represent rotation in terms of unit quaternions, which can be mapped to rotation matrix $R(t)$ as in [56]. A unit quaternion is a four element vector denoted as

$$q(t) \triangleq [q_0(t)q_v^T(t)]^T \in \mathbb{R}^4,$$

where $q_0(t) \in \mathbb{R}$, $q_v(t) \in \mathbb{R}^3$, and $\|q(t)\| = 1$. The unit quaternion is computed from a rotation matrix $R(t)$ as

$$q_0 = \frac{1}{2}\sqrt{1 + \text{tr}(R)} \quad q_v = \frac{1}{2}u\sqrt{3 - \text{tr}(R)}, \quad (67)$$

with the reverse mapping given by

$$R(q) = (q_0^2 - q_v^T q_v)I_3 + 2q_v q_v^T - 2q_0[q_v]_{\times}. \quad (68)$$

To facilitate the subsequent development, let the notation $R_{xj}(q)$, $R_{yj}(q)$, $R_{zj}(q)$ denote the first, second, or third row of $(1/z_j(t))R(q)\tilde{m}_j(t)$. For instance

$$R_{xj} = (q_1^2 - q_2^2 - q_3^2 - q_4^2) \frac{x_j}{z_0} + 2(q_1 q_2 - q_3 q_4) \frac{y_j}{z_0} + 2(q_1 q_3 + q_2 q_4) \frac{z_j}{z_0}.$$

The angular velocity is given by the vector $\omega(t)$, and the derivative of the unit quaternion is given as a function of angular velocity as

$$\dot{q} = \Omega q$$

$$\Omega = \frac{1}{2} \begin{bmatrix} 0 & \omega_z & -\omega_y & \omega_x \\ -\omega_z & 0 & \omega_x & \omega_y \\ \omega_y & -\omega_x & 0 & \omega_z \\ -\omega_x & -\omega_y & -\omega_z & 0 \end{bmatrix}.$$

The state vector for the KF-based IBSE is denoted by $x[k] \in \mathbb{R}^{12+3N}$ and defined as

$$x[k] = \left[m_0^T, \frac{v^T}{z_0}, q^T, \omega^T, \frac{\tilde{m}_1^T}{z_0}, \dots, \frac{\tilde{m}_N^T}{z_0} \right]^T \quad (69)$$

where

$$\tilde{m}_j(t) = [x_j(t), y_j(t), z_j(t)]^T, \quad j \in \{1 \dots N\}$$

for N points, and $m_0(t) \in \mathbb{R}^2$ is the image of the origin given by

$$m_0 = \left[\frac{x_0}{z_0}, \frac{y_0}{z_0} \right]^T.$$

The resulting linear state-space systems can now be expressed as

$$x[k+1] = f(x[k], 0, k) = F[k]x[k]$$

where $F[k]$ is given by

$$F = \begin{bmatrix} F_1 & 0 & 0 \\ 0 & F_2 & 0 \\ F_3 & 0 & F_4 \end{bmatrix} \in \mathbb{R}^{12+3N \times 12+3N}$$

$$F_1 = \begin{bmatrix} -\frac{v_z}{z_0} & 0 & 1 & 0 & -\frac{x_0}{z_0} \\ 0 & -\frac{v_z}{z_0} & 0 & 1 & -\frac{y_0}{z_0} \\ 0 & 0 & -\frac{v_z}{z_0} & 0 & -\frac{v_x}{z_0} \\ 0 & 0 & 0 & -\frac{v_z}{z_0} & -\frac{v_y}{z_0} \\ 0 & 0 & 0 & 0 & -2\frac{v_z}{z_0} \end{bmatrix} \in \mathbb{R}^{5 \times 5}$$

$$\begin{aligned}
F_2 &= \begin{bmatrix} 0 & \omega_z & -\omega_y & \omega_x \\ -\omega_z & 0 & \omega_x & \omega_y \\ \omega_y & -\omega_x & 0 & \omega_z \\ -\omega_x & -\omega_y & -\omega_z & 0 \\ 0_{3 \times 1} & 0_{3 \times 1} & 0_{3 \times 1} & 0_{3 \times 1} \\ -x_1 & q_4 & -q_3 & q_2 \\ -x_2 & q_3 & q_4 & -q_1 \\ -x_3 & -q_2 & q_1 & q_4 \\ -x_4 & -q_1 & -q_2 & -q_3 \\ 0_{3 \times 1} & 0_{3 \times 1} & 0_{3 \times 1} & 0_{3 \times 1} \end{bmatrix} \in \mathbb{R}^{8 \times 7} \\
F_3 &= \begin{bmatrix} 0_{3 \times 4} & \frac{\dot{m}_1^T}{z_0} \\ 0_{3 \times 4} & \vdots \\ 0_{3 \times 4} & \frac{\dot{m}_N^T}{z_0} \end{bmatrix} \in \mathbb{R}^{3N \times 5} \\
F_4 &= -\frac{v_z}{z_0} I_{3N} \in \mathbb{R}^{3N \times 3N} \quad (70)
\end{aligned}$$

where $0_{a \times b}$ is a zero matrix of size $a \times b$ and I_{3N} is a $3N \times 3N$ identity matrix.

The measurement

$$y[k] = h(x[k]) = [m_1^T, \dots, m_N^T]^T$$

is the collection of feature points where the linearization of $h(x[k])$ is denoted by $H[k]$ and given by

$$H = \begin{bmatrix} G_1 & 0 & W_1 & 0 & S_1 & 0 & \dots & 0 \\ G_2 & 0 & W_2 & 0 & 0 & S_2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ G_N & 0 & W_N & 0 & 0 & 0 & \dots & S_N \end{bmatrix} \quad (71)$$

$\in \mathbb{R}^{2N \times 3N}$.

In (71), $G_j(q) \in \mathbb{R}^{2 \times 2}$ is an auxiliary variable defined as

$$G_j = \begin{bmatrix} \frac{1}{1+R_{zj}} & 0 \\ 0 & \frac{1}{1+R_{zj}} \end{bmatrix},$$

$W_j(q) \in \mathbb{R}^{2 \times 4}$ is given elementwise by

$$\begin{aligned}
W_{j1a} &= \frac{1}{(1+R_{zj})^2} \left[(1+R_{zj}) \frac{\partial R_{xj}}{\partial q_a} - (x_0 + R_{zj}) \frac{\partial R_{zj}}{\partial q_a} \right] \\
&\quad a \in \{1, \dots, 4\} \\
W_{j2a} &= \frac{1}{(1+R_{zj})^2} \left[(1+R_{zj}) \frac{\partial R_{yj}}{\partial q_a} - (x_0 + R_{zj}) \frac{\partial R_{zj}}{\partial q_a} \right] \\
&\quad a \in \{1, \dots, 4\}
\end{aligned}$$

and $S_j(q) \in \mathbb{R}^{2 \times 3}$ is given elementwise by

$$S_{jba} = \frac{1}{(1+R_{zj})^2} \left[(1+R_{zj}) R_{ba} - (x_0 + R_{zj}) R_{ba} \right],$$

$b \in \{1, 2\}, \quad a \in \{1, 2, 3\}.$

The definitions of $x[k]$, $F[k]$, $H[k]$ in (69)–(71) can then be used in the recursive estimator defined in equations (62)–(66).

Range Identification

In the previous sections, the distance to the feature points is a pervasive uncertainty in the system. This section provides an example of how the range to feature points can also be estimated. The problem in range identification is to determine the depth of a set of feature points. That is, given a set of image points $m_j(t) = [x_j/z_j, y_j/z_j, 1]^T$, the task in range identification is to determine the terms $z_j(t)$ and thus recover $\tilde{m}(t) = [x_j, y_j, z_j]^T$. In general, additional information about the target or camera motion is required to perform range identification. For example, the methods in this section require known velocity estimation of the camera or viewed target. An example problem is a mobile robot or UAV, where a combination of joint encoders, accelerometers, or inertial measurement units and a camera can be used to determine the velocity and pose relative to objects in the environment.

Several researchers have investigated the range identification problem for conventional imaging systems when the motion parameters are known. In [20], Jankovic and Ghosh developed a discontinuous observer, to exponentially identify range information of features from successive images of a camera where the object model is based on known skew-symmetric affine motion parameters. In [21], Chen and Kano generalized the object motion beyond the skew-symmetric form of [20] and developed a new discontinuous observer that exponentially forced the state observation error to be uniformly ultimately bounded. More recently, a state estimation strategy was developed in [12,24] for affine systems with known motion parameters where only a single homogeneous observation point is provided (i. e., a single image coordinate). In [57], a reduced order observer was developed to yield a semi-global asymptotic stability result for a fixed camera viewing a moving object with known motion parameters for a pinhole camera. In [23], a continuous observer is used to asymptotically identify the range information for a general affine system with known motion parameters. Recent efforts in range estimation (e. g., [12,13]) have extended this concept for use with omnidirectional cameras (i. e., cameras that use curved mirrors to achieve a 360° field of view) and uncertain motion. This section focuses on the case in where a continuous observer is used to asymptotically identify the range information for a general affine system with known motion parameters, using a pinhole camera as in [23].

Nonlinear Estimator for Range Identification

For the problems in this section, consider the scenario of a moving target with a stationary camera or the scenario of moving camera with a stationary target. In both scenarios, the relative motion dynamics can be written in the same form, although the motion parameters have different physical meanings. Specifically, for both scenarios, the affine motion dynamics can be expressed as

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{z} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \quad (72)$$

where $x(t), y(t), z(t) \in \mathbb{R}$ denote the unmeasurable task-space coordinates of the point in an inertial frame attached to the camera. The parameters $a_{i,j}(t) \in \mathbb{R}$ and $b_i(t) \forall i, j = 1, 2, 3$ denote the known motion parameters. To illustrate how the same dynamics apply to both scenarios presented, consider the case of a moving target with linear and angular velocity $v_t(t)$ and $\omega_t(t)$ measured with respect to the camera frame. A point attached to the object has coordinates $\tilde{m}(t)$ in the camera frame, with relative motion to the camera described by

$$\dot{\tilde{m}} = -[\omega_t]_{\times} \tilde{m} - v_t = A\tilde{m} + b. \quad (73)$$

For the moving camera stationary object scenario, consider a feature point attached to a stationary object. The linear and angular velocities of the target with respect to the camera (expressed in \mathcal{F}_c) can be written as

$$v_t = -Rv_c \quad \omega_t = -R\omega_c \quad (74)$$

where $R(t) \in \text{SO}(3)$ denotes the corresponding rotation between the camera and object frame, and $v_c(t)$ and $\omega_c(t)$ denote the linear and angular velocity of the camera, respectively. Based on (74), the relationship in (73) can be rewritten as

$$\dot{\tilde{m}} = [R\omega_c]_{\times} \tilde{m} + Rv_c = A\tilde{m} + b. \quad (75)$$

The normalized Euclidean coordinates of a feature point (which are measurable from the image-space through the pin-hole model in (5)), are denoted by $m(t) \in \mathbb{R}^2$, and defined as

$$m \triangleq [m_1, m_2]^T = \left[\frac{x}{z}, \frac{y}{z} \right]^T. \quad (76)$$

The affine dynamics introduced in (72), and the image-space signal introduced in (76) define the perspective system [21]. After taking the time derivative of (76) and utilizing (72), the image-space trajectory of the object feature

can be obtained as

$$\dot{m}_1 = \frac{a_{11}x + a_{12}y + a_{13}z + b_1}{z} - \frac{x(a_{31}x + a_{32}y + a_{33}z + b_3)}{z^2} \quad (77)$$

$$\dot{m}_2 = \frac{a_{21}x + a_{22}y + a_{23}z + b_2}{z} - \frac{y(a_{31}x + a_{32}y + a_{33}z + b_3)}{z^2}. \quad (78)$$

To facilitate subsequent analysis, the time derivative of the inverse of $z(t)$ is determined as

$$\frac{d}{dt} \left(\frac{1}{z} \right) = \frac{-a_{31}x - a_{32}y - a_{33}z - b_3}{z^2}. \quad (79)$$

By utilizing (76), the expressions given in (77)–(79) can be rewritten as

$$\dot{m}_1 = a_{13} + (a_{11} - a_{33})m_1 + a_{12}m_2 - a_{31}m_1^2 - a_{32}m_1m_2 + f_1 \quad (80)$$

$$\dot{m}_2 = a_{23} + a_{21}m_1 + (a_{22} - a_{33})m_2 - a_{32}m_2^2 - a_{31}m_1m_2 + f_2 \quad (81)$$

$$\frac{d}{dt} \left(\frac{1}{z} \right) = -\frac{1}{z}(a_{31}m_1 + a_{32}m_2 + a_{33}) - \frac{b_3}{z^2} \quad (82)$$

where $f_1(z, m_1), f_2(z, m_2) \in \mathbb{R}$ are unmeasurable signals¹ defined as

$$f_1 \triangleq \frac{1}{z}(b_1 - b_3m_1) \quad (83)$$

$$f_2 \triangleq \frac{1}{z}(b_2 - b_3m_2). \quad (84)$$

The IBSE objective in this section is to determine the unmeasurable state $z(t)$ of the perspective vision system described by (72) and (76). From (76) and the fact that $m_1(t)$ and $m_2(t)$ are measurable, it is clear that if $z(t)$ is identified, then the complete Euclidean coordinate of the feature can be determined. To achieve this objective, an observer is constructed based on the unmeasurable image-space dynamics for $m(t)$. To quantify the performance of the observer, a measurable observer estimation error signal, denoted by $e(t) \in \mathbb{R}^2$, is defined as

$$e \triangleq [e_1, e_2]^T = [m_1 - \hat{m}_1, m_2 - \hat{m}_2]^T, \quad (85)$$

where $\hat{m}(t) \triangleq [\hat{y}_1(t), \hat{y}_2(t)]^T \in \mathbb{R}^2$ denotes a subsequently designed observer signal. To facilitate the ob-

¹The signals $f_1(z, m_1)$, and $f_2(z, m_2)$ are unmeasurable due to a dependence on the unmeasurable state $z(t)$

server design, a filtered observation error signal, denoted by $r(t) \in \mathbb{R}^2$, is designed as

$$r \triangleq [r_1, r_2]^T = [\dot{e}_1 + \alpha_1 e_1, \dot{e}_2 + \alpha_2 e_2]^T, \quad (86)$$

where $\alpha_1, \alpha_2 \in \mathbb{R}$ denote positive constant gains. Based on the dynamics in (80) and (81) and the definitions introduced in (85) and (86), it is clear that $r(t)$ is unmeasurable due to the fact that $\dot{m}(t)$ is a function of the unmeasurable disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$. The subsequent development will follow the strategy given in [23], where the design estimates $\hat{f}_1(z, m_1)$ and $\hat{f}_2(z, m_2)$ based on the strategy that if the mismatch between the estimates and the disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$ can be driven to zero, then $z(t)$ can be identified by exploiting the fact that $b_i(t) \forall i = 1, 2, 3$ and the states $m_1(t)$ and $m_2(t)$ are measurable. Specifically, from (83) and (84), the inverse of the square of $z(t)$ can be determined as

$$\left(\frac{1}{z}\right)^2 = \frac{f_1^2 + f_2^2}{(b_1 - b_3 m_1)^2 + (b_2 - b_3 m_2)^2}. \quad (87)$$

For (87) to be valid, it is clear that the following observability condition must be satisfied

$$(b_1 - b_3 m_1)^2 + (b_2 - b_3 m_2)^2 > 0. \quad (88)$$

That is, $z(t)$ can be identified once the mismatch between the disturbance terms $f_1(z, m_1)$ and $f_2(z, m_2)$ and the respective estimates are driven to zero.

By taking the time-derivative of (85) the following error dynamics can be obtained for $e(t)$

$$\dot{e} = \dot{y} - \dot{\hat{y}}. \quad (89)$$

Based on the structure of (80), (81), and (89), the elements of the observer signal $\hat{y}(t)$ are designed as [23]

$$\begin{aligned} \dot{\hat{m}}_1 = & a_{13} + (a_{11} - a_{33})m_1 + a_{12}m_2 \\ & - a_{31}m_1^2 - a_{32}m_1m_2 + \hat{f}_1 \end{aligned} \quad (90)$$

$$\begin{aligned} \dot{\hat{m}}_2 = & a_{23} + a_{21}m_1 + (a_{22} - a_{33})m_2 \\ & - a_{32}m_2^2 - a_{31}m_1m_2 + \hat{f}_2 \end{aligned} \quad (91)$$

$$\dot{\hat{f}}_1 = -(k_{s1} + \alpha_1)\hat{f}_1 + \gamma_1 \text{sgn}(e_1) + \alpha_1 k_{s1} e_1 \quad (92)$$

$$\dot{\hat{f}}_2 = -(k_{s2} + \alpha_2)\hat{f}_2 + \gamma_2 \text{sgn}(e_2) + \alpha_2 k_{s2} e_2. \quad (93)$$

Under the assumptions that $a_{i,j}(t)$, $b_i(t) \forall i, j = 1, 2, 3$ are known motion parameters; the measured input signals $m_1(t)$ and $m_2(t)$, and that the unmeasurable state $1/z(t)$ are all bounded functions of time; that the parameters $a_{i,j}(t)$ are first order differentiable; and that the parameters $b_i(t)$ are second order differentiable, then the estimate $\hat{z}(t)$ will asymptotically converge to the true value $z(t)$.

Future Directions

The field of IBSE is still an evolving area that has achieved prominence only in the past twenty years. During this time, there have been several major shifts in research, including linear estimation methods, Kalman filtering methods and, more recently, nonlinear estimator/observer methods. However, as new methods are introduced, previous methods continue to be popular and widely used in a variety of problems. It seems likely that these methods will continue to be used, rather than giving way to any entirely new methodology.

There are several open problems attracting attention, mostly rooted in the limitations of IBSE. One open area includes camera field of view issues and moving beyond the pinhole camera model. These two problems are linked in the use of omnidirectional cameras, which have received recent attention. A second open area involves the use of multiple moving cameras and moving objects. Using methods discussed in this chapter, each camera can estimate the relative pose and velocity between itself and visible objects in its FOV, including other cameras and/or the moving platforms (e.g., a vehicle or robot) on which they are mounted. Given overlapping camera FOV's or cameras visible to each other, and communication between cameras, it is possible to relate relative pose of all objects and cameras, even if they are not in the field of view. However, it is uncertain how to maintain such a network of state estimates or how errors might propagate through such a network.

There is also research to reduce the necessary a priori knowledge needed to perform IBSE. Due to the loss of depth information in imaging, some a priori information will always be necessary to recover the correct scale of size and/or motion. Nevertheless, it is desirable to require as little extra information as possible. For example, there are efforts to eliminate complete knowledge of the body dynamics for the range identification problem discussed in Sect. "Range Identification", reducing the necessary knowledge to a subset of the motion parameters.

Bibliography

1. Kruppa E (1913) Zur Ermittlung eines Objektes aus zwei Perspektiven mit innerer Orientierung. Hölder, Wien
2. Longuet-Higgins H (1981) A computer algorithm for reconstructing a scene from two projections. *Nature* 293:133–135
3. Huang T, Faugeras O (1989) Some properties the e matrix in two-view motion estimation. *IEEE Trans Pattern Anal Mach Intell* 11:1310–1312
4. Philip J (1996) A non-iterative algorithm for determining all essential matrices corresponding to five point pairs. *Photogramm Rec* 15:589–99

5. Nister D (2004) An efficient solution to the five-point relative pose problem. *IEEE Trans Pattern Anal Mach Intell* 26:756–770
6. Faugeras O, Lustman F (1988) Motion and structure from motion in a piecewise planar environment. *Int J Pattern Recognit Artif Intell* 2:485–508
7. Spetsakis M, Aloimonos J (1989) Optimal motion estimation. *Proc Workshop Vis Motion*, Irvine, pp 229–237
8. Weng J, Ahuja N, Huang T (1993) Optimal motion and structure estimation. *IEEE Trans Pattern Anal Mach Intell* 15:864–884
9. Ma Y, Košecák J, Sastry S (2001) Optimization criteria and geometric algorithms for motion and structure estimation. *Int J Comput Vis* 44:219–249
10. Caballero F, Merino L, Ferruz J, Ollero A (2006) Improving vision-based planar motion estimation for unmanned aerial vehicles through online mosaicing. *Proc IEEE Int Conf Robotics Autom*, Orlando, pp 2860–2865
11. Kaiser K, Gans N, Dixon W (2007) Localization and control an aerial vehicle through chained, vision-based pose reconstruction. *Proc American Control Conference*, New York, pp 5934–5939
12. Ma L, Chen Y, Moore KL (2005) Range identification for perspective dynamic systems with 3D imaging surfaces. *Proc American Control Conference*, Portland, pp 3671–3675
13. Gupta S, Aiken D, Hu G, Dixon WE (2006) Lyapunov-based range and motion identification for a nonaffine perspective dynamic system. *Proc American Control Conference*, Minneapolis, pp 4471–4476
14. Soatto S, Frezza R, Perona P (1996) Motion estimation via dynamic vision. *IEEE Trans Automat Control* 41:393–413
15. Broida T, Chellappa R (1991) Estimating the kinematics and structure a rigid object from a sequence monocular images. *IEEE Trans Pattern Anal Mach Intell* 13:497–513
16. Azarbayejani A, Pentland AP (1995) Recursive estimation motion, structure, and focal length. *IEEE Trans Pattern Anal Mach Intell* 17:562–575
17. Kano H, Ghosh BK, Kanai H (2001) Single camera based motion and shape estimation using extended kalman filtering. *Math Comput Model* 34:511–525
18. Chiuso A, Favaro P, Jin H, Soatto S (2002) Structure from motion causally integrated over time. *IEEE Trans Pattern Anal Mach Intell* 24:523–535
19. Chitrakaran V, Dawson DM, Dixon WE, Chen J (2005) Identification a moving object's velocity with a fixed camera. *Automatica* 41:553–562
20. Jankovic M, Ghosh B (1995) Visually guided ranging from observations points, lines and curves via an identifier based non-linear observer. *Syst Control Lett* 25:63–73
21. Chen X, Kano H (2002) A new state observer for perspective systems. *IEEE Trans Automat Contr* 47:658–663
22. Ma L, Chen Y, Moore K (2004) Range identification for perspective dynamic systems using linear approximation. *Proc IEEE Int Conf Robotics and Automation*, New Orleans, pp 1658–1663
23. Dixon WE, Fang Y, Dawson DM, Flynn TJ (2003) Range identification for perspective vision systems. *IEEE Trans Automat Contr* 48:2232–2238
24. Ma L, Chen Y, Moore KL (2004) Range identification for perspective dynamic system with single homogeneous observation. *Proc IEEE Int Conf Robotics and Automation*, New Orleans, pp 5207–5212
25. Weng J, Ahuja N, Huang T (1992) *Motion and Structure from Image Sequences*. Springer, New York
26. Ma Y, Soatto S, Kosecká J, Sastry S (2004) *An Invitation to 3-D Vision*. Springer, New York
27. Faugeras O (1993) *Three-Dimensional Computer Vision*. MIT Press, Cambridge
28. Hartley R, Zisserman A (2003) *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge
29. Bouguet J Complete camera calibration toolbox for matlab. http://www.vision.caltech.edu/bouguetj/calib_doc/index.html
30. Sepp W, Fuchs S The dlr camera calibration toolbox. <http://www.dlr.de/rm/desktopdefault.aspx/tabid-1524/>
31. Tsai R (1987) A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Trans Robotics Automat* 3:323–344
32. Tsai R (1989) Synopsis recent progress on camera calibration for 3D machine vision. MIT Press, Cambridge
33. Robert L (1996) Camera calibration without feature extraction. *CVIU: Comput Vis Image Underst* 63:314–325
34. Harris C, Stephens M (1988) A combined corner and edge detector. *Alvey Vision Conference*, vol 15, pp 247–251
35. Smith S, Brady J (1997) Susan – a new approach to low level image processing. *Int J Comput Vis* 23:45–78
36. Lowe D (1999) Object recognition from local scale-invariant features. *Proc IEEE Int Conf Computer Vision*, Toronto, pp 1150–1157
37. Beucher S, Lantuejoul C (1979) Use watersheds in contour detection. *Proc Int Workshop on Image Processing, Real-Time Edge and Motion Detection/Estimation*, Rennes
38. Niethammer M, Tannenbaum A, Angenent S (2006) Dynamic active contours for visual tracking. *IEEE Trans Automat Control* 51:562–579
39. Canny J (1986) A computational approach to edge detection. *IEEE Trans Pattern Anal Mach Intell* 8:679–698
40. Ballard D (1987) *Generalizing the Hough transform to detect arbitrary shapes*. Morgan Kaufmann Publishers, San Francisco
41. Zhang Z, Hanson A (1996) 3D reconstruction based on homography mapping. *Proc ARPA Image Understanding Workshop* Palm Springs CA
42. Luong Q, Faugeras O (1996) The fundamental matrix: Theory, algorithms, and stability analysis. *Int J Comput Vis* 17:43–75
43. Boufama B, Mohr R (1995) Epipole and fundamental matrix estimation using virtual parallax. *Proc Int Conf Computer Vision*, Boston, pp 1030–1036
44. Malis E, Chaumette F (2000) 2 1/2D visual servoing with respect to unknown objects through a new estimation scheme camera displacement. *Int J Comput Vis* 37:79–97
45. Fang Y, Dixon W, Dawson D, Chawda P (2005) Homography-based visual servo regulation mobile robots. *IEEE Trans Syst Man Cybern* 35:1041–1050
46. Dupree K, Gans N, Mackunis W, Dixon W (2007) Euclidean feature tracking for a rotating satellite. *Proc American Control Conference*, New York, pp 3874–3879
47. MacKunis W, Gans N, Kaiser K, Dixon WE (2007) Unified tracking and regulation visual servo control for wheeled mobile robots. *Proc IEEE Multi-Conf Syst Control*, Singapore, pp 88–93
48. Baillard C, Zisserman A (1999) Automatic reconstruction planar models from multiple views. *Proc IEEE Conf Computer Vision and Pattern Recognition*, Toronto, pp 559–565
49. Okada K, Kagami S, Inaba M, Inoue H (2001) Plane segment finder: algorithm, implementation and applications. *Proc IEEE Int Conf Robotics and Automation*, Seoul, pp 2120–2125

50. Zhuang X, Haralick R (1984) Rigid body motion and the optical flow image. *Proc Int Conf Artificial Intelligence Applications*, Denver, pp 366–375
51. Ma Y, Kořecká J, Sastry S (2000) Linear differential algorithm for motion recovery: A geometric approach. *Int J Comput Vis* 36:71–89
52. Malis E, Chaumette F, Boudet S (1999) 2-1/2D visual servoing. *IEEE Trans Robot Automat* 15:238–250
53. Xian B, de Queiroz MS, Dawson DM (2004) A continuous control mechanism for uncertain nonlinear systems. *Optimal Control, Stabilization, and Nonsmooth Analysis. Lecture Notes in Control and Information Sciences*, vol 301. Springer, New York, pp 251–262
54. Sorenson H (1985) *Kalman Filtering: theory and application*. IEEE, New York
55. Grewal M, Andrews A (1993) *Kalman filtering: theory and practice*. Prentice-Hall, Upper Saddle River
56. Shuster M (1993) A survey attitude representations. *J Astronaut Sci* 41:439–518
57. Karagiannis D, Astolfi A (2005) A new solution to the problem range identification in perspective vision systems. *IEEE Trans Automat Control* 50:2074–2077

Immunecomputing

JON TIMMIS^{1,2}

¹ Department of Electronics, University of York,
York, UK

² Department of Computer Science, University of York,
York, UK

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[What Is an Artificial Immune System?](#)

[Current Artificial Immune Systems Biology
and Basic Algorithms](#)

[Alternative Immunological Theories for AIS](#)

[Emerging Methodologies in AIS](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Glossary based on [28]:

Affinity Measure or tightness of the binding between an antigen combining site and an antigenic determinant; the stronger the binding, the higher the affinity.

Antigen Any substance that when introduced into the body, is capable of inducing an immune response.

Antigen presenting cells (APC) B-cells, cells of the monocyte Lineage (including macrophages as well as den-

dritic cells), and various other body cells that present antigen in a form that B- and T-cells can recognize.

Antibody A soluble protein molecule produced and secreted by B-cells in response to an antigen. Antibodies are usually defined in terms of their specific binding to an antigen.

B cell White blood cells expressing immunoglobulin molecules on its surface. Also known as B-lymphocytes, they are derived from the bone marrow and develop into plasma cells that are the main antibody secretors.

Clonal selection theory A theory that states that the specificity and diversity of an immune response are the result of selection by antigen of specifically reactive clones from a large repertoire of preformed lymphocytes, each with individual specificities.

Dendritic cell Set of antigen-presenting cells (APCs) present in lymph nodes, spleen and at low levels in blood, which are particularly active in stimulating T-cells.

Lymph node Small organs of the immune system, widely distributed throughout the body and linked by lymphatic vessels.

Lymphocyte White blood cell found in blood, tissue, and in lymphoid organs.

Major histocompatibility A group of genes encoding polymorphic.

Complex (MHC) Cell-surface molecules (MHC class I and II) that are involved in controlling several aspects of the immune response. MHC genes code for self-markers on all body cells and play a major role in transplantation rejection.

Pathogen A microorganism that causes disease.

T Cell White blood cell that orchestrate and/or directly participate in the immune defenses.

Definition of the Subject

Immunecomputing, or Artificial Immune Systems (AIS), has recently emerged as a computational intelligence approach that show great promise. Inspired by the complexity of the immune system, computer scientists and engineers have created systems that in some way mimic or capture certain computationally appealing properties of the immune system, with the aim of building more robust and adaptable solutions. AIS have been defined by [28] as:

“adaptive systems, inspired by theoretical immunology and observed immune functions, principle and models, which are applied to problem solving”.

However, in order to build AIS an interdisciplinary approach is required that employs modeling of immunology (both mathematical and computational) in order to understand the underlying complexity inherent within the immune system. AIS do not rival their natural counterparts, they do not exhibit the same level of complexity or even perform the same function, but they do capture essential properties of the immune systems that are making them a competitive computational intelligence paradigm.

Introduction

The immune system is a complex system that undertakes a myriad of tasks. The abilities of the immune system have helped to inspire computer scientists to build systems that *mimic*, in some way, various properties of the immune system. This field of research, Artificial Immune Systems (AIS), has seen the application of immune inspired algorithms to problems such as robotic control [72], network intrusion detection [37,67], fault tolerance [5,16], bioinformatics [20,77] and machine learning [70,71,106], to name a few. To many, trying to mimic how the immune system operates in a computer may seem an unusual thing to do, why then would people in computing wish to do this? The answer is that, from a computational point of view, the immune system has many desirable properties that they would like their computer systems to possess. These properties are such things as robustness, adaptability, diversity, scalability, multiple interactions on a variety of timescales and so on.

The origins of AIS has its roots in the early theoretical immunology work of Farmer, Perelson and Varela [35,79,101]. These works investigated a number of theoretical immune network models proposed to describe the maintenance of immune memory in the absence of antigen. Whilst controversial from an immunological perspective, these models began to give rise to an interest from the computing community. The most influential people at crossing the divide between computing and immunology in the early days were Hugues Bersini and Stephanie Forrest. It is fair to say that some of the early work by Bersini [10,11] was very well rooted in immunology, and this is also true of the early work by Forrest [36,53]. It was these works that formed the basis of a solid foundation for the area of AIS. In the case of Bersini, he concentrated on the immune network theory, examining how the immune system maintained its memory and how one might build models and algorithms mimicking that property. With regards to Forrest, her work was focused on computer security (in particular network intrusion detection) [37,55] and formed the basis of a great deal of further research by the

community on the application of immune inspired techniques to computer security.

At about the same time as Forrest was undertaking her work, other researchers began to investigate the nature of learning in the immune system and how that might be used to create *machine learning* algorithms [19]. They had the idea that it might be possible to exploit mechanisms of the immune system (in particular the immune network) in learning systems, so they set about doing a proof of concept [19]. Initial results were very encouraging, and they built on their success by applying the immune ideas to the classification of DNA sequences as either promoter or non-promoter classes, [56] and the detection of potentially fraudulent mortgage applications [57].

The work of Hunt and Cook spawned more work in the area of immune network based machine learning over the next few years, notably in [91] where the Hunt and Cook system was totally rewritten, simplified and applied to unsupervised learning (very similar to cluster analysis). Concurrently, similar work was carried out by [31,32], who developed algorithms for use in function optimization and data clustering (the details of these are described in more details later in the chapter). The work of Timmis on machine learning spawned yet more work in the unsupervised learning domain, in trying to perform dynamic clustering (where the patterns in the input data move over time). This was met with some success in works such as [76,108]. At the same time, using ideas other than the immune network theory, work by [50] used immune inspired associative memory ideas to track moving targets in databases.

In the supervised learning domain, very little happened until work by [102] (later augmented in [106]) developed an immune based classifier known as AIRS. The system developed by Watkins was then adapted into a parallel and distributed learning system in [103], and has shown itself to be one of the real success stories of immune inspired learning [45,46,105].

In addition to the work on machine learning, there has been plenty of other activity in AIS over the years. To outline all the applications of AIS and developments over the past 10 years would take a long time, and there are some good review papers in the literature, thus the reader is directed those [23,28,39,95]. In addition to these works, [52] investigated the application areas AIS have been applied to, and considered the contribution AIS have made to these areas. Their survey of AIS is not exhaustive, but attempts to produce a picture of the general areas to which they have been applied. Of the 97 papers reviewed, 12 categories were identified to reflect the natural groupings of the papers. These were, in the order of most papers first:

clustering/classification, anomaly detection (e. g. detecting faults in engineering systems), computer security, numerical function optimization, combinatoric optimization (e. g. scheduling), learning, bioinformatics, image processing, robotics (e. g. control and navigation), adaptive control systems, virus detection and web mining. Hart and Timmis go on to note that these categories can be summarized into three general application areas of learning, anomaly detection and optimization.

Work in [88] details an alternative approach to the use of immune metaphors and present a deterministic immune network approach, that is in stark contrast to the work presented here. This work has shown to be exceptionally competitive when compared to other computational intelligence approaches [89,90]. Due to a growing amount of work conducted on AIS, the International Conference on Artificial Immune Systems (ICARIS) conference series was started in 2002¹ and has operated in subsequent years [12,58,78,93,98]. This is the best source of reference material to read in order to grasp the variety of application areas of AIS, and also the developments in algorithms and the more theoretical side of AIS.

This remaining article is organized as follows: in Sect. “[What Is an Artificial Immune System?](#)” first discuss the current perception of what AIS are, and we do this in terms of a simple engineering framework; in Sect. “[Current Artificial Immune Systems Biology and Basic Algorithms](#)” we provide a simple overview of the immunology that has served to inspire the development of immune inspired systems to date, and this is coupled with an outline of the basic algorithms that have been use in AIS to date; in Sect. “[Alternative Immunological Theories for AIS](#)” we turn our attention to alternative immunology (away from the more classic described in the previous section) and focus on the danger theory approach which exploits a fundamental different analogy from immunology that other AIS have done to date and discuss the cognitive immune paradigm as an alternative approach to the development of AIS; Sect. “[Emerging Methodologies in AIS](#)” reviews new approaches to the development of AIS in the context of a conceptual framework and finally in Sect. “[Future Directions](#)” we review some of the possible challenges and directions that AIS might follow.

What Is an Artificial Immune System?

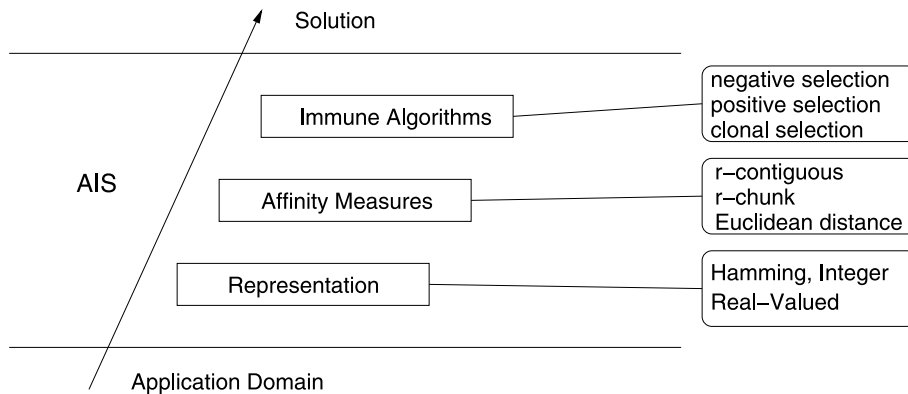
In an attempt to create a common basis for AIS, work in [28] proposed the idea of a framework for engineering AIS. They argued the case for such a framework as the existence of similar frameworks in other biologically inspired

approaches, such as artificial neural networks (ANN) and evolutionary algorithms (EAs), has helped considerably with the understanding and construction of such systems. For example, de Castro and Timmis [28] consider a set of artificial neurons, which can be arranged together to form an artificial neural network. In order to acquire knowledge, these neural networks undergo an adaptive process, known as learning or training, which alters (some of) the parameters within the network. Therefore, they argued that in a simplified form, a framework to design an ANN is composed of: a set of artificial neurons, a pattern of interconnection for these neurons, and a learning algorithm. Similarly, they argued that in evolutionary algorithms, there is a set of artificial chromosomes representing a population of individuals that iteratively suffer a process of reproduction, genetic variation, and selection. As a result of this process, a population of evolved artificial individuals arises. A framework, in this case, would correspond to the genetic representation of the individuals of the population, plus the procedures for reproduction, genetic variation, and selection. Therefore, they proposed that a framework to design a biologically inspired algorithm requires, at least, the following basic elements:

- A representation for the components of the system
- A set of mechanisms to evaluate the interaction of individuals with the environment and each other. The environment is usually simulated by a set of input stimuli, one or more fitness function(s), or other means
- Procedures of adaptation that govern the dynamics of the system, i. e., how its behavior varies over time

This framework can be thought of as a layered approach such as the specific framework for engineering AIS of [28] shown in Fig. 1. This framework follows the three basic elements for designing a biologically inspired algorithm just described, where the set of mechanisms for evaluation are the affinity measures and the procedures of adaptation are the immune algorithms. In order to build a system such as an AIS, one typically requires an application domain or target function. From this basis, the way in which the components of the system will be represented is considered. For example, the representation of network traffic may well be different than the representation of a real time embedded system. In AIS, the way in which something is represented is known as *shape space*. There are many kinds of shape space, such as Hamming, real valued and so on, each of which carries it own bias and should be selected with care [38]. Once the representation has been chosen, one or more affinity measures are used to quantify the interactions of the elements of the system. There are many possible affinity measures (which are partially dependent

¹<http://www.artificial-immune-systems.org>



Immunecomputing, Figure 1

AIS Layered Framework adapted from [28]

upon the representation adopted), such as Hamming and Euclidean distance metrics. Again, each of these has its own bias, and the affinity function must be selected with great care, as it can affect the overall performance (and ultimately the result) of the system [38]. This was also recently shown experimentally in the case of immune networks, where the affinity function affected the overall outcome of the shape of the network [49,51]. The final layer involves the use of algorithms, which govern the behavior (dynamics) of the system. Such algorithms include those based on the following immune processes: negative and positive selection, clonal selection, bone marrow, and immune network algorithms.

Current Artificial Immune Systems Biology and Basic Algorithms

The main developments within AIS, have focused on three main immunological theories: clonal selection, immune networks and negative selection. Researchers in AIS have concentrated, for the most part, on the *learning* and *memory* mechanisms of the immune system inherent in clonal selection and immune networks, and the negative selection principle for the generation of *detectors* that are capable of classifying changes in *self*. In this section, we review the immunology that has been capitalized on by the AIS community. We outline the three main immunological theories noted above that have acted as a source of inspiration. At each stage, we review a simple AIS approach that has extracted some feature from that theory. It is worth noting that, although not covered here, a large effort is currently being made in the AIS community into exploring other immune ideas and mechanisms such as danger theory and innate immunity.

Immunity

The vertebrate immune system (the one which has been used to inspire the vast majority of AIS to date) is composed of diverse sets of cells and molecules. These work in collaboration with other systems, such as the neural and endocrine, to maintain a steady state of operation within the host: this is termed *homeostasis*. The role of the immune system is typically viewed as one of protection from infectious agents such as viruses, bacteria, fungi and other parasites. On the surface of these agents are antigens that allow the identification of the invading agents (pathogens) by the immune cells and molecules, which in turn provoke an immune response. There are two basic types of immunity, innate and adaptive. Innate immunity is not directed towards specific pathogens, but against any pathogen that enter the body. The innate immune system plays a vital role in the initiation and regulation of immune responses, including adaptive immune responses. Specialized cells of the innate immune system evolved so as to recognize and bind to common molecular patterns found only in microorganisms. However, the innate immune system is by no means a complete solution to protecting the body.

Adaptive, or acquired immunity, is directed against specific invaders, with adaptive immune cells being modified by exposure to such invaders. The adaptive immune system mainly consists of lymphocytes, which are white blood cells, more specifically B and T cells. These cells aid in the process of recognizing and destroying specific substances. Any substance that is capable of generating such a response from the lymphocytes is called an antigen or immunogen. Antigens are not the invading microorganisms themselves; they are substances such as toxins or enzymes in the microorganisms that the immune system

considers foreign. Adaptive immune responses are normally directed against the antigen that provoked them and are said to be antigen-specific.

Natural Clonal Selection

The clonal selection theory (CST) [15] is the theory used to explain the basic response of the adaptive immune system to an antigenic stimulus. It establishes the idea that only those cells capable of recognizing an antigenic stimulus will proliferate, thus being selected against those that do not. Clonal selection operates on both T cells and B cells. In the case of B cells, when their antigen receptors (antibodies) bind with an antigen, the B cell becomes activated and begins to proliferate producing new B cell clones that are an exact copy of the parent B cell. The clones then undergo somatic hypermutation and produce antibodies that are specific to the invading antigen [9]. After proliferation, B cells differentiate into *plasma cells* or long-lived B *memory cells*. Plasma cells produce large amounts of *antibodies* which will attach themselves to the antigen and act as a type of *tag* for other immune cells to pick up on and remove from the system. This whole process is known as *affinity maturation*.

Memory cells help the immune system to be protective over periods of time. In the normal course of the evolution of the immune system, an organism would be expected to encounter a given antigen repeatedly during its lifetime. The initial exposure to an antigen that stimulates an adaptive immune response is handled by a small number of B cells, each producing antibodies of different affinity. Storing some high affinity antibody producing cells (memory cells) from the first infection, so as to form a large initial specific B cell sub-population for subsequent encounters, considerably enhances the effectiveness of the immune response to secondary encounters. Such a strategy ensures that both the speed and accuracy of the immune response becomes successively stronger after each infection.

Autoimmunity is the term used to describe the existence of antigen receptors that recognize the body's own molecules, or self-antigens. According to the CST, immune specificity is a property of immune receptors. When a non-self antigen is detected, a suitable immune response is elicited and the antigen is destroyed. Thus, the recognition of self-antigen is forbidden, and self-reacting receptors must be deleted.

Artificial Clonal Selection

Work in [30,32] proposes an optimization algorithm, known as CLONALG, inspired by the clonal selection process, as outlined in the previous section. Given a func-

tion F , a population of candidate solutions (antibodies) are evolved to either minimize or maximize the function. Each member of this population is a vector, in a certain shape space, which maps values to the parameters of the function F . CLONALG exploits the cloning, mutation and selection mechanisms of clonal selection, to effectively evolve a set of memory cells that contain candidate solutions to the function F .

CLONALG operates via the following procedure. A population P is initialized with random vectors, where P is set of candidate solutions for the given function. Each member of P is evaluated against the function, and the highest affinity n number are selected for cloning, where affinity can be measured as the distance to the optimal value. Clones are produced at a rate proportional to the affinity (so the better the affinity, the more clones are produced). Each clone is subject to a mutation rate, which is inversely proportional to the affinity. These clones are added to P and then the n highest affinity are selected to remain in the population. A number of low affinity members are then removed from the population and replaced with the same number of randomly generated members. This process is repeated until some convergence criteria is satisfied, or a fixed number of iterations has been performed.

Experimentally, CLONALG has been shown to perform well on standard benchmark tests for optimization problems [32]. However, it has not been reported in the literature that CLONALG itself outperforms any well known technique. Other algorithms similar to CLONALG exist in the literature, such as [65] and [20], with comparative studies showing that whilst CLONALG is effective, better results can be obtained with more specialized versions of the algorithm [20,77]. Indeed, recent work by [21] has shown that a clonal selection based algorithm using a special aging operator can perform as well as the state-of-art on certain protein folding problems.

Clonal selection based algorithms have also been developed for dynamic environments, reporting good performance [40,66,70]. CLONALG has also been adapted for simple pattern recognition problems, but the results from that work are less conclusive [107]. It has also been adapted for more sophisticated learning systems where results are very encouraging indeed for static learning [45,104,105] and for dynamic learning [80].

Immune Networks

In a landmark paper for its time, [59] proposed that the immune system is capable of achieving immunological memory by the existence of a mutually reinforcing net-

work of B cells. This network of B cells occurs due to the ability of paratopes (molecular portions of an antibody) located on B cells, to match against idiotopes (other molecular portions of an antibody) on other B cells. The binding between idiotopes and paratopes has the effect of stimulating the B cells. This is because the paratopes on B cells react to the idiotopes on similar B cells, as it would an antigen. However, to counter the reaction there is a certain amount of suppression between B cells which acts as a regulatory mechanism. This interaction of B cells due to the network, was said to contribute to a *stable* memory structure, and account for the retainment of memory cells, even in the absence of antigen. This theory was refined and formalized in successive works by [35,79] and combined with work by [13] was very influential in development of the immune network based AIS such as [56,76,96,97].

Artificial Immune Networks

Based on the work of CLONALG, an algorithm known as aiNet was proposed in [31]. aiNet is a simple extension of CLONALG (described above), but exploits interactions between B cells according to the immune network theory. The main difference between the two approaches, is that after new clones are integrated into the population, a network suppression function is employed throughout the population to remove cells that have similar affinities² this facilitates the maintenance of diversity within the population. Recent work by [83] has shown that aiNET performs better on data that has a more uniform underlying distribution, due to the nature of how aiNET performs the suppression. However, this work did focus on the initial version of aiNET, and there are recent variants of aiNET that seem to perform much better.

aiNet was initially designed for data clustering, but has been extended over the years, as a hierarchical clustering tool in [29] and through hybridization with fuzzy systems methods by [14]. In the last paper, aiNet was augmented to take into account an adaptive radius measure instead of a fixed radius for B cell matching. This lead to a much improved version of aiNet, being able to achieve better separation of the data, forming clusters in less time. Work by [27] adapted aiNet for multi-modal function optimization. In that paper, aiNet was also modified to be applied to the same optimization problems as CLONALG, and was shown to have greatly improved performance over CLONALG, but this is not as comparable to other clonal selection based systems [94,99]. However, it was recently identified

that if careful thought was given to the optimization problem, the basic aiNet algorithm can be augmented to give significant gains in performance [3].

Negative Selection

Negative selection is a process of *selection* that takes place in the thymus gland. T cells are produced in the bone marrow and before they are released into the lymphatic system, undergo a maturation process in the thymus gland. The maturation of the T cells is conceptually very simple. T cells are exposed to self-proteins in a binding process. If this binding activates the T cell, then the T cell is killed, otherwise it is allowed into the lymphatic system. This process of *censoring* prevents cells that are reactive to *self* from entering the lymph system, thus endowing (in part) the host's immune system with the ability to distinguish between self and non-self agents.

Artificial Negative Selection

The negative selection principle inspired [36] to propose a negative selection algorithm to detect data manipulation caused by computer viruses. The basic idea is to generate a number of detectors in the complementary space and then to apply these detectors to classify new (unseen) data as self (no data manipulation) or non-self (data manipulation). In the negative selection algorithm as proposed by Forrest et al. we can define self as a set **S** of elements of length *l* in shape-space. Then generate a set **D** of detectors, such that each fails to match any element in **S**. With these detectors, monitor a continual data stream for any changes, by continually matching the detectors in **D** against the stream. This work spawned a great deal of investigations into the use of negative selection for intrusion detection, with early work meeting with some success [37], and this being built on in later years [7,34,54,55]. The work on negative selection has been dominate in AIS. A great deal of work has gone into investigating various alternatives of representations [26,42,44], techniques for estimating detector coverage [43,60,61,62] and applications of said technique [6,24,25,33,63,64,68,69,81] to name only a few. However, later works began to highlight certain limitations of the approach [84,85,86,87] with regards to scalability issues and applicability of the technique to classification. In addition, work in [38] outline the need to consider carefully the application domain when developing AIS, and they give particular attention to negative selection. They review the role AIS have played in the development of a number of machine learning tasks, including that of negative selection. However, Freitas and Timmis point out that there is a lack of appreciation for pos-

²It should be noted that this is a slight departure from the immune network theory, where both suppression and stimulation occur between cells

sible inductive bias within algorithms and positional bias within the choice of representation and affinity measures that comes from not carefully applying not only negative selection, but other algorithms as well.

Alternative Immunological Theories for AIS

Innate Immune System

In recent years there has been a growing interest in the mechanisms of innate immune system in immunology [41]. up to this point, AIS had concentrated solely on the adaptive immune systems, but the Danger theory proposed by [73,74] has caught the interest of the AIS practitioner in recent years as a compliment to the adaptive.

The Danger theory attempts to explain the nature and workings of the immune response in a way different to the more traditional clonal selection view. Matzinger criticizes this idea, as she states that observations demonstrate that it may sometimes be necessary for the body to attack itself and conversely the immune system may not attack cells it knows to be foreign (this is not possible under the classical clonal selection theory). Matzinger argues a more plausible way to describe the triggering of an immune response is a reaction to a stimulus the body considers harmful. This might be seen as a very small change but in reality this is real shift in thinking about how the immune system responds to pathogens. In essence, this model allows for foreign and immune cells to exist together, a situation impossible in the traditional standpoint. When under attack, cells dying unnaturally may release a danger signal, that disperses to cover a small area around that cell: a danger area. It is within this and only within this area that the immune system becomes active and will concentrate its attack against any antigen within it. There is still much debate in the immunological world as to whether the Danger theory is a plausible explanation for observed immune function, but it is proving to be an interesting theory none the less and if affecting the thinking in the AIS world. As we have discussed in the previous section, the idea of using the innate immune systems, in particular the danger theory, has started to become popular. The first to propose the idea was [1] in the context of using such danger theory ideas in the context of network intrusion detection. Here the authors discussed how one might introduce the notion of *danger areas* in networks which might be indicated by unusual behavior without having to define a-priori what those behaviors were (a large departure from the current way of network intrusion detection using AIS, where unusual behavior was used to train the system). This work was then extended to notably [47,48] where the dendritic cell algorithm was proposed and [8] where the idea of ar-

tificial tissue was proposed. Work in [48] describes how dendritic cells (DC), which is a cell considered to be part of the innate immune system, performs a function that in effect controls the adaptive immune response when under attack. What a DC does depends on the signals that it receives within the tissue: these might be danger signals, PAMPS (pathogenic associated molecular patterns), safe signals and inflammatory cytokines. The DC will mature into different states depending on the concentration of these signals, and the state of the DC influences the response of the T cell to which the DC is presenting the antigen. Based on these ideas, a DC inspired algorithm has been developed, and tested within a tissue environment developed in [100]. The application area was that of anomaly detection and the DC had to identify if certain types of behavior on a computer were anomalous or not: no predefined knowledge of what constitutes anomalous is required, but what is required is a definition of what constitutes dangerous behaviors for various variables. Results reported in [48] would seem to indicate that the DC algorithm is capable of identifying anomalous behavior (processes that were considered not be normal) over time. However, this is still quite preliminary work, and the system needs to be baselined against a state-of-art type system in order to fully see the contribution this approach can bring.

Cognitive Immunology

Much like the paradigm shift of danger theory, within AIS recent attention has been paid to the cognitive immune paradigm proposed by [18]. Notably work in [2] discusses how Cohen views the immune system as a cognitive system, capable of detection, cognition and decision making and argues that the primary role of the immune system is not protection (as is considered by say the clonal selection theory), but one of body maintenance. [2] say that in Cohen's view, removal of pathogen, is beneficial to the health of the body, and thus defense against pathogen is considered to be just a special case of body maintenance. In order to carry out body maintenance, the immune system must be able to detect the current state of the body's tissues and elicit an appropriate response.

According to the clonal selection theory, immune specificity is a property of the somatically generated immune receptors of the T and B cells, which both initiates and regulates the immune response. Initiation is achieved via the binding between an antigen and a receptor that is specific to it. As stated by [2], Cohen, however, points out that immune receptors are intrinsically degenerate, i. e. they can bind more than one ligand. Immune specificity,

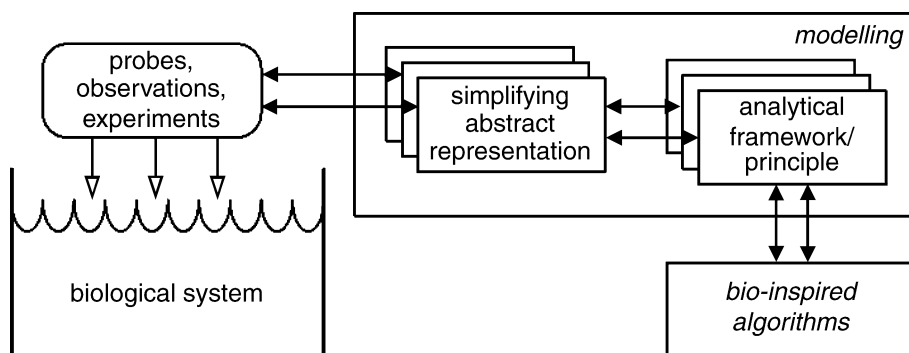
therefore, cannot be purely dependent on molecular binding as no one receptor can be specific to a single antigen. Instead, affinity, the strength of binding between a receptor and its ligand, is a matter of degree. Indeed, as a follow on to their work in [4] outline a simple computational model that demonstrates degenerate recognition in the context of a lymph node, and work in [75] has also investigated degeneracy in the context of pattern recognition and begun to develop high-level AIS as a result.

Emerging Methodologies in AIS

The methodology in which AIS are built has been addressed by work in [82]. This paper proposes a conceptual framework that allows for the development of more biologically grounded AIS, through the adoption of an interdisciplinary approach. As will be clear from the article, metaphors that have been employed have typically been simple, but somewhat effective. However, as proposed in [82], through greater interaction between computer scientists, engineers, biologists and mathematicians, better insights into the workings of the immune system, and the applicability (or otherwise) of the AIS paradigm will be gained. These interactions should be rooted in a sound methodology in order to fully exploit the synergy.

These interactions should be rooted in a sound methodology in order to fully exploit the synergy. The authors argue that rather than going straight from observing the biology and then to the development of an algorithm, a more principled approach is required to adequately capture the required properties of the biological system in the engineered counterpart. The methodology is one of abstraction, as seen in Fig. 2. The first step is observe the biological system through a means of experimentation and analysis. From there it is possible to create a mathematical model of the biosystem: a relatively detailed model of the

system. However, these mathematical models may be too complex to solve, and therefore another level of abstraction is required to gain further insight into the interactions within the system and overall systems dynamics. Therefore, these mathematical models are then used to derive a more abstract computational model: the model can be executed and analyzed for properties that are desired in the engineered system we wish to construct, these can be encompassed into an analytical framework and design principles and high-level abstraction of algorithms and systems can be developed abstract from any application area. This is then instantiated in the application area, being tailored to the specific requirement of that application area. The result is well-grounded bio-inspired algorithm, that is understood better on a theoretical level and captures the *relevant* biological properties for the required application. As part of this process, the authors suggest a second stage of development and that is to create *meta-frameworks* which cut across a number of frameworks that have arisen as part of the initial development. At this stage it is possible to ask unifying questions across these systems that are concerned with complexity of the system, these questions are (1) Openness: how much openness is required in the system, biological systems do not stop computing, therefore should our computations stop? (2) Diversity: How much diversity is required in the system to attain the performance we require, how many different types of actors are needed? (3) Interaction: How should these agents interact? At what timescale and what should they communicate? (4) Structure: biological systems operate on a number of levels, how many levels are needed in our artificial systems, are suitable levels of a hierarchy required? and (5) Scale: Biological systems operate on vast scales, rather different from a typical immune inspired algorithm. How many actors are required in the system to achieve the desired complexity?



Immunecomputing, Figure 2

Conceptual Framework for the development of AIS adapted from [82]

Future Directions

With this may come a better understanding of how to apply AIS, and not fall into the traps highlighted by [38]. A recent paper by [52] highlight the fact that to date, the development of AIS has been *scattergun* i.e. many applications have been tried without a great deal of thought. Indeed, this paper provides a detailed overview of the many application areas that AIS have tried, and this will not be repeated here: the interested reader should consult that paper. The authors go on to propose a number of properties that they feel any AIS should have, and that these properties may help guide the type of application they could be applied to:

- “They will exhibit *homeostasis*.
- They will benefit from interactions between *innate* and *adaptive* immune models.
- They will consist of *multiple, interacting, communicating* components.
- Components can be easily and naturally *distributed*.
- They will be required to perform *life-long learning*” [52].

It apparent that, despite the success of some applications of AIS, all AIS to date fail to fully capture the complex operation of the immune system. What has changed is the increased scope of immunological theories that those working with AIS take inspiration from. For example, In their summaries of the future for AIS, both [39] and [52] point towards an increased emphasis on the innate and homeostatic functions of the immune system as possible areas for AIS exploitation. In addition to the increased scope of AIS, there has been a recent and healthy rise in investigating the theoretical workings of various immune algorithms [17,22,85]. In a recent position paper, [92] argues that the area of AIS has reached something of an impasse. They discusses a number of challenges to the AIS community which they believe will stimulate discussion and help move the area forward:

“Challenge 1: To Develop Novel and Accurate Metaphors and be a Benefit to Immunology. Typically naive approaches to extracting metaphors from the immune system have been taken. This has occurred as an accident of history, and AIS has slowly drifted away from its immunological roots. Time is now ripe for greater interaction with immunologists and mathematicians to undertake specific experimentation and create useful models, all of which can be used as a basis for abstraction into powerful algorithms.

Challenge 2: To Develop a Theoretical basis for AIS. Much work on AIS has concentrated on simple extraction of metaphors and direct application. Despite the creation of a framework for developing AIS, it still lacks significant formal and theoretical underpinning. AIS have been applied to a wide variety of problem domains, but a significant effort is still required to understand the nature of AIS and where they are best applied. For this, a more theoretical understanding is required.

Challenge 3: To Consider the Application of AIS. Work to date in the realm of AIS has mainly concentrated on what other paradigms do, such as simple optimization, learning and the like. This has happened as an accident of history and whilst productive, the time is here to look for the killer application of AIS, or, if not that radical, then applications where the benefit of adopting the immune approach is clear.

Challenge 4: To Consider the Integration of Immune and Other Systems. The immune system does not work in isolation. Therefore, attention should not only be paid to the potential of the immune system as inspiration, but also other systems with which the immune system interacts, in particular the neural and endocrine systems. This will pave the way for a greater understanding of the role and function of the immune system and develop a new breed of immune inspired algorithms.” [92]

To be sure, some of these challenges are now being addressed within the community. The area of AIS is a dynamic and vibrant area of research, it is inherently interdisciplinary in nature and great lessons can be learnt between various communities and they can all benefit from successful interactions.

Bibliography

Primary Literature

1. Aickelin U, Bentley P, Cayzer S, Kim J, McLeod J (2003) Danger theory: The link between AIS and IDS? In: Timmis J, Bentley P, Hart E (eds) (2003) Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 2787. Springer, Berlin, pp 147–155
2. Andrews PS, Timmis J (2005) Inspiration for the next generation of artificial immune systems. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 126–138
3. Andrews PS, Timmis J (2005) On diversity and artificial immune systems: Incorporating a diversity operator into aiNet.

- In: Proceedings of the International Conference on Natural and Artificial Immune Systems (NAIS05). LNCS, vol 391. Springer, Berlin, pp 293–306
4. Andrews PS, Timmis J (2006) A computational model of degeneracy in a lymph node. In: Bersini H, Carneiro J (eds) (2006) Proc of 5th International Conference on Artificial Immune Systems. LNCS. Springer, Berlin, pp 164–177
 5. Ayara M (2005) An immune inspired solution for adaptable error detection in embedded systems. Ph D thesis, University of Kent
 6. Ayara M, Timmis J, de Lemos R, de Castro L, Duncan R (2002) Negative selection: How to generate detectors. In: Proc of the First International Conference on Artificial Immune Systems (ICARIS-2002). University of Kent, Canterbury, pp 89–98
 7. Balthrop J, Forrest S, Glickman M (2002) Revisiting Iisys: Parameters and normal behavior. In: Proceedings of Congress On Evolutionary Computation (CEC). IEEE Press, pp 1045–1050
 8. Bentley PJ, Greensmith J, Ujii S (2005) Two ways to grow tissue for artificial immune systems. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 139–152
 9. Berek C, Ziegner M (1993) The maturation of the immune response. *Immunol Today* 14:200–402
 10. Bersini H (1991) Immune network and adaptive control. In: Proceedings of the 1st European Conference on Artificial Life (ECAL). MIT Press, Cambridge, pp 217–226
 11. Bersini H (1992) Reinforcement and recruitment learning for adaptive process control. In: Proc Int Fuzzy Association Conference (IFAC/IFIP/IMACS) on Artificial Intelligence in Real Time Control, pp 331–337
 12. Bersini H, Carneiro J (eds) (2006) Proc of 5th International Conference on Artificial Immune Systems. LNCS, vol 4163. Springer, Berlin
 13. Bersini H, Varela F (1994) The immune learning mechanisms: Recruitment, reinforcement and their applications. Chapman Hall
 14. Bezerra G, Barra T, de Castro LN, Von Zuben F (2005) Adaptive radius immune algorithm for data clustering. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 290–303
 15. Burnet FM (1959) The clonal selection theory of acquired immunity. Cambridge University Press, Cambridge
 16. Canham RO, Tyrrell AM (2002) A multilayered immune system for hardware fault tolerance within an embryonic array. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 3–11
 17. Clark E, Hone A, Timmis J (2005) A Markov chain model of the B-cell algorithm. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 318–330
 18. Cohen IR (2000) Tending Adam's garden: Evolving the cognitive immune self. Elsevier Academic Press
 19. Cooke D, Hunt J (1995) Recognising promoter sequences using an artificial immune system. In: Proceedings of Intelligent Systems in Molecular Biology. AAAI Press, pp 89–97
 20. Cutello V, Nicosia G, Pavone M (2004) Exploring the capability of immune algorithms: A characterisation of hypermutation operators. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin, pp 263–276
 21. Cutello V, Nicosia G, Pavone M, Timmis J (2007) An immune algorithm for protein structure prediction on lattice models. *IEEE Trans Evol Comp* 11(1):101–117
 22. Cutello V, Nicosia G, Oliveto P, Romeo M (2007) On the convergence of immune algorithms. In: Proc of Foundations of Computational Intelligence. IEEE Press, pp 409–416
 23. Dasgupta D (1999) Artificial Immune Systems and their Applications. Springer, Berlin
 24. Dasgupta D, Forrest S (1995) Tool breakage detection in milling operations using a negative selection algorithm. Tech Rep Report, No CS95-5. Department of Computer Science, University of New Mexico
 25. Dasgupta D, Majumdar NS (2002) Anomaly detection in multidimensional data using negative selection algorithm. In: Proc of Congress on Evolutionary Computation (CEC), Honolulu, Hawaii. IEEE Press, pp 1039–1044
 26. Dasgupta D, Nino F (2000) A comparison of negative and positive selection algorithms in novel pattern detection. In: Proc of the IEEE International Conference on Systems, Man and Cybernetics (SMC), Nashville, 8–11 October
 27. de Castro LN, Timmis J (2002) An artificial immune network for multi modal optimisation. In: Proceedings of the World Congress on Computational Intelligence WCCI, Honolulu, HI. IEEE Press, pp 699–704
 28. de Castro LN, Timmis J (2002) Artificial immune systems: A new computational intelligence approach. Springer, Berlin
 29. de Castro LN, Timmis J (2002) Hierarchy and convergence of immune networks: Basic ideas and preliminary results. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 231–240
 30. de Castro LN, Von Zuben FJ (2000) The clonal selection algorithm with engineering applications. In: GECCO Workshop on Artificial Immune Systems and Their Applications, pp 36–37
 31. de Castro LN, Von Zuben FJ (2001) aiNet: An artificial immune network for data analysis. Idea Group Publishing, pp 231–259
 32. de Castro LN, Von Zuben FJ (2002) Learning and optimization using the clonal selection principle. *IEEE Trans Evol Comp* 6(3):239–251
 33. Ebner M, Breunig H-G, Albert J (2002) On the use of negative selection in an artificial immune system. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). Morgan Kaufman Publishers, San Francisco/New York, pp 957–964
 34. Esponda F, Forrest S, Helman P (2004) A formal framework for positive and negative detection schemes. *IEEE Trans Syst Man Cybern B* 34(1):357–373
 35. Farmer JD, Packard NH, Perelson AS (1986) The immune system, adaptation, and machine learning. *Physica D* 22:187–204
 36. Forrest S, Perelson AS, Allen L, Cherukuri R (1994) Self-nonspecific discrimination in a computer. In: Proc IEEE Symposium on Research Security and Privacy. IEEE Press, pp 202–212
 37. Forrest S, Hofmeyr S, Somayaji A (1997) Computer immunology. *Comm ACM* 40(10):88–96
 38. Freitas A, Timmis J (2003) Revisiting the foundations of artificial immune systems: A problem oriented perspective. In:

- Timmis J, Bentley P, Hart E (eds) (2003) Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 2787. Springer, Berlin, pp 229–241
39. Garrett SM (2005) How do we evaluate artificial immune systems? *Evol Comput* 13(2):145–177
40. Gaspar A, Hirsbrunner B (2002) From optimization to learning in learning in changing environments: The pittsburgh immune classifier system. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 190–199
41. Germain RN (2004) An innately interesting decade of research in immunology. *Nature Medicine* 10:1307–1320
42. González F, Dasgupta D (2003) Anomaly detection using real-valued negative selection. *Genet Program Evolvable Mach* 4(4):383–403
43. González F, Dasgupta D, Kozma R (2002) Combining negative selection and classification techniques for anomaly detection. In: Congress on Evolutionary Computation. IEEE, pp 705–710
44. González F, Dasgupta D, Gómez J (2003) The effect of binary matching rules in negative selection. In: Genetic and Evolutionary Computation – GECCO-2003. Lecture Notes in Computer Science, vol 2723. Springer, Chicago, pp 195–206
45. Goodman D, Boggess L, Watkins A (2002) Artificial immune system classification of multiple-class problems. In: Proc of Intelligent Engineering Systems. ASME, pp 179–184
46. Goodman D, Boggess L, Watkins A (2003) An investigation into the source of power for AIRS, an artificial immune classification system. In: Proc Int Joint Conf Neural Networks. IEEE, pp 1678–1683
47. Greensmith J, Aickelin U, Cayzer S (2005) Introducing dendritic cells as a novel immune-inspired algorithm for anomaly detection. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 153–167
48. Greensmith J, Aickelin U, Twycross J (2006) Articulation and clarification of the dendritic cell algorithm. In: Bersini H, Coutinho A (eds) Proceedings of the 5th International Conference on Artificial Immune Systems. LNCS, vol 4163. Springer, Berlin
49. Hart E (2005) Not all balls are round: An investigation of alternative recognition-region shapes. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 29–42
50. Hart E, Ross P (2002) Exploiting the analogy between immunology and sparse distributed memories: A system for clustering non-stationary data. In: Timmis J, Bentley P (eds) (2002) Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 49–58
51. Hart E, Ross P (2004) Studies on the implications of shape-space models for idiotypic networks. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3239. Springer, Berlin, pp 413–426
52. Hart E, Timmis J (2005) Application areas of AIS: The past, the present and the future. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin, pp 483–497
53. Hightower RR, Forrest SA, Perelson AS (1995) The evolution of emergent organization in immune system gene libraries. In: Proceedings of the 6th International Conference on Genetic Algorithms. Morgan Kaufmann, pp 344–350
54. Hofmeyr S, Forrest S (1999) Immunity by design: An artificial immune system. In: Proc of Genetic and Evolutionary Computation Conference (GECCO), pp 1289–1296
55. Hofmeyr S, Forrest S (2000) Architecture for an artificial immune system. *Evol Comput* 7(1):1289–1296
56. Hunt J, Cooke D (1996) Learning using an artificial immune system. *J Netw Comput Appl* 19:189–212
57. Hunt J, Timmis J, Cooke D, Neal M, King C (1998) JISYS: Development of an artificial immune system for real-world applications. In: Dasgupta D (ed) Artificial Immune Systems and their Applications. Springer, Berlin, pp 157–186
58. Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) Proc of the 4th International Conference on Artificial Immune Systems (ICARIS). LNCS, vol 3627. Springer, Berlin
59. Jerne NK (1974) Towards a network theory of the immune system. *Ann Immunol (Inst Pasteur)* 125C:373–389
60. Ji Z, Dasgupta D (2004) Augmented negative selection algorithm with variable-coverage detectors. In: Congress on Evolutionary Computation. IEEE, pp 1081–1088
61. Ji Z, Dasgupta D (2004) Real-valued negative selection algorithm with variable-sized detectors. In: Genetic and Evolutionary Computation – GECCO-2004, Part I. Lecture Notes in Computer Science, vol 3102. Springer, Seattle, pp 287–298
62. Ji Z, Dasgupta D (2005) Estimating the detector coverage in a negative selection algorithm. In: Proceedings of Genetic and Evolutionary Computation Conference (GECCO). ACM Press, pp 281–288
63. Ji Z, Dasgupta D (2006) Applicability issues of the real-valued negative selection algorithms. In: Proceedings of Genetic and Evolutionary Computation Conference (GECCO). ACM Press, pp 111–118
64. Ji Z, Dasgupta D, Yang Z, Teng H (2006) Analysis of dental images using artificial immune systems. In: Proceedings of Congress On Evolutionary Computation (CEC). IEEE Press, pp 528–535
65. Kelsey J, Timmis J (2003) Immune inspired somatic contiguous hypermutation for function optimisation. In: Proc of Genetic and Evolutionary Computation Conference (GECCO). LNCS, vol 2723. Springer, Berlin, pp 207–218
66. Kelsey J, Timmis J, Hone A (2003) Chasing chaos. In: Proc of Congress on Evolutionary Computation (CEC). IEEE, Canberra, pp 89–98. <http://www.cs.ukc.ac.uk/pubs/2002/1504>
67. Kim J (2002) Integrating artificial immune algorithms for intrusion detection. Ph D thesis, UCL
68. Kim J, Bentley PJ (2001) An evaluation of negative selection in an artificial immune system for network intrusion detection. In: Proc of Genetic and Evolutionary Computation Conference (GECCO), San Francisco, USA. Morgan Kaufmann, pp 1330–1337
69. Kim J, Bentley PJ (2001) Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with negative selection operator. In: Proc of Congress on Evolutionary Computation (CEC), Seoul, Korea. Morgan Kaufmann, pp 1244–1252
70. Kim J, Bentley PJ (2002) Immune memory in the dynamic clonal selection algorithm. In: Timmis J, Bentley P (eds) (2002)

- Proc of the 1st International Conference on Artificial Immune Systems (ICARIS). University of Kent, Canterbury, pp 59–67
71. Knight T, Timmis J (2003) A multi-layered immune inspired machine learning algorithm. In: Lotfi A, Garibaldi M (eds) *Applications and Science in Soft Computing*. Springer, Berlin, pp 195–202. <http://www.cs.kent.ac.uk/pubs/2003/1760>
 72. Krohling R, Zhou Y, Tyrrell A (2002) Evolving FPGA-based robot controllers using an evolutionary algorithm. In: Timmis J, Bentley P (eds) (2002) *Proc of the 1st International Conference on Artificial Immune Systems (ICARIS)*. University of Kent, Canterbury, pp 41–46
 73. Matzinger P (1997) An innate sense of danger. *Semin Immunol* 10(5):399–415
 74. Matzinger P (2002) The danger model: A renewed sense of self. *Science* 296:301–305
 75. Mendao M, Timmis J, Andrews PS, Davies M (2007) The immune system in pieces: Computational lessons from degeneracy in the immune system. In: Fogel DB (ed) *Proc of Foundations of Computational Intelligence*. IEEE Press, pp 394–400
 76. Neal M (2002) An artificial immune system for continuous analysis of time-varying data. In: Timmis J, Bentley P (eds) (2002) *Proc of the 1st International Conference on Artificial Immune Systems (ICARIS)*. University of Kent, Canterbury, pp 76–85
 77. Nicosia G (2004) Immune algorithms for optimization and protein structure prediction. Ph D thesis, University of Catania
 78. Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) *Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS)*. LNCS, vol 3239. Springer, Berlin
 79. Perelson AS (1989) Immune network theory. *Immunol Rev* 110:5–36
 80. Secker A, Freitas A, Timmis J (2003) AISEC: An artificial immune system for email classification. In: *Proc of Congress on Evolutionary Computation (CEC)*. IEEE Press, pp 131–139
 81. Singh S (2002) Anomaly detection using negative selection based on the r-contiguous matching rule. In: Timmis J, Bentley PJ (eds) *Proceedings of the 1st International Conference on Artificial Immune Systems ICARIS*. University of Kent at Canterbury, University of Kent at Canterbury Printing Unit, pp 99–106. <http://www.aber.ac.uk/icaris-2002>
 82. Stepney S, Smith R, Timmis J, Tyrrell A, Neal M, Hone A (2006) Conceptual frameworks for artificial immune systems. *Int J Unconv Comput* 1(3):315–338
 83. Stibor T, Timmis J (2007) An investigation into the compression quality of ainet. In: Fogel D (ed) *Proc of Foundations of Computational Intelligence*. IEEE Press
 84. Stibor T, Bayarou KM, Eckert C (2004) An investigation of R-chunk detector generation on higher alphabets. In: *Proc of Genetic and Evolutionary Computation Conference (GECCO)*. LNCS, vol 3102. Springer, Berlin, pp 299–307
 85. Stibor T, Timmis J, Eckert C (2005) A comparative study of real-valued negative selection to statistical anomaly detection techniques. In: Jacob C, Pilat M, Bentley P, Timmis J (eds) (2005) *Proc of the 4th International Conference on Artificial Immune Systems (ICARIS)*. LNCS, vol 3627. Springer, Berlin, pp 262–275
 86. Stibor T, Mohr P, Timmis J, Eckert C (2005) Is negative selection appropriate for anomaly detection? In: *Proc of Genetic and Evolutionary Computation Conference (GECCO)*. ACM Press
 87. Stibor T, Timmis J, Eckert C (2006) Generalization regions in hamming negative selection. In: *Intelligent Information Processing and Web Mining. Advances in Soft Computing*. Springer, Berlin, pp 447–456
 88. Tarakanov AO, Skormin VA, Sokolova SP (2003) *Immunocomputing: Principles and Applications*. Springer, New York
 89. Tarakanov AO, Goncharova LB, Tarakanov OA (2005) A cytokine formal immune network. In: *Advances in Artificial Life, 8th European Conference, ECAL 2005, Canterbury, UK, 5–9 September 2005*, pp 510–519
 90. Tarakanov AO, Kvachev SV, Sukhorukov AV (2005) A formal immune network and its implementation for on-line intrusion detection. In: *MMM-ACNS*, pp 394–405
 91. Timmis J (2000) Artificial immune systems: A novel data analysis technique inspired by the immune system. Ph D thesis, University of Wales
 92. Timmis J (2007) Artificial immune systems: Today and tomorrow. *Natural Comput* 6(1):1–18
 93. Timmis J, Bentley P (eds) (2002) *Proc of the 1st International Conference on Artificial Immune Systems (ICARIS)*. University of Kent, Canterbury
 94. Timmis J, Edmonds C (2004) A comment on opt-AINet: An immune network algorithm for optimisation. In: *Proc of Genetic and Evolutionary Computation Conference (GECCO)*. LNCS, vol 3102. Springer, Berlin, pp 308–317
 95. Timmis J, Knight T (2001) Artificial immune systems: Using the immune system as inspiration for data mining. In: Abbas H, Ruhul A, Sarker A, Newton S (eds) *Data Mining: A Heuristic Approach*. Idea Group, pp 209–230
 96. Timmis J, Neal M (2001) A resource limited artificial immune system for data analysis. *Knowl Based Syst* 14(3–4):121–130
 97. Timmis J, Neal M, Hunt J (2000) An artificial immune system for data analysis. *Biosystems* 55(1/3):143–150
 98. Timmis J, Bentley P, Hart E (eds) (2003) *Proc of the 2nd International Conference on Artificial Immune Systems (ICARIS)*. LNCS, vol 2787. Springer, Berlin
 99. Timmis J, Edmonds C, Kelsey J (2004) Assessing the performance of two immune inspired algorithms and a hybrid genetic algorithm for function optimisation. In: *Proc of Congress on Evolutionary Computation (CEC)*, vol 1. IEEE, pp 1044–1051
 100. Twycross J, Aickelin U (2006) libtissue: Implementing innate immunity. In: *Proc Congress on Evolutionary Computation*. IEEE Press, pp 499–506
 101. Varela F, Coutinho A, Dupire B, Vaz N (1988) Cognitive networks: Immune, neural and otherwise. *J Theor Imm* 2:359–375
 102. Watkins A (2001) AIRS: A resource limited artificial immune classifier. Master's thesis, Mississippi State University
 103. Watkins A (2005) Exploiting immunological metaphors in the development of serial, parallel and distributed learning algorithms. Ph D thesis, University of Kent
 104. Watkins A, Timmis J (2004) Exploiting parallelism inherent in AIRS, an artificial immune classifier. In: Nicosia G, Cutello V, Bentley P, Timmis J (eds) (2004) *Proc of the 3rd International Conference on Artificial Immune Systems (ICARIS)*. LNCS, vol 3239. Springer, Berlin, pp 427–438
 105. Watkins A, Xintong B, Phadke A (2003) Parallelizing an immune-inspired algorithm for efficient pattern recognition. In: *Intelligent Engineering Systems through Artificial Neural Networks: Smart Engineering System Design: Neural Networks*,

Fuzzy Logic, Evolutionary Programming, Complex Systems and Artificial Life. ASME Press, pp 224–230

106. Watkins A, Timmis J, Boggess L (2004) Artificial immune recognition system (AIRS): An immune inspired supervised machine learning algorithm. *Genet Program Evolvable Mach* 5(3):291–318. <http://www.cs.kent.ac.uk/pubs/2004/1634>
107. Whitesides GM, Boncheva M (2002) Beyond molecules: Self-assembly of mesoscopic and macroscopic components. *Proc Natl Acad Sci USA* 99(8):4769–4774
108. Wierzbichon S, Kuzelewska U (2002) Stable clusters formation in an artificial immune system. In: Timmis J, Bentley P (eds) (2002) *Proc of the 1st International Conference on Artificial Immune Systems (ICARIS)*. University of Kent, Canterbury, pp 68–75

Books and Reviews

- Dasgupta D (1999) *Artificial Immune Systems and Their Applications*. Springer, Berlin
- Cohen I, Segal L (2001) *Design Principles for the Immune System and Other Distributed Autonomous Systems*. SFT, Oxford University Press
- de Castro LN, Timmis J (2002) *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer, Berlin
- Tarakanov AO, Skormin VA, Sokolova SP (2003) *Immunocomputing: Principles and Applications*. Springer, New York
- Ishida Y (2004) *Immunity-Based Systems: A Design Perspective*. Springer, New York

Implementation Theory

LUIS C. CORCHÓN

Departamento de Economía, Universidad Carlos III, Madrid, Spain

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Brief History of Implementation Theory
 The Main Concepts
 The Main Insights
 Unsolved Issues and Further Research
 Answers to the Questions
 Acknowledgments
 Bibliography

Glossary

Type of an agent All the information possessed by this agent. It may refer to the preferences of this agent and/or to the knowledge of this agent of the preferences of other agents.

State of the world Description of all information possessed by all agents.

Social choice rule A correspondence mapping the set of states of the world in the set of allocations. It represents the social objectives that the society or its representatives want to achieve.

Mechanism A list of message spaces and an outcome function mapping messages into allocations. It represents the communication and decision aspects of the organization.

Equilibrium concept A mapping (or a collection of them) from the set of states of the world into allocations yielded by equilibrium messages. This equilibrium is a game-theoretical notion of how agents behave, e.g. Nash Equilibrium, Bayesian Equilibrium, Dominant Strategies, etc.

Implementable social choice rule in an equilibrium concept (e.g. nash equilibrium) A Social Choice Rule is implementable in an equilibrium concept (e.g. Nash Equilibrium) if there is a mechanism such that for each state of the world the allocations prescribed by the Social Choice Rule and those yielded by the equilibrium concept coincide.

Definition of the Subject

Implementation theory studies which social objectives (i.e. Social Choice Rules) are compatible with the incentives of the agents (i.e. are implementable). In other words it is the systematic study of the social goals that can be achieved when agents behave strategically.

Introduction

Dear colleague;

I wrote this survey with you in mind. You are an economist doing research who would like to know why implementation is important. And by this I do not mean why some people won the Nobel Prize working in this area. I mean, what are the deep insights found by implementation theory and what applications are delivered by these tools. I propose a simple game: try to answer the following questions. If you cannot answer them, but you think they are important, read the survey. At the end of this survey, I will give you the answers. I will also tell you why I like implementation theory so much!

1. Why are agents price-takers? Is price-taking possible in economies with a finite number of agents?
2. Suppose two firms wish to merge. They claim that the merger will bring large cost reductions but some people fear that the firms just want to avoid competition. What would be your advice?

3. How should a monopoly be regulated when regulators do not know the cost function or the demand function of the monopolist?
4. How should it be determined whether or not a public facility – a road, a bridge, a stadium – should be constructed and who should pay for it?
5. Is justice possible in this world? Can we reconcile justice and self-interest?
6. Can an uninformed planner achieve better allocations than those produced by completely-informed agents in an unregulated market?
7. In competitive ice skating, the highest and lowest marks awarded by judges are discarded and the remaining are averaged. Do you think that this procedure eliminates incentives to manipulate votes?
8. What kind of policies would you advocate to fight Global Warming?

The answers to these questions are found in Sect. “[Answers to the Questions](#)” The rest of this paper goes as follows. Sect. “[Brief History of Implementation Theory](#)” is a historical introduction that can be skipped. Sect. “[The Main Concepts](#)” explains the basic model. Sect. “[The Main Insights](#)” explains the main results. Sect. “[Unsolved Issues and Further Research](#)” offers some thoughts about the future direction of the topic.

Brief History of Implementation Theory

From, at least Adam Smith on, we have assumed that agents are motivated by self-interest. We also assumed that agents interact in a market economy where prices match supply and demand. This tradition crystallized in the Arrow–Debreu–McKenzie model of General equilibrium in the 1950s. But it was quickly discovered that this model had important pitfalls other than focusing on a narrow class of economic systems: On the one hand, an extra agent was needed to set prices, the auctioneer. On the other hand agents follow rules, i. e. to take prices as given, which are not necessarily consistent with self-interest. An identical question had arisen earlier when Taylor [129] and Lange [70], following Barone [17], proposed a market socialism, where socialist managers maximize profits: Why would socialist managers choose output in the way prescribed to them (or who will provide and preserve capital in a system where the private property of such items is forbidden?)? Samuelson [113] voiced identical concern about the Lindahl solution to allocate public goods: “It is in the selfish interest of each person to give false signals”. This concern gave rise later on to the golden rule of incentives – as stated by Roger Myerson [91]: “An organization must give its members the correct incentives to share in-

formation and act appropriately”. Earlier, it had aroused the interest of Leonid Hurwicz, the father of Implementation theory, in economic systems other than the market. In any case it was clear that an important ingredient was missing in the theory of economic systems. This element was that not all information needed for resource allocation was transmitted by prices: Some vital items have to be transmitted by agents.

Several proposals arose to fill the gap: On the one hand, models of markets under asymmetric information, Vickrey [135], Akerlof [3], Spence [128] and Rothchild and Stiglitz [106]. On the other hand models of public intervention, like optimal taxation, Mirless [82], and mechanisms for allocating public goods, Clarke [27] and Groves [48], with the so-called Principal-Agent models somewhere in the middle. The key word was “Truthful Revelation” or “Incentive Compatibility”: Truthful revelation of information must be an equilibrium strategy, either a dominant strategy, as in Clarke and Groves, or a Bayesian equilibrium as in Arrow [7] and D’Aspremont and Gerard-Varet [10]. A motivation for this procedure was provided by the “Revelation Principle”, Gibbard [45], Myerson [89], Dasgupta, Hammond and Maskin [38] and Harris and Townsend [50]: If a mechanism yields certain allocations in equilibrium, telling the truth about one’s characteristics must be an equilibrium as well (however, telling the truth may not be an equilibrium in the original mechanism you might have to use an equivalent direct mechanism). This result is of utmost importance and it will be thoroughly considered in Sect. “[The Main Concepts](#)” However, it was somehow misread as “there is no loss of generality in focusing on incentive compatibility”. But what the revelation principle asserts is that truthful revelation is *one* of the, possibly, many equilibria. It does not say that truthful revelation is *the only* equilibrium. As we will see in some cases it is a particularly unsatisfactory way of selecting equilibria.

The paper by Hurwicz [53], popularized by Reiter [102], presented a formal structure for the study of economic mechanisms which has been followed by all subsequent papers. Maskin [76], whose first version circulated in 1977, is credited as the first paper where the problem of multiple equilibria was addressed as a part of the model and not as an afterthought, see the report of the Nobel Prize Committee [101]. Maskin studied implementation in Nash equilibrium (see Glossary). Later his results were generalized to Bayesian Equilibrium by Postlewaite and Schmeidler [100] and Palfrey and Srivastava [94,95].

Finally, Moulin [87] studied Dominance Solvability and Moore and Repullo [87] Subgame Perfect Equilibrium. The century closed with several characterizations on

what can be implemented in other equilibrium concepts: Moore and Repullo [85] in Nash Equilibrium, Palfrey and Srivastava [96] in Undominated Nash Equilibrium, Jackson [58] in Bayesian Equilibrium, Dutta and Sen [39] in Strong Equilibrium and Sjöström [125] in Trembling Hand Equilibria. With all these papers in mind, the basic aspects of implementation theory are now well understood.

The interested reader may complement the previous account with the surveys by Maskin and Sjöström [78] and Serrano [122] which cover the basic results and by Baliga and Sjöström [13] for new developments including experiments. See also Maskin [75], Moore [84], Corchón [29], Jackson [60] and Palfrey [93]. Several important applications of Implementation Theory are not surveyed here: Auctions, see Krishna [67], Contract theory, see Laffont and Martimort [69], Matching, see Roth (forthcoming) and Moral Hazard see Ma, Moore and Turnbull [73].

The Main Concepts

We divide this section into four subsections: The first describes the environment, the second deals with social objectives, the third revolves around the notion of a mechanism and the last defines the equilibrium concepts that we will use here.

The Environment

Let $I = \{1, \dots, n\}$ be the set of agents. Let θ_i be the type of i . This includes all the information in the hands of i . Let Θ_i be agent i 's type set. The set $\Theta \subset \prod_{i=1}^n \Theta_i$ is the set of states of the world. For each $\theta \in \Theta$ we have a feasible set $A(\theta)$ and a preference profile $R(\theta) = (R_1(\theta), \dots, R_n(\theta))$. $R_i(\theta)$ is a complete, reflexive and transitive binary relation on $A(\theta)$. $I_i(\theta)$ denotes the corresponding indifference relation. Set $A \equiv \bigcup_{\theta \in \Theta} A(\theta)$. Let $a = (a_1, a_2, \dots, a_n) \in A$ be an allocation, also written (a_i, a_{-i}) , where $a_{-i} \equiv (a_1, a_2, \dots, a_{i-1}, a_{i+1}, \dots, a_n)$.

The standard model of an exchange economy is a special case of this model: θ is an economy. $X_i(\theta) \subset \mathbb{R}^k$ is the consumption set of i . $w_i(\theta) \in \text{int} X_i(\theta)$ are the endowments in the hands of i . The preferences of i are defined on $X_i(\theta)$. The set of allocations $A(\theta)$ is defined as

$$A(\theta) = \left\{ a \mid \sum_{i=1}^n (a_{ij} + w_{ij}(\theta)) \leq 0, \quad j = 1, 2, \dots, k, \right. \\ \left. (a_{i1}, a_{i2}, \dots, a_{ik}) \in X_i(\theta), \quad \forall i \in I \right\}.$$

A special case of an exchange economy is bilateral trading: Here there are two agents, the seller and the buyer. The

seller has a unit of an indivisible good and both agents are endowed with an infinitely divisible good ("money"). Preferences are representable by linear utility functions. The type of each agent, also called her valuation, is the marginal rate of substitution between both goods. Finally, the set of types is a closed interval of the real line.

Another example is the social choice model where the set of states of the world is the Cartesian product of individual type sets, $\Theta = \prod_{i=1}^n \Theta_i$. The set of feasible allocations is constant. The preferences of each agent only depend on her type, for all $\theta \in \Theta$, $R_i(\theta) = R_i(\theta_i)$ all $i \in I$.

The model of public goods is a hybrid of the social choice and the exchange economy models. For a subset of goods, say $1, 2, \dots, l$, agents receive the same bundle (these are the public goods). For goods $l + 1, \dots, k$, agents can consume possibly different bundles.

Social Objectives

Implementation begins by asking what allocations *we* want to achieve. In this sense, implementation theory reverses the usual procedure, namely, fix a mechanism and see what the outcomes are. The theory is rather agnostic as to who is behind *we*: It could be a democratic society, it could be a dictator, a benevolent planner, etc. Formally, a correspondence $F: \Theta \rightarrow A$ such that $F(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$ will be called a *Social Choice Rule* (SCR). Under risk or uncertainty, allocations are state-dependent (recall the concept of contingent commodities in General Equilibrium). Thus an allocation is a single-valued function $f: \Theta \rightarrow A$. The notion of a SCR is replaced by that of a *Social Choice Set* (SCS) defined as a collection of functions mapping Θ into A . Examples of SCR are the Pareto rule, which maps every state into the set of Pareto efficient allocations for this state, the Walrasian SCR which maps every economy in the set of allocations that are a Walrasian Equilibrium for this economy, etc.

If states of the world were contractible, i.e. if they could be written in an enforceable contract specifying the allocations in each state, SCR or SCS would be directly achieved, assuming that those not complying could be punished harshly. Unfortunately, states of the world are a description of preferences and productive capabilities, being those difficult to describe and therefore easy to manipulate. Thus, we have to find another method to reach the desired allocations.

Mechanisms

If the information necessary to judge the desirability of allocations is in the hands of agents, it seems that the only way of retrieving this information is by asking them. But,

of course, agents cannot be trusted to reveal truthfully their information because they might lose by doing so. Thus the owner of a defective car will think twice about revealing the true state of the car if the price of defective cars is less than the price of reliable cars. But perhaps we may design ways in which the messages sent by different agents are checked one against the other. We may also design ways in which agents send information by indirect means, say by raising flags, making gestures, and so on and so forth. This is the idea behind the concept of a mechanism (also called a game form).

Formally, a *mechanism* is a pair (M, g) where $M \equiv \prod_1^n M_i$ is the *message space* and $g: M \rightarrow A$ is the outcome function. M_i denotes agent i 's *message space with typical element* m_i . In some cases, i. e. when goods are indivisible, the outcome function maps M into the set of lotteries on A , denoted by $\mathcal{L}A$. In this case the outcome function yields the probability of obtaining an object. Let $m = (m_1, \dots, m_n) \in M$, be a list of messages, also written (m_i, m_{-i}) where m_{-i} is a list of all messages except those sent by i .

Another interpretation of a mechanism, more in tune with decentralized systems, is that messages describe contracts among agents and the outcome function is a legal system that converts contracts into allocations.

If feasible sets are state dependent we have a problem: Suppose that at θ we want to achieve allocation $a \in A(\theta)$. So there must be a message, say m such that $g(m) = a$. But what if there is another state, say θ' for which $a \notin A(\theta')$? In this case $g(m) \notin A(\theta')$. In other words, since mechanisms are not state dependent they may yield unfeasible allocations. We will postpone the discussion of this problem until Sect. “Unsolved Issues and Further Research” For the time being, let us assume that feasible sets are not state dependent.

Equilibrium

Since the messages sent by agents are tied to their incentives, it is clear that we have to use an equilibrium concept borrowed from game theory. Thus, given $\theta \in \Theta$, a mechanism (M, g) induces a game in normal form (M, g, θ) . There are many “solutions” to what would constitute an equilibrium. Let us begin by considering the notion of a Nash equilibrium:

Definition 1 A message profile $m^* \in M$ is a Nash equilibrium for (M, g, θ) if, for all $i \in I$ $g(m^*)R_i(\theta)g(m_i, m_{-i}^*)$ for all $m_i \in M_i$.

Let $NE(M, g, \theta)$ be the set of allocations yielded by all Nash equilibria of (M, g, θ) . We now ask, given a SCR, what mechanism, if any, would produce outcomes identi-

cal to the SCR. In this sense, the mechanism is the variable of our analysis i. e. the mechanism “solves” the equation $NE(M, g, \theta) = F(\theta)$, for all $\theta \in \Theta$. Formally,

Definition 2 The SCR F is implementable in Nash equilibrium if there is a mechanism (M, g) such that, for all $\theta \in \Theta$, $NE(M, g, \theta) \neq \emptyset$ and:

1. $F(\theta) \subseteq NE(M, g, \theta)$.
2. $NE(M, g, \theta) \subseteq F(\theta)$.

The previous concept can be easily generalized. Given a mechanism (M, g) an equilibrium concept is a mapping, say $E_{(M, g)}: \Theta \rightarrow A$ such that $E_{(M, g)}(\theta) \subseteq A(\theta)$ for all $\theta \in \Theta$. For instance $E_{(M, g)}(\theta)$ may be the set of allocations arising from dominant strategy profiles in θ when the mechanism (M, g) is in place. The notion of implementation in an equilibrium concept easily follows. See Thomson [132] for a discussion of other concepts of implementation.

The problem is that some equilibrium concepts can not be written in the way we just described because the actions to be taken in state, say θ' , depend on preferences in states other than θ' . To see this, suppose that agents attach a vector of probabilities to each possible type of the other agents, Harsanyi [51]. Denote by $q(\theta_{-i}/\theta_i)$ the vector of probabilities attached by i that other agents have types θ_{-i} given that she is of type θ_i . For simplicity assume that it is a strictly positive vector. Suppose that preferences are representable by a von Neumann-Morgenstern utility index $V_i(a, \theta)$. In this framework, (as first noticed by Vickrey [135]) a strategy for i , denoted by s_i , is no longer a message but a function from the set of types of i in the set of messages of i , namely, $s_i: \Theta_i \rightarrow M_i$. A strategy profile, s , is a collection of strategies, one for each agent, $s = (s_1, \dots, s_n)$ also written as (s_i, s_{-i}) . For simplicity, the next definition assumes that type sets are finite.

Definition 3 A Bayesian Equilibrium (BE) for $(M, g, R(\cdot))$ is a s^* such that for all $i, \theta \in \Theta$, and $m_i \in M_i$,

$$\begin{aligned} \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} | \theta_i) V_i(g(s^*(\theta)), \theta) \\ \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} | \theta_i) V_i(g(m_i, s_{-i}^*(\theta_{-i})), \theta) \end{aligned}$$

Thus, an equilibrium concept – given a mechanism – is a collection of functions, denoted by $H_{(M, g)}$, such that for all $h_{(M, g)} \in H_{(M, g)}$ $h_{(M, g)}: \Theta \rightarrow A$. Finally, the definition of implementable SCS in BE follows.

Definition 4 The mechanism (M, g) implements a SCS F in BE if:

1. For any BE s there exists $x \in F(\theta)$, such that $g(s(\theta)) = x(\theta)$ for all $\theta \in \Theta$.
2. For any $x \in F$, there is a BE s such that $g(s(\theta)) = x(\theta)$ for all $\theta \in \Theta$.

Looking at our definitions of an implementable SCR or SCS we see that the first requirement is that all equilibria yield “good” allocations. The second requirement is that given an allocation to be implemented, there is an equilibrium “sustaining” this allocation. These two requirements bear some resemblance to the two fundamental theorems of welfare economics, namely that competitive equilibrium is efficient and that any efficient allocation can be achieved as a competitive equilibrium with the appropriate endowment redistribution. Notice that endowment redistribution is not used in the definition of implementation.

The Main Insights

We group our results here under three headings: The Revelation Principle and its consequences, Monotonicity and how to avoid it and the limits of design. We will discuss them each in turn.

The Revelation Principle and its Consequences

The definition of a mechanism is extremely abstract. No conditions have been imposed on what might constitute a message space or an outcome function. And since implementation theory considers the mechanism the variable to be found, this is an unhappy situation: we are asked to find something whose characteristics we do not know! Fortunately the revelation principle comes to the rescue by stating a necessary condition for implementation: If a single valued SCR, which we will call a Social Choice Function (SCF) is implementable, there is a *revelation mechanism* for which telling the truth is an equilibrium. A revelation mechanism (associated with a SCF) is a mechanism in which the message space for each agent is her set of types and the outcome function is the SCF. We say that a SCF is truthfully implementable or incentive compatible if truth-telling is a Bayesian equilibrium (or a dominant strategy) of the direct mechanism associated with it. The following result formally states the revelation principle:

Theorem 1 *If f is a Bayesian (resp. dominant strategy) implementable SCF, f is incentive compatible.*

Proof Let f be Bayesian implementable. Therefore, there exists a mechanism (M, g) and a Bayesian equilibrium s^*

such that $g(s^*(\theta)) = f(\theta)$ for every $\theta \in \Theta$. Since $s^*(\cdot)$ is a Bayesian equilibrium, $\forall \theta \in \Theta, \forall m_i \in M$

$$\begin{aligned} & \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(s^*(\theta)), \theta) \\ & \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(g(m_i, s_{-i}^*(\theta_{-i})), \theta). \end{aligned}$$

Which implies that $\forall \theta'_i \in \Theta_i, \forall \theta_{-i} \in \Theta_{-i}$,

$$\begin{aligned} & \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(f(\theta), \theta) \\ & \geq \sum_{\theta_{-i} \in \Theta_{-i}} q(\theta_{-i} \mid \theta_i) V_i(f(\theta'_i, \theta_{-i}), \theta). \end{aligned}$$

The proof for the case of dominant strategies is identical. \square

Theorem 1 (T.1 in the sequel) can be explained in terms of a mediator, i. e. somebody to whom you say “who you are” and who chooses the strategy that maximizes your payoffs on your behalf. Would you try to fool such a person? If you do so, you are fooling yourself because the mediator would choose a strategy that is not the best for you. Thus, the best thing for you to do is to tell the truth (providing an unexpected backing to the aphorism “honesty is the best policy”!).

Consider now the following results, due to Hurwicz [54] (who proved it for the case of $n = 2$) and to Gibbard [45] and Satterthwaite [115] respectively.

Theorem 2 *In exchange economies environments there is no SCF such that:*

1. *It is truthfully implementable in dominant strategies.*
2. *It selects individually rational allocations.*
3. *It selects efficient allocations.*
4. *Its domain includes all economies with convex and continuous preferences.*

Theorem 2' *In social choice environments there is no SCF such that:*

1. *It is truthfully implementable in dominant strategies.*
2. *It is non-dictatorial.*
3. *Its range is A , with $\#A > 2$.*
4. *Its domain includes all possible preference profiles.*

It is clear that there are trivial SCF in which any three conditions in T.2–2' are compatible. But T.2–2' are very robust in the sense that they hold for small domains of economies [15,16], for weaker notions of individual rationality [110,119] and in public goods domains [71]. Moreover, assuming quasi-linear utility functions, Hurwicz and

Walker [56], building on a previous paper by Walker, proved that the set of economies for which conditions 1–3 in T.2 are incompatible is open and dense. Beviá and Corchón [20] show that these conditions are incompatible for *any* economy where utility functions are quasi-linear, strictly concave, differentiable and fulfill a very mild regularity condition. These results show that Vickrey–Clarke–Groves mechanisms fail to achieve efficient allocations in general (Vickrey–Clarke–Groves mechanisms are revelation mechanisms that work in public good economies where utility functions are quasi-linear in “money”). The outcome function selects the level of public good that maximizes the sum of utilities announced by agents and the money received by each individual is the sum of the utility functions announced by all other agents. For an exposition of these mechanisms, see Green and Laffont [46]).

A proof of T.2 can be found in Serizawa [118]. Simple proofs of T.2' can be found in Barberá [14], Benoit [18]) and Sen [117].

T.1 and 2–2' imply that there is no mechanism implementing an efficient and individually rational (resp. non-dictatorial) SCF in dominant strategies when the domain of the SCF is large enough. In other words, the revelation principle implies that the restriction to mechanisms where agents announce their own characteristic is not important when considering negative results. Thus, the Revelation principle is an appropriate tool for producing negative results. But we will see that to rely entirely on this principle when trying to implement a SCF may yield disastrous results.

A natural question to ask is what happens with the above impossibility results when we weaken the requirement of implementation in dominant strategies to that of implementation in Bayesian equilibrium. The following result, due to Myerson and Satterthwaite [92], answers this question.

Theorem 2'' *In the bilateral trading environment there is no SCF such that:*

1. *It is truthfully implementable in Bayesian Equilibrium.*
2. *It selects individually rational allocations once agents learn their types.*
3. *It selects ex-post efficient allocations.*
4. *Its domain includes all linear utility functions with independent types distributed with positive density and the sets of types have a nonempty intersection.*

Proof (Sketch, see Krishna and Perry [66] for details) By the revenue equivalence theorem (see Klemperer [64], Appendix A), all mechanisms fulfilling conditions 2) and 3) above raise identical revenue. So it is sufficient to con-

sider the Vickrey–Clarke–Groves which, as we remarked before, is not efficient. \square

Again the weakening of any condition in T.2'' may produce positive results [138] (Table 1 presents an illuminating discussion of this issue). For instance, suppose seller valuations are 1 or 3, and buyer valuations are 0 or 2. The mechanism fixes the price at 1.5 and a sale occurs when the valuation of the buyer is larger than the valuation of the seller. This mechanism implements truthfully a SCF satisfying 2) and 3) above. Unfortunately, it does not work when valuations are drawn from a common interval with positive densities.

But unlike T.2–2', there are robust examples of SCF truthfully implementable in Bayesian Equilibrium when conditions 2) or 4) are relaxed. Also, inefficiency converges to zero very quickly when the number of agents increases (see [47]). This is because the equilibrium concept is now weaker and we are approaching a land where incentive compatibility has no bite, as we will see in T.3 below.

First, d'Aspremont and Gerard-Varet [9,10], and Arrow [7] showed that conditions 1)-3)-4) are compatible with individual rationality *before* agents learn their types in the domain of public goods with quasi-linear utility functions. They proposed the “expected externality mechanism” in which each agent is charged the expected externality she creates on the remaining players. Later on, Myerson [90] and Makowski and Mezzetti [74] presented incentive compatible SCF yielding ex-post efficient and individually rational allocations in the domain of exchange economies with quasi-linear preferences and more than two buyers. In Myerson [90], agents have correlated valuations. Buyers are charged even if they do not obtain the object or they may receive money and no object or even receive the object plus some money. Makowski and Mezzetti [74] assume no correlation and that the highest possible valuation for a buyer is larger than the seller's highest possible valuation. They consider a family of mechanisms, called Second Price Auction With Seller (SPAWS), in which the highest bidder obtains the object, the seller receives the first bid and the winning buyer pays the second price. These mechanisms not only induce truthful behavior and yield ex-post efficient and individually rational allocations: For any other mechanism with these properties we can find a SPAWS mechanism yielding the same allocation.

Suppose now that information is *Non-Exclusive* in the sense that the type of each player can be inferred from the knowledge of all the other players' type. Intuition suggests that in this case, incentive compatibility has no bite whatsoever (i. e. T.2'' does not apply) since the behavior of each

player can be “policed” by the remaining players. In order to prove this, we will concentrate on an extreme, but illuminating, case of non-exclusive information, namely Nash equilibrium. In this framework, since information is complete, a direct mechanism is one where each agent announces a state of the world.

Consider the following assumption:

- (W) $\exists z \in A$ such that $\forall \theta \in \Theta, \forall a \in A, aR_i(\theta)z, \forall i \in I$.

This assumption will be called “universally worst outcome” because it postulates the existence of an allocation which is unanimously deemed as the worst. In an exchange economy this allocation would be zero consumption for everybody. Now we have the following result (Repullo [103], Matsushima [79]):

Theorem 3 *If $n = 2$ and W holds, any SCF is truthfully implementable in Nash Equilibrium. If $n > 2$, any SCF is truthfully implementable in Nash Equilibrium.*

Proof When $n = 2$ consider the following outcome function: $g(\theta', \theta') = f(\theta') \forall \theta' \in \Theta, g(\theta', \theta'') = z$ for all $\theta' \neq \theta''$. Clearly, truth is an equilibrium. When $n > 2$, consider the following outcome function: If m is such that $n - 1$ agents announce state θ' , then $g(m) = f(\theta')$. Otherwise, $g(\cdot)$ is arbitrary. Clearly truth is an equilibrium as well in this case. \square

The first thing to notice is the difference between the cases of two and more than two individuals. We will have more to say about this in the next section. The second is that the construction in Theorem 3 produces a large number of equilibria, and that there seems to be no good reason for individuals to coordinate in the truthful equilibria.

For instance, suppose workers can be either fit or unfit. When a profit-maximizing firm asks its employees about their characteristics, and all workers are fit, a unanimous announcement such as “we are all unfit” is an equilibrium. If fit workers are required hard work and unfit workers are asked to light work, do you think it is reasonable that workers coordinate in the truthful equilibrium? A more elaborate example was produced by Postlewaite and Schmeidler [100]: There are three agents. The first agent has no information and agents 2 and 3 are perfectly informed. The ranking of agent 1 over alternatives is the opposite of agents 2 and 3 who share the same preferences. The SCF is the top alternative of agent 1 in each state. It is intuitively clear that besides the truthful equilibria, there is another untruthful equilibrium where both informed agents lie and they are strictly better off than under truthful behavior. Again, coordination in the truthful equilibrium seems very unlikely. Thus, we have to recog-

nize that we have a problem here. The next section will tell you how we can solve it.

Summing up, what do we learn from the results in this section?

1. When looking for an implementable SCF, a useful first test is whether this SCF yields incentives for the agents to tell the truth, see T.1. But this test is incomplete because of the existence of equilibria other than the truthful one, see T.3. These untruthful equilibria sometimes sound more plausible than the truthful one.
2. All impossibility theorems – T.2–2'–2'' – have the same structure: Truthful implementation, individual rationality/non-dictatorship, efficiency/large range of the SCR and large domain. Usually in social choice environments conditions 2 and 3 are weaker than in economic environments but the condition on the domain is stronger.
3. The classic story of the market making possible efficient allocation of resources under private information has to be revised. Private information in many cases precludes the existence of *any* mechanism achieving efficient and individually rational allocations under informational decentralization, see T.2–2'–2''.
4. The same remarks apply to naive applications of the Coase theorem where agents are supposed to achieve Pareto efficient allocations just because they have contractual freedom (ditto about Bargaining Theory). In the parlance of Coase, private information is an important transaction cost.
5. When mechanisms with adequate properties exist, like those proposed by Arrow, d'Aspremont and Gerard-Varet, Myerson and Makowski and Mezzetti, they are not of the kind that we see in the streets. Careful design is needed. These mechanisms are tailored to specific assumptions on valuations, thus their range of applicability may be limited.

Monotonicity and How to Avoid It

We have seen that equilibria other than the truthful one are likely to arise. We have also seen that these equilibria cannot be disregarded a priori. So we have to find a way of getting rid of equilibria that do not yield desirable allocations. Under Dominant Strategies, clearly, if all preference orderings are strict, implementation and truthful implementation becomes identical, see [38], Corollary 4.1.4 ([68] presents other conditions under which this result holds. See [103] for the case where implementation and truthful implementation in dominant strategies do not coincide). For the ease of exposition we consider next Nash equilibria.

It turns out that the key to this issue in the case of Nash Equilibrium is the following monotonicity property, sometimes called Maskin monotonicity because Maskin [76] established its central relevance to implementation.

- (M) A SCR F is Monotonic if

$$\{a \in F(\theta), aR_i(\theta)b \rightarrow aR_i(\theta')b \ \forall i \in I\} \rightarrow a \in F(\theta')$$

Monotonicity says that if an allocation is chosen in state θ and this allocation doesn't fall in anybody's ranking in state θ' , this allocation must also be chosen in θ' . We will also speak of a "monotonic transformation of preferences at θ " when the requirement $aR_i(\theta)b \rightarrow aR_i(\theta')b \ \forall i \in I$ is satisfied. This requirement simply says that the set of preferred allocations shrinks when we go from θ to θ' .

Monotonicity looks like a not unreasonable property, even though as we will see in a moment, there are cases in which it is incompatible with other very desirable properties. In any case the importance of monotonicity comes from the fact that it is a necessary condition for implementation in Nash Equilibrium, as proved by Maskin [76].

Theorem 4 *If a SCR is implementable in Nash Equilibrium it is Monotonic.*

Proof If F is Nash implementable, there must be a mechanism (M, g) such that $\forall a \in F(\theta)$, there is a Nash equilibrium m^* , such that $g(m^*) = a$. Since $aR_i(\theta)b \rightarrow aR_i(\theta')b \ \forall i \in I$, m^* is also a Nash Equilibrium at θ' . Since F is implementable, $a \in F(\theta')$. \square

Let us now discuss the concept of monotonicity. First, the bad news. Popular concepts in voting, like Plurality, Borda Scoring and Majority Rule are not monotonic, neither is the Pareto correspondence, see Palfrey and Srivastava [96], p. 484. Even the venerable Walrasian correspondence is not monotonic! The failure of the Pareto and the Walrasian SCR to be monotonic can be amended: If preferences are strictly increasing in all goods, the Pareto SCR is monotonic in economic environments. The *Constrained Walrasian* SCR – in which consumers maximize with respect to the budget constraint and the availability of resources – is also monotonic. More serious is a result due to Hurwicz [55] that uses two weak conditions on a SCR defined in the domain of exchange economies.

- (L) The domain of F contains all preferences representable by linear utility functions.
- (ND) If $a \in F(\theta)$ and $aI_i(\theta)b \ \forall i \in I$, then $b \in F(\theta)$.

The first condition is a rather modest requirement on the richness of the domain of F . The second is a non-discrimination property which says that if everybody considers two

allocations to be indifferent and one allocation belongs to the SCR then it must be the other. Now we have the following:

Theorem 5 *Let F be a SCR satisfying L and ND and such that:*

1. *It is Nash implementable.*
2. *It selects individually rational allocations.*
3. *Then, if x is a Walrasian allocation at θ , $x \in F(\theta)$.*

Proof (Sketch, see [130] for details) Take an economy θ . Let x be a Walrasian allocation for θ . Consider a new economy, called θ^L , where the marginal rates of substitution among goods are constant and equal to a vector of Walrasian prices. By individual rationality, F must select an allocation which is indifferent to x . By ND, $x \in F(\theta^L)$. Since F is Nash implementable, it satisfies M. Now since $xR_i(\theta^L)b \rightarrow xR_i(\theta)b \ \forall i \in I$, by M, $x \in F(\theta)$. \square

Thus under weak conditions, Walrasian allocation are always in the set of those selected by a Monotonic SCR. And these allocations may fail to satisfy properties of fairness or justice as pointed out by the critics of the market. Under stronger assumptions, the converse is also true, i.e. only Walrasian allocations can be selected by a Nash implementable SCR [55]. Also, T.5 has the following unpleasant implication.

Theorem 6 *There is no SCF in exchange economies such that:*

1. *It is Nash implementable.*
2. *It selects individually rational allocations.*
3. *ND holds.*
4. *It is defined on all exchange economies.*

Proof T.5 implies that any Walrasian allocation belongs to the allocations selected by F . Since Walrasian equilibrium is not unique for some economies in the domain, hence the result. \square

T.6 has a counterpart in social choice domains, Muller and Satterthwaite [88].

Theorem 6' *There is no SCF in a social choice domain such that:*

1. *It is monotonic.*
2. *It is not dictatorial.*
3. *Its range is A with $\#A > 2$.*
4. *It is defined on all possible preferences.*

An implication of T.6–6' is that single valued SCR are still problematic. But the consideration of multivalued SCR,

brings a new problem: The existence of several Nash equilibria. For instance, if $a, b \in F(\theta)$ with a and b being efficient allocations, agents play a kind of “Battle of the Sexes” game with no clear results. Moreover the Nash Equilibrium in mixed strategies may yield allocations outside $F(\theta)$ (the concern about mixed-strategy equilibria was first raised by Jackson [59]).

Now let us come to the good news. Firstly, the ND condition, which is essential for T.5 to hold, is not as harmless as it appears to be. For instance, it is not satisfied by the Envy-Free SCR, see Thomson [131] for a discussion. Secondly, there are perfectly reasonable SCR which are monotonic: we have already encountered the Constrained Walrasian SCR. Also any SCR selecting interior allocations in $\mathcal{L}A$ when preferences are von Neumann–Morgenstern is monotonic. In the domain of exchange economies with strictly increasing preferences, the Core and the Envy-Free SCR are also monotonic. In domains where indifference curves only cross once – the single-crossing condition – monotonicity vacuously holds. So Monotonicity, restrictive as it is, is worth a try. But before this, let us introduce a new assumption

- (NVP) A SCR f satisfies No Veto Power if $\forall \theta \in \Theta$, $\{aR_i(\theta)b, \forall b \in A, \text{ for at least } n-1 \text{ agents}\} \rightarrow a \in F(\theta)$

In other words, if there is an allocation which is top-ranked by, at least, $n-1$ agents, NVP demands that this allocation belongs to the SCR. This sounds like a reasonable property for large n . Also in exchange economies with strictly increasing preferences and more than two agents, NVP is vacuously satisfied because there is no top allocation for $n-1$ agents.

The following positive result, a relief after so many negative results, was stated and proved by Maskin [76], although his proof was incomplete.

Theorem 7 *If a SCR satisfies M and NVP is Nash implementable when $n > 2$.*

Proof (Sketch) Consider the following mechanism. $M_i = \Theta \times A \times \mathbb{N}$ where \mathbb{N} is the set of natural numbers. The outcome function has three parts.

Rule 1 (Unanimity). If m is such that all agents announce the same state of the world, θ , the same allocation a with $a \in F(\theta)$ and the same integer, then $g(m) = a$.

Rule 2 (One Dissident). If there is only one agent whose message is different from the rest, this agent can choose any allocation that leaves her worse off, according to her preference as announced by others.

Rule 3 (Any other case). $a \in g(m)$ iff a was announced by the agent who announced the highest integer (ties are broken by an arbitrary rule).

Let us show that such a mechanism implements any SCR with the required conditions. Clearly if the true state is $\tilde{\theta}$, $m_i = (\tilde{\theta}, a, 1)$ with $a \in F(\tilde{\theta})$ is a Nash Equilibrium since no agent can gain by saying otherwise, so Condition 1 in the definition of Nash implementation holds. Let us now prove that Condition 2 there also holds. Suppose we have a Nash Equilibrium in Rule 1. Could it be an “untruthful” equilibrium? If so we have two cases. Either the announced preferences are a monotonic transformation of preferences at $\tilde{\theta}$, in which case, M implies that the announced allocation is also optimal at $\tilde{\theta}$. If they are not, there is an agent who can profitably deviate. Clearly, if equilibrium occurs in Rule 2, with, say, agent i as the dissident, any agent other than i can drive the mechanism to Rule 3, so it must be that all these agents are obtaining their most preferred allocation, which by NVP belongs to $F(\tilde{\theta})$. An equilibrium in Rule 3 implies that all agents are obtaining their most preferred allocation which, again by NVP belongs to $F(\tilde{\theta})$. \square

The interpretation of the mechanism given in the proof of T.7 is that if everybody agrees on the state and the allocation is what the planner wants, this allocation is selected. If there is a dissident (a term due to Danilov [37]) she can make her case by choosing an allocation (a “test allocation”) in her lower contour set, as announced by others. Finally, with more than one dissident, it’s the jungle! Any agent can obtain her most preferred allocation by the choice of an integer. Typically, there is no equilibrium in this part of the mechanism. Notice that (M) is just used to eliminate unwanted equilibria.

The mechanism is an “augmented” revelation mechanism (a term due to Mookherjee and Reichelstein [83]), where the announcement of the state is complemented with the announcement of an allocation – this can be avoided if the SCR is single valued – and an integer. The final proof of T.7 was done independently by Williams [137], Repullo [104], Saijo [109] and McKelvey [95].

The case of two agent is more complicated because when an agent deviates from a common announcement and becomes a dissident, she converts the other agent into another dissident! As in T.3, W does the job, i.e. any SCR satisfying M, NVP and W is Nash implementable, see Moore and Repullo [85] and Dutta and Sen [40] for a full characterization. Again, the cases of two agents and more than two agents are different. In some areas of mathematics, such as statistics and differential equations the cases of two dimensions and more than two dimensions are also different. The relationship of these with the findings of implementation is not yet fully explored, see [108].

Under asymmetric information M is substituted by a – rather ugly – *Bayesian Monotonicity* (BM) condition which is a generalization of M to these environments. BM is again necessary and in conjunction with some technical conditions plus incentive compatibility, sufficient for implementation in BE. The interested reader can do no better than to read the account of these matters in [93]. It must be remarked that many well-known SCR – including Arrow–Debreu contingent commodities and some efficient SCR – do not satisfy BM and thus cannot be implemented in BE. However, the Rational Expectations Equilibria and the (interim) Envy-Free SCR satisfy BM, see Palfrey and Srivastava [94].

T.7 was the first positive finding of implementation theory. And it prompted researchers to be more ambitious: Can we implement without Monotonicity? An interesting observation, due to Matsushima [79] and Abreu and Sen [2], is that if agents have preferences representable by von Neumann–Morgenstern utility functions, *any* SCR can be “virtually implemented” in the sense that the set of allocations yielded by Nash equilibria is arbitrarily close to the set of desired allocations. This is because, as we saw before, any SCR mapping in the interior of $\mathcal{L}A$ is Monotonic. Thus allocations in the boundary can be arbitrarily approximated by allocations in the interior.

A more satisfying approach was introduced by Moore and Repullo [86] by considering subgame perfection as the solution concept. It is not possible to explain fully this approach here because it would take us too far; in particular the notion of a mechanism must be generalized to “stage mechanism”. Instead, we give a result that conveys the force of subgame perfect implementation. It refers to public good economies with quasi-linear utility functions – where under dominant strategies the set of economies with inefficient outcomes is large – and with two individuals – where Nash implementability is harder to obtain.

Suppose that utility functions read $U_i = V(y, \theta_i) + m_i$ where $y \in Y \subseteq \Re$, $\theta_i \in \Theta_i$ with $\#\Theta_i < \infty$ and $m_i \in \Re$, $i = 1, 2$. The set of allocations $A = \{(y, m_1, m_2) \in Y \times \Re^2 / m_1 + m_2 \leq \omega\}$ where ω are the endowments of “money”. Moore and Repullo [86] proved the following:

Theorem 8 *Any SCF is implementable in Subgame Perfect Equilibrium in the domain of economies explained above.*

Moore and Repullo proved that many SCR which could not be implemented in Nash Equilibrium can be implemented in Subgame Perfect Equilibrium. This is because subgames can be designed to kill unwanted equilibria without using monotonicity. Their result was improved upon by Abreu and Sen [2]. The problem with this approach is that the concept of subgame perfection

is problematic because it requires that, no matter what has happened in past, in the remaining subgame, players are rational, even if this subgame was attained because some players made irrational choices.

The Moore–Repullo result was not only important by itself but it opened the way to the consideration of other equilibrium concepts that allow very permissive results. For instance, Palfrey and Srivastava [96] proved the following result

Theorem 8' *Any SCR satisfying NVP is implementable in Undominated Nash Equilibrium.*

At this point, it seemed that by invoking the adequate refinement of Nash equilibrium, any SCR could be implemented. But the implementing mechanisms were getting weird and some people were beginning to get suspicious. Why and how is discussed in the next section.

Summing up the results obtained here, we have the following:

1. (Maskin) Monotonicity is a necessary and in many cases sufficient condition for implementation in Nash Equilibrium, see T.4 and 7. Similar results are obtained with Bayesian Monotonicity in Bayesian Equilibrium.
2. The Monotonicity requirements are not harmless. Many solution concepts do not satisfy it. Even worse, Monotonicity has some unpalatable consequences, see T.5–6.
3. Monotonicity can be avoided by considering stage games or refinements of Nash Equilibrium. Practically, any reasonable SCR can be implemented in this way, see T.8–8'.

The Limits of Design

So far we have assumed that there are no limits to what the designer can do. She can pick up any mechanism with no restrictions on its shape. This procedure, indeed, pushes the possibilities of design to the limit. But by doing this, we have learned a good deal about the limitations of the theory of implementation. It is fair to say that today the consensus is that there are *some extra properties* which should be considered when designing an implementing mechanism. We review here five approaches to this question.

Game-Theoretical Concerns. Jackson [59] was the first to point out that some mechanisms had unusual features from the point of view of game theory: Some subgames have no Nash equilibrium. Message spaces, which in the corresponding game become strategy spaces, are unbounded or open. Thus, in the integer game con-

sidered in T.7, if agents eliminate dominated strategies, each integer is dominated by the next highest one and no integer is undominated: Those agents who eliminate dominated strategies are unable to make a choice. These constructions eliminate unwanted equilibria, which as we saw before, is the problem with Nash implementation. Jackson illustrates his point by showing that under no restrictions on mechanisms, any SCR can be implemented in undominated strategies, a weak solution concept. Then he requires that the mechanism be *Bounded* in the following sense: whenever a strategy m_i is dominated, there is another strategy dominating m_i and which is undominated. He shows that implementation in undominated strategies with bounded mechanisms result many times in incentive compatibility, which as we saw in Sect. “[The Main Concepts](#)” is a hard requirement. This shows the bite of the boundedness assumption. However, in the case of implementation with undominated Nash equilibrium, the boundedness assumption has little impact, see [61,126]. The first of these papers introduced a related requirement, the *Best Response Property*: for every strategy played by the other agents, each agent has a best response.

Natural Mechanisms Given that we have run so far from the kind of mechanisms we are used to, it seems reasonable to ask what can be implemented by mechanisms that resemble real-life mechanisms. These mechanisms must be simple too because simplicity is an important characteristic in practice. Let us call them *Natural Mechanisms*. Dutta, Sen and Vohra [41] consider mechanisms in which messages are prices and quantities and thus resemble market mechanisms. Their approach was refined by Saijo, Tatamitani and Yamato [111] who demanded the best response property as well. They showed that several well-known SCR such as the (constrained) Walrasian, are implementable in Nash equilibrium. Beviá, Corchón and Wilkie [21] showed that in Bertrand-like market games, the Walrasian SCR is implementable in Nash and Strong equilibrium, showing that the fear of coalitions destabilizing market outcomes is, at least, partially unwarranted. Sjöström [127] considered quantity mechanisms, reminiscent of those used by Soviet planners, with negative results about what these mechanisms can achieve. In public good economies, Corchón and Wilkie [31] and Peleg [97] introduced a market mechanism implementing Lindahl allocations in Nash and Strong equilibrium. The mechanism works because Lindahl prices have to add up to the marginal cost. If an agent pretends to free ride she decreases the quantity of the public good. Here, contrary to Samuelson’s dictum it is in the selfish interest of each per-

son to give true signals. Pérez-Castrillo and Wettstein [98] offered a bidding mechanism that implements efficient allocations when choosing between a finite number of public projects. They also applied these ideas to back up Shapley value.

Credibility Another implicit assumption is that once the mechanism is in place, there is no way to stop it. Thus, if for some m , $g(m)$ is a “universally worst outcome”, the planner has to deliver this allocation even if she is trying to implement a Pareto Efficient allocation. Is this a credible procedure? In many cases, if the planner is a real person it seems that she would do her best to avoid $g(m)$! Here we have two possibilities: Either we identify additional constraints on the planner that look reasonable or we jump to model the planner as a full-fledged player. The first road leads us to identify a subset of allocations of A , say X , which can never be used by the mechanism. For instance, in Chakravorty, Corchón and Wilkie [25] X is the set of allocations that are never selected by the SCR for some state of the world, i.e. $X = \{a \in A/\exists \theta \in \Theta, a \in F(\theta)\}$. The motivation for this definition is that it hardly seems credible that the planner can choose an allocation that is never intended to be implemented. Redefining the allocation set as $A \equiv A \setminus X$ the definitions of a mechanism and an implementable SCR can be easily translated in this framework. However, depending on the domain, SCR that are monotonic when defined on A are no longer monotonic when defined on A : For instance, the (constrained) Walrasian SCR. Thus, these SCR can not be implemented when the planner can only use allocations in A . A weakness of this approach is that the list of reasonable constraints on allocations may be large. The second possibility drives us to model implementation as a signaling game where the planner receives signals – messages – from the agents, updates her beliefs and then chooses an allocation which maximizes her expected utility [11]. Again, some SCR that are Nash implementable, are not implementable in this framework. However, in this case there are SCR that are not Nash implementable but are implementable in this framework. This is because the model takes a basic assumption of game theory to the limit, namely, that agents know the strategies of other players. In this case, the planner knows if a report on agents’ types is truthful or not before the allocation is delivered!

Renegotiation Another strong assumption is that the mechanism prescribes actions that can not be changed by agents. This contradicts experiences such as black markets where agents trade on the existing goods (Hammond [49]). A way of modeling this is to assume that

agents are able to renegotiate some allocations (Maskin and Moore [77]. Renegotiation in a different context was considered by Rubinstein and Wolinsky [107]). Assuming that agents have complete information, this is formalized by means of the concept of a reversion function. This function, say r , maps each allocation and each state of the world into a new allocation, i.e. $r: A \times \Theta \rightarrow A$. The reversion function induces new preferences, called *reverted preferences* (this is the “translation principle” in Maskin and Moore [77]). Notice that reverted preferences are state dependent even if preferences are not. Formally, given a reversion function r , the *reversion* of $R(\theta)$, denoted by $R'(\theta)$ is defined as $aR'_i(\theta)b \Leftrightarrow r(a, \theta)R_i(\theta)r(b, \theta)$, $\forall a, b \in A, \forall i \in I$. Given a reversion function r , we can interpret that agents’ preferences are the reverted preferences. Then, all definitions given before can be adapted to this case. Again, SCR that were monotonic there, are not so in this framework and vice versa. See Jackson and Palfrey [62] for applications. An extension to the case where there are several renegotiation functions is given by Amorós [6]. A weakness of this approach is that models renegotiation as a “black box”.

Multiple Implementation Maskin [75] was the first to realize that the notion of implementation requires the planner to know the solution concept used by the agents to analyze the game. He proposed the notion of “Double Implementation” where a SCR was implemented at the same time in Nash and Strong equilibria. He showed that many Nash implementable SCR indeed are doubly implementable. We have seen in Point 2 above that the (constrained) Walrasian and Lindahl SCR are doubly implementable by natural mechanisms. They are also doubly implementable by abstract mechanisms, Schmeidler [116]. Double implementation also occurs with several solutions to the problem of the commons, Shin and Suh [124] and Pigouvian Taxes, Alcalde, Corchón and Moreno [4]. Yamato [139] introduced another type of double implementation by requiring implementation in Nash and Undominated Nash Equilibria [139]. He showed that in a large class of exchange economies with at least three agents, monotonicity is necessary and sufficient for double implementation. Saijo, Sjöström and Yamato [112] considered implementation in Dominant Strategies and Nash Equilibrium. Clearly, other variations of the idea of Double Implementation are possible, see Point 4 in Sect. “[Unsolved Issues and Further Research](#)” below.

Summing up, it is now clear that implementing mechanisms can not be just “anything”. Their features matter. Demanding that mechanisms satisfy the best response property, be simple, not use extreme allocations, be robust

to the possibility of renegotiation and implement in several equilibrium concepts makes our lives more difficult but makes our models a great deal better.

Unsolved Issues and Further Research

Implementation with State Dependent Feasible Sets

A motivation of implementation theory was to study the possibility of socialism. However, *all* the results presented in this survey refer to environments where the feasible set is given, a far cry from any kind of planning procedure. In fact, there are only a handful of papers dealing with implementation when the feasible set is unknown: Postlewaite [99] and Sertel and Sanver [123] studied manipulation of endowments. Hurwicz, Maskin and Postlewaite [57] studied implementation assuming that endowments/production possibilities can be hidden or destroyed but never exaggerated. Instead of a mechanism we have a collection of state dependent mechanisms each meant for an economy. After the mechanism is played, production capabilities are shown, e.g. endowments are put on the table. This idea was worked out in a series of papers by Hong on private good economies, Hong [52], and by Tian on public good economies, Tian and Li [134]. Serrano and Vohra [120] worked out implementation of the core and Dagan, Serrano and Volij [36] of taxation methods. And that’s all folks! Why has such an important issue been almost neglected? My explanation is that the proposed mechanisms are difficult to understand. Another approach has been tried by Corchón and Triossi [35] where a reversion function takes care of restoring feasibility when messages lead to unfeasible allocations. The approach is tractable and simpler but relies on the black box of the renegotiation function.

Sociological Factors/Bounded Rationality

So far, all the solution concepts describing the behavior of agents are game-theoretical. In recent years, we have seen a host of equilibrium concepts based on “irrational” agents. It would be interesting to see what SCR can be implemented with these forms of behavior. Eliaz [42] considers “Fault Tolerant” implementation where a subset of players (“faulty players”) fail to achieve their optimal strategies. Under complete information, No Veto Power and a strong form of Monotonicity are sufficient for implementation when the number of faulty players is less than $n/2 - 1$, $n > 2$. Matsushima [80] shows that a small preference for honesty is sufficient to knock down unwanted equilibria, see also Corchón and Herrero [34].

Dynamic Implementation

The theory presented here is static but there are some papers dealing with implementation in dynamic set-ups. We mention a few: Freixas, Guesnerie and Tirole [43] studied the “Ratchet Effect”, where firms underproduce for fear of being asked to do too much in the future. Kalai and Ledyard [63] showed that if the planner is sufficiently patient, every SCR is dominant-strategy implementable. Burguet [22] showed that the revelation principle does not hold when outcomes are chosen in several periods. Candel [24] proved a revelation principle in a model where a public good is produced in two periods. Finally, Cabrales [23] and Sandholm [114] studied implementation in an evolutionary setting. A related topic is that of complexity, see Conitzer and Sandholm [28]. It seems likely that a dynamic theory of incentives will bring new insights and will need new analytical tools.

Robustness Under Incomplete Information

When designing a mechanism, sometimes the planner does not know the structure of information. In this case a mechanism must implement regardless the structure of information, i.e. priors of agents, type spaces, etc. Corchón and Ortuño-Ortín [30] approached the problem by assuming that the economy is composed of “islands” and that there is complete information inside each island. A mechanism robustly implements a SCR if it does it in BE for every possible prior (compatible with the island assumption) and in Uniform Nash Equilibrium. The latter requires that an equilibrium strategy for an agent must be the best reply to what other agents in the island play and to any possible message sent by agents outside the island when they follow their equilibrium strategies (D’Aspremont and Gerard-Varet [10]). They showed that any SCR satisfying M and NVP is robustly implementable (a later contribution by Yamato [140] showed that Robust and Nash Implementation coincide in this framework). The same concern has been approached in a series of papers by Bergemann and Morris (see e.g. [19]) where they ask SCR to be implemented whatever the players’ beliefs and higher order beliefs about other players’ types. Artemov, Kunimoto and Serrano [8] require implementation for the payoff type space and the space of first-order beliefs about other agents’ payoff types. They obtain very permissive results.

In a different vein, Koray [65], has argued that, since priors are not contractible, the regulator needs to be regulated in order to stop her from manipulating the priors. He shows that the outcomes of this game vary over a wide

spectrum. Again the need of prior-free implementation is clear.

Answers to the Questions

1. Yes. We already saw in 4.3, Point 2, that “Bertrand-like” mechanisms implement the Constrained Walrasian SCR in Nash and Strong equilibrium. But this is not all: Schmeidler [116] exploited the connection between price taking – which underlies Walrasian equilibrium – and “strategy taking”, which underlies Nash and Strong equilibrium and obtained double implementation by a mechanism which does not resemble the market. Implementation of the Lindahl SCR by an abstract mechanism was obtained by Walker [136] building on previous papers by Groves and Ledyard and Hurwicz. Unfortunately, these positive results turn negative when we consider Arrow–Debreu contingent commodities, Chattopadhyay, Corchón and Naeve [26] and Serrano and Vohra [121].
2. A merger affects social welfare in two ways: Positively, from cost savings and negatively, from restricting competition. The first effect is uncertain and, by now, I do not have to convince you that we should take with utmost caution all announcements made by firms concerning cost savings. Corchón and Faulí-Oller [33] show that under a condition that is fulfilled in several standard IO models, the SCR that maximize social surplus can be implemented by a dominance solvable mechanism with budget balance.
3. There is a very simple mechanism which attains maximum surplus, Loeb and Magath [72]. But in this mechanism the monopolist receives all the surplus and the demand function must be known by the planner. These points were worked out by subsequent contributions from Baron and Myerson, Lewis and Sappington, Sibley and others.
4. By now the reader should know the difficulties of implementing efficient public decisions. When information is exclusive this is impossible, even though an approximate efficient decision can be obtained when the number of agents is large. When information is complete, we have seen several examples of mechanisms implementing efficient outcomes.
5. There is no difference between implementing market or fair outcomes. Both have to pass the same tests, i.e. incentive compatibility, monotonicity and simplicity/credibility of design. In exchange economies, Thomson [133] presents a simple and elegant mechanism that implements envy-free allocations in Nash Equilibrium. In cooperative production, Corchón and Puy [32] pre-

sented a family of mechanisms that implement in Nash Equilibrium any efficient SCR where the distribution of rewards is a continuous function of efforts.

6. Yes! An uninformed planner can set up a mechanism that yields efficient outcomes in circumstances where the market yields inefficient allocations, i. e. under externalities or public goods see Point 5 in 4.3 above. All we need is non-exclusive information and that the SCR be Monotonic, the latter requirement can be skipped under refinements of Nash Equilibrium.
7. Not completely. Suppose complete information among three or more judges and that they all perceive the same quality of a given performance. Clearly, truth is an equilibrium, because if all judges minus you tell the truth you cannot change the outcome by saying something different. Unfortunately, any unanimous announcement is also an equilibrium by the same reason. Thus we are in a situation akin to T.3. Fortunately, if preferences of judges fulfill certain restrictions, full implementation of the true ranking of ice skaters is possible, because Monotonicity and No Veto Power hold so T.7 applies, Amorós, Corchón and Moreno [5]. If judges have differential information, the truth is no longer implementable as suggested by T.2''. See Gerardi, McLean and Postlewaite [44] for further insights and references on this problem.
8. ????? Do you think that we have all answers? This is just economics!! In any case, the application of Implementation to environmental policies may be a topic of great importance in the future, see Baliga and Maskin [12]. Finally I will tell you why I like implementation theory so much. Firstly, the implementation model solves the problems of the General Equilibrium model mentioned in Sect. "Brief History of Implementation Theory", namely: 1: It models a general economic system. 2: All variables are endogenously determined by the interaction of agents. 3: Agents incentives are carefully modeled and are taken fully into account. Secondly, the theory is not based on assumptions like convexity or continuity/differentiability which, no matter how much we are used to them, are very stringent. By the way, a beautiful paper by Laffont and Maskin (referenced in their [68] survey) developed incentive compatibility in a differentiable framework.

Acknowledgments

I am grateful to Pablo Amorós, Claude d'Aspremont, Carmen Beviá, Luis Cabral, Eric Maskin, Carlos Pimienta, Sororro Puy, Tömas Sjöström, William Thomson, Matteo Triossi, Galina Zudenkova and an anonymous referee for

helpful suggestions and to the Spanish Ministry of Education for financial support under grant SEJ2005-06167. I also thank the Department of Economics, Stern School of Business, NYU, for their hospitality while writing this survey. This survey is dedicated to Leo Hurwicz to celebrate his 90th birthday and his Nobel Prize and to the memory of those who contributed to the area and are no longer with us: Louis-André Gerard-Varet, Jean-Jacques Laffont, Richard McKelvey and Murat Sertel.

Bibliography

1. Abreu D, Sen A (1990) Virtual implementation in Nash equilibrium. *Econometrica* 59:997–1021
2. Abreu D, Sen A (1991) Subgame Perfect Implementation: A Necessary and Almost Sufficient Condition. *J Econ Theory* 50:285–299
3. Akerlof G (1970) The Market for Lemons: Qualitative Uncertainty and the Market Mechanism. *Q J Econ* 84:488–500
4. Alcalde J, Corchón L, Moreno B (1999) Pigouvian Taxes: A Strategic Approach. *J Public Econ Theory* 1 2:271–281
5. Amorós P, Corchón L, Moreno B (2002) The Scholarship Assignment Problem. *Games Econ Behav* 38:1–18
6. Amorós P (2004) Nash Implementation and Uncertain Renegotiation. *Games Econ Behav* 49:424–434
7. Arrow K (1979) The Property Rights Doctrine and Demand Revelation under Incomplete Information. Technical report No. 243, IMSSS, Stanford University, Stanford
8. Artemov G, Kunimoto T, Serrano R (2007) Robust virtual implementation with incomplete information: Towards a reinterpretation of the Wilson doctrine. W.P. 2007–06, Brown University, Providence
9. d'Aspremont C, Gerard-Varet L-A (1975) Individual Incentives and Collective Efficiency for an Externality Game with Incomplete Information. CORE DP 7519. Université Catholique de Louvain, Louvain-la-Neuve
10. d'Aspremont C and L-A Gerard-Varet (1979) Incentives and Incomplete Information. *J Public Econ* 11:25–45
11. Baliga S, Corchón L, Sjöström L, Sjöström T (1997) The theory of implementation when the planner is a player. *J Econ Theory* 77:15–33
12. Baliga S, Maskin E (2003) Mechanism Design for the Environment. In: Maler K, Vincent JR (eds) *Handbook of Environmental Economics*, vol 1, ch 7, pp 305–324
13. Baliga S, Sjöström T (2007) Mechanism Design: Recent Developments. In: Blume L, Durlauf S (eds) *The New Palgrave Dictionary of Economics*, 2nd edn. Palgrave Macmillan, New York
14. Barberá S (1983) Strategy-proofness and pivotal voters: A direct proof of the Gibbard–Satterthwaite theorem. *Int Econ Rev* 24:413–417
15. Barberá S, Peleg B (1990) Strategy-proof voting schemes with continuous preferences. *Soc Choice Welf* 7:31–38
16. Barberá S, Sonnenschein H, Zhou L (1991) Voting by Committees. *Econometrica* 59(3):595–609
17. Barone E (1935) The Ministry of Production in a Collectivist State. Translated from the Italian and reprinted in F. A. von Hayek *Collectivist Economic Planning*. Routledge and Keegan, London

18. Benoit JP (2000). The Gibbard–Satterthwaite theorem: a simple proof. *Econ Lett* 69:319–322
19. Bergemann D, Morris S (2005) Robust Mechanism Design. *Econometrica* 73:1771–1813
20. Beviá C, Corchón L (1995) On the Generic Impossibility of Truthful Behavior. *Econ Theory* 6:365–371
21. Beviá C, Corchón L, Wilkie S (2003) Implementation of the Walrasian Correspondence by Market Games. *Rev Econ Des* 7:429–442
22. Burguet R (1990) Revelation in Informational Dynamic Settings. *Econ Lett* 33:237–239. Corrigendum (1994) 44:451–452
23. Cabrales A (1999) Adaptive Dynamics and the Implementation Problem with Complete Information. *J Econ Theory* 86:159–184
24. Candel F (2004) Dynamic Provision of Public Goods. *Econ Theory* 23:621–641
25. Chakravorty B, Corchón L, Wilkie S (2006) Credible Implementation. *Games Econ Behav* 57:18–36
26. Chattopadhyay S, Corchón L, Naeve J (2000) Contingent Commodities and Implementation. *Econ Lett* 68:293–298
27. Clarke E (1971) Multipart pricing of Public Goods. *Public Choice* 11:17–33
28. Conitzer V, Sandholm T (2002) Complexity of Mechanism Design. In: *Proceedings of the 18th Annual Conference on Uncertainty in Artificial Intelligence (UAI-02)*, Edmonton, Canada
29. Corchón L (1996) *The Theory of Implementation of Socially Optimal Decisions in Economics*. St. Martin's Press, New York
30. Corchón L, Ortuño-Ortín I (1995) Robust Implementation under Alternative Information Structures. *Econ Des* 1:159–171
31. Corchón L, Wilkie S (1996) Doubly Implementing the Ratio Correspondence by a Market Mechanism. *Rev of Econ Des* 2:325–337
32. Corchón L, Puy S (2002) Existence and Nash Implementation of Efficient Sharing Rules for a Commonly Owned Technology. *Soc Choice Welf* 19:369–379
33. Corchón L, Faulí-Oller R (2004) To Merge or not to Merge: That is the Question. *Rev Econ Des* 9:11–30
34. Corchón L, Herrero C (2004) A Decent Proposal. *Spanish. Econ Rev* 6:107–125
35. Corchón L, Triossi M (2005) Implementation with State Dependent Feasible Sets: A Renegotiation Approach. WP, Carlos Alberto Notebooks, #24, Torino
36. Dagan N, Serrano R, Volij O (1999) Feasible implementation of taxation methods. *Rev Econ Des* 4:57–72
37. Danilov V (1992) Implementation via Nash Equilibria. *Econometrica* 60(1):43–56
38. Dasgupta P, Hammond P, Maskin E (1979) The implementation of Social Choice Rules: Some Results on Incentive Compatibility. *Rev Econ Stud* 46:185–216
39. Dutta B, Sen A (1991) Implementation under strong equilibrium. A complete characterization. *J Math Econ* 20:49–67
40. Dutta B, Sen A (1991) A Necessary and Sufficient Condition for Two-Person Nash Implementation. *Rev Econ Stud* 58:121–128
41. Dutta B, Sen A, Vohra R (1995) Nash Implementation through Elementary Mechanisms in Economic Environments. *Econ Des* 1:173–204
42. Eliaz K (2002) Fault Tolerant Implementation. *Rev Econ Stud* 69:589–610
43. Freixas X, Guesnerie R, Tirole J (1985) The Ratchet Effect. *Rev Econ Stud* 52:173–191
44. Gerardi D, McLean R, Postlewaite A (2005) Aggregation of Expert Opinions. Cowles Foundation Discussion Paper # 1503, Yale
45. Gibbard A (1973) Manipulation of Voting Schemes: A General Result. *Econometrica* 41:587–602
46. Green J, Laffont JJ (1979) *Incentives in Public Decision Making*. Elsevier, Amsterdam
47. Gresik T, Satterthwaite M (1989) The Rate at Which a Simple Market Converges to Efficiency as the Number of Traders Increases. *J Econ Theory* 48:304–332
48. Groves T (1973) Incentives in Teams. *Econometrica* 41:617–631
49. Hammond P (1987) Markets as Constraints: Multilateral Incentive Compatibility in Continuum Economies. *Rev Econ Stud* 54:399–412
50. Harris M, Townsend R (1981) Resource Allocation Under Asymmetric Information. *Econometrica* 49:33–64
51. Harsanyi J (1967). Games with Incomplete Information Played by 'Bayesian' Players. Parts I, II and III. *Management Science* 14:159–182, 320–334 and 486–502
52. Hong L (1998) Feasible Bayesian Implementation with State Dependent Feasible Sets. *J Econ Theory* 80:201–221
53. Hurwicz L (1959) Optimality and Informational Efficiency in Resource Allocation Processes. In: Arrow KJ (ed) *Mathematical Methods in the Social Sciences*. Stanford University Press, Stanford, pp 27–46
54. Hurwicz L (1972) On Informationally Decentralized Systems. In: Radner R, McGuire CB (eds) *Decision and Organization: A Volume in Honor of Jacob Marslak*. Elsevier, Amsterdam, pp 297–336
55. Hurwicz L (1979) On Allocations Attainable Through Nash Equilibria. *J Econ Theory* 21:40–65
56. Hurwicz L, Walker M (1990) On the Generic Non-Optimality of Dominant Strategy Mechanisms. *Econometrica* 58(3):683–704
57. Hurwicz L, Maskin E, Postlewaite A (1995) Feasible Nash Implementation of Social Choice Rules when the Designer does not know Endowments or Production set. In: Ledyard J (ed) *The Economics of Informational Decentralization: Complexity, Efficiency and Stability*. Kluwer, Amsterdam
58. Jackson M (1991) Bayesian Implementation. *Econometrica* 59:461–477
59. Jackson M (1992) Implementation in Undominated Strategies: A Look to Bounded Mechanisms. *Rev Econ Stud* 59:757–775
60. Jackson MO (2001) A Crash Course in Implementation Theory. *Soc Choice Welf* 18:655–708
61. Jackson MO, Palfrey T, Srivastava S (1994) Undominated Nash Implementation in Bounded Mechanisms. *Games Econ Behav* 6:474–501
62. Jackson MO, Palfrey T (2001) Voluntary implementation. *J Econ Theory* 98:1–25
63. Kalai E, Ledyard J (1988) Repeated Implementation. *J Econ Theory* 83:308–317
64. Klemperer P (1999) Auction theory, a guide to the literature. *J Econ Surv* 13:227–268
65. Koray S (2005) The Need of Regulating a Bayesian Regulator. *J Regul Econ* 28:5–21

66. Krishna V, Perry M (1997) Efficient mechanism design. Unpublished paper, Penn State University, Pennsylvania
67. Krishna V (2002) Auction Theory. Academic Press, San Diego
68. Laffont JJ, Maskin E (1982) The theory of incentives: An overview. In: Hildenbrand W (ed) *Advances in Economic Theory*. 4th World Congress of the Econometric Society. Cambridge University Press, Cambridge
69. Laffont JJ, Martimort D (2001) *The Theory of Incentives: The Principal-Agent Model*. Princeton University Press, New Jersey
70. Lange O (1936). On the Economic Theory of Socialism. *Rev Econ Stud* 4:53–71, 123–142
71. Ledyard J, Roberts J (1974) On the Incentive Problem with Public Goods. Discussion Paper 116, Centre for Mathematical Studies in Economics and Management Science. Northwestern University
72. Loeb M, Magath W (1979) A Decentralized Method for Utility Regulation. *J Law Econ* 22:399–404
73. Ma A, Moore J, Turnbull S (1988) Stop Agents from Cheating. *J Econ Theory* 46:355–372
74. Makowski L, Mezzetti C (1993) The possibility of efficient mechanisms for trading an indivisible object. *J Econ Theory* 59:451–465
75. Maskin E (1985) The theory of implementation in Nash equilibrium: A Survey. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) *Social Goals and Social Organization*. Cambridge University Press, Cambridge, pp 173–204
76. Maskin E (1999) Nash Equilibrium and Welfare Optimality. (Circulating in working paper version since 1977) *Rev Econ Stud* 66(1):23–38
77. Maskin E, Moore J (1999) Implementation with Renegotiation. *Rev Econ Stud* 66:39–56
78. Maskin E, Sjöström T (2002) Implementation Theory. In: Arrow KJ, Sen AK (ed) Chapter 5 in *Handbook of Social Choice Welfare*. Elsevier, Amsterdam
79. Matsushima H (1988) A New Approach to the Implementation Problem. *J Econ Theory* 45:128–144
80. Matsushima H (2008) Role of honesty in full implementation. *J Econ Theory* 139:353–359
81. McKelvey R (1989) Game Forms for Nash Implementation of General Social Choice Correspondences. *Soc Choice Welf* 6:139–156
82. Mirless J (1971) An Exploration in the Theory of Optimum Income Taxation. *Rev Econ Stud* 38:175–208
83. Mookherjee D, Reichelstein S (1990) Implementation via augmented revelation mechanisms. *Rev Econ Stud* 57:453–475
84. Moore J (1992) Implementation, contracts and renegotiation in environments with complete information in *Advances in Economic Theory*. In: Laffont JJ (ed) 4th World Congress of the Econometric Society, vol I. Cambridge University Press, Cambridge, pp 182–282
85. Moore J, Repullo R (1990) Nash Implementation: A Full Characterization. *Econometrica* 58:1083–1089
86. Moore J, Repullo R (1988) Subgame Perfect Implementation. *Econometrica* 56:1191–1220
87. Moulin H (1979) Dominance Solvable Voting Schemes. *Econometrica*, 47(6):1337–1351
88. Muller E, Satterthwaite M (1977) The equivalence of strong positive association and strategy-proofness. *J Econ Theory* 14:412–418
89. Myerson R (1979) Incentive Compatibility and the Bargaining Problem. *Econometrica* 47:61–73
90. Myerson R (1981): Optimal auction design. *Math Oper Res* 6:58–73
91. Myerson R (1985) Bayesian Equilibrium and Incentive Compatibility: An Introduction. In: Hurwicz L, Schmeidler D, Sonnenschein H (eds) *Social Goals and Social Organization*, Chapter 8. Cambridge University Press, Cambridge
92. Myerson R, Satterthwaite MA (1983): Efficient mechanisms for bilateral trading. *J Econ Theory* 29:265–281
93. Palfrey TR (2002) Implementation Theory. In: Aumann RJ, Hart S (eds) *Handbook of Game Theory with Economic Applications*, vol III. Elsevier Science, New York, pp 2271–2326
94. Palfrey T, Srivastava S (1987) On Bayesian Implementable Allocations. *Rev Econ Stud* 54:193–208
95. Palfrey T, Srivastava S (1989) Implementation with Incomplete Information in Exchange Economies. *Econometrica* 57:115–134
96. Palfrey T, Srivastava S (1991) Nash Implementation using Undominated Strategies. *Econometrica* 59:479–501
97. Peleg B (1996) Double implementation of the Lindahl equilibrium by a continuous mechanism. *Rev Econ Des* 2(1):311–324
98. Pérez-Castrillo D, Wettstein D (2002) Choosing Wisely: A Multi-Bidding Approach. *Am Econ Rev* 92:1577–1587
99. Postlewaite A (1979) Manipulation via endowments. *Rev Econ Stud* 46:255–262
100. Postlewaite A, Schmeidler D (1986) Implementation in Differential Information Economies. *J Econ Theory* 39:14–33
101. Prize Committee of the Royal Swedish Academy of Sciences (2007) *Mechanism Design Theory*. Royal Swedish Academy of Sciences, Stockholm, October 15
102. Reiter S (1977) Information and Performance in the (New) Welfare Economics. *Am Econ Rev* 67:226–234
103. Repullo R (1986) On the Revelation Principle under Complete and Incomplete Information. In: Binmore K, Dasgupta P (eds) *Economics Organizations as Games*. Basil Blackwell, Oxford
104. Repullo R (1987) A simple proof of Maskin theorem on Nash implementation. *Soc Choice Welf* 4:39–41
105. Roth AE (2008) What have we learned from market design? Hahn Lecture, *Econ J* 118:285–310
106. Rothchild M, Stiglitz J (1976) Equilibrium in Competitive Insurance Markets: An essay on the Economics of Imperfect Information. *Q J Econ* 90:629–650
107. Rubinstein A, Wolinsky A (1992) Renegotiation-proof implementation and time preferences. *Am Econ Rev* 82:600–614
108. Saari D (1987) The Source of Some Paradoxes from Social Choice and Probability. *J Econ Theory* 41:1–22
109. Saijo T (1988) Strategy space reduction in Maskin's theorem. *Econometrica* 56:693–700
110. Saijo T (1991) Incentive Compatibility and Individual Rationality in Public Good Economies. *J Econ Theory* 55:103–112
111. Saijo T, Tatamitani TY, Yamato T (1996) Toward Natural Implementation. *Int Econ Rev* 37(4):949–980
112. Saijo T, Sjöström T, Yamato T (2007) Secure Implementation. *Theor Econ* 2:203–229
113. Samuelson PA (1954) The Pure Theory of Public Expenditure. *Rev Econ Stat* 36:387–9
114. Sandholm W (2007) Pigouvian Pricing and Stochastic Evolutionary Implementation. *J Econ Theor* 132:367–382

115. Satterthwaite MA (1975) Strategy-Proofness and Arrows conditions: Existence and Correspondence Theorems for Voting procedures and Social Choice Functions. *J Econ Theory* 10:187–217
116. Schmeidler D (1980) Walrasian Analysis via Strategic Outcome Functions. *Econometrica* 48:1585–1593
117. Sen A (2001) Another direct proof of the Gibbard–Satterthwaite Theorem. *Econ Lett* 70:381–385
118. Serizawa S (2002) Inefficiency of strategy-proof rules for pure exchange economies. *J Econ Theory* 106:219–241
119. Serizawa S, Weymark J (2003) Efficient Strategy-Proof Exchange and Minimum Consumption Guarantees. *J Econ Theory* 109:246–263
120. Serrano R, Vohra R (1997) Non Cooperative Implementation of the Core. *Soc Choice Welf* 14:513–525
121. Serrano R, Vohra R (2001) Some Limitation of Bayesian Virtual Implementation. *Econometrica* 69:785–792
122. Serrano R (2004) The Theory of Implementation of Social Choice Rules. *SIAM Rev* 46:377–414
123. Sertel M, Samver R (1999) Equilibrium outcomes of Lindahl-endowment pretension games. *Eur J Politi Econ* 15:149–162
124. Shin S, Suh S-C (1997) Double Implementation by a Simple Game Form in the Commons Problem. *J Econ Theory* 77:205–213
125. Sjöström T (1993) Implementation in Perfect Equilibria. *Soc Choice Welf* 10:97–106
126. Sjöström T (1994) Implementation in Undominated Nash Equilibrium without Integer Games. *Games Econ Behav* 6:502–511
127. Sjöström T (1996) Implementation by Demand Mechanisms. *Econ Des* 1:343–354
128. Spence M (1973) Job Market Signalling. *Q J Econ* 87:355–374
129. Taylor FM (1929) The Guidance of Production in a Socialist State. *Am Econ Rev* 19:1–8
130. Thomson W (1985) Manipulation and Implementation in Economics. Unpublished manuscript. Rochester, New York
131. Thomson W (1987) The vulnerability to manipulative behavior of economic mechanisms designed to select equitable and efficient outcomes. In: Groves T, Radner R, Reiter S (eds) *Information, Incentives and Economic Mechanisms*. University of Minnesota Press, Minnesota, pp 375–396
132. Thomson W (1996) Concepts of Implementation. *Jpn Econ Rev* 47:133–143
133. Thomson W (2005) Divide and Permute. *Games Econ Behav* 52:186–200
134. Tian G, Li Q (1995) On Nash-Implementation in the Presence of Withholding. *Games Econ Behav* 9:222–233
135. Vickrey W (1961) Counterspeculation, Auctions and Competitive Sealed Tenders. *J Finance* 16:8–37
136. Walker M (1981) A Simple Incentive Compatible Scheme for Attaining Lindahl Allocations. *Econometrica* 49:65–73
137. Williams S (1986) Realization of Nash implementation: Two Aspects of Mechanism Design. *Econometrica* 54:139–151
138. Williams S (1999) A Characterization of Efficient, Bayesian Incentive Compatible Mechanisms. *Econ Theory* 14:155–180
139. Yamato T (1993) Double Implementation in Nash and Undominated Nash Equilibria. *J Econ Theory* 59(2):311–323
140. Yamato T (1994) Equivalence of Nash Implementability and Robust Implementability with Incomplete Information. *Soc Choice Welf* 11:289–303

Infinite Dimensional Controllability

OLIVIER GLASS

Laboratoire Jacques-Louis Lions,
Université Pierre et Marie Curie, Paris, France

Article Outline

Glossary
 Definition of the Subject
 Introduction
 First Definitions and Examples
 Linear Systems
 Nonlinear Systems
 Some Other Problems
 Future Directions
 Bibliography

Glossary

Infinite dimensional control system A infinite dimensional control system is a dynamical system whose state lies in an infinite dimensional vector space—typically a Partial Differential Equation (PDE)—and depending on some parameter to be chosen, called the control.

Exact controllability The exact controllability property is the possibility to steer the state of the system from any initial data to any target by choosing the control as a function of time in an appropriate way.

Approximate controllability The approximate controllability property is the possibility to steer the state of the system from any initial data to a state arbitrarily close to a target by choosing a suitable control.

Controllability to trajectories The controllability to trajectories is the possibility to make the state of the system join some prescribed trajectory by choosing a suitable control.

Definition of the Subject

Controllability is a mathematical problem, which consists of determining the targets to which one can drive the state of some dynamical system, by means of a control parameter present in the equation. Many physical systems such as quantum systems, fluid mechanical systems, wave propagation, diffusion phenomena, etc., are represented by an infinite number of degrees of freedom, and their evolution follows some partial differential equation (PDE). Finding active controls in order to properly influence the dynamics

of these systems generate highly involved problems. The control theory for PDEs, and among this theory, controllability problems, is a mathematical description of such situations. Any dynamical system represented by a PDE, and on which an external influence can be described, can be the object of a study from this point of view.

Introduction

The problem of controllability is a mathematical description of the general following situation. We are given an evolution system (typically a physical one), on which we can exert a certain influence. Is it possible to use this influence to make the system reach a certain state? More precisely, the system takes generally the following form:

$$\dot{y} = F(t, y, u), \quad (1)$$

where y is a description of the state of the system, \dot{y} denotes its derivative with respect to the time t , and u is the control, that is, a parameter which can be chosen in a suitable range. The standard problem of controllability is the following. Given a time $T > 0$, an initial state y_0 and a target y_1 , is it possible to find a control function u (depending on the time), such that the solution of the system, starting from y_0 and provided with this function u reaches the state y_1 at time T ?

If the state of the system can be described by a finite number of degrees of freedom (typically, if it belongs to an Euclidean space or to a manifold), we call the problem finite dimensional. The present article deals with the case where y belongs to an infinite-dimensional space, typically a Banach space or a Hilbert space. Hence the systems described here have an infinite number of degrees of freedom. The potential range of the applications of the theory is extremely wide: the models from fluid dynamics (see for instance [52]) to quantum systems (see [6,54]), networks of structures (see [23,44]), wave propagation, etc., are countless.

In the infinite dimensional setting, the Eq. (1) is typically a partial differential equation, where F acts as a differential operator on the function y , and the influence of u can take multiple different forms: typically, u can be an additional (force) term in the right-hand side of the equation, localized in a part of the domain; it can also appear in the boundary conditions; but other situations can clearly be envisaged (we will describe some of them).

Of course the possibilities to introduce a control problem for partial differential equations are virtually infinite. The number of results since the beginning of the theory in the 1960s has been constantly growing and has reached

huge dimensions: it is, in our opinion, hopeless to give a fair general view of all the activity in the theory. This paper will only try to present some basic techniques connected to the problem of infinite-dimensional controllability.

As in the finite dimensional setting, one can distinguish between the linear systems, where the partial differential equation under view is linear (as well as the action of the control), and the nonlinear ones.

Structure of the Paper

In Sect. “First Definitions and Examples”, we will introduce the problems that are under view and give some examples. The main parts of this paper are Sects. “Linear Systems” and “Nonlinear Systems”, where we consider linear and nonlinear systems, respectively.

First Definitions and Examples

General Framework

We define an infinite-dimensional control system as the following data:

1. an evolution system (typically a PDE)

$$\dot{y} = F(t, y, u),$$

2. the unknown y is the state of the system, which is a function depending on time: $t \in [0, T] \mapsto y(t) \in \mathcal{Y}$, where the set \mathcal{Y} is a functional space (for instance a Banach or a Hilbert space), or a part of a functional space,
3. a parameter u called the control, which is a time-dependent function $t \in [0, T] \mapsto u(t) \in \mathcal{U}$, where the set \mathcal{U} of admissible controls is again some part of a functional space.

As a general rule, one expects that for any initial data $y|_{t=0}$ and any appropriate control function u there exists a unique solution of the system (at least locally in time). In some particular cases, one can find problems “of controllability” type for stationary problems (such as elliptic equations), see for instance [56].

Examples

Let us give two classical examples of the situation. These are two types of acting control frequently considered in the literature: in one case, the control acts as a localized source term in the equation, while in the second one, the control acts on a part of the boundary conditions. The ex-

amples below concern the wave and the heat equations with Dirichlet boundary conditions: these classical equations are reversible and irreversible, respectively, which, as we will see, is of high importance when considering controllability problems.

Example 1 Distributed control for the wave/heat equation with Dirichlet boundary conditions. We consider:

- Ω a regular domain in \mathbb{R}^n , which is in general required to be bounded,
- ω a nonempty open subdomain in Ω ,
- the wave/heat equation is posed in $[0, T] \times \Omega$, with a localized source term in ω :

wave equation

$$\begin{cases} \square v := \partial_{tt}^2 v - \Delta v = \mathbf{1}_\omega u, \\ v|_{\partial\Omega} = 0, \end{cases}$$

heat equation

$$\begin{cases} \partial_t v - \Delta v = \mathbf{1}_\omega u, \\ v|_{\partial\Omega} = 0. \end{cases}$$

- In the first case, the state y of the system is given by the couple $(v(t, \cdot), \partial_t v(t, \cdot))$, for instance considered in the space $H_0^1(\Omega) \times L^2(\Omega)$ or in $L^2(\Omega) \times H^{-1}(\Omega)$.
- In the second case, the state y of the system is given by the function $v(t, \cdot)$, for instance in the space $L^2(\Omega)$.
- In both cases, the control is the function u , for instance considered in $L^2([0, T]; L^2(\omega))$.

Example 2 Boundary control for the wave/heat equation with Dirichlet boundary conditions. We consider:

- Ω a regular domain in \mathbb{R}^n , typically a bounded one,
- Σ an open nonempty subset of the boundary $\partial\Omega$,
- the heat/wave equation in $[0, T] \times \Omega$, with nonhomogeneous boundary conditions inside Σ :

wave equation

$$\begin{cases} \square v := \partial_{tt}^2 v - \Delta v = 0, \\ v|_{\partial\Omega} = \mathbf{1}_\Sigma u, \end{cases}$$

heat equation

$$\begin{cases} \partial_t v - \Delta v = 0, \\ v|_{\partial\Omega} = \mathbf{1}_\Sigma u. \end{cases}$$

The states are the same as in the previous example, but here the control u is imposed on a part of the boundary. One can for instance consider the set of controls as $L^2([0, T]; L^2(\Sigma))$ in the first case, as $C_0^\infty((0, T) \times \Sigma)$ in the second case.

Needless to say, one can consider other boundary conditions than Dirichlet's. Let us emphasize that, while these two types of control are very frequent, these are not by far the only ones: for instance, one can consider the following "affine control": the heat equation with a right hand side

$u(t)g(x)$ where the control u depends only on the time, and g is a fixed function:

$$\begin{cases} \partial_t v - \Delta v = u(t)g(x), \\ v|_{\partial\Omega} = 0, \end{cases}$$

see an example of this below. Also, one could for instance consider the "bilinear control," which takes the form:

$$\begin{cases} \partial_t v - \Delta v = g(x)u(t)v, \\ v|_{\partial\Omega} = 0. \end{cases}$$

Main Problems

Now let us give some definitions of the typical controllability problems, associated to a control system.

Definition 1 A control system is said to be exactly controllable in time $T > 0$ if and only if, for all y_0 and y_1 in \mathcal{Y} , there is some control function $u: [0, T] \rightarrow \mathcal{U}$ such that the unique solution of the system

$$\begin{cases} \dot{y} = F(t, y, u) \\ y|_{t=0} = y_0, \end{cases} \quad (2)$$

satisfies

$$y|_{t=T} = y_1.$$

Definition 2 We suppose that the space \mathcal{Y} is endowed with a metric d . The control system is said to be approximately controllable in time $T > 0$ if and only if, for all y_0 and y_1 in \mathcal{Y} , for any $\varepsilon > 0$, there exists a control function $u: [0, T] \rightarrow \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$d(y|_{t=T}, y_1) < \varepsilon.$$

Definition 3 We consider a particular element 0 of \mathcal{Y} . A control system is said to be zero-controllable in time $T > 0$ if and only if, for all y_0 in \mathcal{Y} , there exists a control function $u: [0, T] \rightarrow \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$y|_{t=T} = 0.$$

Definition 4 A control system is said to be controllable to trajectories in time $T > 0$ if and only if, for all y_0 in \mathcal{Y} and any trajectory \bar{y} of the system (typically but not necessarily satisfying (2) with $u = 0$), there exists a control function $u: [0, T] \rightarrow \mathcal{U}$ such that the unique solution of the system (2) satisfies

$$y|_{t=T} = \bar{y}(T).$$

Definition 5 All the above properties are said to be fulfilled locally, if they are proved for y_0 sufficiently close to the target y_1 or to the starting point of the trajectory $\bar{y}(0)$; they are said to be fulfilled globally if the property is established without such limitations.

Remarks

We can already make some remarks concerning the problems that we described above.

1. The different problems of controllability should be distinguished from the problems of optimal control, which give another viewpoint on control theory. In general, problems of optimal control look for a control u minimizing some functional

$$J(u, y(u)),$$

where $y(u)$ is the trajectory associated to the control u .

2. It is important to notice that the above properties of controllability depend in a crucial way on the choice of the functional spaces \mathcal{Y} , \mathcal{U} . The approximate controllability in some space may be the exact controllability in another space. In the same way, we did not specify the regularity in time of the control functions in the above definitions: it should be specified for each problem.
3. A very important fact for controllability problems is that when a problem of controllability has a solution, it is almost never unique. For instance, if a time-invariant system is controllable regardless of the time T , it is clear that one can choose u arbitrarily in some interval $[0, T/2]$, and then choose an appropriate control (for instance driving the system to 0) during the interval $[T/2, T]$. In such a way, one has constructed a new control which fulfills the required task. The number of controls that one can construct in this way is clearly infinite. This is of course already true for finite dimensional systems.
4. Formally, the problem of interior control when the control zone ω is the whole domain Ω is not very difficult, since it suffices to consider the trajectory

$$v(t) := v_0 + \frac{t}{T}(v_1 - v_0),$$

to compute the left-hand side of the equation with this trajectory, and to choose it as the control. However, by doing so, one obtains in general a control with very low regularity. Note also that, on the contrary, as long as the boundary control problem is concerned, the case when Σ is equal to the whole boundary $\partial\Omega$ is not that simple.

5. Let us also point out a “principle” which shows that interior and boundary control problems are not very different. We give the main ideas.

Suppose for instance that controllability holds for any domain and subdomain Ω and ω . When considering the controllability problem on Ω via boundary control localized on Σ , one may introduce an extension $\tilde{\Omega}$ of Ω , obtained by gluing along Σ an “additional” open set Ω_2 , so that

$$\tilde{\Omega} = \Omega \cup \Omega_2, \quad \overline{\Omega} \cap \overline{\Omega_2} \subset \Sigma \quad \text{and} \quad \tilde{\Omega} \text{ is regular.}$$

Consider now $\omega \subset \Omega_2$, and obtain a controllability result on $\tilde{\Omega}$ via interior control located in ω (one has, of course, to extend initial and final states from Ω to $\tilde{\Omega}$). Consider y a solution of this problem, driving the system from y_0 to y_1 , in the case of exact controllability, for instance. Then one gets a solution of the boundary controllability problem on Ω , by taking the restriction of y on Ω , and by fixing the trace of y on Σ as the corresponding control (in the case of Dirichlet boundary conditions), the normal derivative in the case of Neumann boundary conditions, etc.

Conversely, when one has some boundary controllability result, one can obtain an interior control result in the following way. Consider the problem in Ω with interior control distributed in ω . Solve the boundary control problem in $\Omega \setminus \omega$ via boundary controls in $\partial\omega$. Consider y the solution of this problem. Extend properly the solution y to Ω , and as previously, compute the left hand side for the extension, and consider it as a control (it is automatically distributed in ω).

Of course, in both situations, the regularity of the control that we obtain has to be checked, and this might need a further treatment.

6. Let us also remark that for linear systems, there is no difference between controllability to zero and controllability to trajectories. In that case it is indeed equivalent to bring y_0 to $\bar{y}(T)$ or to bring $y_0 - \bar{y}(0)$ to zero. Note that even for linear systems, on the contrary, approximate controllability and exact controllability differ: an affine subspace of an infinite dimensional space can be dense without filling all the space.

Linear Systems

In this section, we will briefly describe the theory of controllability for linear systems. The main tool here is the duality between controllability and observability of the adjoint system, see in particular the works by Lions, Russell, and Dolecki and Russell [24,50,51,63,65]. This duality

is also of primary importance for finite-dimensional systems.

Let us first describe informally the method of duality for partial differential equations in two cases given in the examples above.

Two Examples

The two examples that we wish to discuss are the boundary controllability of the wave equation and the interior controllability of the heat equation. The complete answers to these problems have been given by Bardos, Lebeau and Rauch [9] for the wave equation (see also Burq and Gérard [14] and Burq [13] for another proof and a generalization), and independently by Lebeau and Robbiano [48] and Fursikov and Imanuvilov, see [35] for the heat equation. The complete proofs of these deep results are clearly out of the reach of this short presentation, but one can explain rather easily with these examples how the corresponding controllability problems can be transformed into some observability problems. These observability problems consist of proving a certain inequality. We refer for instance to Lions [51] or Zuazua [76] for a more complete introduction to these problems.

First, we notice an important difference between the two equations, which will clearly have consequences concerning the controllability problems. It is indeed well-known that, while the wave equation is a reversible equation, the heat equation is on the contrary irreversible and has a strong regularizing effect. From the latter property, one sees that it is not possible to expect an exact controllability result for the heat equation: outside the control zone ω , the state $u(T)$ will be smooth, and in particular one cannot attain an arbitrary state. As a consequence, while it is natural to seek the exact controllability for the wave equation, it will be natural to look either for approximate controllability or controllability to zero as long as the heat equation is concerned.

For both systems we will introduce the adjoint system (typically obtained via integration by parts): it is central in the resolution of the control problems of linear equations. In both cases, the adjoint system is written backward in time form. We consider our two examples separately.

Wave Equation We first consider the case of the wave equation with boundary control on Σ and Dirichlet boundary conditions on the rest of the boundary:

$$\begin{cases} \partial_{tt}^2 v - \Delta v = 0, \\ v|_{\partial\Omega} = \mathbf{1}_\Sigma u, \\ (v, v_t)|_{t=0} = (v_0, v'_0). \end{cases}$$

The problem considered is the exact controllability in $L^2(\Omega) \times H^{-1}(\Omega)$ (recall that the state of the system is (v, v_t)), by means of boundary controls in $L^2((0, T) \times \Sigma)$.

For this system, the adjoint system reads:

$$\begin{cases} -\partial_{tt}^2 \psi - \Delta \psi = 0, \\ \psi|_{\partial\Omega} = 0, \\ (\psi, \psi_t)|_{t=T} = (\psi_T, \psi'_T). \end{cases} \quad (3)$$

Notice that this adjoint equation is well-posed: here it is trivial since the equation is reversible.

The key argument which connects the controllability problem of the equation with the study of the properties of the adjoint system is the following duality formula. It is formally easily obtained by multiplying the equation with the adjoint state and in integrating by parts. One obtains

$$\begin{aligned} \left[\int_{\Omega} \psi(\cdot, x) v_t(\cdot, x) dx - \int_{\Omega} \psi_t(\cdot, x) v(\cdot, x) dx \right]_0^T \\ = - \iint_{(0,T) \times \Sigma} \frac{\partial \psi}{\partial n} \mathbf{1}_\Sigma u dt d\sigma. \end{aligned} \quad (4)$$

In other words, this central formula describes in a simple manner the jump in the evolution of the state of the system in terms of the control, when measured against the dual state.

To make the above computation more rigorous, one can consider for dual state the ψ “classical” solutions in $C^0([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ (these solutions are typically obtained by using a diagonalizing basis for the Dirichlet laplacian, or by using evolution semigroup theory), while the solutions of the direct problem for $(v_0, v_1, u) \in L^2(\Omega) \times H^{-1}(\Omega) \times L^2(\Sigma)$ are defined in $C^0([0, T]; L^2(\Omega)) \cap C^1([0, T]; H^{-1}(\Omega))$ via the transposition method; for more details we refer to the book by Lions [51].

Now, due to the linearity of the system and because we consider the problem of exact controllability (hence things will be different for what concerns the heat equation), it is not difficult to see that it is not restrictive to consider the problem of controllability starting from 0 (that is the problem of reaching any y_1 when starting from $y_0 := (v_0, v'_0) = 0$). Denote indeed $R(T, y_0)$ the affine subspace made of states that can be reached from y_0 at time T for some control. Then calling $y(T)$ the final state of the system for $y|_{t=0}$ and $u = 0$, then one has $R(T, y_0) = y(T) + R(T, 0)$. Hence $R(T, y_0) = \mathcal{Y} \Leftrightarrow R(T, 0) = \mathcal{Y}$.

From (4), we see that reaching (v_T, v'_T) from $(0, 0)$ will be achieved if and only if the relation

$$\begin{aligned} \int_{\Omega} \psi(T, x) v'_T dx - \int_{\Omega} \psi'(T, x) v_T dx \\ = - \iint_{(0, T) \times \Sigma} \frac{\partial \psi}{\partial n} \mathbf{1}_{\Sigma} u dt d\sigma \end{aligned} \quad (5)$$

is satisfied for all choices of (ψ_T, ψ'_T) .

On the left-hand side, we have a linear form on (ψ_T, ψ'_T) in $H_0^1(\Omega) \times L^2(\Omega)$, while on the right hand side, we have a bilinear form on $((\psi_T, \psi'_T), u)$. Suppose that we make the *choice* of looking for a control in the form

$$u = \frac{\partial \bar{\psi}}{\partial n} \mathbf{1}_{\Sigma}, \quad (6)$$

for some $\bar{\psi}$ solution of (3).

Then one sees, using Riesz' theorem, that to solve this problem for $(v_T, v'_T) \in L^2(\Omega) \times H^{-1}(\Omega)$, it is sufficient to prove that the map $(\psi_T, \psi'_T) \mapsto \|\partial \psi / (\partial n) \mathbf{1}_{\Sigma}\|_{L^2((0, T) \times \Sigma)}$ is a norm equivalent to the $H_0^1(\Omega) \times L^2(\Omega)$ one: for some $C > 0$,

$$\|(\psi_T, \psi'_T)\|_{H_0^1(\Omega) \times L^2(\Omega)} \leq C \left\| \frac{\partial \psi}{\partial n} \mathbf{1}_{\Sigma} \right\|_{L^2((0, T) \times \Sigma)}. \quad (7)$$

This is the observability inequality to be proved to prove the controllability of the wave equation. Let us mention that the inequality in the other sense, that is, the fact that the linear map $(\psi_T, \psi'_T) \mapsto \partial \psi / (\partial n) \mathbf{1}_{\Sigma}$ is well-defined and continuous from $H_0^1(\Omega) \times L^2(\Omega)$ to $L^2(\partial \Omega)$ is true but not trivial: it is a "hidden" regularity result, see [66].

When this inequality is proved, a constructive way to select the control is to determine a minimum $(\bar{\psi}_T, \bar{\psi}'_T)$ of the functional

$$\begin{aligned} (\psi_T, \psi'_T) \mapsto J(\psi_T, \psi'_T) := \frac{1}{2} \iint_{\Sigma} \left| \frac{\partial \psi}{\partial n} \right|^2 dt d\sigma \\ + \langle \phi_T, v_1 \rangle_{H_0^1 \times H^{-1}} - \langle \psi'_T, v_0 \rangle_{L^2 \times L^2}, \end{aligned} \quad (8)$$

then to associate to $(\bar{\psi}_T, \bar{\psi}'_T)$ the solution $\bar{\phi}$ of (5), and finally to set u as in (6).

The way described above to determine a particular control—it is clear that not all controls are in the form (6)—is referred as Lions's HUM method (see [51]). This particular control can be proved to be optimal in the L^2 norm, that is, any other control answering to the controllability problem has a larger norm in $L^2((0, T) \times \Sigma)$. As a matter of fact, looking for the L^2 optimal control among those which answer to the problem is a way to justify the choice (6), see [51].

Heat Equation Now let us consider the heat equation with Dirichlet boundary conditions and localized distributed control:

$$\begin{cases} \partial_t v - \Delta v = \mathbf{1}_{\omega} u, \\ v|_{\partial \Omega} = 0, \\ v|_{t=0} = v_0. \end{cases}$$

In this case, we consider in the same way the dual problem:

$$\begin{cases} -\partial_t \phi - \Delta \phi = 0, \\ \phi|_{\partial \Omega} = 0, \\ \phi(T) = \phi_T. \end{cases} \quad (9)$$

Notice that this adjoint equation is well-posed. Here it is very important that the problem is formulated in a backward way: the backward in time setting compensates the opposite sign before the time derivative. It is clear that the above equation is ill-posed when considering initial data at $t = 0$.

In the same way as for the wave equation, multiplying the equation by the adjoint state and integrating by parts yields, at least formally:

$$\int_{\Omega} \phi_T v|_{t=T} dx - \int_{\Omega} \phi(0) v_0 dx = \iint_{(0, T) \times \omega} \phi u dt dx. \quad (10)$$

Note that standard methods yield regular solutions for both direct and adjoint equations when v_0 and ϕ_T belong to $L^2(\Omega)$, and u belongs to $L^2((0, T) \times \omega)$.

Now let us discuss the approximate controllability and the controllability to zero problems separately.

Approximate controllability. Because of linearity, the approximate controllability is equivalent to the approximate controllability starting from 0. Now the set $R(0, T)$ of all final states $v(T)$ which can be reached from 0, via controls $u \in L^2((0, T) \times \omega)$, is a vector subspace of $L^2(\Omega)$. The density of $R(0, T)$ in $L^2(\Omega)$ amounts to the existence of a nontrivial element in $(R(0, T))^{\perp}$. Considering ϕ_T such an element, and introducing it in (10) together with $v_0 = 0$, we see that this involves the existence of a nontrivial solution of (9), satisfying

$$\phi|_{(0, T) \times \omega} = 0.$$

Hence to prove the approximate controllability, we have to prove that there is no such nontrivial solution, that is, we have to establish a unique continuation result. In this case, this can be proved by using Holmgren's unique continuation principle (see for instance [40]), which establishes the result. Note in passing that Holmgren's theorem is a very general and important tool to prove unique continuation

results; however it requires the analyticity of the coefficients of the operator, and in many situations one cannot use it directly.

Let us also mention that as for the exact controllability of the wave equation above, and the zero-controllability for the heat equation below, one can single out a control for the approximate controllability by using a variational approach consisting of minimizing some functional as in (8): see [26,53].

Controllability to zero. Now considering the problem of controllability to zero, we see that, in order that the control u brings the system to 0, it is necessary and sufficient that for all choice of $\phi_T \in L^2(\Omega)$, we have

$$-\int_{\Omega} \phi(0)v_0 dx = \iint_{(0,T) \times \omega} \phi u dt dx .$$

Here one would like to reason as for the wave equation, that is, make the *choice* to look for u in the form

$$u := \phi \mathbf{1}_{\omega} ,$$

for some solution ϕ of the adjoint system. But here the application

$$N: \phi_T \in L^2(\Omega) \mapsto \left(\iint_{(0,T) \times \omega} \phi^2 dt dx \right)^{\frac{1}{2}} .$$

determines a norm (as seen from the above unique continuation result), but this norm is no longer equivalent to the usual $L^2(\Omega)$ norm (if it was, one could establish an exact controllability result!). A way to overcome the problem is to introduce the Hilbert space X obtained by completing $L^2(\Omega)$ for the norm N . In this way, we see that to solve the zero-controllability problem, it is sufficient to prove that the linear mapping $\phi_T \mapsto \phi(0)$ is continuous with respect to the norm N : for some $C > 0$,

$$\|\phi(0)\|_{L^2(\Omega)} \leq C \|\phi \mathbf{1}_{\omega}\|_{L^2((0,T) \times \omega)} . \quad (11)$$

This is precisely the observability inequality which one has to prove to establish the zero controllability of the heat equation. It is weaker than the observability inequality when the left-hand side is $\|\phi(T)\|_{L^2(\Omega)}$ (which as we noticed is false).

When this inequality is proven, a constructive way to determine a suitable control is to determine a minimum $\bar{\phi}_T$ in X of the functional

$$\phi_T \mapsto \frac{1}{2} \iint_{(0,T) \times \omega} |\phi|^2 dt dx + \int_{\Omega} \phi(0)v_0 dx ,$$

then to associate to $\bar{\phi}_T$ the solution $\bar{\phi}$ of (3), and finally to set

$$u := \bar{\phi} \mathbf{1}_{\omega} .$$

(That the mapping $\bar{\phi}_T \mapsto \bar{\phi} \mathbf{1}_{\omega}$ can be extended to a mapping from X to $L^2(\omega)$ comes from the definition of X .)

So in both situations of exact controllability and controllability to zero, one has to establish a certain inequality in order to get the result; and to prove approximate controllability, one has to establish a unique continuation result. This turns out to be very general, as described in Paragraph Subject. “[Abstract Approach](#)”.

Remarks Let us mention that concerning the heat equation, the zero-controllability holds for any time $T > 0$ and for any nontrivial control zone ω , as shown by means of a spectral method—see Lebeau and Robbiano [48], or by using a global Carleman estimate in order to prove the observability property—see Fursikov and Imanuvilov [35]. That the zero-controllability property does not require the time T or the control zone ω to be large enough is natural since parabolic equations have an infinite speed of propagation; hence one should not require much time for the information to propagate from the control zone to the whole domain. The zero controllability of the heat equation can be extended to a very wide class of parabolic equations; for such results global Carleman estimates play a central role, see for instance the reference book by Fursikov and Imanuvilov [35] and the review article by Fernández-Cara and Guerrero [28]. Note also that Carleman estimates are not used only in the context of parabolic equations: see for instance [35] and Zhang [70]. Let us also mention two other approaches for the controllability of the one-dimensional heat equation: the method of moments of Fattorini and Russell (see [27] and a brief description below), and the method of Laroche, Martin and Rouchon [45] to get an explicit approximate control, based on the idea of “flatness” (as introduced by Fliess, Lévine, Rouchon and Martin [33]).

On the other hand, the controllability for the wave equation does not hold for any T or any Σ . Roughly speaking, the result of Bardos, Lebeau and Rauch [9] states that the controllability property holds if and only if every ray of the geometric optics in the domain (reflecting on the boundary) meets the control zone during the time interval $[0, T]$. In this case, it is also natural that the time should be large enough, because of the finite speed of propagation of the equation: one has to wait for the information coming from the control zone to influence the whole domain. Let us emphasize that in some cases, the geometry

of the domain and the control zone is such that the controllability property does not hold, no matter how long the time T is. An example of this is for instance a circle in which some antipodal regions both belong to the uncontrolled part of the boundary. This is due to the existence of Gaussian beams, that is, solutions that are concentrated along some ray of the geometrical optics (and decay exponentially away from it); when this ray does not meet the control zone, this contradicts in particular (7). The result of [9] relies on microlocal analysis. Note that another important tool used for proving observability inequalities is a multiplier method (see in particular [42,51,60]); however in the case of the wave equation, this can be done only in particular geometric situations.

Abstract Approach

The duality between the controllability of a system and the observability of its adjoint system turns out to be very general, and can be described in an abstract form due to Dolecki and Russell [24]. For more recent references see [46,69]. Here we will only describe a particular simplified form of the general setting of [24]. Consider the following system

$$\dot{y} = Ay + Bu, \quad (12)$$

where the state y belongs to a certain Hilbert space H , on which A , which is densely defined and closed, generates a strongly continuous semi-group of bounded operators. The operator B belongs to $\mathcal{L}(U; H)$, where U is also a Hilbert space. The solutions of (12) are given by the method of variation of constants so that

$$y(T) = e^{tA}y(0) + \int_0^T e^{(T-\tau)A}Bu(\tau)d\tau.$$

We naturally associate with the control equation (12), the following observation system:

$$\begin{cases} \dot{z} = -A^*z, \\ z(T) = z_T, \\ c := B^*z. \end{cases} \quad (13)$$

The dual operator $-A^*$ generates a strongly continuous semi-group for negative times, and c in $L^2(0, T; U)$ is obtained by

$$c(t) = B^*e^{(T-t)A^*}z_T.$$

In the above system, the dynamics of the (adjoint) state z is free, and c is called the observed quantity of the system. The core of the method is to connect the controllability properties of (12) with observability properties of (13) such as described below.

Definition 6 The system (13) is said to satisfy the unique continuation property if and only if the following implication holds true:

$$c = 0 \text{ in } [0, T] \implies z = 0 \text{ in } [0, T].$$

Definition 7 The system (13) is said to be observable if and only if there exists $C > 0$ such that the following inequality is valid for all solutions z of (13):

$$\|z(T)\|_H \leq C\|c\|_{L^2(0,T;U)}. \quad (14)$$

Definition 8 The system (13) is said to be observable at time 0 if and only if there exists $C > 0$ such that the following inequality is valid for all solutions z of (13):

$$\|z(0)\|_H \leq C\|c\|_{L^2(0,T;U)}. \quad (15)$$

The main property is the following.

Duality property

1. The exact controllability of (12) is equivalent to the observability of (13).
2. The zero controllability of (12) is equivalent to the observability at time 0 of (13).
3. The approximate controllability of (12) in H is equivalent to the unique continuation property for (13).

Brief explanation of the duality property. The main fact is the following duality formula

$$\begin{aligned} \langle y(T), z(T) \rangle_H - \langle y(0), z(0) \rangle_H \\ = \int_0^T \left\langle u(\tau), B^*e^{(T-\tau)A^*}z(T) \right\rangle_U d\tau, \end{aligned} \quad (16)$$

which in fact can be used to define the solutions of (12) by transposition.

1 and 3. Now it is rather clear that by linearity, we can reduce the problems of exact and approximate controllability to the ones when $y(0) = 0$. Now the property of exact controllability for system (12) is equivalent to the surjectivity of the operator

$$S: u \in L^2(0, T; U) \mapsto \int_0^T e^{(T-\tau)A}Bu(\tau)d\tau \in H,$$

and the approximate controllability is equivalent to the density of its range. By (16) its adjoint operator is

$$S^*: z_T \in H \mapsto \int_0^T B^*e^{(T-\tau)A^*}z_T d\tau \in L^2(0, T; U).$$

Hence the equivalences 1 and 3 are derived from a classical result from functional analysis (see for instance the book of Brezis [12]): the range of S is dense if and only if S^* is one-to-one; S is surjective if and only if S^* satisfies for some $C > 0$:

$$\|z_T\| \leq C\|S^* z_T\| ,$$

that is, when (14) is valid.

2. In this case, still due to linearity, the zero-controllability for system (12) is equivalent to the following inclusion:

$$\text{Range}(e^{TA}) \subset \text{Range}(S) .$$

In this case there is also a functional analysis result (generalizing the one cited above) which asserts the equivalence of this property with the existence of $C > 0$ such that

$$\|(e^{TA})^* h\| \leq C\|S^* h\| ,$$

see Dolecki and Russell [24] and Douglas [25] for more general results.

It follows that in many situations, the exact controllability of a system is proved by establishing an observability inequality on the adjoint system. But this general method is not the final point of the theory: not only is this only valid for linear systems, but importantly as well, it turns out that these types of inequalities are in general very difficult to establish. In general when a result of exact controllability is established by using this duality, the largest part of the proof is devoted to establishing the observability inequality.

Another information given by the observability is an estimate of the size of the control. Indeed, following the lines of the proof of the correspondence between observability and controllability, one can see that one can find a control u which satisfies:

- Case of exact controllability:

$$\|u\|_{L^2(0,T;U)} \leq C_{\text{obs}}\|y_1 - e^{TA}y_0\|_H ,$$

- Case of zero controllability:

$$\|u\|_{L^2(0,T;U)} \leq C_{\text{obs}}\|y_0\|_H ,$$

where in both cases C_{obs} determines a constant for which the corresponding observability inequality is true. (Obviously, not *all* the controls answering to the question satisfy this estimate). This gives an upper bound on the size of a possible control. This can give some precise estimates on the cost of the control in terms of the parameters of the problem, see for instance the works of Fernández-Cara and Zuazua [30] and Miller [58] and Seidman [66] concerning the heat equation.

Some Different Methods

Herein we will discuss two methods that do not rely on controllability/observability duality. This does not pretend to give a general vision of all the techniques that can be used in problems of controllability of linear partial differential equations. We just mention them in order to show that in some situations, duality may not be the only tool available.

Characteristics First, let us mention that in certain situations, one may use a characteristics method (see e.g. [22]). A very simple example is the first-dimensional wave equation,

$$\begin{cases} v_{tt} - v_{xx} = 0 , \\ v_{x|_{x=0}} = 0 , \\ v_{x|_{x=1}} = u(t) , \end{cases}$$

where the control is $u(t) \in L^2(0, T)$. This is taken from Russell [62]. The problem is transformed into

$$\begin{aligned} \frac{\partial w}{\partial t} &= \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial w}{\partial x} \quad \text{with} \\ w &:= \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} := \begin{pmatrix} v_t \\ v_x \end{pmatrix} , \\ w_2(t, 0) &= 0 \quad \text{and} \quad w_2(t, 0) = u(t) . \end{aligned}$$

Of course, this means that $w_1 - w_2$ and $w_1 + w_2$ are constant along characteristics, which are straight lines of slope 1 and -1 , respectively (this is d'Alembert's decomposition).

Now one can deduce an explicit appropriate control for the controllability problem in $[0, 1]$ for $T > 2$, by constructing the solution w of the problem directly (and one takes the values of w_2 at $x = 1$ as the "resulting" control u).

The function w is completely determined from the initial and final values in the domains of determinacy D_1 and D_2 :

$$\begin{aligned} D_1 &:= \{(t, x) \in [0, T] \times [0, 1] / x + t \leq 1\} \quad \text{and} \\ D_2 &:= \{(t, x) \in [0, T] \times [0, 1] / t - x \geq T - 1\} . \end{aligned}$$

That $T > 2$ means that these two domains do not intersect. Now it suffices to complete the solution w in $D_3 := [0, T] \times [0, 1] \setminus (D_1 \cup D_2)$. For that, one chooses $w_1(0, t)$ *arbitrarily* in the part ℓ of the axis $\{0\} \times [0, 1]$ located between D_1 and D_2 , that is for

$$x = 0 \quad \text{and} \quad 1 \leq t \leq T - 1 .$$

Once this choice is made, it is not difficult to solve the Goursat problem consisting of extending w in D_3 : using

the symmetric role of x and t , one considers x as the time. Then the initial condition is prescribed on ℓ , as well as the boundary conditions on the two characteristic lines $x + t = 1$ and $x - t = T - 1$. One can solve elementarily this problem by using the characteristics, and this finishes the construction of w , and hence of u .

Note that as a matter of fact, the observability inequality in this (one-dimensional) case is also elementary to establish, by relying on Fourier series or on d'Alembert's decomposition, see for instance [23].

The method of characteristics described above can be generalized in broader situations, including for instance the problem of boundary controllability of one-dimensional linear hyperbolic systems

$$v_t + A(x)v_x + B(x)v = 0,$$

where A is a real symmetric matrix with eigenvalues bounded away from zero, and A and B are smooth; see for instance Russell [65]. As a matter of fact, in some cases this method can be used to establish the observability inequality from the controllability result (while in most cases the other implication in the equivalence is used).

Of course, the method of characteristics may be very useful for transport equations

$$\frac{\partial f}{\partial t} + v \cdot \nabla f = g \quad \text{or} \quad \frac{\partial f}{\partial t} + \operatorname{div}(vf) = g.$$

An example of this is the controllability of the Vlasov-Poisson equation, see [36]. Let us finally mention that this method is found also to tackle directly several nonlinear problems, as we will see later.

Moments Another method which we would like to briefly discuss is the method of moments (see for instance Avdonin and Ivanov [4] for a general reference, see also Russell [65]), which can appear in many situations; in particular this method was used by Fattorini and Russell to prove the controllability of the heat equation in one space dimension, see [27]. Consider for instance the problem of controllability of the one-dimensional heat equation

$$\begin{cases} v_t - v_{xx} = g(x)u(t), \\ v|_{[0,T] \times \{0,1\}} = 0. \end{cases}$$

Actually in [27], much more general situations are considered: in particular it concerns more general parabolic equations and boundary conditions, and boundary controls can also be included in the discussion.

It is elementary to develop the solution in the $L^2(0, 1)$ orthonormal basis $(\sin(k\pi x))_{k \in \mathbb{N}^*}$. One obtains that the

state zero is reached at time $T > 0$ if and only if

$$\begin{aligned} \sum_{k>0} e^{-k^2 T} v_k \sin(k\pi x) \\ = - \sum_{k>0} \int_0^T e^{-k^2(T-t)} g_k u(t) \sin(k\pi x) dt, \end{aligned}$$

where v_k and g_k are the coordinates of v_0 and g in the basis. Clearly, this means that we have to find u such that for all $k \in \mathbb{N}$,

$$\int_0^T e^{-k^2(T-t)} g_k u(t) dt = -e^{-k^2 T} v_k.$$

The classical Muntz-Szász theorem states that for an increasing family $(\lambda_n)_{n \in \mathbb{N}^*}$ of positive numbers, the family $\{e^{-\lambda_n t}, n \in \mathbb{N}^*\}$ is dense in $L^2(0, T)$ if and only if

$$\sum_{n \geq 0} \frac{1}{\lambda_n} = +\infty,$$

and in the opposite case, the family is independent and spans a proper closed subspace of $L^2(0, T)$. Here the exponential family which we consider is $\lambda_n = n^2$ and we are in the second situation. The same method applies for other problems in which λ_n cannot be completely computed but is a perturbation of n^2 ; this allows us to treat a wider class of problems. Now in this situation (see e.g. [65]), one can construct in $L^2(0, T)$ a biorthogonal family to $\{e^{-\lambda_n t}, n \in \mathbb{N}^*\}$, that is a family $(p_n(t))_{n \in \mathbb{N}^*} \in (L^2(0, T))^{\mathbb{N}}$ satisfying

$$\int_0^T p_n(t) e^{-\lambda_k^2 t} dt = \delta_{kn}.$$

Once such a family is obtained, one has formally the following solution, under the natural assumption that $|g_k| \geq ck^{-\alpha}$:

$$u(t) = \sum_{k \in \mathbb{N}^*} -\frac{e^{-\lambda_k T} v_k}{g_k} p_k(t).$$

To actually get a control in $L^2(0, T)$, one has to estimate $\|p_k\|_{L^2}$, in order that the above sum is well-defined in L^2 . In [27] it is proven that one can construct p_k in such a way that

$$\|p_k\|_{L^2(0,T)} \leq K_0 \exp(K_1 \omega_k), \quad \omega_k := \sqrt{\lambda_k},$$

which allows us to conclude.

Nonlinear Systems

The most frequent method for dealing with control problems of nonlinear systems is the natural one (as for the Cauchy problem): one has to linearize (at least in some sense) the equation, try to prove some controllability result on the linear equation (for instance by using the duality principle), and then try to pass to the nonlinear system via typical methods such as inverse mapping theorem, fixed point theory, iterative schemes...

As for the usual inverse mapping theorem, it is natural to hope for a local result from the controllability of the linear problem. Here we find an important difference between linear and nonlinear systems: while for linear systems, no distinction has to be made between local and global results, concerning the nonlinear systems, the two problems are really of different nature.

One should probably not expect a very general result indicating that the controllability of the linearized equation involves the local controllability of the nonlinear system. The linearization principle is a general approach which one has to adapt to the different situations that one can meet. We give below some examples that present to the reader some existing approaches.

Some Linearization Situations

Let us first discuss some typical situations where the linearized equation has good controllability properties, and one can hope to get a result (in general local) from this information. In some situations where the nonlinearity is not too strong, one can hope to get global results. As previously, the situation where the underlying linear equation is reversible and the situation when it is not have to be distinguished. We briefly describe this in two different examples.

Semilinear Wave Equation Let us discuss first an example with the wave equation. This is borrowed from the work of Zuazua [71,72], where the equation considered is

$$\begin{cases} v_{tt} - v_{xx} + f(v) = u \mathbf{1}_\omega \text{ in } [0, 1], \\ v|_{x=0} = 0, v|_{x=1} = 0, \end{cases}$$

where $\omega := (l_1, l_2)$ is the control zone and the nonlinearity $f \in C^1(\mathbb{R}; \mathbb{R})$ is at most linear at infinity (see however the remark below) in the sense that for some $C > 0$,

$$|f(x)| \leq C(1 + |x|) \text{ on } \mathbb{R}. \quad (17)$$

The global exact controllability in $H_0^1 \times L^2$ (recall that the state of the system is (v, v_t)) by means of a control in

$L^2((0, T) \times \omega)$ is proved by using the following linearization technique. In [72], it is proven that the following linearized equation:

$$\begin{cases} v_{tt} - v_{xx} + a(x)v = u \mathbf{1}_\omega \text{ in } [0, 1], \\ v|_{x=0} = 0, v|_{x=1} = 0. \end{cases}$$

is controllable in $H_0^1(0, 1) \times L^2(0, 1)$ through $u(t) \in L^2((0, T) \times \omega)$, for times $T > 2 \max(l_1, 1 - l_2)$, for any $a \in L^\infty((0, T) \times (0, 1))$. The corresponding observability inequality is

$$\|(\psi_T, \psi'_T)\|_{L^2(\Omega) \times H^{-1}(\Omega)} \leq C_{\text{obs}} \|\psi \mathbf{1}_\omega\|_{L^2((0, T) \times \omega)}.$$

Moreover, the observability constant that one can obtain can be bounded in the following way:

$$C_{\text{obs}} \leq \alpha(T, \|a\|_\infty), \quad (18)$$

where α is nondecreasing in the second variable. One would like to describe a fixed-point scheme as follows. Given v , one considers the linearized problem around v , solves this problem and deduces a solution \hat{v} of the problem of controllability from (v_0, v'_0) to (v_1, v'_1) in time T . More precisely, the scheme is constructed in the following way.

We write

$$f(x) = f(0) + xg(x),$$

where g is continuous and bounded. The idea is to associate to any $v \in L^\infty((0, T) \times (0, 1))$ a solution of the linear control problem

$$\begin{cases} \hat{v}_{tt} - \hat{v}_{xx} + \hat{v}g(v) = -f(0) + u \mathbf{1}_\omega \text{ in } [0, 1], \\ \hat{v}|_{x=0} = 0, \hat{v}|_{x=1} = 0. \end{cases}$$

(Note that the “drift” term $f(0)$ on the right hand side can be integrated in the final state—just consider the solution of the above system with $u = 0$ and $\hat{v}|_{t=0} = 0$ and withdraw it—or integrated in the formulation (5)).

Here an issue is that, as we recalled earlier, a controllability problem has almost never a unique solution. The method in [72] consists of selecting a particular control, which is the one of smallest L^2 -norm. Taking the fact that g is bounded and (18) into account, this particular control satisfies

$$\begin{aligned} \|u\|_{L^2(0, T)} &\leq C(\|v_0\|_{H_0^1} + \|v'_0\|_{L^2} + \|v_1\|_{H_0^1} + \|v'_1\|_{L^2} + |f(0)|), \end{aligned}$$

for some $C > 0$ independent of v .

Using the above information, one can deduce estimates for \hat{v} in $C^0([0, T]; L^2(0, 1)) \cap L^2(0, T; H_0^1)$ independently of v , and then show by Schauder's fixed point theorem that the above process has a fixed point, which shows a global controllability result for this semilinear wave equation.

Remark 1 As a matter of fact, [72] proves the global exact controllability for f satisfying the weaker assumption

$$\lim_{|x| \rightarrow +\infty} \frac{|f(x)|}{(1 + |x|) \log^2(|x|)} = 0.$$

This is optimal since it is also proven in [72] that if

$$\liminf_{|x| \rightarrow +\infty} \frac{|f(x)|}{(1 + |x|) \log^p(|x|)} > 0$$

for some $p > 2$ (and $\omega \neq (0, 1)$) then the system is not globally controllable due to blow-up phenomena. To obtain this result one has to use Leray–Schauder's degree theory instead of Schauder's fixed point theorem. For analogous conditions on the semilinear heat equation, see Fernández-Cara and Zuazua [31].

Burgers Equation Now let us discuss a parabolic example, namely the local controllability to trajectories of the viscous Burgers equation:

$$\begin{cases} v_t + (v^2)_x - v_{xx} = 0 & \text{in } (0, T) \times (0, 1), \\ v|_{x=0} = u_0(t) \text{ and } v|_{x=1} = u_1(t), \end{cases} \quad (19)$$

controlled on the boundary via u_0 and u_1 . This is taken from Fursikov and Imanuvilov [34]. Now the linear result concerns the zero-controllability via boundary controls of the system

$$\begin{cases} v_t + (zv)_x - v_{xx} = 0 & \text{in } (0, T) \times (0, 1), \\ v|_{x=0} = u_0(t) \text{ and } v|_{x=1} = u_1(t). \end{cases} \quad (20)$$

Consider X the Hilbert space composed of functions z in $L^2(0, T; H^2(0, 1))$ such that $z_t \in L^2(0, 1; L^2(0, 1))$. In [34] it is proved that given $v_0 \in H^1(0, 1)$ and $T > 0$, one can construct a map which to any z in X , associates $v \in X$ such that v is a solution of (20) which drives v_0 to 0 during time interval $[0, T]$, and moreover this map is compact from X to X . As in the previous situation, the particular control (u_0, u_1) has to be singled out in order for the above mapping to be single-valued (and compact). But here, the criterion is not quite the optimality in L^2 -norm of the control. The idea is the following: first, one transforms the controllability problem for (20) from v_0 to 0 into

a problem of “driving” 0 to 0 for a problem with right-hand side:

$$\begin{cases} w_t + (zw)_x - w_{xx} = f_0 & \text{in } (0, T) \times (0, 1), \\ w|_{x=0} = u_0(t) \text{ and } w|_{x=1} = u_1(t), \end{cases} \quad (21)$$

for some f_0 supported in $(T/3, 2T/3) \times (0, 1)$. For this, one introduces $\chi \in C^\infty([0, T]; \mathbb{R})$ such that $\chi = 1$ during $[0, T/3]$ and $\chi = 0$ during $[2T/3, T]$, and \hat{v} the solution of (20) starting from v_0 with $u_0 = u_1 = 0$, and considers $w := v - \chi \hat{v}$.

Now the operator mentioned above is the one which to z associates the solution of the controllability problem which minimizes the L^2 -norm of w among all the solutions of this controllability problem. The optimality criterion yields a certain form for the solution. That the corresponding control exists (and is unique) relies on a Carleman estimate, see [34] (moreover this allows estimates on the size of w). As a matter of fact, to obtain the compactness of the operator, one extends the domain, solves the above problem in this extended domain, and then uses an interior parabolic regularity result to have bounds in smaller spaces, we refer to [34] for more details.

Once the operator is obtained, the local controllability to trajectories is obtained as follows. One considers \bar{v} a trajectory of the system (19), belonging to X . Withdrawing (19) for \bar{v} to (19) for the unknown, we see that the problem to solve through boundary controls is

$$\begin{cases} y_t - y_{xx} + [(2\bar{v} + y)y]_x = f_0 & \text{in } (0, T) \times (0, 1), \\ y|_{x=0} = v_0 - \bar{v}(0) \text{ and } y|_{t=T} = 0. \end{cases}$$

Now consider $v_0 \in H^1(0, 1)$ such that

$$\|v_0 - \bar{v}(0)\|_{H^1(0,1)} < r,$$

for $r > 0$ to be chosen. To any $y \in X$, one associates the solution of the controllability problem (21) constructed above, driving $v_0 - \hat{v}(0)$ to 0, for $z := (2\hat{v} + y)$. The estimates on the solution of the control problem allow one to establish that the unit ball of X is stable by this process provided r is small enough. The compactness of the process is already proved, so Schauder's theorem allows one to conclude.

Note that it is also proved in [34] that the global approximate controllability does not hold.

Some Other Examples Let us also mention two other approaches that may be useful in this type of situation.

The first is the use of Kakutani–Tikhonov fixed-point theorem for multivalued maps (cf. for instance [68]), see in particular [38, 39] and [26]. Such a technique is particularly

useful, because it avoids the selection process of a particular control. One associates to v the set $T(v)$ of all \hat{v} solving the controllability problem for the equation linearized around v , with all possible controls (in a suitable class). Then under appropriate conditions, one can find a fixed point in the sense that $v \in T(v)$.

Another approach that is very promising is the use of a Nash–Moser process, see in particular the work [10] by Beauchard. In this paper the controllability of a Schrödinger equation via a bilinear control is considered. In that case, one can solve some particular linearized equation (as a matter of fact, the return method described in the next paragraph is used), but with a loss of derivative; as a consequence the approaches described above fail, but the use of Nash–Moser’s theorem allows one to get a result. Note finally that in certain other functional settings, the controllability of this system fails, as shown by using a general result on bilinear control by Ball, Marsden and Slemrod [5].

The Return Method

It occurs that in some situations, the linearized equation is not systematically controllable, and one cannot hope by applying directly the above process to get even local exact controllability. The return method was introduced by Coron to deal with such situations (see in particular [18]). As a matter of fact, this method can be useful even when the linearized equation is controllable. The principle of the method is the following: find a particular trajectory \bar{y} of the nonlinear system, starting at some base point (typically 0) and returning to it, such that the linearized equation around this is controllable. In that case, one can hope to find a solution of the nonlinear local controllability problem close to \bar{y} .

A typical situation of this is the two-dimensional Euler equation for incompressible inviscid fluids (see Coron [16]), which reads

$$\begin{cases} \partial_t y + (y \cdot \nabla) y = -\nabla p & \text{in } \Omega, \\ \operatorname{div} y = 0 & \text{in } \Omega, \\ y \cdot n = 0 & \text{on } \partial\Omega \setminus \Sigma, \end{cases} \quad (22)$$

where the unknown is the velocity field $y: \Omega \rightarrow \mathbb{R}^2$ (the pressure p can be eliminated from the equation), Ω is a regular bounded domain (simply connected to simplify), n is the unit outward normal on the boundary, and $\Sigma \subset \partial\Omega$ is the control zone. On Σ , the natural control which can be assigned is the normal velocity $y \cdot n$ and the vorticity $\omega := \operatorname{curl} y := \partial_1 y^2 - \partial_2 y^1$ at “entering” points, that is points where $y \cdot n < 0$.

Let us discuss this example in an informal way. As noticed by J.-L. Lions, the linearized equation around the null

state

$$\begin{cases} \partial_t y = -\nabla p & \text{in } \Omega, \\ \operatorname{div} y = 0 & \text{in } \Omega, \\ y \cdot n = 0 & \text{on } \partial\Omega \setminus \Sigma, \end{cases}$$

is trivially not controllable (even approximately). Now the main goal is to find the trajectory \bar{y} such that the linearized equation near \bar{y} is controllable. It will be easier to work with the vorticity formulation

$$\begin{cases} \partial_t \omega + (y \cdot \nabla) \omega = 0 & \text{in } \Omega, \\ \operatorname{curl} y = \omega, \operatorname{div} y = 0. \end{cases} \quad (23)$$

In fact, one can show that assigning $y(T)$ is equivalent to assigning both $\omega(T)$ in Ω and $y(T) \cdot n$ on Σ , and since this latter is a part of the control, it is sufficient to know how to assign the vorticity of the final state.

We can linearize (23) in the following way: to y one associates \hat{y} through

$$\begin{cases} \partial_t \omega + (y \cdot \nabla) \omega = 0 & \text{in } \Omega, \\ \operatorname{curl} \hat{y} = \omega, \operatorname{div} \hat{y} = 0. \end{cases} \quad (24)$$

Considering Eq. (24), we see that if the flow of y is such that any point in $\bar{\Omega}$ at time T “comes from” the control zone Σ , then one can assign easily ω through a method of characteristics. Hence one has to find a solution \bar{y} of the system, starting and ending at 0, and such that in its flow, all points in $\bar{\Omega}$ at time T , come from Σ at some stage between times 0 and T . Then a simple Gronwall-type argument shows that this property holds for y in a neighborhood of \bar{y} , hence Eq. (24) is controllable in a neighborhood of \bar{y} . Then a fixed-point scheme allows one to prove a controllability result locally around 0.

But the Euler equation has some time-scale invariance:

$$\begin{aligned} y(t, x) \text{ is a solution on } [0, T] &\Rightarrow y^\lambda(t, x) \\ &:= \lambda^{-1} y(\lambda^{-1} t, x) \text{ is a solution on } [0, \lambda T]. \end{aligned}$$

Hence given y_0 and y_1 , one can solve the problem of driving λy_0 to λy_1 for λ small enough. Changing the variables, one sees that it is possible to drive y_0 to y_1 in time λT , that is, in *smaller* times. Hence one deduces a global controllability result from the above local result.

As a consequence, the central part of the proof is to find the function \bar{y} . This is done by considering a special type of solution of the Euler equation, namely the potential solution: any $\bar{y} := \nabla \theta(t, x)$ with θ regular satisfying

$$\begin{aligned} \Delta_x \theta(t, x) &= 0 & \text{in } \Omega, \quad \forall t \in [0, T], \\ \partial_n \theta &= 0 & \text{on } \partial\Omega, \quad \forall t \in [0, T], \end{aligned}$$

satisfies (22). In [16] it is proven that there exists some θ satisfying the above equation and whose flow makes all points at time T come from Σ . This concludes the argument.

This method has been used in various situations; see [18] and references therein. Let us underline that this method can be of great interest even in the cases where the linearized equation is actually controllable. An important example obtained by Coron concerns the Navier–Stokes equation and is given in [17] (see also [19]): here the return method is used to prove some global approximate result, while the linearized equation is actually controllable but yields in general a local controllability result (see in particular [29,35,41]).

Some Other Methods

Let us finally briefly mention that linearizing the equation (whether using the standard approach or the return method) is not systematically the only approach to the controllability of a nonlinear system. Sometimes, one can “work at the nonlinear level.” An important example is the control of one-dimensional hyperbolic systems:

$$v_t + A(v)v_x = F(v), \quad v: [0, T] \times [0, 1] \rightarrow \mathbb{R}^n, \quad (25)$$

via the boundary controls. In (25), A satisfies the hyperbolicity property that it possesses at every point n real eigenvalues; these are moreover supposed to be strictly separated from 0.

In the case of regular C^1 solutions, this was approached by a method of characteristics to give general local results, see in particular the works by Cirinà [15] and Li and Rao [49]. Interestingly enough, the linear tool of duality between observability and controllability has some counterpart in this particular nonlinear setting, see Li [55]. In the context of weak entropy solutions, some results for particular systems have been obtained via the ad hoc method of front-tracking, see in particular Ancona, Bressan and Coclite [2], the author [37] and references therein.

Other nonlinear tools can be found in Coron’s book [18]. Let us mention two of them. The first one is power series expansion. It consists of considering, instead of the linearization of the system, the development to higher order of the nonlinearity. In such a way, one can hope to attain the directions which are unreachable for the linearized system. This has for instance applications for the Korteweg–de Vries equation, see Coron and Crépeau [20] and the earlier work by Rosier [61]. The other one is quasi-static deformations. The general idea is to find an (explicit) “almost trajectory” $(\bar{y}(\varepsilon t), \bar{u}(\varepsilon t))$ during $[0, T/\varepsilon]$ of

the control system $\dot{y} = f(y, u)$, in the sense that

$$\frac{d}{dt}[\bar{y}(\varepsilon t)] - f(\bar{y}(\varepsilon t), \bar{u}(\varepsilon t))$$

is of order ε (due to the “slow” motion). Typically, the trajectory (\bar{y}, \bar{u}) is composed of equilibrium states (that is $f(\bar{y}(\cdot), \bar{u}(\cdot)) = 0$). In such a way, one can hope to connect $\bar{y}(0)$ to a state close to $\bar{y}(T)$, and then to exactly $\bar{y}(T)$ via a local result. This was used for instance by Coron and Trélat in the context of a semilinear heat equation, see [21].

Finally, let us cite a recent approach by Agrachev and Sarychev [1] (see also [67]), which uses a generalization of the “Lie bracket” approach (a standard approach from finite-dimensional nonlinear systems), to obtain some global approximate results on the Navier–Stokes equation with a finite-dimensional (low modes) affine control.

Some Other Problems

As we mentioned earlier, many aspects of the theory have not been referred to herein. Let us cite some of them. Concerning the connection between the problem of controllability and the problem of stabilization, we refer to Lions [50], Russell [63,64,65] and Lasiecka and Triggiani [47].

A very important problem which has recently attracted great interest is the problem of numerics and discretization of distributed systems (see in particular [74,75] and references therein). The main difficulty here comes from the fact that the operations of discretizing the equation and controlling it do not commute.

Other important questions considered in particular in the second volume of Lions’s book [51] are the problems of singular perturbations, homogenization, thin domains in the context controllability problems. Here the questions are the following: considering a “perturbed” system, for which we have some controllability result, how does the solution of the controllability problem (for instance associated to the L^2 -optimal control) behave as the system converges to its limit? This kind of question is the source of numerous new problems.

Another subject is the controllability of equations with some stochastic terms (see for instance [8]). One can also consider systems with memory (see again [51] or [7]). Finally, let us mention that many partial differential equations widely studied from the point of view of Cauchy theory, still have not been studied from the point of view of controllability.

The reader looking for more discussion on the subject can consider the references below.

Future Directions

There are many future challenging problems for control theory. Controllability is one of the possible approaches to try to construct strategies for managing complex systems. On the road towards systems with increasing complexity, many additional difficulties have to be considered in the design of the control law: one should expect the control to take into account the possible errors of modelization, of measurement of the state, of the control device, etc. All of these should be included in the model so some robustness of the control can be expected, and to make it closer to the real world. Of course, numerics should play an important role in the direction of applications. Moreover, one should expect more and more complex systems, such as environmental or biological systems (how are regulation mechanisms designed in a natural organism?), to be approached from this point of view. We refer for instance to [32,59] for some discussions on some perspectives of the theory.

Bibliography

1. Agrachev A, Sarychev A (2005) Navier–Stokes equations: control lability by means of low modes forcing. *J Math Fluid Mech* 7(1):108–152
2. Ancona F, Bressan A, Coclite GM (2003) Some results on the boundary control of systems of conservation laws. Hyperbolic problems: theory, numerics, applications. Springer, Berlin, pp 255–264
3. Ancona F, Marson A (1998) On the attainable set for scalar nonlinear conservation laws with boundary control. *SIAM J Control Optim* 36(1):290–312
4. Avdonin SA, Ivanov SA (1995) Families of exponentials. The method of moments in controllability problems for distributed parameter systems. Cambridge Univ Press, Cambridge
5. Ball JM, Marsden JE, Slemrod M (1982) Controllability for distributed bilinear systems. *SIAM J Control Optim* 20:575–597
6. Bandrauk AD, Delfour MC, Le Bris C (eds) (2003) Quantum control: mathematical and numerical challenges. Papers from the CRM Workshop held at the Université de Montréal, Montréal, 6–11 October 2002. CRM Proceedings and Lecture Notes, vol 33. American Mathematical Society, Providence
7. Barbu V, Iannelli M (2000) Controllability of the heat equation with memory. *Differ Integral Equ* 13(10–12):1393–1412
8. Barbu V, Răşcanu A, Tăşitire G (2003) Null controllability of stochastic heat equations with a multiplicative noise. *Appl Math Optim* 47(2):97–120
9. Barbos C, Lebeau G, Rauch J (1992) Sharp sufficient conditions for the observation, control and stabilisation of waves from the boundary. *SIAM J Control Optim* 30:1024–1065
10. Beauchard K (2005) Local controllability of a 1D Schrödinger equation. *J Math Pures Appl* 84:851–956
11. Bensoussan A, Da Prato G, Delfour MC, Mitter SK (1993) Representation and control of infinite-dimensional systems, vol I, II. Systems and Control: Foundations and Applications. Birkhäuser, Boston
12. Brezis H (1983) Analyse fonctionnelle, Théorie et applications; Collection Mathématiques Appliquées pour la Maîtrise. Masson, Paris
13. Burq N (1997) Contrôle de l'équation des ondes dans des ouvertures peu réguliers. *Asymptot Anal* 14:157–191
14. Burq N, Gérard P (1997) Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes. *C R Acad Sci Paris Sér. I Math* 325(7):749–752
15. Cirinà M (1969) Boundary controllability of nonlinear hyperbolic systems. *SIAM J Control* 7:198–212
16. Coron JM (1996) On the controllability of 2-D incompressible perfect fluids. *J Math Pures Appl* 75:155–188
17. Coron JM (1996) On the controllability of the 2-D incompressible Navier–Stokes equations with the Navier slip boundary conditions. *ESAIM Control Optim Calc Var* 1:35–75
18. Coron JM (2007) Control and Nonlinearity. Mathematical Surveys and Monographs, vol 136. American Mathematical Society, Providence
19. Coron JM, Fursikov AV (1996) Global exact controllability of the 2-D Navier–Stokes equations on a manifold without boundary. *Russian J Math Phys* 4:1–19
20. Coron JM, Crépeau E (2004) Exact boundary controllability of a nonlinear KdV equation with critical lengths. *JEMS J Eur Math Soc* 6(3):367–398
21. Coron JM, Trélat E (2004) Global steady-state controllability of one-dimensional semilinear heat equations. *SIAM J Control Optim* 43(2):549–569
22. Courant R, Hilbert D (1989) Methods of mathematical physics, vol II. Partial differential equations, Wiley Classics Library. Wiley, New York
23. Dàger R, Zuazua E (2006) Wave propagation, observation and control in 1-d flexible multi-structures. *Mathématiques and Applications*, vol 50. Springer, Berlin
24. Dolecki S, Russell DL (1977) A general theory of observation and control. *SIAM J Control Optim* 15(2):185–220
25. Douglas RG (1966) On majorization, factorization, and range inclusion of operators on Hilbert space. *Proc Amer Math Soc* 17:413–415
26. Fabre C, Puel JP, Zuazua E (1995) Approximate controllability of the semilinear heat equation. *Proc Roy Soc Edinburgh Sect A* 125(1):31–61
27. Fattorini HO, Russell DL (1971) Exact controllability theorems for linear parabolic equation in one space dimension. *Arch Rat Mech Anal* 43:272–292
28. Fernández-Cara E, Guerrero S (2006) Global Carleman inequalities for parabolic systems and applications to controllability. *SIAM J Control Optim* 45(4):1399–1446
29. Fernández-Cara E, Guerrero S, Imanuvilov OY, Puel JP (2004) Local exact controllability to the trajectories of the Navier–Stokes equations. *J Math Pures Appl* 83:1501–1542
30. Fernández-Cara E, Zuazua E (2000) The cost of approximate controllability for heat equations: The linear case. *Adv Differ Equ* 5:465–514
31. Fernández-Cara E, Zuazua E (2000) Null and approximate controllability for weakly blowing up semilinear heat equations. *Ann Inst H Poincaré Anal Non Linéaire* 17(5):583–616
32. Fernández-Cara E, Zuazua E (2003) Control Theory: History, mathematical achievements and perspectives. *Boletín SEMA* 26:79–140
33. Fliess M, Lévine J, Martin P, Rouchon P (1995) Flatness and de-

- fect of non-linear systems: introductory theory and examples. *Int J Control* 61(6):1327–1361
34. Fursikov A, Imanuvilov OY (1995) On controllability of certain systems simulating a fluid flow, Flow control, Minneapolis, 1992. IMA Math Appl, vol 68. Springer, New York, pp 149–184
 35. Fursikov A, Imanuvilov OY (1996) Controllability of evolution equations. Lecture Notes Series, vol 34. Seoul National University Research Institute of Mathematics Global Analysis Research Center, Seoul
 36. Glass O (2003) On the controllability of the Vlasov-Poisson system. *J Differ Equ* 195(2):332–379
 37. Glass O (2007) On the controllability of the 1-D isentropic Euler equation. *J Eur Math Soc* 9:427–486
 38. Henry J (1978) Controllability of some non linear parabolic equations. In: Ruberti A (ed) Distributed Parameter Systems: Modelling and Identification. Proceedings of the IFIP working conference, Rome, 21–24 June 1976. Lecture Notes in Control and Information Sciences, vol 1. Springer, Berlin
 39. Henry J (1977) Étude de la contrôlabilité de certaines équations paraboliques non linéaires. Thèse, Paris VI
 40. Hörmander L (1983) The analysis of linear partial differential operators, vol I. Grundlehren der mathematischen Wissenschaften, vol 256. Springer, Berlin
 41. Imanuvilov OY (2001) Remarks on exact controllability for the Navier-Stokes equations. *ESAIM Control Optim Calc Var* 6:39–72
 42. Komornik V (1994) Exact controllability and stabilization, The multiplier method. RAM: Research in Applied Mathematics. Masson, Paris
 43. Komornik V, Loret P (2005) Fourier series in control theory. Springer Monographs in Mathematics. Springer, New York
 44. Lagnese JE, Leugering G, Schmidt EJPG (1994) Modeling, analysis and control of dynamic elastic multi-link structures. Systems and Control: Foundations and Applications. Birkhäuser, Boston
 45. Laroche B, Martin P, Rouchon P (2000) Motion planning for the heat equation. *Int J Robust Nonlinear Control* 10(8):629–643
 46. Lasiecka I, Triggiani R (1983) Regularity of hyperbolic equations under $L^2(0, T; L^2(\Gamma))$ -Dirichlet boundary terms. *Appl Math Optim* 10(3):275–286
 47. Lasiecka I, Triggiani R (2000) Control theory for partial differential equations: continuous and approximation theories, vol I. Abstract parabolic systems and II, Abstract hyperbolic-like systems over a finite time horizon. Encyclopedia of Mathematics and its Applications, vols 74, 75. Cambridge University Press, Cambridge
 48. Lebeau G, Robbiano L (1995) Contrôle exact de l'équation de la chaleur. *Comm PDE* 20:335–356
 49. Li TT, Rao BP (2003) Exact boundary controllability for quasilinear hyperbolic systems. *SIAM J Control Optim* 41(6):1748–1755
 50. Lions JL (1988) Exact controllability, stabilizability and perturbations for distributed systems. *SIAM Rev* 30:1–68
 51. Lions JL (1988) Contrôlabilité exacte, stabilisation et perturbations de systèmes distribués, Tomes 1, 2. RMA 8, 9. Masson, Paris
 52. Lions JL (1990) Are there connections between turbulence and controllability? In: Bensoussan A, Lions JL (eds) Analysis and Optimization of Systems. Lecture Notes Control and Inform Sci, vol 144. Springer, Berlin
 53. Lions JL (1992) Remarks on approximate controllability. *J Analyse Math* 59:103–116
 54. Le Bris C (2000) Control theory applied to quantum chemistry: some tracks, Contrôle des systèmes gouvernés par des équations aux dérivées partielles, Nancy, 1999. *ESAIM Proc* 8:77–94
 55. Li TT (2008) Exact boundary observability for quasilinear hyperbolic systems. *ESAIM Control Optim Calc Var* 14:759–766
 56. Lions JL (1971) Optimal control of systems governed by partial differential equations. Grundlehren der mathematischen Wissenschaften, vol 170. Springer, Berlin
 57. López A, Zuazua E (2002) Uniform null controllability for the one dimensional heat equation with rapidly oscillating periodic density. *Ann IHP Analyse linéaire* 19(5):543–580
 58. Miller L (2006) On exponential observability estimates for the heat semigroup with explicit rates. *Atti Accad Naz Lincei CI Sci Fis Mat Natur Rend Lincei*, vol 9. *Mat Appl* 17(4):351–366
 59. Murray RM (ed) (2003) Control in an information rich world. Report of the Panel on Future Directions in Control, Dynamics, and Systems. Papers from the meeting held in College Park, 16–17 July 2000. Society for Industrial and Applied Mathematics, Philadelphia
 60. Osses A (2001) A rotated multiplier applied to the controllability of waves, elasticity, and tangential Stokes control. *SIAM J Control Optim* 40(3):777–800
 61. Rosier L (1997) Exact boundary controllability for the Korteweg-de Vries equation on a bounded domain. *ESAIM Control Optim Calc Var* 2:33–55
 62. Russell DL (1967) On boundary-value controllability of linear symmetric hyperbolic systems. *Mathematical Theory of Control. Proc Conf Los Angeles*, pp 312–321
 63. Russell DL (1973) A unified boundary controllability theory for hyperbolic and parabolic partial differential equations. *Int Math Res Notices* 52:189–221
 64. Russell DL (1974) Exact boundary value controllability theorems for wave and heat processes in star-complemented regions. *Differential games and control theory. Proc NSFCBMS Regional Res Conf Univ. Rhode Island, Kingston, 1973. Lecture Notes in Pure Appl Math*, vol 10. Dekker, New York, pp 291–319
 65. Russell DL (1978) Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions. *SIAM Rev* 20:639–739
 66. Seidman TI (1988) How violent are fast controls? *Math Control Signal Syst* 1(1):89–95
 67. Shirikyan A (2006) Approximate control lability of three-dimensional Navier-Stokes equations. *Comm Math Phys* 266(1):123–151
 68. Smart DR (1974) Fixed point theorems, Cambridge Tracts in Mathematics, No 66. Cambridge University Press, New York
 69. Tucsnak M, Weiss G (2009) Observation and Control for Operator Semigroups. Birkhäuser Advanced Texts. Birkhäuser, Basel
 70. Zhang X (2001) Explicit observability inequalities for the wave equation with lower order terms by means of Carleman inequalities. *SIAM J Cont Optim* 39:812–834
 71. Zuazua E (1990) Exact controllability for the semilinear wave equation. *J Math Pures Appl* 69(1):1–31
 72. Zuazua E (1993) Exact controllability for semilinear wave equations in one space dimension. *Ann Inst H Poincaré Anal Non Linéaire* 10(1):109–129
 73. Zuazua E (1998) Some problems and results on the controllability of Partial Differential Equations. Proceedings of the Second European Conference of Mathematics, Budapest, July

1996. Progress in Mathematics, vol 169. Birkhäuser, Basel, pp 276–311
74. Zuazua E (2002) Controllability of Partial Differential Equations and its Semi-Discrete Approximations. *Discret Continuous Dyn Syst* 8(2):469–513
 75. Zuazua E (2005) Propagation, observation, and control of waves approximated by finite difference methods. *SIAM Rev* 47(2):197–243
 76. Zuazua E (2006) Controllability and observability of partial differential equations: some results and open problems. In: Dafermos C, Feireisl E (eds) *Handbook of differential equations: evolutionary differential equations*, vol 3. Elsevier/North-Holland, Amsterdam

Information Theoretic Complexity Measures

DANAIL G. BONCHEV

Virginia Commonwealth University, Richmond, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Information Content of Atoms](#)

[Information Content of Atomic Nuclei](#)

[Information Content of Molecules](#)

[Information Content of Networks](#)

[Future Directions](#)

[Bibliography](#)

Glossary

Adjacency matrix A square $V \times V$ matrix, where V is the number of vertices of graph G . The matrix entries are $a_{ij} = 1$ for i, j – neighbors, and $a_{ij} = 0$, otherwise. In **undirected graphs**, the matrix is symmetric with respect to the main diagonal.

Branch A linear (path) **subgraph**, beginning with a vertex of degree higher than 2, and ending with a vertex of degree 1.

Branching The (topological) property of a graph to have branches, as well the patterns of branch connectivity.

Centrality The (topological) property of vertex (or edge) organization in a graph with respect to central vertex(es) or edge(s).

Complexity measures Quantitative assessments of systems complexity, obtained mainly by applying information theory and graph theory.

Connected graph A graph is connected, if there is a **walk** between any pair of vertices.

Cycle A **path**, the initial and end vertices of which coincide.

Cyclicity The (topological) property of a graph to have cycles, as well as the patterns of cycle connectivity.

Distance The number of edges connecting two vertices along the shortest path between them.

Eccentricity – see **Vertex eccentricity**.

Graph A mathematical structure composed of points (vertices), connected by lines (edges).

Graph center Vertex(es) in a graph with the smallest **eccentricity**.

Graph distance The sum of all **vertex-vertex distances** in the graph. Also: The sum of all **vertex distances** in the graph.

Information content A quantitative characteristic of a system calculated by using information theory. Three major types of information content have been of use: compositional, structural (mainly topological), and functional ones.

Information theory Created initially as theory of communications by Shannon in 1949, it extends the statistical thermodynamics of Boltzmann to processes involving information. Other versions of this theory have been proposed; the best known one being the Kolmogorov nonprobabilistic theory of information. One of the main applications of information theory is to define **information content** of systems.

Information theoretical descriptor (information index)

A number, which is a quantitative characteristic of a system, calculated by using **information theory**.

Graph path A sequence of edges connecting two graph vertices without a repeated visit of a vertex or edge. The length of the path is equal to the number of path edges.

Molecular graph A graph, the vertices of which stand for atoms of a molecule, and edges represent chemical bonds between the atoms.

Network An interconnected system of elements of any physical nature, and the graph representing the system. The terminology used in network theory is similar to that of graph theory, vertices and edges being called *nodes* and *links*, respectively.

Subgraph A graph $G' = (V', E')$ is called a subgraph of graph $G = (V, E)$ if V' belongs to the set of vertices V and E' belongs to the edges of E .

Total adjacency of a graph The sum of all adjacency matrix entries. Alternatively, the sum of all **vertex degrees**. In undirected graphs, the total adjacency is equal to the doubled number of edges.

Undirected graph A graph in which the binary adjacency relation $a_{ij} = a_{ji} = 1$ exists for all adjacent vertices i, j . If in a graph there is at least one adjacency rela-

tion $a_{ij} = 1$, the symmetric relationship a_{ji} for which does not exist, the graph is **directed (a Digraph)**.

Vertex degree The number of the nearest neighbors of a vertex. Also, the number of edges incident to the vertex.

Vertex distance The sum of the distances from a given vertex to all other vertices in the graph.

Vertex eccentricity The largest distance from a vertex to any other vertex in the graph.

Walk A walk in a simple graph (having no multiple edges or loops) is a sequence of consecutive vertices and edges with repetitions allowed. The length of a walk is the number of edges (including repetitions).

Weighted distribution The ordered set of all values of given quantity.

Definition of the Subject

Complexity is a multifaceted concept, related to the degree of organization of systems. Patterns of complex organization and behavior are identified in all kinds of systems in nature and technology. Essential for the characterization of complexity is its quantification, the introduction of complexity measures or descriptors, following Lord Kelvin's words that science begins when we can use numbers. Historically, the first attempt to quantify complexity was based on Shannon's information theory [1], and it involved the information content as a measure of molecular complexity [2]. Fifty years later, the complexity of molecules and their interactions is assessed by a variety of methods, with information theory preserving its leading role. This article aims to review the vast area of complexity measures, based on information theory as applied to chemical and biochemical systems. Many of these measures have found application for predicting physico-chemical properties and biological activities of chemical compounds, contributing thus to the development of new drugs and chemical products. The expertise accumulated has recently found a new vast area of application, the networks of biomolecules performing the basic functions of life in cells and organisms. The essence of life itself has been reformulated to incorporate as an essential component the processing of information.

Introduction

The notion of information has been first used in scientific literature in 1894 by Boltzmann, who stated that "Every piece of information obtained for a physical system is related to the decrease in the number of its possible states; therefore the increase of entropy means 'loss of information'" [3]. One may speculate that only the early death of

Boltzmann, who was far ahead of his contemporaries, prevented the development of information theory at the very beginning of 20th century. Indeed, Boltzmann's statistical thermodynamics was a direct predecessor of Shannon's information theory. The entropy S of a system of N particles, distributed over t states, having N_1, N_2, \dots, N_t particles with energies E_1, E_2, \dots, E_t , respectively, was related by Boltzmann to the total number W of physical states of the system, where k is a constant:

$$S = k \ln W = k \frac{N!}{N_1! N_2! \dots N_t!} . \quad (1)$$

For $N \gg 1$, the approximation of Stirling turns Eq. (1) into

$$S \approx k \left(N \ln N - \sum_{i=1}^t N_i \ln N_i \right) . \quad (2)$$

Equation (2) after an elementary substitution of the constant k is identical to the Shannon equation for the entropy of information H of a message transmitted through information channels:

$$H = N \log_2 N - \sum_{i=1}^t N_i \log_2 N_i , \quad \text{bits} . \quad (3)$$

Shannon's information theory regards such a message as a specific set of symbols (an "outcome") selected from an ensemble of all t such sets containing the same total number of symbols N . Probabilities p_1, p_2, \dots, p_t are assigned to each outcome, the probability of the i th outcome being proportional to the number of symbols N_i it contains: $p_i = N_i/N$. Shannon's entropy of information H characterizes the uncertainty of the expected outcome. Upon a totally random transmission all outcomes are equiprobable and the entropy of information is maximal. Conversely, in case of a single outcome, $H = 0$. In the intermediate cases, the amount of information transmitted is the difference between the maximum entropy and the specific value the Shannon H -function has for the system of interest. Thus, information emerges as a measure for the eliminated outcome uncertainty.

Another more popular form of the Shannon equation defines the average entropy of information \bar{H} per communication symbol:

$$\begin{aligned} \bar{H} &= - \sum_{i=1}^t p_i \log_2 p_i \\ &= - \sum_{i=1}^t \frac{N_i}{N} \log_2 \frac{N_i}{N} , \quad \text{bits/symbol} . \end{aligned} \quad (4)$$

One bit of information is obtained when learning the outcome of a process eliminating the uncertainty of a choice between two equally probable options. The values of the two Shannon's entropies introduced by Eqs. (3) and (4) vary within the following ranges in bits:

$$0 \leq H \leq N, \quad (5a)$$

$$0 \leq \bar{H} \leq 1. \quad (5b)$$

Shannon's theory soon exceeded the narrow limits of a communication theory, and was considered as an extension of Boltzmann's theory. This view was advocated by Brillouin [4], who's negentropy principle of information views information as the negative component of entropy. The second law of thermodynamics thus generalized allows only such spontaneous processes in closed systems that increase entropy and lose information.

Another important extension was the view on *structure* as a message that carries certain amount of information. That was how the notion of *information content* of a (chemical) structure emerged in the early 1950s [5,6,7,8,9]. A radical reinterpretation of the meaning of Shannon's equations (3) and (4) was proposed by Mowshowitz [10] in 1968. When applied to molecules, the *H*-function does not measure the average uncertainty for selecting a molecule from the ensemble of all molecules having the same number of atoms. Rather, it is the information content of the structure relative to a system of symmetry transformations that leaves the structure invariant. Bonchev [11] supported this interpretation by the argument that entropy is transformed into information by the mere process of the structure formation from its constituent elements. This structural information is conserved until the structure is destroyed, when it turns back into entropy.

Mowshowitz [10] presented his approach as a *finite probabilistic scheme*, applicable to any system having symmetry elements. The system of *N* interacting elements is treated as a graph, the vertices of which are partitioned into *k* equivalence classes, according to symmetry operations (graph automorphisms), which exchange vertices while preserving the graph adjacency.

Equivalence classes	$1, 2, \dots, k$
Partition of elements	N_1, N_2, \dots, N_k
Probability distribution	p_1, p_2, \dots, p_k

Here, $p_i = N_i/N$ is the probability of a randomly chosen element to belong to class *i*, which has N_i elements, and $N = \sum N_i$. The analysis of Eqs. (3) and (4), expressing now the information content *I* of the system, shows

that information has the maximum value when each element is in a separate class, i. e., when the system has no symmetry. The information content is zero when all elements belong to a single equivalence class, i. e., when the system has no structure, due to its high symmetry. One might infer that the total and average information content, *I* and \bar{I} , could be used as complexity measures, which relate high complexity to low symmetry and larger diversity of system's elements. Low complexity (simplicity) is characterized by uniformity, resulting from high symmetry and lack of diversity.

In what follows till the end of this article, it will be shown how this information-theoretic formalism can be used to characterize the structure of atoms, molecules, and (molecular) networks. It will be demonstrated that symmetry-based information-theoretic descriptors cannot always be good complexity measures, because symmetry is a simplifying factor. A better approach will be introduced, proceeding from a weighted version of the original Mowshowitz' scheme.

Information Content of Atoms

Information Content of Chemical Elements and Nuclides

One may define the information content of atoms and their nuclei in a variety of ways proceeding from the total number of protons *z*, neutrons *n*, and electrons $e = z$ (in neutral atoms), and their different distributions [12]. A good starting point is considering the atomic structure partitioned into two substructures – a nucleus and an electron shell. The *total information content of a nuclide* is thus defined [13] as:

$$I_{\text{nuclide}} = (A + z) \log_2(A + z) - A \log_2 A - z \log_2 z, \quad (6)$$

where the mass number $A = z + n$. In atomic mass units, *A* (an integer) is approximately equal to the atom's mass (a real number).

The information content of a chemical element can then be defined [13] as the average information content of all naturally occurring nuclides of this element:

$$I_{\text{chem.element}} = \sum_i c_i I_{\text{nuclide},i}, \quad (7)$$

c_i being the abundance fraction of nuclide *i*.

The next step is to analyze the distribution of electrons of the atoms of chemical elements into electron shells (I_n), subshells (I_{nl}), atomic orbitals (I_{nlm}), and spin-orbitals (I_{nlmms}) [11,14,15]. In all these cases, *z* electrons are considered distributed into *k* equivalence classes having

N_1, N_2, \dots, N_k electrons, respectively. Denoting the combination of quantum numbers that defines the specific type of information content by x , one finds the corresponding total and average information content, $I(x)$ and $\bar{I}(x)$:

$$I(x) = z \log_2 z - \sum_i z_i \log_2 z_i, \quad (8)$$

and

$$\bar{I}(x) = - \sum_i \frac{z_i}{z} \log_2 \frac{z_i}{z}. \quad (9)$$

A third information function, termed *differential information content*, $\Delta I_x(z)$, is defined by the difference between the information content of the chemical element with atomic number z and combination of quantum numbers x , and those of the element with atomic number $z - 1$:

$$\Delta I_x(z) = I_x(z) - I_x(z - 1). \quad (10)$$

The differential information content was shown to be a sensitive descriptor for the periodicity of chemical elements. As shown in Fig. 1, $\Delta I_x(z)$ has a sharp maximum in the first element of each period, or s -, p -, d -, and f -subperiod, followed by a gradual decrease to a minimum in the corresponding last element. This regular trend is best demonstrated by periods II and III. In periods IV to VII, the filling of the d - and f -subshells with delay produces the inequality:

$$\Delta I_n((n-2)f) < \Delta I_n((n-1)d) < \Delta I_n(np) < \Delta I_n(ns). \quad (11)$$

The violations in the “ideal order” of filling electron f - and d -subshells, caused by the accelerated adding of $(n-1)d$ - or $(n-2)f$ -electrons at the cost of the decreased

population of the ns -subshells, are also captured in the information function by sharp minima as seen in Fig. 1 for the extra d -electron in Cr, Cu, Nb, Pd, etc.

Information Equations for Periods and Groups of Chemical Elements

Such equations are derived proceeding from several atomic information descriptors [16,17]. Equation (12) is based on the information for electron distribution over an nl -subshell. It contains the period constant P , which is equal for periods I through VII to 0, 2, 19.51, 37.02, 87.75, 138.48, and 242.51 bits, and a group constants $k_n l = 1$ or 2 for groups 1 and 2, $k_n l = 1$ to 10 for groups 3 to 12, and $k_n l = 1$ to 6 for groups 13 to 18, respectively:

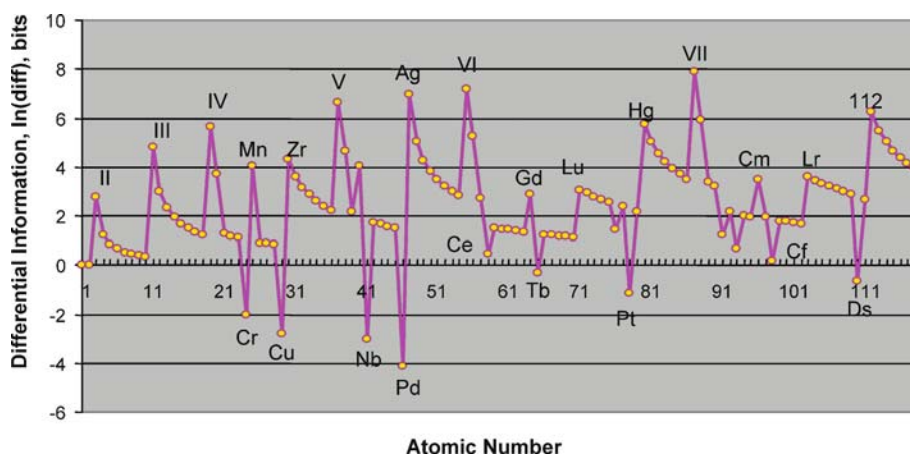
$$I_{nl} = z \log_2 z - P - \sum_l k_{nl} \log_2 k_{nl}. \quad (12)$$

Another equation for the periods and groups in the periodic table is obtained with the total information on electron distribution over atomic orbitals:

$$I_{AO} = (z_0 + a + b) \log_2(z_0 + a + b) - z_0 - b. \quad (13)$$

Here, b is the number of paired electrons in the atomic orbitals of the chemical element. In the ground state of the atoms of elements of groups 1, 2, and 13 through 18, the group constant a is equal to the lowest valence of the element (1, 0, 1, 2, 3, 2, 1, 0, respectively), whereas in the highest valence state it equals the old numbering of the eight main groups (1 to 8). The period constant z_0 is the atomic number of the noble gas that ends the preceding period ($z_0 = 2, 10, 18, 36, 54, 86$).

The information on electron distribution over the values of the magnetic spin quantum number ($m_s = +1/2$



Information Theoretic Complexity Measures, Figure 1

Differential information on the distribution of atomic electrons over electron shells

and $m_s = -1/2$) provides yet another period/group equation for chemical elements:

$$\begin{aligned} \bar{I}_{ms} &= -\frac{z+a}{2z} \log_2 \frac{z+a}{2z} - \frac{z-a}{2z} \log_2 \frac{z-a}{2z} \\ &\approx 1 - \frac{k}{(z_0 + a + b)^2}. \end{aligned} \quad (14)$$

Here, b and the period constant z_0 are those from Eq. (13). The group constant $k = a^2/(2 \ln 2)$ includes the number of unpaired electrons a . The error introduced by the approximation used decreases with the fourth power of z and is very small.

The Pauli Exclusion Principle and Hund's First Rule Maximize the Atomic Information Content

The information equations for electron distributions in atoms provided the basis for a reinterpretation of the physical principles and rules controlling the building of the atomic electronic structure [18]. Hund's first rule, which requires maximum filling of atomic orbitals in s -, p -, d -, and f -subshells with unpaired electrons, may be interpreted as a rule demanding maximum information on atomic orbitals, I_{nlm} . This follows directly from our Eq. (13), which maximizes when the number of paired electrons is $b = 0$. The absolute maximum of the atomic information content according to Eq. (8) is reached when all $z_i = 1$. This case corresponds to electron distribution over spin-orbitals, defined by the four quantum numbers n , l , m , and m_s , and required by the Pauli exclusion principle. The Hund rule and the Pauli exclusion principle, thus, emerge as different manifestations of a more general trend requiring the maximum information content of atoms, thus characterizing atoms as structures of maximum complexity. Proceeding from quantum mechanical and group theory analysis, this trend was shown to extend to any fermionic system.

Atomic Information Descriptors as Tools for Calculating and Predicting the Properties of Atoms and Chemical Elements

Encoding a detailed description of the electronic structure of atoms, the information theoretic indices were shown to be an effective tool in quantifying the periodicity in the properties of chemical elements. They provide very high correlation with almost all atomic characteristics and physico-chemical properties of chemical elements. As an example, 21 of 23 examined properties of alkali metals have been found to correlate with the atomic information descriptors with an average correlation coefficient of 0.997, a degree of correlation far better than that with atomic

number. The models derived have been applied to the prediction of a number of atomic characteristics and properties of the transactinide elements 113–120 [12,19].

Information Content of Atomic Nuclei

Information on Proton–Neutron Composition

Atomic nuclei have been characterized by the distribution of protons z and neutrons n in shells and subshells in a manner similar to the one described for electron distribution in Sect. "Information Content of Atoms". The information index on the proton–neutron composition of atomic nuclei, I_{pn} , has been found to exhibit interesting properties [20]:

$$I_{pn} = A \log_2 A - z \log_2 z - n \log_2 n, \quad \text{bits}, \quad (15)$$

$$\bar{I}_{pn} = -\frac{z}{A} \log_2 \frac{z}{A} - \frac{n}{A} \log_2 \frac{n}{A}. \quad (16)$$

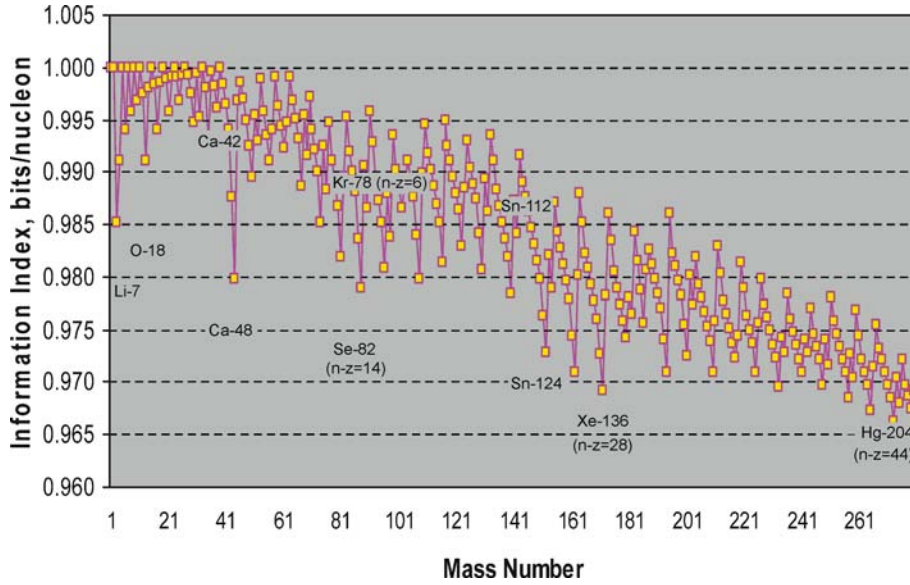
These equations identify the symmetric nuclei having equal number of protons and neutrons, $p = z = n$, as nuclei with the highest complexity, due to their maximum information on the proton–neutron composition. The total information content of these nuclei is equal to the mass number A , whereas the mean information is exactly 1 bit. Such are the symmetric stable nuclei of the elements at the beginning of the periodic table: ^2H , ^4He , ^6Li , etc., up to ^{40}Ca . With the increase of the atomic number of the element, the filling of the proton energy levels in the nucleus proceeds with delay. The resulting excess of neutrons is termed *isotopic number*: $\beta = n - z = A - 2z$.

The basic patterns in the behavior of the proton–neutron information content are demonstrated in Fig. 2 for all stable nuclei up to ^{204}Hg . The symmetric light nuclei are located on the line of maximum mean information content of exactly 1 bit. With the increase in atomic mass, I_{pn} diverges more and more from its maximum, in correspondence with the delay in filling the proton energy levels. The minima in the figure correspond to the largest excess of neutrons for the series of several isotopes of the same chemical element. Conversely, the maxima after ^{40}Ca refer to the least possible excess of neutrons.

The mean I_{pn} index, defined by Eq. (16) is approximated with a sufficient accuracy (the relative error does not exceed 0.02% with exception of ^3He with 0.15%) by

$$\bar{I}_{pn} \approx 1 - \frac{1}{2 \ln 2} \times \frac{\beta^2}{A^2} = 1 - \frac{1}{2 \ln 2} \times \frac{(A - 2z)^2}{A^2}. \quad (17)$$

The information index on the proton–neutron composition of atomic nuclei is thus shown to increase with the increase in mass number A and atomic number z , and



Information Theoretic Complexity Measures, Figure 2

The average information on the proton–neutron composition of the naturally occurring atomic nuclei. The horizontal line of 1 bit information content describes the symmetric light nuclei having the same number of protons and neutrons, whereas the minima and maxima stand for the isotopes of certain elements having the largest and the smallest neutron excess, respectively

to decrease with the increase in the isotopic number and the number of neutrons.

The Concept for “Defect” of Information and the Systematics of Nuclides

The total information on the proton–neutron composition of a nucleus, I_{pn} , expressed in bits according to Eq. (15), is very close to the mass number A . The difference ΔI_{pn}^* between the two quantities has been introduced [20] as “defect” of information by analogy with the defect of mass upon formation of atomic nuclei:

$$\Delta I_{pn}^* = A - I_{pn} \approx \frac{1}{2 \ln 2} \times \frac{(A - 2z)^2}{A}. \quad (18)$$

Since this deviation results in a decrease of the binding energy, it has been conjectured that the defect of information can be regarded as a negative component of the binding energy, E_b . In fact, ΔI_{pn}^* coincides (with a transition coefficient $k = 25.1$ MeV/bit) with the parameter of the relative symmetry of the nucleus, δ , in Weizsäcker’s equation for nuclear binding energy:

$$\delta = -18.1 \times \frac{(A - 2z)^2}{A} \text{ MeV}. \quad (19)$$

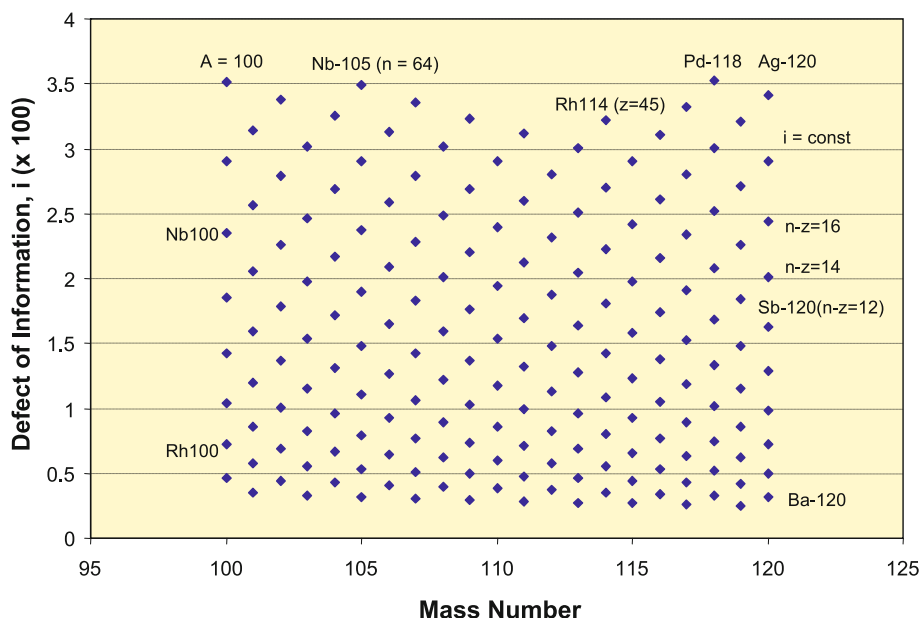
The remarkable equivalence of Eqs. (18) and (19) affords evidence for the usefulness of information theory in

the study of atomic nuclei. Correlation of I_{pn} and ΔI_{pn}^* with the total nuclear binding energy has also been reported for series of nuclides with a constant isotopic number β .

The concept for a defect of information has also found application in the systematics of nuclides, introducing the new series of isodefective nuclides having the same defect of information per nucleon i [21]:

$$\begin{aligned} i &= \frac{\Delta I_{pn}^*}{A} = \frac{1}{2 \ln 2} \times \left(\frac{\beta}{A} \right)^2 \\ &= \frac{1}{2 \ln 2} \times \left(1 - 2 \frac{z}{A} \right)^2, \quad \text{bits/nucleon}. \end{aligned} \quad (20)$$

It can be seen from Eq. (20) that nuclei having the same relative charge z/A also have the same mean defect of information i . The new class of nuclides (“isodefectants”) thus introduced adds a fifth dimension to the Chart of the Nuclides, traditionally based on the classes of isotopes ($z = \text{const}$), isotones ($n = \text{const}$), isobars ($A = \text{const}$), and isodifferent nuclides ($\beta = n - z = \text{const}$). A portion of the Chart of the Nuclides represented with five classes of nuclides is shown in Fig. 3 with defect of information/mass number coordinates (i/A diagram). The illustration includes nuclides having mass numbers between 100 and 120, and atomic numbers between 39 (Y) and 56 (Ba).



Information Theoretic Complexity Measures, Figure 3

A fragment from the representation of the Chart of the Nuclides with five classes of nuclides: isobars (e. g., Rh-100 to Nb-100), isotopes (e. g., Rh-100 to Rh-116), isotones (e. g., Nb-105 to Ba-120), isodifferents (e. g., Nb-100 to Sb-120), and isodefectants (e. g., the line with $i = 3$)

The behavior of the series of isobars, isotopes, isotones, and isodifferent nuclides in the i/A plot is described by equations for the mean defect of information i :

$$i(\text{isobars}) = c_0(1 - c_1 z)^2, \quad (21a)$$

$$i(\text{isotopes}) = c_0 \left(1 - \frac{c_2}{A}\right)^2, \quad (21b)$$

$$i(\text{isotones}) = c_0 \left(\frac{c_3 - z}{c_3 + z}\right)^2, \quad (21c)$$

$$i(\text{isodifferents}) = c_0 \frac{c_4}{A^2}, \quad (21d)$$

where $c_0 = 1/2 \ln 2$, $c_1 = 2/A$, $c_2 = 2z$, $c_3 = n$, $c_4 = \beta = A - 2z$, respectively.

Predictions of Nuclear Binding Energies of Nuclides of Elements #101–108

A considerable similarity has been found in the information equations describing the nuclear and electronic structure of chemical elements [12]. This finding, along with the observed correlations of the information on the proton–neutron composition and the defect of this information, with nuclear binding energy, has prompted the search for direct relationships between nuclear and electronic energies. Equations have been derived that relate fairly well the ground state electronic energy E_e of chemical elements to

the nuclear binding energy $E_b(id)$ of isodifferent series of nuclides [12]. Predictions have been made on this basis in the year 1980 for the nuclear binding energy of 45 not synthesized isotopes of elements 101–108 [22]. In 2003, these predictions have been confirmed [12] with relative error of 0.1% for 41 isotopes, which have meanwhile been synthesized. This result has confirmed the usefulness of information content as a measure of structural similarity and complexity on atomic level, and has indicated the potential existence of correlated motions of nucleons and electrons.

This concludes the analysis of information theoretic descriptors of atomic and nuclear complexity. In the remaining sections the focus will be shifted to the description of molecular structures and the biochemical networks they form in living things.

Information Content of Molecules

Early History

Several years after Shannon published his fundamental work on information theory [1] a group of US biologists proposed to define the information content of molecules and use it in assessing the complexity of living cells and organisms [2]. Several alternative definitions have been proposed. Dancoff and Quastler introduced in 1953 the *information for the kind of atoms in a molecule* [5] (renamed

Information Theoretic Complexity Measures, Table 1

Comparison of the predicted and measured values of the nuclear binding energies of some nuclides of chemical elements #101 to 108 [12]

β	z	A	E_b (pred.)	E_b (exper.)	β	z	A	E_b (pred.)	E_b (exper.)
49	103	255	1889.235	1887.465	53	106	265	1945.166	1943.199
	104	257	1899.300	1896.954		107	267	1954.800	1952.342
	105	259	1909.259	1906.121	54	102	258	1913.639	1911.128
	106	261	1919.112	1915.447		103	260	1924.599	1919.621
	107	263	1928.862	1924.336		104	262	1935.477	1930.934
50	104	258	1903.592	1904.563		105	264	1946.285	1939.257
	105	260	1912.916	1912.603		106	266	1957.026	1950.468
	106	262	1922.085	1923.259	55	102	259	1916.513	1916.569
	107	264	1931.101	1930.932		103	261	1926.384	1926.418
	108	266	1939.969	1941.345		104	263	1936.086	1936.563
51	104	259	1912.138	1910.716		105	265	1945.581	1946.224
	105	261	1921.943	1920.042		106	267	1954.913	unknown
	106	263	1931.617	1929.620	56	101	258	1910.768	1911.701
	107	265	1941.163	1938.568		102	260	1920.146	1923.139
	108	267	1950.582	1947.803		103	262	1929.240	1931.927
52	103	258	1911.461	1906.916		104	264	1938.056	1943.295
	104	260	1922.186	1918.037		105	266	1946.616	unknown
	105	262	1932.856	1926.206	57	101	259	1918.088	1917.837
	106	264	1943.456	1937.123		102	261	1928.501	1928.317
	107	266	1953.973	1944.952		103	263	1938.762	1938.413
53	103	259	1915.474	1913.955		104	265	1948.874	unknown
	104	261	1925.505	1923.949		105	267	1958.840	unknown
	105	263	1935.472	1933.417					

later as *information for atomic composition* I_{ac} , and *information for chemical composition*, I_{cc}). Atoms are not the best structural units in describing complex biomolecules like DNA and proteins, and Branson [6] calculated the *information on aminoacid composition*, I_{AAC} , of 26 proteins. Augenstine [23] added the *configurational information* of polypeptide chains in order to calculate the total information content of proteins. Rashevsky [7] has shown that even molecules composed of the same kind of atoms can have a large information content based on the atom-atom connectivity, which he termed *topological information*. All these attempts enabled the first quantitative estimates [5,24] of the information content of a living cell and a human to be about 10^{11} and 10^{25} bits, respectively. It was argued that a certain lower limit of complexity must exist even for the simplest organisms. The topological information of Rashevsky was more specifically defined by classifying the vertices of a *molecular graph* (a two-dimensional representation of the molecule) into layers of first, second, etc. neighborhoods (an idea developed in more detail much later [25,26] as neighborhood complexity of first, second, etc. order). Trucco [8] has pointed out that the equivalence neighborhood criterion of Rashevsky cannot

always provide topological equivalence of two graph vertices, and redefined the topological information descriptor in terms of the vertex orbits of the automorphisms group of the graph. Trucco [9] also extended the topological characterization of the molecule by partitioning the graph edges into the edge orbits of the graph. Later, these two types of topological descriptors of molecules were more concretely renamed as vertex and edge orbit's information, respectively [27]. Other structural aspects have been taken into account by Morovitz [28], who combined the information on the chemical composition of a molecule to that on the *possible valence bonds* between the atoms in the molecule. The relation between the vertex orbit information descriptor and the complexity of graphs has been studied in detail by Mowshowitz [10], who also introduced the chromatic information content of graphs [29]. This concludes the brief early history of molecular information theoretic descriptors. Since the beginning of the 1970s, the explosive development of chemical graph theory [30,31] contributed greatly to the corresponding variety of contributions to chemical information theory [27], which will be reviewed in the rest of this section.

Application of the Shannon Equation to Sets of Elements with Complex Equivalence Criteria, and to Weighted Sets of Elements

Three major types of complexity are considered in natural sciences: compositional, structural, and functional ones. The functional complexity is essential for the characterization of biological systems ► [Biological Complexity and Biochemical Information](#), [32], but has not been a subject of studies in chemistry. The behavior of small molecules is described sufficiently accurately not only by quantum mechanics, but also by simple concepts like functional groups, polarity, acidic and basic properties, etc. For this reason, the present article deals with the information theoretic description of molecular compositional and structural complexity. It has been shown that the finite probability scheme of Mowshowitz [10], introduced in Sect. “[Introduction](#)”, provides excellent descriptors of compositional complexity. However, for structural complexity this scheme does not work well, producing predominantly information descriptors of low sensitivity, which cannot match well structural patterns of different complexity [33]. Examples illustrating this conclusion will be given further in this section.

The search for a better solution resulted in two different approaches. The insufficiency of the equivalence criterion for partitioning the set of system elements into equivalence classes has prompted the simultaneous use of two or more such criteria. A typical situation is to group the atoms in a molecule first according to their chemical nature, and to partition further the classes thus formed according to some structural criterion, e.g., atoms of the same kind having different neighborhood of atoms [7].

A more general approach [34] is based on weighted graphs (vertex-, edge-, and vertex-edge-weighted ones). Weights can represent any measured or calculated quantitative characteristics of the atoms and bonds in the molecule, such as electronegativity, bond orders, and different graph theoretical invariants like vertex and edge degrees, distances, etc. The Mowshowitz finite probabilistic scheme has, thus, been extended so as to include not only the partition of elements of the system into k classes with N_i elements in class i , and equivalence-based probabilities $p_i = N_i / \sum_i N_i = N_i / N$, but also the weight w_i of type α of this class, and the “weighted” probability ${}^w p_i = w_i / \sum_i w_i = w_i / W$:

Equivalence classes	$1, 2, \dots, k$
Partition of elements	N_1, N_2, \dots, N_k
Probability distribution	p_1, p_2, \dots, p_k
Weights	w_1, w_2, \dots, w_k
Probability w -distribution	${}^w p_1, {}^w p_2, \dots, {}^w p_k$

The Shannon equations (3) and (4) are thus modified to express the information on the weighted distribution α of the system:

$${}^w I(\alpha) = W \log_2 W - \sum_{i=1}^k N_i w_i \log_2 w_i, \quad (22)$$

$${}^w \tilde{I}(\alpha) = - \sum_{i=1}^k N_i \frac{w_i}{W} \log_2 \frac{w_i}{W}. \quad (23)$$

The two weighted information descriptors from Eqs. (22) and (23) are defined within the ranges:

$$0 \leq {}^w I(\alpha) \leq W \log_2 W; \quad 0 \leq {}^w \tilde{I}(\alpha) \leq \log_2 W, \quad (24)$$

where the lower bound corresponds to a system without structure ($w_i = W$), and the upper bound can be attained by a system having the maximum number of classes with a single element of unit weight in each class ($k = N = W$). Typical weights that might be associated with vertices of molecular graphs are vertex degree (the sum of all edges emanating from the vertex) [27] or vertex distance (the sum of distances from that vertex to all other ones), as the weighted probabilistic scheme has been introduced for the first time [34].

Information Descriptors of Compositional Complexity

As mentioned above, the first information descriptor introduced is the *information on chemical composition*, I_{CC} [5], based on the distribution of atoms into subsets containing atoms of the same chemical element. This descriptor is a measure for the molecular complexity derived from its *compositional diversity*. An example illustrating this approach to molecular complexity is the increase in diversity within the series CH_4 , CH_3F , CH_2FCl , CHFClBr . The atomic compositional distributions of these molecules $\{1,4\}$, $\{1,1,3\}$, $\{1,1,1,2\}$, and $\{1,1,1,1,1\}$ produce the increase in I_{CC} of this series from the simplest molecule of CH_4 to the most complex one, CHFClBr (3.61, 6.85, 9.61, 11.61 bits, respectively).

A similar approach was used to characterize the composition of complex molecules like proteins, proceeding from the distribution of their aminoacid residues [6]. Thus, the insulin molecule having 51 aminoacid residues is characterized by the following composition: Cys (6), Leu (6), Val (5), Glu (4), Gly (4), Tyr (4), Ala (3), Asn (3), Glu (3), Phe (3), Ser (3), His (2), Arg (1), Ile (1), Lys (1), Pro (1), Thr (1). The compositional distribution $51\{6,6,5,4,4,4,3,3,3,3,3,2,1,1,1,1\}$ produces the aminoacid information content of insulin, $I_{AAC} = 196.89$ bits. Information theory has been widely applied in a variety of versions to assess the complexity and evolution of

proteins [35,36,37,38,39,40,41]. Mostly, these applications have been based on the original interpretation of Shannon's equations by selecting the aminoacid sequence of a protein from the set of all possible sequences of the same length. The evolution of a protein characterized quantitatively in this manner is considered also as a measure of evolution of the corresponding DNA, which encodes that protein [42,43].

The complexity of DNA and its evolution has been itself a subject of very broad studies applying information theory ► **Biological Complexity and Biochemical Information**, [44,45,46,47,48], which could be a subject of a separate article. Different types of DNA information content have been defined. The simplest one is the DNA nucleotide composition, proceeding from the frequency of adenine (A), cytosine (C), guanine (G) and thymine (T) in protein-coding and -noncoding sequences. The DNA and individual genes' exon/intron composition, and codon composition has also been defined. The length of exons and introns can also be taken into account either in the weighted type information descriptors (Eqs. (22) and (23)) or by using the Jensen–Shannon divergence of sequence S , $J(S)$ [49,50]

$$J(S) = H(S) - \sum_{i=1}^k \frac{l_i}{L} H(S_i), \quad (25)$$

where $H(S)$ is the Shannon sequence entropy, and $H(S_i)$ is the entropy of the i th sequence segment of length l_i . These two approaches are applicable to all possible genome segmentations.

Another approach widely used for determining the information content of symbolic sequences is Kolmogorov's algorithmic information [51], which has been defined as the length of the sequence shortest description in some fixed **description language** or as the length of the shortest code generating the given sequence. A variety of constructive realizations of such coding have been developed [52,53,54] and applied to DNA sequences [55,56,57]. The mutual information has been used as a tool for predicting protein coding regions in DNA [58]. More details on complexity of DNA and proteins can be found in the article ► **Biological Complexity and Biochemical Information**.

Information Measures of Topological Complexity of Molecules

Adjacency-Based Information Descriptors Graph theory [59,60] offers very effective graphical and quantitative tools for representing molecular structure. Molecules are three-dimensional structures; however the two-dimen-

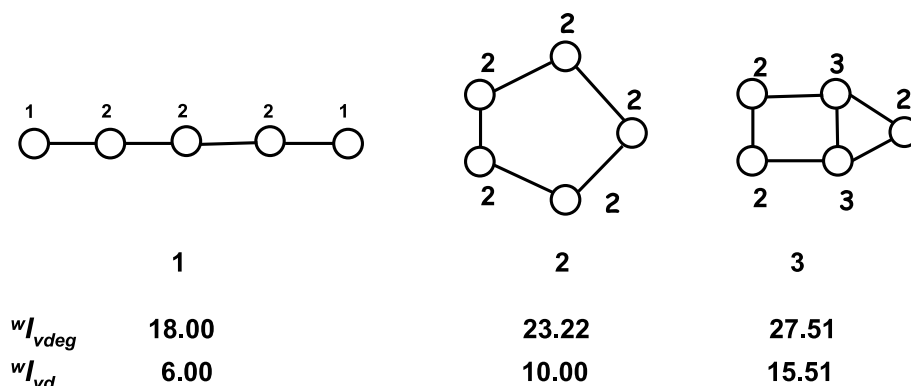
sional representation of molecular graphs conserves a very important part of the structural information. Molecular graphs are composed of points (termed *vertices*) representing atoms, and connecting lines (called *edges*) representing covalent chemical bonds. Atom-atom connectivity, which is the fundament of molecular topology, is characterized by the *adjacency matrix*, a square matrix with a size determined by the number of vertices V . The entries a_{ij} of this matrix are equal to one for any pair of adjacent vertices i and j , and zero otherwise. The sum of the entries in a matrix row is in fact the number of nearest neighbors of the vertex i associated with this row, and is called *vertex degree*, a_i . The sum of all vertex degrees defines the graph *total adjacency*, $A(G)$:

$$a_i(G) = \sum_{j=1}^V a_{ij}; \quad (26a)$$

$$A(G) = \sum_{i=1}^V a_i = \sum_{i=1}^V \sum_{j=1}^V a_{ij}. \quad (26b)$$

Equations (26a), (26b) allow the generation of two distributions. The first one is the distribution of vertices in equivalence classes, according to their degrees, whereas the second one partitions the total adjacency into classes of vertex degree values. The unweighted information index for vertex degrees, which is based only on their equivalence but not the values of vertex degrees, *cannot* serve as a complexity index, in contrast to its weighted counterpart, obtained from the second distribution. This is illustrated in Fig. 4 by the comparison of the complexity of graphs 1, 2, and 3. Having two elementary cycles and higher vertex connectivity, graph 3 is topologically the most complex one, whereas the linear (path) graph 1 is the least complex one. However, the *unweighted* vertex degree distributions of graphs 1 and 3, constructed by considering only the equivalence of vertices with the same degree, are the same (one class of three and one class of two vertices), which results in equal information content of $I_{vd} = 4.85$ bits. Even worse is the complexity estimate of graph 2, all five vertices of which are equivalent, and its information content is zero, while a cyclic graph is considered more complex than an acyclic one. The *weighted* distributions take into account the different values of vertices in these three graphs, and the partitioning of their total adjacencies: $8\{2,2,2,1,1\}$, $10\{2,2,2,2,2\}$, and $12\{3,3,2,2,2\}$, respectively. The basic equation (22) for total weighted information content has thus been specified [27,33] for the vertex degree distribution as:

$$^w I_{v \text{ deg}} = A \log_2 A - \sum_i a_i \log_2 a_i. \quad (26c)$$



Information Theoretic Complexity Measures, Figure 4

Three graphs ordered according to their increasing topological complexity, as measured by two information theoretic complexity descriptors (Eqs. (26c), (26d)) based on the distribution of the vertex degree values. The latter are shown at each graph vertex

The information content of graphs 1–3 calculated from these distributions increases from the acyclic to the monocyclic to the bicyclic graph, and can be used as an approximate measure of graph complexity. Later study [61] has shown that the second term in Eq. (26c) is a more adequate measure of graph complexity. Denoted as $^wI_{vd}$, and calculated by Eq. (26d),

$$I_{vd} = \sum_{i=1}^V a_i \log_2 a_i ; \quad (26d)$$

its values are also shown in Fig. 4 and Fig. 7 (*vide infra*).

Different schemes have been applied in the attempts to make the unweighted information more sensitive to complex patterns in molecular structures. In the search for a better definition of vertex equivalence, Rashevsky [7] requested two equivalent vertices to have not only their first neighbors, but also their second, third, ..., k th neighbors to have the same degree. Trucco [8] has used the most rigorous definition of equivalence based on the orbits of the automorphisms group of the graph. (two vertices belong to the same graph orbit, if they can be interchange while preserving the adjacency). The vertex orbit information also fails to order correctly graphs 1–3, according to their complexity. Thus, graph 1 has three orbits and the vertex distribution is $5\{2,2,1\}$, in graph 2 all five vertices continue to belong the same equivalence class (orbit), and graph 3 has also three orbits like graph 1, and they are with the same cardinality, $5\{2,2,1\}$. Thus, the vertex orbit information, shows graphs 1 and 3 as equally complex, and graph 2 as the least complex, in contrast with the intuitive idea of topological complexity, which increases with the number of branches and cycles. Another way to a more sensitive description of molecular topology is to take into account structural elements of graphs that are more complex than

the vertex. Using the equivalence of edges, [9] offers only a slight improvement, which results from the larger number of edges exceeding that of vertices in polycyclic graphs. Basak [25,26] considered the symmetry of sets of vertex neighborhoods of 1-, 2-, etc. order.

Bertz [62] made use of the equivalency of subgraphs of two adjacent edges, which he called “connections”, and designed a complexity measure (termed later after his name) which has a better sensitivity to complexity features. Gordon and Kennedy [63] first used the number of such subgraphs as a measure of degree of branching of molecular skeleton. Before them, Platt [64] introduced a twice larger index obtained as the sum of all edge neighbors of each edge in the graph. In order to resolve the case with monocyclic graphs, which would always give zero equivalence-based information content no matter how large subgraphs are used, Bertz added a “size” term $n \log_2 n$ in his BI index:

$$BI = 2n \log_2 2n - \sum_{i=1}^k n_i \log_2 n_i \text{ bits} , \quad (27)$$

where n is the total number of two-edge subgraphs, k is the number of equivalence classes, and n_i is the number of such subgraphs in the i th equivalence class (connections orbits). In calculating the BI index for the three graphs, one finds for graph 1 that there are a couple of two-edge subgraphs with vertex degrees 1,2,2, and one two-edge subgraph with degrees 2,2,2, thus obtaining from Eq. (27) $BI(1) = 7.55$. Graph 2 contains five identical two-edge subgraphs with vertex degrees 2,2,2, which gives $BI(2) = n \log_2 n = 11.61 > BI(1)$. The most complex graph 3 contains nine such subgraphs: four of them with degrees 2,3,3, two with degrees 2,3,2, two with degrees 3,2,2, and one with degree 3,2,3, thus producing $BI(3) = 45.06$. If a *weighted* version were used for the

graph “connections”, Eq. (22) would suffice without any additional size term to produce the correct complexity ordering of these three graphs ($I_{\text{conn}}(1) = 16 \log_2 16 - 2 \times 5 \log_2 5 - 6 \log_2 6 = 9.29$, $I_{\text{conn}}(2) = 30 \log_2 30 - 5 \times 6 \log_2 6 = 69.66$, $I_{\text{conn}}(3) = 68 \log_2 68 - 5 \times 8 \log_2 8 - 4 \times 7 \log_2 7 = 215.34$).

Distance-Based Information Descriptors The first *weighted information theoretic indices* have been introduced by Bonchev and Trinajstić [34] in 1977 to characterize molecular branching [65,66,67,68,69,70,71] and, later, molecular cyclicity patterns [72,73,74,75] as basic components of molecular skeleton’s complexity. The weighted distribution of vertex distances has been used. Graph distances are integers equal to the number of edges along the shortest path between pairs of nodes. The total *graph distance*, $D(G)$, is calculated as the sum of all entries (distances d_{ij} between pairs of vertices i and j) of the distance matrix $\mathbf{D}(G)$. The latter is symmetric with respect to its main diagonal. For this reason, the sum of distances in molecular graphs is frequently presented by the Wiener number [65,66], $W(G) = D(G)/2$. The sum of distances d_i from a vertex i to all other graph vertices is calculated as the sum of distance matrix’ i th row entries d_{ij} :

$$d_i(G) = \sum_{j=1}^V d_{ij}; \quad (28a)$$

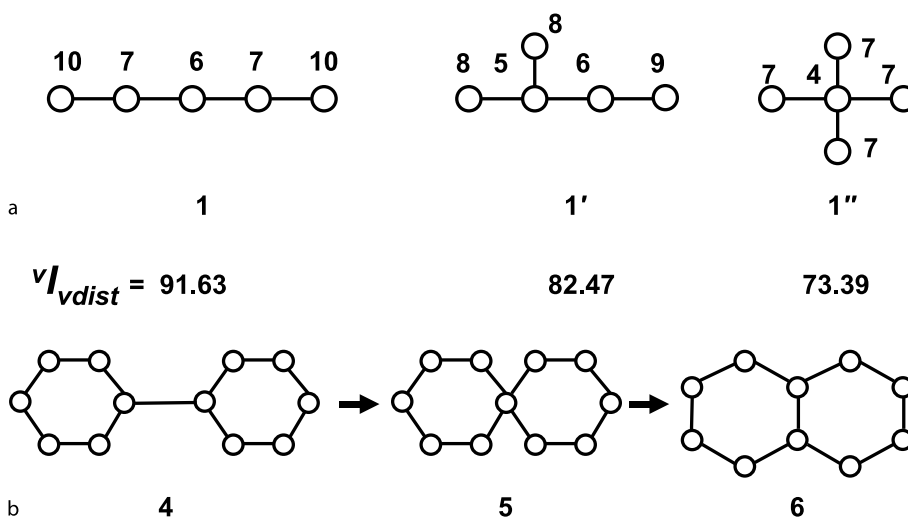
$$D(G) = \sum_{i=1}^V d_i = \sum_{i=1}^V \sum_{j=1}^V d_{ij}. \quad (28b)$$

where the sums run over all graph vertices V ($d_{ii} = 0$). The vertex distances thus calculated form a distribution $\{D\} \equiv \{d_{\text{max}}, d_{\text{max}} - 1, \dots, 2, 1\}$, which is used to define the *weighted information on graph vertex distances*, ${}^w I_{\text{vdist}}(G)$ (also termed *information on graph distance magnitude*, ${}^m I_{\text{vdist}}$):

$${}^w I_{\text{vdist}}(G) = D(G) \log_2 D(G) - \sum_{i=1}^{d(\text{max})} N_i d_i(G) \log_2 d_i(G), \quad (29)$$

$${}^w \bar{I}_{\text{vdist}}(G) = - \sum_{i=1}^{d(\text{max})} N_i \frac{d_i(G)}{D(G)} \log_2 \frac{d_i(G)}{D(G)}. \quad (30)$$

The ${}^w I_{\text{vdist}}$ indices increase with the number of vertices, and decrease with the number, size, and more central position of the graph branches and cycles. The two opposing trends prevent the usage of these indices as more general complexity measures. Yet, they are very useful in assessing (in a reverse order) the relative complexity of isomeric molecules or graphs having the same number of vertices. Such studies have been performed to define a number of branching [65,71] and cyclicity [72,73,74,75] complexity patterns. One of the branching patterns rigorously proved, shown in Fig. 5a, is defined as follows: “Branched trees are more complex than path graphs, and less complex than star graphs”. Here, the term *tree* is used for acyclic graphs, *path* is a tree without branches, and *star graphs* with V vertices have a single central point, and



Information Theoretic Complexity Measures, Figure 5

Patterns of increasing complexity in series of **a** acyclic, and **b** cyclic graphs, as measured in a reverse order by the values of the weighted information on vertex distance distribution Eq. (30). The vertex distances shown are used in the calculations

$N - 1$ branches of a unit length. Figure 5b presents a series of bicyclic structures 4–6 with complexity increasing with the stronger type of cycle connectivity, from a bridge, to a shared vertex, to a shared edge.

Centrality Information Descriptors Another topological complexity factor is *graph centrality*, the graph organization with respect to a certain central point, which can be defined in a variety of ways. Most of these have been introduced in network analysis, and will be discussed in Sect. “Information Content of Networks”. Here, the analysis is focused on distance-based centrality. The general definition for a graph center [59] states that the center is the vertex with the lowest *eccentricity* $e(i)$, eccentricity being defined as the longest distance from a given vertex to any other vertex in the graph. This definition frequently classifies as central several nonequivalent graph vertices. A hierarchical definition based on several criteria has been proposed, including Harary’s definition as the first criterion [76,77]. The second criterion reduces the set of central vertices to those, which have the lowest eccentricity and the smallest vertex distance, d_i . When several vertices have the same minimal eccentricity and the same minimal vertex distance, then the third criterion, requiring a minimal occurrence of the largest distance, $n_{ij}(\max)$, is used:

$$\text{Criterion 1: } e_i = \max(d_{ij}) = \min, \quad (31a)$$

$$\text{Criterion 2: } d_i = \sum_j d_{ij} = \min, \quad (31b)$$

$$\text{Criterion 3: } n_{ij}(\max) = \min. \quad (31c)$$

Hierarchical criteria 1–3 reduce considerably the number of nonequivalent graph centers, yet, in some cases some nonequivalent central vertices still exist. The final solution found [78] resolves effectively the problem for centric ordering of the vertex and edge orbits of the automorphism group of the graph on the basis of the iterative vertex–edge centricity (IVEC) concept:

Central are those vertices that are incident to the most central edges and, vice versa, central are those edges that are adjacent to the most central vertices.

A similar approach has been later applied in the Google search engine algorithm [79].

Once the graph center is defined, a vertex (as well as an edge) centric distribution can be constructed, proceeding from the distance from each vertex to the center. (In case of several equivalent central vertices, the distance to the closest one is used.) The vertex centric organization is

considered more complex when it is more compact, i. e., when vertices are organized in a smaller number of layers around the center or, at the same number of layers, when the outer layer(s) incorporate a smaller number of vertices. This complexity pattern matches the fact that the highly organized networks in living systems and technology are “small-world” ones [80,81], i. e., they have a rather small diameter.

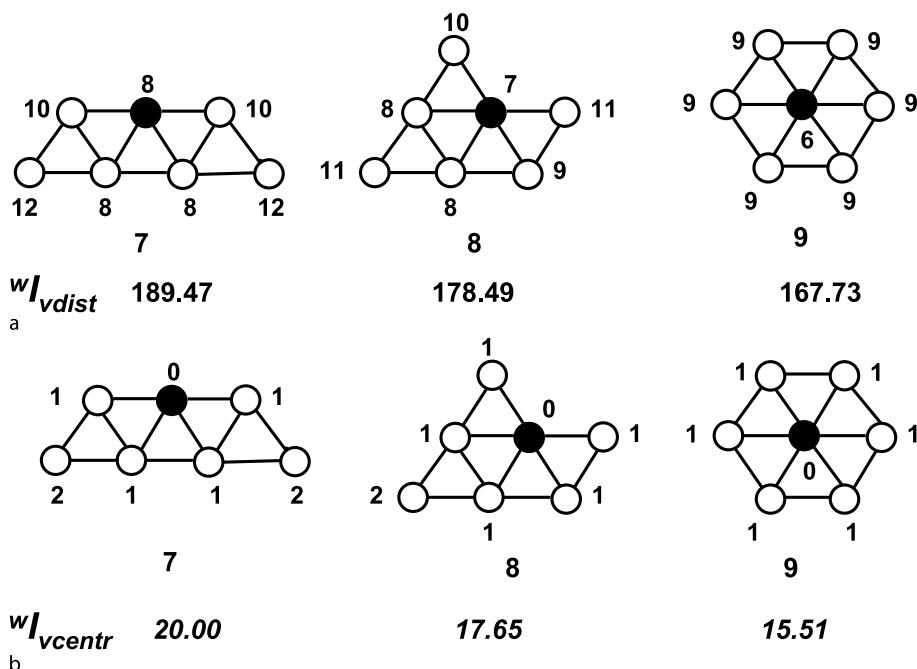
The centrality patterns of complexity are illustrated in Fig. 6 with three polycyclic graphs having seven vertices each. Their ordering from the least complex to the most complex central organization is inversely proportional to the values of their distance-based information descriptors, the weighted information for the vertex distances distribution, $^w I_{v\text{dist}}$, and the weighted information for the centric vertex distribution, $^w I_{v\text{centr}}$. Graph 7 has the most compact structure with eccentricity equal to one for the central vertex. The several central points in graphs 7 and 8 have all the same eccentricity equal to two. However, the smallest vertex distances criterion (Fig. 6a) allows one to reduce the centers to a single vertex in both graphs. The layers of centric ordering of graphs 7–9 can be traced from Fig. 6b, where vertices are labeled by their distances to the center. As seen, graph 9 is centrally the most complex having only one layer of centric neighborhood. Graphs 7 and 8 have both two layers of vertices around the center; however, graph 8 is considered more complex than graph 7, because of the smaller number of vertices in the outer layer (1 vs. 2).

Combined Information Complexity Descriptors Based on Vertex Degrees and Distances

It was shown in the foregoing that while the information measure of topological complexity based on vertex degree distribution increases in parallel with the increase in such complexity elements as branches and cycles, the measure based on the distribution of vertex distances decreases. It is logical to expect that the ratio A/D of the total adjacency and total distance of a graph could combine the two measures in a way to vary in direct proportionality of increasing complexity. A more sensitive complexity measure based on the same idea proceeds from calculating the sum of vertex ratios $b_i = a_i/d_i$, as well as from calculating the information on the distribution of these ratios, $I_{a/d}$. Thus, three complexity descriptors have been defined, and called Bourgas indices $B1$, $B2$, and $B3$, respectively [61,82,83]:

$$B1 = \frac{A}{D}, \quad (32a)$$

$$b_i = \frac{a_i}{d_i}, \quad (32b)$$



Information Theoretic Complexity Measures, Figure 6

a Three polycyclic graphs ordered according to their increasing centric complexity, and decreasing weighted information on the vertex distance distribution. The centers are presented by *full points*, and the numbers at each vertex are the vertex distances (the total distance from the vertex to all other vertices); **b** The same graph ordering corresponds also the decreasing weighted information on the distribution of distances from each vertex to the center (shown at each vertex)

$$B2 = \sum_{i=1}^V b_i = \sum_{i=1}^V \frac{a_i}{b_i}, \quad (32c)$$

$$B3 = B2 \log_2 B2 - \sum_{i=1}^V b_i \log_2 b_i. \quad (32d)$$

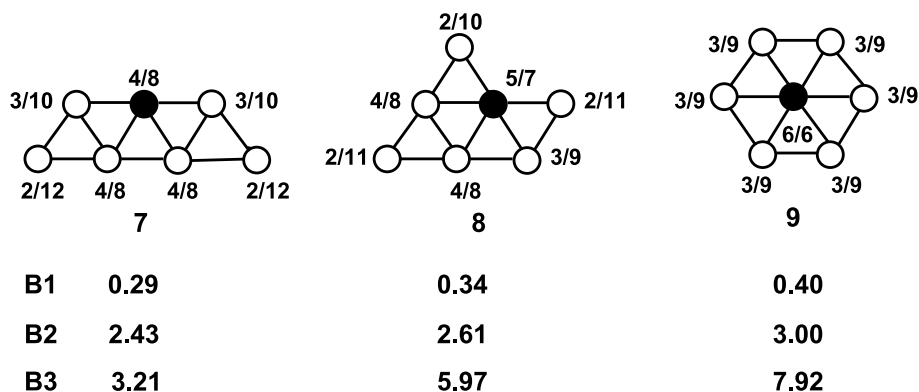
The Bourgas indices are shown in Fig. 7 to increase from graphs 7 to graph 8 to graph 9, in agreement with the complexity ordering of these graphs shown in Fig. 6.

Subgraph-Based Information Complexity Descriptors

The idea of using fragments of molecular graphs for a more complete representation of molecular structure emerged in the 1960s [84]. It has been developed in detail by Gordon and Kennedy [85], and has found a full realization in the molecular connectivity concept of Kier and Hall [86,87], widely applied in drug design. The design of subgraph-based complexity descriptors has been reported at the end of the 1990s [88,89,90,91,92,93,94,95,96,97,98,99,100], although an idea for such a complexity/similarity measure was proposed a decade earlier [101].

The *subgraph count*, SC, introduced initially under different names [88,89,90,91,92,93], proceeds from the logi-

cal assumption that the larger the number of subgraphs, the more complex the graph. This concept has been enriched by weighting each of the subgraphs with its total adjacency: *the larger the adjacency of the graph and its subgraphs, the higher the complexity*. The sum of the total adjacencies of all subgraph was termed *overall connectivity index*, OC (also initially named differently) [88,90,91,93,94]. Other weighting schemes, using the subgraph total distance and two simple functions of vertex degrees, have also been applied, and named as *overall Wiener index* [100] and *overall Zagreb indices* [102,103,104], respectively. Rucker and Rucker [96,97,98] used another graph invariant as a complexity measure, *the total walk count*, TWC, which is the total number of graph walks of all possible lengths l allowed by the graph size (A *walk* of length l on a graph is a sequence of vertices and edges such that edges can be traversed repeatedly). As in the case with subgraph count, *the larger the total walk count, the more complex the graph*. All these sophisticated measures of topological complexity have also been used to define an information-type of complexity indices [33]. For this purpose, they are considered to be partitioned into e classes, according to the number of edges e in the subgraph (eSC , eOC , eOW) and for TWC,



Information Theoretic Complexity Measures, Figure 7

Three polycyclic graphs in increasing complexity order, well matched by the three Bourgas complexity indices (Eqs. (32a), (32b), (32c), (32d)), combining vertex degrees and distances

into classes of walks of different length l , lTWC . All these cases can be summarized by the general equations:

$$X(G) = \sum_{e=1}^E {}^eX; \quad \{X\} \equiv \{{}^1X, {}^2X, \dots, {}^EX\}, \quad (33)$$

$$I_X(G) = X \log_2 X - \sum_{e=1}^E {}^eX \log_2 {}^eX. \quad (34)$$

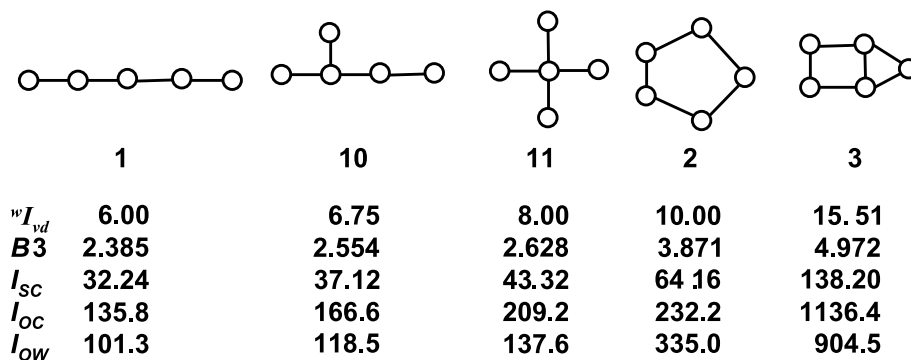
Here, the overall topological complexity descriptor is defined as $X \equiv SC, OC, OW$, and TWC ; E stands for the total number of graph edges and, for TWC , e and E are replaced by l and L_{\max} , respectively. The subgraph-based information descriptors are designed to increase with the increasing number of structurally complexifying elements like branches and cycles, and other, more subtle topological factors. In Fig. 8, they are compared to the information on vertex degree distribution ${}^wI_{vd}$ and the one for the

combined $B3$ descriptor of vertex degree and distance distribution introduced earlier in the text.

This concludes the analysis of information theoretic descriptors for the topological complexity of molecules. More information about topological complexity descriptors not using information theory can be found in another article of this volume ► [Topological Complexity of Molecules](#).

Information Content of Networks

Networks composed of biologically important molecules, such as DNA, RNA, proteins, and metabolites, turned in the beginning of the 21st century into a focus of major interest in computational biology and bioinformatics, as a universal language to describe the systems approach in biology. The links (edges) connecting the nodes



Information Theoretic Complexity Measures, Figure 8

The increase in topological complexity with the increase in the number of branches and cycles is precisely matched by the five information theoretic descriptors, based on vertex degrees, combination of vertex degrees and distances, subgraph count, overall connectivity, and overall Wiener subgraph distributions

(vertices) of these networks can represent a variety of interactions, from physical and chemical interaction (protein-protein interaction; posttranslational protein modification) to regulation (gene-gene, protein-gene) to co-expression (gene-gene), molecular transport (proteins), and others. The topological characterization of these dynamic evolutionary networks [61,83,105,106] proceeds from the same basic features as in characterizing molecules (which are in fact atomic networks): connectivity, distances, and subgraphs. There are, however, considerable differences in the distributions of the underlying graph invariants. While vertex degrees in molecular graphs rarely exceed the value of four, in biological networks the maximal node degree may be very high. Such highly connected nodes, termed *hubs*, are essential for the existence of the living cell; their deletion or malfunctioning is usually lethal. The distribution of vertex degrees in the complex networks in nature and technology seems to obey a scale-free law [107,108], as a consequence of which there are few hubs and many low-connected nodes. Another important topological feature of these networks is their very small diameter; the networks are “small-world” type [80]. This very compact organization is the key to life’s resilience against attacks of different kinds. A signature of a complex network is its specific composition of subgraphs, called *motifs*, which are much more abundant than in random networks of the same size [109]. Larger subgraphs, termed *complexes*, *modules* and *pathways*, are responsible for the performance of specific biological functions.

The *compositional complexity* of biochemical networks is determined similarly to molecular one. One may consider, for example, a protein-protein interaction network as composed of protein complexes, which in turn include

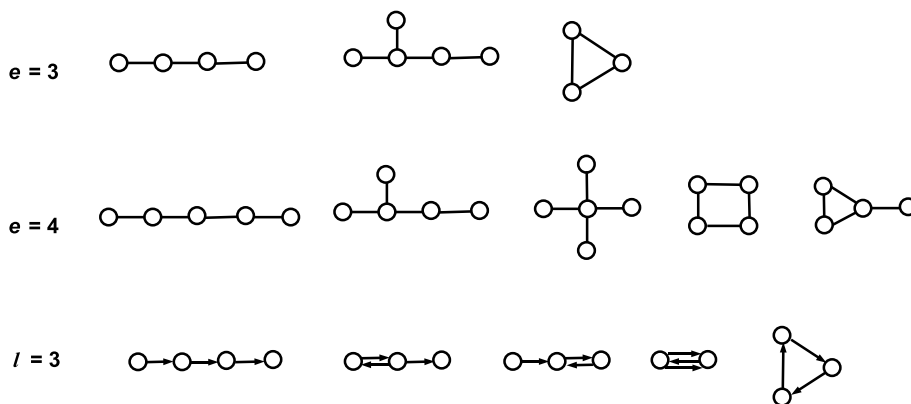
individual proteins. Similarly, a metabolic network may be regarded as being built of metabolic pathways containing individual metabolites and enzymes. The larger the diversity of network’s constituents, the higher the network compositional complexity. Equations (3), (4) are of use for calculation of the compositional information content, as used for molecules.

Networks *topological complexity* also makes use of the tools introduced for assessing the topological complexity of molecules. The information descriptor of vertex degree distribution, ${}^w I_{vd}$, and the Bourgas information index $B3$ are applied directly as defined by Eqs. (26d) and (32d), respectively, proceeding from the network connectivity and distance data. Some adaptation is needed to apply the information descriptors for network subgraphs. Due to the size of intracellular networks, which may exceed in some cases 10,000 nodes, the calculation of the subgraph-based complexity descriptors would lead to combinatorial explosion. However, it suffices for sufficiently accurate assessments of network complexity to use the distribution of small subgraphs having only three or four edges or the distribution of walks of length 3 (Fig. 9).

Equation (34) has to be slightly modified to include summing up not over all subgraphs with e edges or all walk lengths l , but at a given size (e or l) the sum has to be taken over all types i of subgraphs or walks, as illustrated in Fig. 9:

$${}^{e(l)} I_X(G) = {}^{e(l)} X \log_2 {}^{e(l)} X - \sum_i {}^{e(l)} X_i \log_2 {}^{e(l)} X_i. \quad (35)$$

Here, ${}^{e(l)} X = \sum_i {}^{e(l)} X_i$, and X can be each one of subgraph count (SC), overall connectivity (OC), overall Wiener (OW), and total walk count (TWC), as defined



Information Theoretic Complexity Measures, Figure 9

All subgraphs with 3 and 4 edges, and all walks of length 3, recommended for assessing complexity of large size networks by Eq. (35)

in Subsect. "Subgraph-Based Information Complexity Descriptors".

The five information descriptors of network complexity shown in Fig. 8 can be applied to both undirected and directed networks. The specifics of directed graphs, however, require replacing the single-value vertex degree by in-degree and out-degree, which are defined as the total number of directed edges incoming to the vertex and outgoing from it, respectively. The *B3* descriptor, combining vertex degrees and distances, cannot be directly applied for directed graphs, because there is no path between some pairs of vertices and the distance between such vertices is equal to infinity. A procedure has been proposed to recalculate the total distance in directed networks by taking into account the vertex accessibility [61,110].

Future Directions

The use of information theory in characterizing molecular structure and composition has been strongly influenced by the pioneering work of mathematical biologists in the 1950s [2]. In turn, the considerable expertise accumulated in chemical information theory during the last 30 years is providing nowadays its feedback to bioinformatics and computational biology by offering not only a variety of specific information descriptors for topological complexity, but also the much more sensitive weighted version of Shannon's basic equations. Along with the other developed graph theoretic complexity measures, the information complexity indices are expected to find a broad application for large-scale comparative and evolutionary studies in biology. Finding the most adequate quantitative measure of evolution is a great challenge, along with elucidating the minimum compositional, topological and functional complexity needed for the emergence of life. Practical results may be expected soon from network topology/biological function quantitative studies in the area of biomedical research, in which the information measures of network complexity will play an important role. Theoretical chemistry is also expected to benefit greatly from the exchange of ideas with network analysis in biology. The network modular and clique structure could be of interest for studies in atomic clusters, nanotechnology and crystallography. The concept of centrality, which has been very broadly developed in bioinformatics, might prove of importance to some areas of organic chemistry. The recent application of the bipartivity measure of networks [111] to the stability of fullerenes [112] and branching of molecular skeletons [113] are only among the first signs of such future developments in theoretical chemistry.

Bibliography

1. Shannon C, Weaver W (1949) Mathematical theory of communications. University of Illinois Press, Urbana
2. Quastler H (ed) (1953) Essays on the use of information theory in biology. University of Illinois Press, Urbana
3. Boltzmann L (1866) Über die mechanische Bedeutung der zweiten Hauptsatzes der Wärmetheorie. Wien Ber 53:195
4. Brillouin L (1956) Science and information theory. Academic Press, New York
5. Dancoff SM, Quastler H (1953) The information content and error rate of living things. In: Quastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
6. Branson HR (1953) Information theory and the structure of proteins. In: Quastler H (ed) Essays on the use of information theory in biology. University of Illinois Press, Urbana
7. Rashevsky N (1955) Life, information theory, and topology. Bull Math Biophys 17:229–235
8. Trucco E (1956) A note on the information content of graphs. Bull Math Biophys 18:129–135
9. Trucco E (1956) On the information content of graphs: Compound symbols; Different states for each point. Bull Math Biophys 18:237–253
10. Mowshowitz A (1968) Entropy and the complexity of graphs. I. An index of the relative complexity of a graph. Bull Math Biophys 30:175–204
11. Bonchev D (1979) Information indices for atoms and molecules. Commun Math Comput Chem (MATCH) 7:65–113
12. Bonchev D (2006) Periodicity of chemical elements and nuclides. Information theoretic analysis. In: Rouvray DH, King RB (eds) Mathematics of the periodic table. Nova Science, New York, pp 161–188
13. Bonchev D, Peev T (1973) Information theoretic study of chemical elements. Mean information content of a chemical element. God Vissh Khim–Technol Inst Bourgas 10:561–574
14. Bonchev D, Kamenska V, Kamenski D (1977) Informationsgehalt chemischer Elemente. Monatsh Chem 108:477–487
15. Bonchev D, Kamenska V (1978) Information theory in describing the electronic structure of atoms. Croat Chem Acta 51:19–27
16. Dimov D, Bonchev D (1976) Spin-information equations of the groups and periods in the periodic table of chemical elements. Commun Math Comput Chem (MATCH) 2:111–115
17. Bonchev D, Kamenska V, Tashkova C (1976) Equations for the elements in the periodic table based on information theory. Commun Math Comput Chem (MATCH) 2:117–122
18. Bonchev D (1981) Information theory interpretation of the Pauli principle and Hund rule. Intern J Quantum Chem 19:673–679
19. Bonchev D, Kamenska V (1981) Predicting the properties of the 113–120 transactinide elements. J Phys Chem 85:1177–1186
20. Bonchev D, Peev T, Rousseva B (1976) Information study of atomic nuclei. Information on proton–neutron composition. Commun Math Comput Chem (MATCH) 2:123–137
21. Rousseva B, Bonchev D (1978) A theoretic-information variant of nuclide systematics. Commun Math Comput Chem (MATCH) 4:173–192
22. Rousseva B, Bonchev D (1980) Prediction of the nuclear bind-

- ing energies of the nuclides of period VII. *Radiochem Radioanal Lett* 45:341–346
23. Augenstine L (1953) Remarks on Pauling's protein models. In: Kastler H (ed) *Essays on the use of information theory in biology*. University of Illinois Press, Urbana
 24. Linshitz H (1953) The information content of a bacterial cell. In: Kastler H (ed) *Essays on the use of information theory in biology*. University of Illinois Press, Urbana
 25. Magnusson VR, Harris DK, Basac SC (1983) Topological indices based on neighborhood symmetry. In: King RB (ed) *Chemical applications of topology and graph theory*. Elsevier, Amsterdam, pp 178–191
 26. Basak SC (1999) Information theoretic indices of neighborhood complexity and their applications. In: Devillers J, Balaban AT (eds) *Topological indices and related descriptors in QSAR and QSPR*. Gordon and Breach Science Publishers, Chichester, pp 563–593
 27. Bonchev D (1983) *Information-theoretic indices for characterization of chemical structures*. Research Studies Press, Chichester
 28. Morovitz H (1955) Some order-disorder considerations in living systems. *Bull Math Biophys* 17:81–86
 29. Mowshowitz A (1968) Entropy and the complexity of graphs: IV. Entropy measures and graphical structure. *Bull Math Biophys* 30:533–546
 30. Balaban AT (ed) (1976) *Chemical applications of graph theory*. Academic Press, London
 31. Trinajstić N (1983) *Chemical graph theory*. CRC Press, Boca Raton
 32. Szostak JW (2003) Functional information: Molecular messages. *Nature* 423:689
 33. Bonchev D (2003) Shannon's information and complexity. In: Bonchev D, Rouvray DH (eds) *Complexity in chemistry*. Mathematical Chemistry Series, vol 7. Taylor and Francis, Boca Raton, pp 155–187
 34. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. *J Chem Phys* 67:4517–4533; (1982) Chemical information theory. Structural aspects. *Intern J Quantum Chem Symp* 16:463–480
 35. Yockey HP (1977) On the information content of cytochrome C. *J Theor Biol* 67:345–376
 36. Yockey HP (1992) *Information theory and molecular biology*. University Press, Cambridge
 37. Eigen M (1992) *Steps toward life*. Oxford University Press, Oxford
 38. Kauffman SA (1992) *Applied molecular evolution*. *J Theor Biol* 157:1–7
 39. Kauffman SA (1993) *The origins of order*. Oxford University Press, New York
 40. Volkenstein MV (1994) *Physical approaches to biological evolution*. Springer, Berlin
 41. Adami C (2004) Information theory in molecular biology. *Phys Life Rev* 1:3–22
 42. Hasegawa M, Yano T-A (1975) Entropy of the genetic information and evolution. *Orig Life Evol Biosph* 6:219–227
 43. Strait BJ, Dewey TG (1996) The Shannon information entropy of protein sequences. *Biophys J* 71:148–155
 44. Gatlin LL (1972) *Information theory and the living system*. Columbia University Press, New York
 45. Li W (1997) The complexity of DNA. *Complexity* 3:33–37
 46. Schneider TD (1997) Information content of individual genetic sequences. *J Theor Biol* 189:427–441
 47. Schneider TD (2000) *Nucl Acids Res* 28:2794–2799
 48. Adami C (2002) What is complexity? *BioEssays* 24:1085–1094
 49. Lin J (1991) Divergence measures based on the Shannon entropy. *IEEE Trans Inf Theor* 37:145–151
 50. Bernaola-Galvan P, Roman-Roldan R, Oliver J (1996) Compositional segmentation and long-range fractal correlations in DNA sequences. *Phys Rev E* 53:5181–5189
 51. Kolmogorov A (1965) Three approaches to the quantitative definition of information. *Probl Inf Transm* 1:4
 52. Lempel A, Ziv J (1976) On the complexity of finite sequences. *IEEE Trans Inf Theory* IT-22:75–81
 53. Gusev VD, Kulichkov VA, Chupahina OM (1991) Complexity analysis of genomes. I. Complexity and classification methods of detected structural regularities. *Mol Biol (Mosk)* 25:825–834
 54. Chen X, Kwong S, Li MA (1999) Compression algorithm for DNA sequences and its applications to genome comparison. *Genome Inform Ser Workshop Genome Inform* 10:51–61
 55. Ebeling W, Jimenez-Montano MA (1980) On grammars, complexity, and information measures of biological macromolecules. *Math Biosci* 52:53–71
 56. Grumbach S, Tahi F (1994) A new challenge for compression algorithms: genetic sequences. *J Inf Process Manag* 30:875–886
 57. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng C-K, Simons M, Stanley HE (1994) Linguistic features of noncoding DNA sequences. *Phys Rev Lett* 73:3169–3172
 58. Grosse I, Herzel H, Buldyrev SV, Stanley HE (2000) Species independence of mutual information in coding and noncoding DNA. *Phys Rev E* 61:5624–5629
 59. Harary F (1971) *Graph theory*, 2nd printing. Addison-Wesley, Reading
 60. Gross JK, Yellen J (eds) (2004) *Handbook of graph theory*. CRC Press, Boca Raton
 61. Bonchev D, Buck GA (2005) Quantitative measures of network complexity. In: Bonchev D, Rouvray DH (eds) *Complexity in chemistry, biology and ecology*. Springer, New York, pp 191–235
 62. Bertz S (1981) The first general index of molecular complexity. *J Am Chem Soc* 103:3599–3601
 63. Gordon M, Scantlebury GR (1964) Non-random polycondensation: Statistical theory of the substitution effect. *Trans Faraday Soc* 60:604–621
 64. Platt JR (1952) Prediction of isomeric differences in paraffin properties. *J Phys Chem* 56:328–336
 65. Wiener H (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69:17–20
 66. Wiener H (1948) Relation of the physical properties of the isomeric alkanes to molecular structure. *J Phys Chem* 52:1082–1089
 67. Hosoya H (1971) Topological index: A newly proposed quantity characterizing the topological nature of structured isomers of saturated hydrocarbons. *Bull Chem Soc Jpn* 44:2332–2339
 68. Lovasz L, Pelikan J (1973) On the eigenvalues of trees. *Period Math Hung* 3:175–182
 69. Gutman I, Rušćić B, Trinajstić N, Wilcox CW Jr (1975) Graph theory and molecular orbitals. 12. Acyclic polyenes. *J Chem Phys* 62:3399–3405

70. Randić M (1975) On characterization of molecular branching. *J Am Chem Soc* 97:6609–6615
71. Bonchev D (1995) Topological order in molecules 1. Molecular branching revisited. *J Mol Struct (Theochem)* 336:137–156
72. Bonchev D, Mekenyan O, Trinajstić N (1980) Topological characterization of cyclic structures. *Intern J Quantum Chem* 17:845–893
73. Mekenyan O, Bonchev D, Trinajstić N (1979) Topological rules for spiro-compounds. *Commun Math Comput Chem (MATCH)* 6:93–115
74. Mekenyan O, Bonchev D, Trinajstić N (1981) On algebraic characterization of bridged polycyclic compounds. *Intern J Quantum Chem* 19:929–955
75. Mekenyan O, Bonchev D, Trinajstić N (1981) A topological characterization of cyclic structures with acyclic branches. *Commun Math Comput Chem (MATCH)* 11:145–168
76. Bonchev D, Balaban AT, Mekenyan O (1980) Generalization of the graph center concept, and derived topological indexes. *J Chem Inf Comput Sci* 20:106–113
77. Bonchev D (1989) The concept for the center of a chemical structure and its applications. *Theochem* 185:155–168
78. Bonchev D, Mekenyan O, Balaban AT (1989) An iterative procedure for the generalized graph center in polycyclic graphs. *J Chem Inf Comput Sci* 29:91–97
79. Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web search engine. *Proceedings of the seventh international conference on World Wide Web, Brisbane, April 1998*, pp 107–117
80. Watts DJ, Strogatz SH (1998) Collective dynamics of “small-world” networks. *Nature* 393:440–442
81. Newman MEJ, Strogatz SH, Watts DJ (2001) Random graphs with arbitrary degree distributions and their applications. *Phys Rev E* 64:026118
82. Bonchev D (2005) My life-long journey in mathematical chemistry. *Intern Electr J Mol Design* 4:434–490
83. Bonchev D, Buck G (2007) From molecular to biological structure and back. *J Chem Inform Model* 7:909–917
84. Smolenski EA (1964) Graph-theory application to the calculations of structural-additive properties of hydrocarbons. *Zh Fiz Khim* 38:1288–1291
85. Gordon M, Kennedy JW (1973) The graph-like state of matter. Part 2. LCGI schemes for the thermodynamics of alkanes and the theory of inductive inference. *J Chem Soc Faraday Trans* 2 69:484–504
86. Kier LB, Hall LH (1976) *Molecular connectivity in chemistry and drug research*. Academic Press, New York
87. Kier LB, Hall LH (1986) *Molecular connectivity in structure-activity analysis*. Research Studies Press, Letchworth
88. Bonchev D (1995) Kolmogorov’s information, Shannon’s entropy, and topological complexity of molecules. *Bulg Chem Commun* 28:567–582
89. Bertz SH, Sommer TJ (1997) Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices. *J Chem Soc, Chem Commun* 2409–2410
90. Bonchev D (1997) Novel indices for the topological complexity of molecules. *SAR QSAR Environ Res* 7:23–43
91. Bonchev D, Seitz WA (1997) The concept of complexity in chemistry. In: Rouvray DH (ed) *Concepts in chemistry: a contemporary challenge*. Wiley, New York, pp 353–381
92. Bertz SH, Wright WF (1998) The graph theory approach to synthetic analysis: definition and application of molecular complexity and synthetic complexity. *Graph Theory Notes New York Acad Sci* 35:32–48
93. Bonchev D (1999) Overall connectivity and molecular complexity. In: Devillers J, Balaban AT (eds) *Topological indices and related descriptors*. Gordon and Breach, Reading, pp 361–401
94. Bonchev D (2000) Overall connectivities/topological complexities: a new powerful tool for QSPR/QSAR. *J Chem Inf Comput Sci* 40:934–941
95. Nikolić S, Tolić M, Trinajstić N, Baučić I (2000) On the Zagreb indices as complexity indices. *Croat Chem Acta* 73:909–921
96. Rücker G, Rücker C (2000) Automatic enumeration of molecular substructures. *MATCH – Commun Math Comput Chem* 41:145–149
97. Rücker G, Rücker C (2000) Walk count, labyrinthicity and complexity of acyclic and cyclic graphs and molecules. *J Chem Inf Comput Sci* 40:99–106
98. Rücker G, Rücker C (2001) Substructure, subgraph and walk counts as measures of the complexity of graphs and molecules. *J Chem Inf Comput Sci* 41:1457–1462
99. Bonchev D (2001) Overall connectivity – a next generation molecular connectivity. *J Mol Graph Model* 5271:1–11
100. Bonchev D (2001) The overall Wiener index – a new tool for characterization of molecular topology. *J Chem Inf Comput Sci* 41:582–592
101. Bertz SH, Herndon WC (1986) The similarity of graphs and molecules. In: Pierce TH, Hohne BA (eds) *Artificial intelligence applications to chemistry*. ACS, Washington, pp 169–175
102. Bonchev D, Trinajstić N (2001) Overall molecular descriptors. 3. Overall Zagreb indices. *SAR QSAR Environ Res* 12:213–235
103. Nikolić S, Trinajstić N, Tolić M, Rücker G, Rücker C (2003) On molecular complexity indices. In: Bonchev D, Rouvray DH (eds) *Complexity in chemistry*. Mathematical chemistry series, vol 7. Taylor and Francis, London, pp 29–89
104. Bonchev D (2005) The overall topological complexity indices. *Lecture series on computer and computational sciences*, vol 4. Koninklijke Brill NV, Leiden, pp 1554–1557
105. Bonchev D (2003) Complexity of protein-protein interaction networks, complexes and pathways. In: Conn M (ed) *Handbook of proteomics methods*. Humana, New York, pp 451–462
106. Bonchev D (2004) Complexity analysis of yeast proteome network. *Chem Biodivers* 1:312–326
107. Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512
108. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi A-L (2000) The large-scale organization of metabolic networks. *Nature* 407:651–654
109. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D et al (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824–827
110. Bonchev D (2003) On the complexity of directed biological networks. *SAR QSAR Environ Res* 14:199–214
111. Estrada E, Rodríguez-Velázquez JA (2005) Spectral measures of bipartivity in complex networks. *Phys Rev E* 72:055510
112. Došlić T (2005) Bipartivity of fullerene graphs and fullerene stability. *Chem Phys Lett* 412:336–340
113. Estrada E, Rodríguez-Velázquez JA, Randić M (2006) Atomic branching in molecules. *Int J Quantum Chem* 106:823–832

Infrasound from Earthquakes, Tsunamis and Volcanoes

MILTON GARCES¹, ALEXIS LE PICHON²

¹ Infrasound Laboratory, HIGP, SOEST, University of Hawaii, Manoa, Kailua-Kona, USA

² CEA/DASE/LD, Bruyères-le-Châtel, France

Article Outline

Glossary

Definition of the Subject

Introduction

Infrasound Arrays

Earthquake Infrasound

Tsunami Infrasound

Volcano Infrasound

Future Directions

Concluding Remarks

Acknowledgments

Bibliography

Glossary

Infrasound atmospheric sound waves with frequencies lower than the 20 Hz hearing threshold of the human ear.

Infrasound array four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range.

Trace velocity apparent horizontal phase velocity of an acoustic arrival measured by an array.

Celerity effective propagation speed of a signal, measured from the ratio of the total range over the total travel time along the great circle path from a source to a receiver.

Tremor volcanic signal consisting of a nearly continuous oscillation of the ground and atmosphere, with durations of minutes to years. Harmonic tremor may have multiple distinct spectral peaks.

LP Long Period event. Transient volcanic signal with durations of tens of seconds to minutes and distinct spectral peaks.

Definition of the Subject

Infrasound may be radiated by earthquakes, tsunamis, and volcanoes through the displacement or rupture of Earth's surface and the subsequent flow and excitation of fluids. These complex and sometimes cataclysmic phenomena share some common physics, yet have different ways of converting energy into atmospheric sound. Sig-

nals from earthquakes, tsunamis, and volcanoes captured by the present generation of infrasound arrays are introduced in this chapter through case studies. Contemporary methods used in the analysis, interpretation, and modeling of these diverse signatures are discussed and some of the associated geophysical problems that remain unsolved are considered.

Introduction

The human ear may perceive sound in the frequency band of 20 to 20,000 cycles per second (Hz). Infrasound consists of acoustic waves in the atmosphere with frequencies lower than the 20 Hz hearing threshold of the human ear. Because of reduced acoustic attenuation and scattering in the atmosphere at long infrasonic wavelengths and the large spatial scales of the physical processes driving earthquakes, tsunamis, and volcanoes, the infrasound frequency band is well suited to the remote monitoring of these events. The ambient infrasound field at any location is rich and diverse, with sources originating from the solid Earth, the ocean, the atmosphere, space-born objects, and human activity [4,39]. These acoustic pressure waves co-exist with non-acoustic atmospheric pressure fluctuations associated with meteorological changes such as wind and frontal passages (e.g. [9]). Sound propagation paths are controlled primarily by the temperature and wind stratification in the lower, middle, and upper atmosphere [16,19,24]. The effective sound velocity of the atmosphere at a given height above the ground may be approximated as the sum of the scalar sound speed, which is proportional to the square root of temperature, and the vector wind velocity, which typically may reach magnitudes of 15–20% of the sound speed in the upper atmosphere. An acoustic waveguide may efficiently direct sound to ground-based stations, and is defined by a high sound velocity layer at the upper boundary and a lower sound velocity layer near the ground. The high temperature of the mesosphere and lower thermosphere (MLT) would always refract infrasound back to the ground, but severe attenuation above ~110 km can suppress thermospheric returns. Waveguides in the troposphere and stratosphere are expected to only transmit primarily along the downwind direction. However, observations suggest that the elevated acoustic waveguide formed by the low temperature zone in the stratosphere may routinely leak energy back to the ground through scattering and diffraction [27,28,29]. Ground cooling and low altitude winds may also produce a stable waveguide in the boundary layer near the ground surface [20,42]. A new generation of global atmospheric specifications designed

for the study of infrasound propagation has been developed by integrating multiple meteorological and upper atmosphere models [19]. Validation and further refinement of these atmospheric models is ongoing [3,43].

Infrasound Arrays

Background

Modern infrasound array technology emerged at the turn of the 21st century after the 1996 adoption of the Comprehensive Nuclear-Test-Ban Treaty and the subsequent growth of the International Monitoring System (IMS), which was designed for the detection of clandestine nuclear test explosions [55,61]. An infrasound array generally consists of four or more horizontally separated identical microphones or microbarometers with precisely known locations that optimize the reception of a specified wavelength range. Wind is the most pernicious source of incoherent noise, and infrasound's greatest vulnerability. An array that is sheltered from the wind by forest cover, snow cover, or topographical blocking will have a low noise floor, and provide high sensitivity and robust measurements. If a wind-sheltered site is not found, wavelength-specific wind noise reducing filters have to be designed for the boundary layer conditions at the array site [1,40]. IMS-type infrasound arrays have a maximum distance between sensors of 1–3 km and use sensors with over 100 dB of dynamic range and a flat frequency response between 0.02 and 20 Hz. Portable arrays often have apertures of ~ 100 m or less, and are thus optimized for smaller wavelengths (frequencies >1 Hz). Infrasonic sensors for portable array applications often operate within the 0.1–100 Hz frequency band, and may overlap into the audio band. Infrasound data are typically recorded with GPS time-stamped 24-bit digitizers and are often sent via digital communications to a central processing facility for real time analysis.

Basic Principles

A number of calibrated microphones, precisely timed and arranged in optimal spatial configurations as arrays, present the best design for recording infrasound. A wavefront is the spatial surface containing sound emitted from a source at the same time. An array relies on the principle that sound along a (possibly curved) wavefront has recognizable features, so that as the wavefront passes through the multiple sensors in an array it is possible to determine the time of passage of a specific waveform feature (for example, the peak of an explosive pulse). From the time of arrival of a waveform feature at each known sensor location, the direction of propagation of the incident wavefront as

well as its apparent propagation speed across the array can be inferred. Once the arrival direction and speed are determined, it is possible to digitally apply time-delays or phase shifts to temporally align all the microphone waveforms along a beam of energy (beamform) to improve the ratio of the signal amplitude to the ambient noise. Based on these fundamental principles, more advanced contemporary techniques permit the extraction of waveform features with a very small signal to noise ratio [12,65].

Array Design

If two identical microphones spaced some distance from each other record identical waveforms, the two observed signals are perfectly coherent. As the microphone spacing increases beyond the characteristic wavelength of a pulse, the coherence decreases until the waveforms no longer resemble each other. Microphone arrays are designed to detect coherent signals within a specific wavelength range, and the ability of an array to accurately determine the speed and angle of arrival of coherent sound signals depends largely on the sensor distribution relative to the wavelengths of interest. A minimum of three sensors, deployed as an L, are required to discriminate incidence angle. However, a three element array has a very poor angular resolution and leaves no margin for error, as failure of a single sensor will substantially degrade its detection capability. Four or more sensors are preferable, yielding a broader frequency response and better measurement precision.

Detection, Location, and Signal Identification

Infrasonic signals measured a few kilometers from the source can vary from powerful explosions ($>10^2$ Pa) that dominate the recorded time series, to a background rumble (10^{-3} Pa) buried within the ambient sound field. A single infrasonic array can discriminate between a coherent signal arriving from the direction and height of the target source and a competing signal coming from a different angle and elevation [10]. Thus, arrays can 1) separate the coherent sound field from the incoherent ambient noise field, 2) identify and extract a specific infrasonic signal within the coherent ambient sound field (clutter) and 3) separate acoustic arrivals propagating through different waveguides yet originating from a signal source which may be stationary or moving. If two arrays are available, the target signal would be recorded at each array with an arrival angle pointing to the source and an arrival time consistent with the propagation path from the source to the receiver. The two array beams would intersect at the source, thus providing a geographic loca-

tion which would be confirmed and refined with the signal travel time information [25]. Thus two properly sited infrasonic arrays can unambiguously locate a source region and discriminate between sources. By optimizing detection with arrays, performing locations with two or more arrays, and characterizing preexisting sources of clutter, it is possible to acoustically recognize a source with high confidence. If other detection (such as seismic or imaging) and identification [36]) technologies are available, uncertainties in source location and identification drop substantially. This process of signal detection, location, and identification is routinely used successfully by the international community in the monitoring of natural and man-made events [27].

Acoustic Speed and Velocity

Infrasound studies may refer to the acoustic phase velocity, group velocity, effective sound speed, trace velocity, and the celerity of an acoustic arrival. The first three quantities depend on the sound speed and wind speed in the atmosphere along the source-receiver path. The trace velocity of a signal may be measured directly by observing the apparent horizontal speed and direction of propagation of a signal across an array. The celerity of an arrival is defined as the ratio of range over the travel time, and may be conceived as the effective, or average propagation speed over the complete propagation path. Given a known source-receiver configuration, the celerity may be computed directly. Although it is tempting to use the trace velocity for identifying a given arrival, measurement and calibration uncertainties can make this procedure rather inaccurate for distant sources. Celerity estimates from various infrasonic events with known locations suggests that 1) our knowledge of the atmosphere is presently insufficient to reliably predict all infrasonic arrivals to a station under all atmospheric conditions, and, 2) diffraction and/or scattering can feed acoustic energy to waveguides that are elevated above the ground, and these elevated waveguides may also leak sound back to the ground [10].

Analysis Method and Scope

Array processing methods allow us to extract coherent signals from the incoherent ambient sound field. One efficient and popular technique for estimating the infrasonic wave parameters is the *Progressive Multi-Channel Correlation* method (PMCC) [12]. This method, originally designed for seismic arrays, is well adapted for analyzing low-amplitude coherent waves within incoherent noise and efficient for differentiating signals of interest from background clutter [22,26]. The PMCC method was

used for array processing of the signals from earthquakes, tsunamis, and volcanoes presented in this chapter.

Acoustic-gravity waves form a class of propagating atmospheric pressure signals which were studied in detail during the early Megaton-yield atmospheric nuclear tests (e. g. [32]). These waves are affected by buoyancy, have periods longer than 50 s, and have unique source and propagation characteristics that are beyond the scope of this paper. The IMS and portable arrays discussed in this chapter are tuned to higher frequencies, so are generally not designed to process acoustic gravity waves with high precision. The reader may refer to [33] for an excellent introduction to gravity and acoustic-gravity waves.

This chapter will also omit the acoustics of structural collapses such as pyroclastics flows from volcanoes, avalanches, landslides, or rockfalls. Although such signals may portend hazardous events (e. g. [79]), this family of acoustic signals are poorly understood and even less well modeled.

Earthquake Infrasound

Ground vibrations can produce sound in manifold ways. Relative to the atmosphere, earthquakes can act as distributed, supersonic acoustic sources moving at seismic or fault rupture velocities. When seismic surface waves travel through mountainous regions, the predominant source of infrasound is attributed to the re-radiation of pressure waves by topography [48,60,78]. In this case, the earthquake-induced displacement perpendicular to the ground surface can locally generate ground-coupled air waves. The local conversion from seismic waves to the sound pressure has been observed on microbarometers at regional and teleseismic distances [14,15,18,69]. Seismic to acoustic wave coupling at the ground-air interface will be enhanced when the horizontal phase velocity, or trace velocity, of the infrasonic waves and the seismic waves are matched. This type of seismoacoustic coupling can be particularly efficient in sediments and loosely consolidated volcanic environments with a low shear wave velocity [24].

The generation of infrasonic waves from the epicenter region has also been postulated [59,62]. At large infrasonic periods (50–100 s) acoustic-gravity waves from the sudden strong vertical ground displacements have been detected at distances of thousands kilometers from the origin [6,57]. This mechanism would also apply for large submarine earthquakes [30,58]. In all the aforementioned cases, an actual pressure wave is radiated from the ground into the atmosphere. This process is distinguished from microphonics, where the recorded signal is due to the sensor response to accelerations and ground level elevations

independently of any pressure changes occurring in the atmosphere [2,4,45]. In this case, the microphone acts as a seismometer.

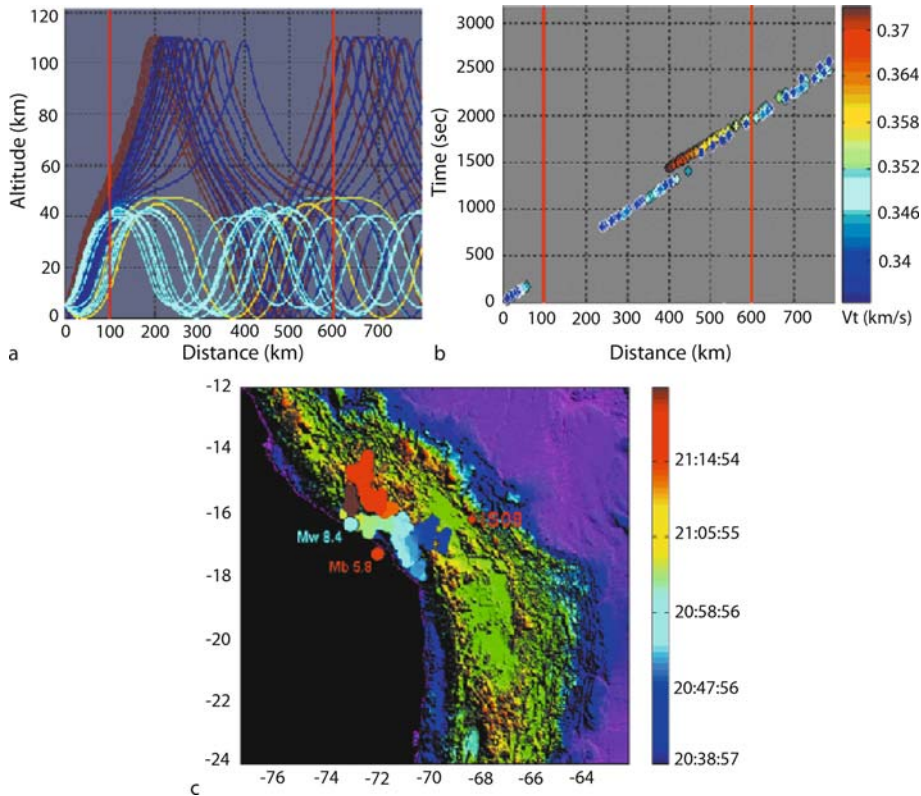
The principles introduced in the discussion of the generation, propagation, measurement, and interpretation of earthquake infrasound are applied to the progressive case studies of the Arequipa earthquake of June 23, 2001, recorded at a range of ~ 500 km, the Mongolia earthquake of November 14, 2001, recorded at a range of ~ 1800 km, and the Chile earthquake of June 13, 2005, detected by multiple infrasonic arrays to a maximum range of 2300 km.

Case Study 1: M_w 8.4 Arequipa Earthquake Detected by a Single Station

On June 23, 2001, at 20:33:13 UTC, a strong earthquake measuring M_w 8.4 (NEIC) ripped along the coast of south-central Peru. The earthquake origin (16.15°S , 73.40°W , fo-

cal depth ~ 30 km) was centered along the Peruvian coast about 600 km southeast of Lima and 110 km northwest of Camana (Fig. 1). The Pacific Tsunami Warning Center reported a moderate tsunami struck the Peruvian coast. Infrasonic waves associated with this event were detected for more than an hour at IMS infrasonic station IS08 in La Paz, Bolivia (16.26°S , 68.45°W) [48].

Due to the relative proximity of the station (~ 500 km from the epicenter), it was possible to perform relatively direct analyses of the apparent horizontal propagation speed (trace velocity) of the incident wavefield and the arrival angle of the different wave types at the array. The estimated trace velocity ranges from several kilometers per second to the sound velocity, as expected from the arrival of both seismic and infrasonic waves. Although the aperture of the IS08 array is designed for infrasonic wavelengths, array processing also yields the arrival characteristics of seismic waves. The azimuth variation for the seismic waves indicates that the rupture propagated from



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 1
Ray traces **a** and travel time curves **b** for infrasonic waves launched almost horizontally from the epicenter area at 5 km height. The red line indicates the propagation range of the rays from the secondary sources to IS08. The color scale indicates the horizontal trace velocity of each ray. **c** Location of the sources of distant generation of infrasonic waves measured from 20:39 to 21:28 due to the M_w 8.4 earthquake and the m_b 5.8 aftershock (Topography data: USGS DEM & Cornell Andes Project). The colored dots indicate the arrival times (UTC) of the infrasonic waves at the station

the northwestern to the southeastern part of the fault at a speed of 3.3 km/s. However, the predominant source of infrasound is attributed to pressure waves radiated by the Andean Cordillera. This is consistent with the theory that the vibration of mountains can generate infrasonic waves which travel to the station along atmospheric waveguides. By performing basic inversions, the azimuth variation of the infrasonic waves can then be interpreted as a distribution of secondary sources along the highest mountain ranges. Using the azimuth and arrival time determination, the infrasonic radiation zone was estimated to be ~ 100 by 400 km long.

Case Study 2: M_s 8.1 Mongolia Earthquake Detected by a Single Station

On November 14, 2001, at 09:26:10 UTC, a magnitude M_s 8.1 earthquake rattled the mountainous western Chinese region near the Qinghai-Xinjiang border. The earthquake origin (36.0°N , 90.5°E , focal depth ~ 5 km) was centered along the northern margin of the Tibetan Plateau at the foot of the Kunlun Mountains where substantial surface fault ruptures have occurred before. Coherent infrasonic waves associated with this event were detected for more than one hour at a distance of 1800 km from the epicenter by IMS station I34MN in Mongolia [49].

Building on the conclusions from the Arequipa earthquake analysis, both an inverse location procedure and a complete simulation of the radiated pressure field are used to locate the distant source regions. The input parameters of the location procedure include the measured signal azimuths and arrival times as well as the origin time and coordinates of the main shock. The propagation model is based on a constant velocity of 3.3 km/s for seismic surface waves propagating from the epicenter area. The atmosphere is specified by sound velocity and wind speed profiles obtained from the time-varying MSISE-90 and HWM-93 empirical reference models [25,38], and the infrasonic wave propagation was performed using 3D ray theory [73,74].

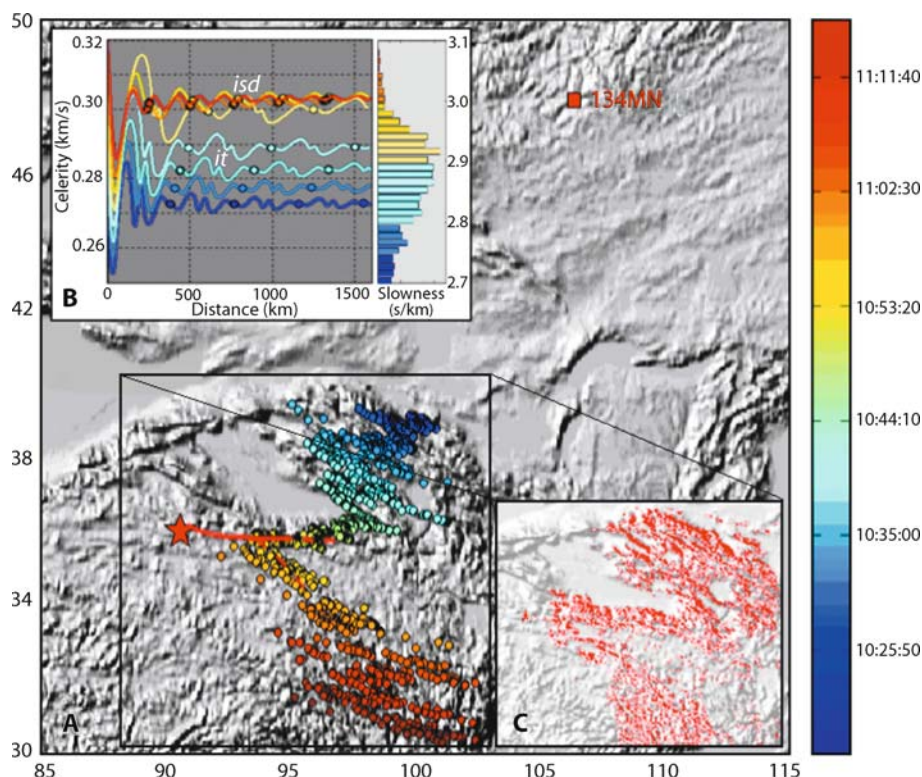
As shown by Fig. 2, two dominant guided wave groups are predicted: (i) thermospheric arrivals refracted below 120 km for slowness ranging from 2.7 to 2.9 s/km (trace velocities of 0.345–0.37 km/s), and (ii) stratospheric ducted waves refracted below 45 km for slowness values ranging from 2.9 to 3.1 s/km (0.32–0.345 km/s). Such trapped waves can be observed when the source is located above the station [75].

The slowness distribution derived from the measured trace velocity presents a maximum between 2.85 and 2.95 s/km. These values correspond to a celerity of 0.28–

0.30 km/s. The component of the wind transverse to the propagation direction deflects the rays from the original launch azimuth by $\sim 2^\circ$. This deviation is taken into account by correcting the measured azimuths. Figure 3a reconstructs the distant source regions using a celerity of 0.29 km/s. The spatial extent of the radiating zone is estimated to be 9° in latitude and 10° in longitude. The source distributions fall into line with the Qilian range, then borders the eastern part of the Qaidam basin and join the Kunlun range. To the south of Qaidam basin, more scattered source distributions follow the Bayan Har mounts.

To verify these locations, a simulation of the radiated pressure field was performed. First, the inversion for the rupture propagation along the fault uses a slip patches model developed by [7]. Using this extended model of rupture, synthetic seismograms of surface waves are computed using a discrete wavenumber method [8,70,72] with a one-dimensional regional crust model. The source modeling displays a strong directivity, with most of the seismic energy radiated along the main strike-slip of the fault and a maximum ground velocity placed ~ 300 km to the east of the epicenter [53]. To compute the acoustic radiation of the topography surrounding the fault, the topography is divided into adjacent strip-line sources radiating energy proportional to the simulated ground velocity [41].

Compared to the wavelength of the seismic surface waves (~ 60 km), the area of each source element ($3 \times 3 \text{ km}^2$) is small enough to consider isophase vibration. Due to the low frequencies of interest ($kL > 1$, k and L defining the acoustic wavenumber and the side of each cell, respectively), source elements radiate essentially simultaneously with a pronounced directivity. Based on this assumption, the topography is divided in adjacent strip-line sources of length L radiating energy proportional to the simulated ground velocity V_l normal to each surface element l . Considering a distance of observation R_l significantly greater than L , the Fraunhofer approximation of the Helmholtz–Huygens integral yields: $p_k(t) = iL(k\rho c)/2\pi \sum_{l=1}^N V_l(t_l) \Delta h_l(e^{-ikR_l})/R_l [\sin(k\hat{x}_l L/2)/(k\hat{x}_l L/2)] e^{[-ikc_o(t-t_l)]}$, where t_l is the origin time of each source element, $p_k(t)$ is the predicted pressure at the arrival time $t(t = t_l + R_l/c_{\text{eff}})$, c_{eff} is the celerity in the atmosphere corresponding to the predicted wave guides (0.29 km/s), ρ is the air density, c_o is the sound speed, Δh_l is the height difference of the radiating surface and $\hat{x}_l = \sin(\theta_l)$ is given by the angle θ_l between the outward unit normal and the source/receiver vector. Note that the sinc function in the predicted pressure reaches its peak when the surface normal and the source-receiver vector are aligned. Thus the Fraunhofer approximation provides a means of evaluating a discretized source pressure term



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 2

Propagation modeling and location of distant source regions of infrasonic waves (Topography data: ETOPO30). **A** The colored dots indicate the sources location according the detected arrival times (UTC) of the infrasonic waves. Taking into account uncertainties due to the measurements and the propagation modeling, a maximum location error of 20 km is estimated for each dot. **B** Predicted celerity models versus slowness and propagation range for a source located at the main shock epicenter. The definition range of the celerity is given by the maximum of the slowness distribution derived from the measured trace velocities (Fig. 2). The circles indicate the locations of the ground reception of both ducted stratospheric (*isd* arrivals) and thermospheric paths (*it* arrivals). **C** Normalized surface pressure distribution along the Kunlun fault

for radiating strip lines of topography, where each strip will yield its largest pressure contribution when its displacement is along the source-receiver direction.

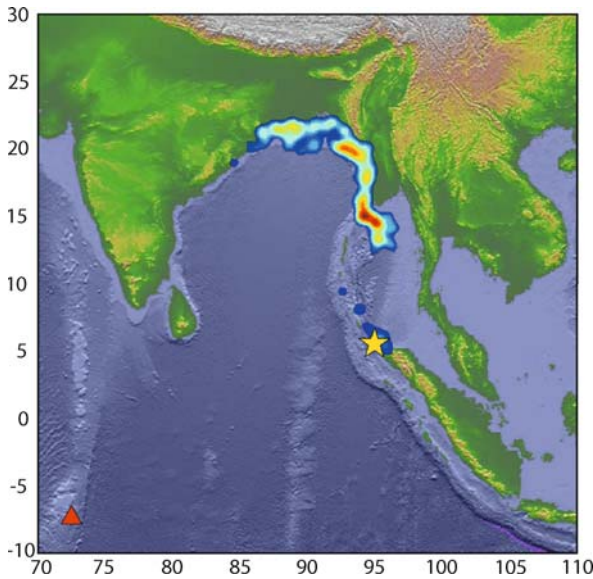
Using the simulated ground velocity and this approximation of the integral formulation, the predicted surface pressure distribution matches reasonably well with those obtained with the inverse location procedure (Fig. 2). Thus, the azimuth variations and the expansion of the signal duration suggest that the Kunlun Mountains acted as sources of infrasonic waves over a radiating zone of $\sim 1000 \times 1000$ km. The maximum seismo-acoustic coupling is found to the east of the main shock epicenter, which is consistent with the seismic radiation pattern.

Case Study 3: M 7.8 Chile Earthquake Detected by Multiple Stations

On June 13, 2005, a major earthquake occurred in the mountainous section of the Tarapaca Province (North

Chile) (19.93°S – 69.03°W at 22:44:33 UTC, M7.8, focal depth 117 km, USGS). The epicenter was located deep under the Andes mountain range, near Chile's border with Bolivia. At large distances from the epicenter, coherent infrasonic waves were detected by IMS infrasound stations I08O-Bolivia, I09BR-Brazilia, and I41PA-Paraguay (410 km, 2300 km, and 1420 km from the epicenter, respectively). The multiple station recordings at different ranges and azimuths from the epicenter allowed a more complete reconstruction of the infrasound source regions compared to what was obtained using one single station [52].

To invert for the main source regions of infrasound, the location procedure requires the signal azimuths and arrival times measured independently by each station and the origin time and coordinates of the epicenter. A joint inversion for the source area using data from all three stations simultaneously was not used due to the pronounced directivity of the radiation pattern. Infrasound



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 3 Location distribution of distant source regions of infrasonic waves generated by the December 26, 2004 Sumatra earthquake. Color scales are normalized. The red triangle indicates the location of the IS52 station. Source locations are computed for seismic surface waves and tsunami waves originating from the maximum of coseismic slip

propagation was simulated using a spherical coordinate 3D ray theory formulation which accounts for topography and the spatio-temporal variations of the horizontal wind terms along the ray paths [17]. Atmospheric absorption is integrated using altitude dependent attenuation coefficients [66,67]. The atmospheric conditions of June 13, 2005 are described by the sound velocity and wind speed profiles provided by the time-varying Ground to Space (G2S) atmospheric specifications [19].

As in the previous case study, the slip patches model developed by [72] is used to check the association of regions radiating infrasound with areas of strong ground motion. The simplest extended elliptic source model able to explain the teleseismic seismograms is found, and the first and second order characteristics of the event (location, depth, duration, focal mechanism, and refined kinematic parameters such as spatial slip distribution on the fault and rupture velocity) are calculated from teleseismic body waves. Then, from the resulting extended source model, low frequency synthetic seismograms (period lower than 10 s) are computed on a grid in the vicinity of the epicenter using the discrete wavenumber method and a one-dimensional regional crust model [8]). Finally, the root mean square of the maximum velocity of the vertical and horizontal components of the surface waves is used to reconstruct the areas of strong ground motion.

The reconstructed source regions confirm that most of the energy is radiated by the vibration of land masses near the epicenter, which is consistent with the predicted areas of strong ground motion. No clear signal originates from the Altiplano. Southern high mountain ranges, even far from the epicenter, also generated infrasound. The Central Cordillera extending to altitudes greater than 5000 m efficiently produced infrasound in the direction of I09BR, although the predicted seismic movement is low in this region. Consistent with previous observations, these results suggest an amplification of the ground displacement caused by the topography surrounding the Altiplano. Such site effect could not be predicted from our seismic source modeling since the topography is not considered. The reconstructed source regions extend over ~ 800 km from the Central Cordillera to the Occidental Cordillera. The spatial extent of the radiating zones differs from one station to another, which confirms the influence of shadow zone effects for nearby stations and the directivity of the radiation. As in the previous case study, the topography is modeled as a succession of adjacent strip-line sources, so that mountain ranges radiate energy essentially simultaneously with a pronounced directivity and may generate infrasound arrivals with different azimuths. This suggests that the amount of energy radiated in the direction of the receiver and the duration of the signals also depends on the orientation of the highest mountain ranges around the station.

Summary of Earthquake Infrasound

The three case studies presented in this section build in complexity and sophistication while producing consistent results. High mountain chains rattled by large earthquakes reliably radiate infrasound and have acoustic radiation patterns that depend on the chain's orientation. Substantial contributions to the sound field are expected from steep topography, which would primarily radiate perpendicular to exposed faces. For large earthquakes occurring in mountainous regions, infrasonic measurements are valuable for the analysis of the remote effects of earthquakes and site effects over broad areas. In remote regions where there is a lack of surface motion instrumentation, infrasonic observations could lead to a rapid determination of the regions where the seismic movements are the largest.

Tsunami Infrasound

Previous study has shown that significant infrasound is produced by breaking waves (surf) and the complex interaction of open-ocean swells (microbaroms, e. g. [76]).

However, interest in tsunami infrasound emerged when IMS infrasound arrays in the Pacific and Indian Oceans recorded distinct signatures associated with the December 26, 2004 Great Sumatra-Andaman earthquake and tsunami. As in the case of mountains stirred by continental earthquakes, islands which undergo significant surface displacements during submarine earthquakes can also produce infrasound. It also appears that the initiation and propagation of a tsunami may produce low frequency sound near the epicenter as well as along coastlines and basins. In some environments, precursory sound could potentially be used for confirmation and early warning for tsunamis. This field of research is still in its infancy, and our interpretations leave much room for further development. Substantial complexity is involved in the separation of the earthquake-generated infrasound from the sound that may be produced by the genesis and propagation of the tsunami.

Sumatra Earthquake and Tsunami

The magnitude 9.1 Great Sumatra-Andaman earthquake of December 26, 2004 [63] is the largest earthquake since the 1964 magnitude 9.2 Great Alaska earthquake, and produced the deadliest tsunami in recorded history. In contrast to the Great Alaska earthquake, the Great Sumatra earthquake and tsunami were recorded globally by multiple digital sensor networks in the ground, ocean, atmosphere, and space. Further, the resulting signals were analyzed and interpreted using far more advanced computational capabilities than were available 40 years ago. Although the study of tsunami infrasound rose in the wake of the Sumatra-Andaman event, many fundamental questions on the generation of these deep sounds remain unanswered.

The clearest infrasonic signatures associated with the Sumatra event were captured by the station in Diego Garcia (IS52GB, Fig. 3), which recorded (1) seismic arrivals from the earthquake, (2) tertiary arrivals (T-phases) that propagated along sound channels in the ocean and coupled back into the ground, (3) infrasonic arrivals associated with either the tsunami generation mechanism near the seismic source or the motion of the ground above sea level, and (4) deep infrasound (with a dominant frequency lower than 0.06 Hz) coinciding with the propagation of the tsunami into the Bay of Bengal [30,51]. These signals were all recorded by the pressure sensors in the arrays. The seismic and T-phase recordings are a result of the sensitivity of the microphones to ground vibration (microphonics), whereas the infrasound arrivals correspond to sound propagating through atmospheric waveguides.

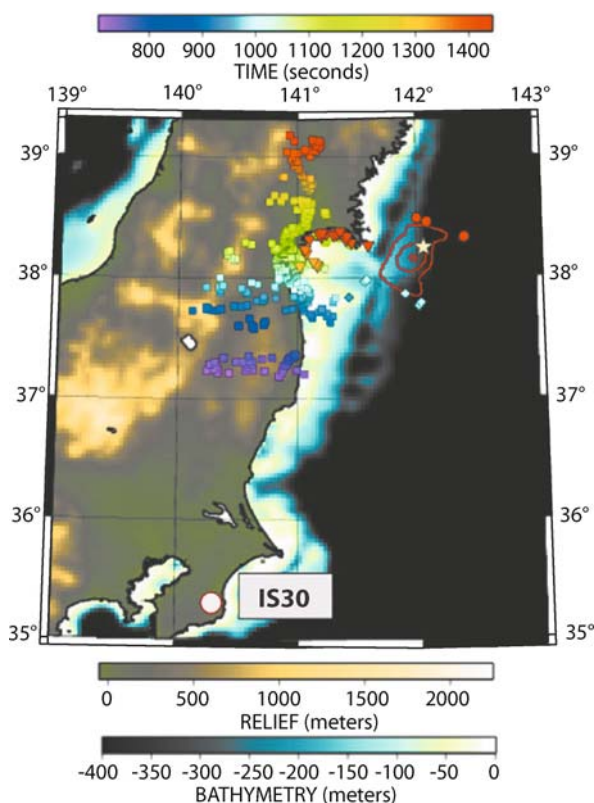
Similar, but not identical arrivals were observed at Diego Garcia (range of ~ 2900 km) during the March 28, 2005 Nias earthquake (M8.7, which produced a non-destructive tsunami). Yet only infrasonic arrivals were observed from the April 10, 2005 Mentawai earthquakes (M6.7 and 6.5, no reported tsunami), indicating that above-water ground motion from submarine earthquakes can produce sound in the Sumatra region. The deep infrasound from the Bay of Bengal region following the Sumatra earthquake suggests that sound can be produced by the interaction of a tsunami with coastal bathymetry.

To reconstruct the main source regions of infrasound recorded at IS52 for this event, the input parameters of the location procedure included the measured signal azimuths and arrival times, and the origin time and coordinates of the epicenter. As described in the previous section, standard velocity models were used to describe the propagation of the seismic surface waves and the propagation of infrasound through the atmosphere in the direction of IS52. However, the speed of propagation of the tsunami from the epicenter needs to be added to the inversion procedure. For the December 26, 2004 earthquake, a velocity of 3.3 km/s is used for the seismic surface waves, and a speed related to the square root of the water depth for the tsunami waves.

Using the infrasonic arrivals to IMS station in Diego Garcia and Palau, the location inversion for the high frequency infrasonic component corresponded to sound originating from the mountains and possibly from the epicenter. However, a unique characteristic of the Sumatra earthquake is that it generated large amplitude coherent waves with a dominant period of ~ 30 s over four hours. The infrasonic source locations derived from the inversion procedure indicate that the tsunami also created infrasonic waves at lower frequencies when it propagated in shallower water as it reached the Bay of Bengal (Fig. 3). Even lower-frequency acoustic gravity waves, with periods of hundreds of seconds, may have been produced by the underwater ground displacement [58]. For wavelengths which are much greater than the water depth, the ocean surface displacement nearly matches the submarine ground displacement above the fault plane (e. g. [44]) and may efficiently radiate long-period pressure waves into the atmosphere.

Miyagi-Oki Earthquake and Tsunami

The events off the coast of Sumatra were ~ 3000 km from the closest infrasound station in Diego Garcia. The longer ranges, coupled with the fact that all infrasound stations used in those studies were transverse to the axis of Suma-



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 4
 Estimated infrasonic source locations associated with ground vibration, tsunami genesis, and the interaction of the tsunami with the coastline. The squares represent stratospheric arrivals with a celerity of 0.3 km/s. The diamonds are also stratospheric arrivals but with the celerity of 0.32 km/s predicted for that azimuth. The circles are thermospheric arrivals with a celerity of 0.27 km/s. The triangles are stratospheric arrivals with a celerity of 0.3 km/s, but with a delay time attributed to seismic seiche formation. The color of the symbols indicates the arrival time in seconds since the earthquake's origin time. The topography is from NOAA ETOPO2 data

tra, caused uncertainty in the ability to discriminate between sounds potentially produced during tsunami genesis at the ocean surface and the sounds produced by the earthquake-induced vibration of mountains and islands. In contrast, IMS infrasound station IS30 in Japan (Fig. 4) is optimally situated to recognize the different source regions of infrasound associated with the Miyagi-Oki earthquake and tsunami.

The magnitude 7.2 Miyagi-Oki earthquake occurred on August 16th, 2005 at 02:46:28 UTC. The epicenter was off the coast of Japan near Honshu (38.251°N, 142.059°E), with an estimated depth of 36 km. The earthquake caused landslides, ~60 injuries, as well as power and transportation disruptions. A local, nondestructive tsunami was observed on the coast of northern Japan with a wave height

of ~10 cm. Because of its auspicious location, station IS30 can use the angle of arrival information derived from array processing to identify infrasonic arrivals originating from mountain chains, the earthquake epicenter (which is used as the tsunami epicenter), and the coastline of the Bay of Sendai.

In contrast with the Sumatra event, where substantial energy was observed in the deep infrasound band (0.002–0.1 Hz), most of the infrasonic energy for the smaller Miyagi-Oki event was above 0.5 Hz. The arrival azimuths at IS30 range from -5° to 28° , and do not have a well-defined temporal sequence, suggesting multiple sources, propagation paths, and possible wave types.

Using the ground to space (G2S) atmospheric profiles [19] specific to the station location and time of the event, source locations [30,51] were estimated for the infrasonic arrivals shown in Fig. 4. The first infrasonic arrivals would correspond to acoustic waves coupled to the atmosphere from the seismic vibration of land masses [49]. For ranges less than 330 km and northerly arrivals between -5° and 18° (measured clockwise from north), the time- and site-specific G2S specifications do not support thermospheric arrivals at the station, so the observed arrivals would propagate in stratospheric waveguides with a celerity of 0.3 km/s. In contrast, for arrival azimuths of $\sim 18^\circ$ to 28° originating offshore, thermospheric arrivals are also supported for ranges greater than 330 km, although the celerity of stratospheric arrivals is faster.

Multiple wave propagation paths are invoked to produce the temporal and frequency distribution of the infrasonic signals observed at IS30. The squares shown in Fig. 4 correspond to seismic vibrations radiating acoustic energy into stratospheric waveguides, thereafter propagating with a celerity of 0.3 km/s, the diamonds to stratospheric arrivals with a celerity of 0.32 km/s, and the circles correspond to thermospheric arrivals with a celerity of 0.27 km/s. Inclusion of seismic speeds made a negligible difference, as it offsets the locations by <6 km. As expected, many of the arrivals originate from land, and are consistent with the predicted seismic intensity [13,71]. Of more relevance to this paper, clear stratospheric and thermospheric arrivals arrive from the perimeter of the epicenter, which is assumed to be the region of tsunami genesis. These results are consistent with the proposed infrasonic source locations obtained by [30] for the Sumatra earthquakes, and support the idea that ocean surface displacements associated with submarine earthquakes produce infrasound [58]. The possibility that vertical ocean displacements near the epicenter may radiate sound suggests that infrasound signals may be potential discriminants for tsunami genesis.

The later arrivals (shown as triangles in Fig. 4) originate from the mountain regions to the north of the Bay of Sendai when stratospheric and thermospheric arrivals are assumed. Yet, analyses of the Great Sumatra earthquake suggest that the Bay of Bengal produced deep infrasound, and the possibility that the Bay of Sendai may also act as an acoustic source is considered. The triangles in Fig. 3 contour the coastline of the northern Bay of Sendai when a combination of a delay time and stratospheric apparent propagation speeds are considered. For a celerity of 0.3 km/s, the inferred delay time is ~ 250 s. This would require an acoustic generation process that requires ~ 4 min to be established over a coastal region ~ 100 km long. The earthquake occurred in a shallow region, which corresponds to slow tsunami propagation speeds of ~ 0.05 km/s. This would mean the tsunami would have traveled at most 12 km in 250 s, which is not enough time for the tsunami to reach the coastline and produce these signals. Thus the interaction of the small tsunami with the coastline is eliminated as a possible source.

A plausible ensonification mechanism is the local production of coastal waves by the earthquake, as in the generation of seismic seiches [5,56]. Low frequency radiation from enclosed bodies of water have been proposed by other authors [9,46]. For the Miyaki-Oki earthquake, substantial water displacement in shallow regions of the bay is strongly suggested by acoustic sources above the water near the coastline (blue and green arrivals in Fig. 4). Satellite imagery shows that the northern part of the Bay of Sendai has an abundance of lakes, bays, man-made harbor structures and rivers which may sustain seiches. Lower order seiche modes would take minutes to be established, but higher-order modes and coupled oscillations associated with the narrower of the volume dimension could sustain an ensonification process akin to microbarom generation from the ocean (e. g. [76]. Alternatively, the triangular symbols in Fig. 4 would should be shifted ~ 75 km North of their shown location.

Summary of Tsunami Infrasound

These two case studies strongly suggest that submarine earthquakes and the water level changes they induce can produce low-frequency sound. The sound may be radiated from the ocean surface during the tsunami genesis, produced by the vibration of land masses near the epicenter, or be excited by the interaction of seismic and water waves with the coastline, shallow bathymetry, and harbors.

There is some potential for using infrasound in conjunction with other technologies for remote tsunami monitoring. The effective propagation speeds of tsunami

(~ 50 – 200 m/s) and sound waves (~ 300 m/s) yield an advance warning time of at least 1.7 s/km. At 100 km, sound leads the tsunami by at least 170 s, but some of this time would be taken up by signal transmission, processing and identification, leaving less than one minute to issue an alert. However, infrasound may provide early warning in areas within shallow basins or further than a few hundred kilometers from the tsunami source region.

Volcano Infrasound

The previous sections discussed how earthquake infrasound acoustically couples solids and gases, and tsunami infrasound couples solids, liquids, and gases. Similarly, a volcano can transmit information about its current eruptive state through vibrations induced in the ground, volcanic fluids, and atmosphere. However, there is a great deal of ambiguity about the composition of the acoustically active volcanic fluids, which may vary from vesiculated magma, through gas-ash mixtures, to steam. Although a sealed volcanic conduit can produce faint sounds through earthquakes, once the volcanic plumbing breaches the surface it can broadcast infrasound quite unambiguously.

Basic Principles

The birth of volcano infrasound research may be traced to the cataclysmic 1883 eruption of Krakatoa, which triggered a series of interdisciplinary, international geophysical studies of the pressure signals produced by the volcano [77]. Barometric records observed throughout the US, Europe, and Russia, and reports of cannon-like sounds in surrounding islands (as far as Diego Garcia and Rodrigues Islands) demonstrated for the first time the ability of low frequency sound to propagate for thousands of kilometers.

Infrasonic signals originating inside magma conduits contain information about the pressure fluctuations driving eruption processes. These signals may be broadly categorized as explosions, long-period events, and tremor. Very large eruptions can also produce acoustic gravity waves [34,68]. Explosions are impulsive and have durations of seconds, long period events are more emergent than explosions, have distinct spectral peaks, and may last seconds to minutes, and tremor signals are sustained atmospheric vibrations that can persist from minutes to years. Decades ago, volcano seismology identified tremor and long-period events as signals indicative of near-surface intrusion of volcanic fluids (e. g. [11]). In contrast, the first infrasonic measurements of tremor from Sakurajima

were published in [64]. As the sensitivity of instrumentation and sophistication of analysis methods increased with advancing technology, the catalogue of known volcanic sounds has expanded and the ability to remotely detect hazardous eruptions has extended. Ongoing monitoring efforts suggest that relatively gentle Hawaiian and mild Strombolian activity may be consistently observed from a range of ~ 10 km [20] and strong Strombolian, Vulcanian and Plinian eruptions from distances of tens to hundreds of kilometers [31,50,54,68].

The difficulty with modeling and interpreting volcanic sounds is that there is a lot of ambiguity in the geometry, composition, thermodynamics, and phase (solid, liquid, gas, or a mixture) of the volcanic interior preceding and during an eruption. In addition, there are many possible ways of exciting volcanic fluids into oscillation through unsteady and transient flow. As examples of these complications, some possible ways of exciting tremor signals in a magma conduit are described. Tremor signals are endemic to volcanoes, and are generally attributed to magma intrusion. Some plausible driving mechanisms of tremor signals and the physical properties of the materials inside volcanic pipes are discussed. In the following example, it is postulated that the harmonic spectrum of tremor events is caused by the acoustic resonance of the fluid within the magma conduit of Arenal volcano, Costa Rica [23]. The possible relationships between observed signals and the gas content of the melt, the physical conditions inside the volcano, and the flow dynamics of the volcanic fluids are discussed.

Case Study: Tremor Signal at Arenal Volcano

This section hopes to provide an overview of the complexities of modeling volcanic sounds by considering the frequency changes in a harmonic tremor signal recorded at Arenal volcano, Costa Rica. Figure 5 shows a section of the infrasonic tremor signal shown in [23]. This signal is illustrative of most of the tremor recorded during April–May 1997 by a three-element array of infrasonic sensors and a five-element array of seismometers [35].

The prominent features of the Arenal tremor are the harmonic character of the spectra and the oscillation of the spectral bands with time. This oscillation of the spectral bands about an equilibrium value is referred to as gliding. Since steady flow cannot generate acoustic waves, it must be the pressure, volume, and flow fluctuations of the magma injection process that are driving the infrasonic signals. It is possible that the gliding of the spectral bands is caused by time-varying changes in the flow regime or the physical properties of the material in the conduit. The

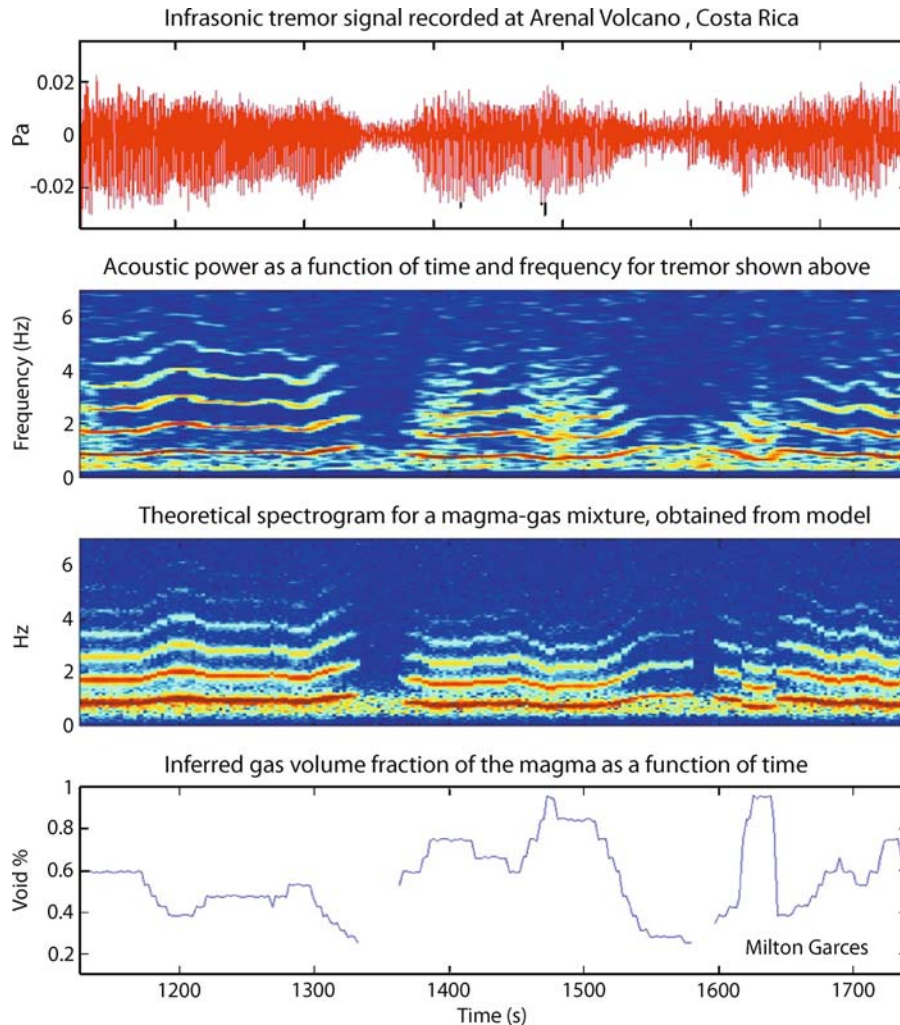
resonant magma conduit model of [21] is used to explore the ramifications of having changes in the conduit length, flow velocity of a gas and in the void fraction of a magma-gas mixture.

The tremor signal shown in Fig. 5 lasts over ten minutes and is not a transient event. Attempts to model the spectra with a random continuous source, as successfully done for Pavlof volcano [24], failed at Arenal because of the sharpness of the spectral peaks and the rapid rolloff at higher frequencies. The tremor signal appears to be sustained by a source mechanism that is triggering repeatedly in time. The source region may consist of a compliant gas volume within a magma-filled conduit. A bubble may form around a corner or in a constriction in a magma-filled conduit, where fluid acceleration may generate relatively stable gas-rich pockets. These cavities will be sensitive to flow fluctuations, and may act as effective acoustic sources [21].

Lava flows at Arenal and incandescent pieces ejected during explosions strongly suggests an open magma conduit associated with the infrasonic recordings. However, there is some ambiguity on whether the acoustically active part of the conduit is filled with a magma-gas or an ash-gas mixture. The source region is placed at the lower part of the magma conduit, and it is assumed that it ensonifies the magma-gas mixture above it. Although for the Arenal signal there is evidence for a buried tremor source [35], it may also be possible to have efficient atmospheric excitation from jet flow above the vent during more powerful eruptions [31]. For the purposes of our discussion, a two-layer, open vent solution representing a slow-velocity, gas rich magma floating atop a less vesiculated layer is used, as developed in Sect. 4 of [21]. The source region at the bottom of the conduit could correspond to a constriction in the conduit, where cavitation may occur due to fluid acceleration.

The sound speed of a liquid gas mixture is a strong function of the amount of bubbles, or the void fraction, in the mixture. For high void fractions, the sound speed of a magma-gas mixture may reach sound speeds of tens of meters/second, and sound would be heavily attenuated. For a resonant fluid column of length L , $f_n = n(1 - M_2)c/(2L)$, where f_n is the resonant frequency of the n th spectral peak, M is the Mach number, or ratio of the flow speed to the sound speed, and c is the sound speed of the fluid.

Assuming a gas-rich magma inside a resonant magma conduit, the changes in the spectral peaks of the tremor signal may be attributed to changes in the amount of exsolved gas in the melt (Fig. 5), which would dramatically change the sound speed c . The tremor signal would cease



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 5

Recorded acoustic signal, spectrogram, synthetic spectrogram, and one possible interpretation for a harmonic tremor signal observed at Arenal Volcano, Costa Rica

if the mass flux stops or becomes steady. It is unlikely that there is only gas inside the conduit because the acoustic attenuation in the gas would not be sufficient to explain the rapid amplitude decay with frequency. However, it is possible to have a gas-ash mixture, in which case it may be plausible to also explain the gliding with a change in the Mach number of the flow or a variation in the length of the resonant portion of the conduit [21].

Attenuation coefficients for dusty gases introduce an additional number of poorly characterized physical parameters, adding further ambiguity to our interpretations. It is due to these and many other complexities in modeling volcanic systems that the science of volcanology has evolved towards the synergy of geophysical observations

and geological parameters [37]. Some of the primary goals of volcanic studies are to permit reliable eruption forecasting, hazard assessment, and early warning. Although it is difficult to forecast an eruption without a clear understanding of the physical processes associated with precursory activity, it is possible to rapidly identify a powerful hazardous eruption and provide early warning.

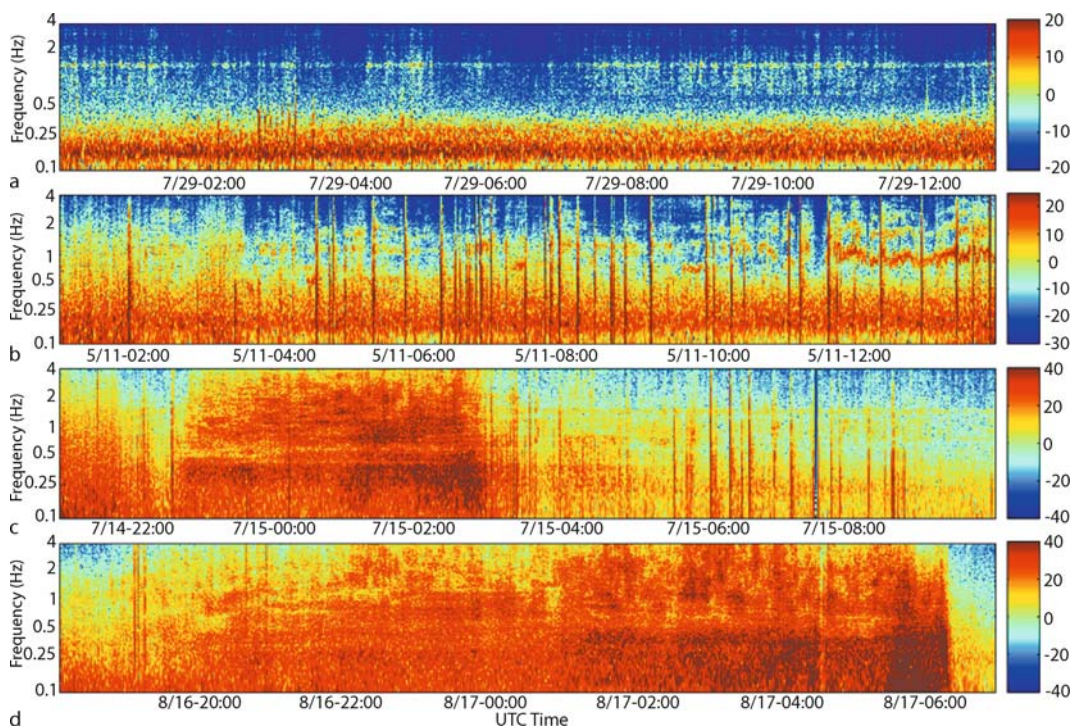
Prototype Operational System: ASHE

The Acoustic Surveillance for Hazardous Eruptions (ASHE) proof-of-concept project seeks to develop and evaluate the potential for robust, operational infrasonic remote sensing of volcanic eruptions. In contrast to

other ground-based volcano surveillance systems, the autonomous ASHE arrays are sited tens to hundreds of kilometers away from the devastation zones of erupting volcanoes. The diverse eruption signals produced by Tungurahua Volcano and captured by the ASHE arrays illustrate how acoustic remote sensing may complement seismic observations and satellite remote sensing to improve continuous monitoring of wide regions of potential eruption hazard.

The prototype ASHE stations consist of a four-element infrasound array with an aperture of ~ 100 m, a broadband seismic sensor, and a wind sensor [54]. A satellite dish sends the digital data in real time from the field to the Geological Survey of Canada data center in Ottawa, Canada, which distributes the data in real-time to other collaborating parties. In January of 2006 the ASHE team deployed two infrasound arrays in Ecuador [31]. These two arrays, separated by ~ 250 km, were sited to detect volcanic eruption in Ecuador or Southern Colombia by one of the arrays within 15 min of the eruption, and as early within 5 min for Tungurahua and Sangai volcanoes, which were within 40 km of an array.

Three distinct types of eruption signatures can be identified at Tungurahua (Fig. 6). The first and most common is low ash-producing background tremor with a dominant frequency of 1.4 Hz (Fig. 6a). During mid-May 2006, volcanic activity changed and was temporarily characterized by large explosions followed by harmonic tremor (Fig. 6b), reminiscent of the aforementioned tremor at Arenal. However, no significant ash release was observed, and this second type of eruptive regime was characterized as ash-poor. On July 14th, 2006 a large Vulcanian to sub-Plinian eruption (Fig. 6c) produced dangerous pyroclastic flows, substantial stratospheric ash clouds, and a significant increase in tremor. Between August 16th–17th, 2006 a larger Vulcanian to Plinian eruption occurred (Fig. 6d) and was characterized by lethal pyroclastic flows, a larger stratospheric ash cloud, and considerable infrasonic tremor. As the eruption progressed, the majority of acoustic energy shifted to lower frequencies (<0.5 Hz). The acoustic signatures of the two major eruptions are comparable and easily identifiable in infrasonic records due to the vast amount of energy present over a broader band of frequencies. These two eruptions injected substantial ash



Infrasound from Earthquakes, Tsunamis and Volcanoes, Figure 6

Spectrograms for different eruption styles at Tungurahua volcano. The horizontal axis represents 13 hours of data, and the vertical scale shows frequency in a logarithmic scale from 0.1 to 4 Hz. The color denotes acoustic power in decibels referenced to 1 Pa/Hz. The ocean microbarom shows up in the upper panel two panels as a red band centered about 0.2 Hz, and is not of volcanic origin. Explosive events appear as intense vertical bands

into the stratosphere, suggesting this type of acoustic signature may be used for remote infrasonic monitoring of hazardous eruptions.

Automatic analysis techniques for eruption detection prototyped by the ASHE project produce automated low-latency email notifications coupled with more detailed real-time data products which can be used by responsible operational agencies to disseminate updated information. Based on the acoustic records captured during the Tungurahua eruptions of July and August 2006, source parameters that may be estimated during large eruptions include (but may not be limited to) the start time and duration of an ash cloud injection that could pose a hazard to international aircraft at cruising altitudes.

Summary of Volcano Infrasound

Volcanic signals incorporate many of the complexities found in earthquake and tsunami signals, and then add some because of the inherently unstable thermodynamic and hydrodynamic conditions that lead to and accompany eruptions. Because infrasound measures the excess pressure change that drives an eruption, it can be used to help unravel the dynamic physical processes driving eruptive activity. Although infrasound may be used with additional technologies in forecasting and hazard assessment, its clearest contribution is in the identification of the intensity and timing of a powerful eruption. The technology and methodology to use infrasound for acoustic remote sensing of hazardous eruptions is mature and ready to implement into operations.

Future Directions

Although infrasound has manifold applications in near-source studies, recent advances in sensor and analysis techniques have strengthened the potential to develop infrasound as a robust remote sensing tool that may cover observational gaps in distant and inhospitable environments. Although operational acoustic monitoring systems that can help notify aircraft of hazardous eruptions have been demonstrated as viable, much work is still needed to implement these systems and establish clear relationships between ash heights, infrasonic source parameters, and atmospheric conditions. Many parts of Earth's oceans remain undersampled, and there is a great potential for the application of infrasound to the detection of tsunami-genes, the identification and tracking of severe ocean weather, and the monitoring of breaking ocean wave intensity and other coastal hazards. Much fundamental research lies ahead in the construction of functional models for these various source processes. One of the most am-

bitious projects the community has initiated is the passive acoustic tomography of the atmosphere using natural sources [50]. These studies, refined and extended over the next decades, may contribute to our long-term assessment of global climate change.

Concluding Remarks

There is beauty in complexity, as there is perplexity in beauty. The oscillatory interaction between the solid and fluid phases of our Earth reveals a wealth of information that can be turned into knowledge by the meticulous and persistent use of observation and modeling. Infrasound captures this solid-fluid interaction at the temporal scales of tens of seconds to tens of hertz, offering alluring glimpses into the inner workings of the natural world. The heaving ground and the frothing magmas, borne of the dark depths of Earth, oft conspire with the surging oceans to pervade the skies with deep sound. The field of infrasound strives to interpret the imperceptible yet omnipresent soundscape that permeates the air about us, so as to assuage uncertainty and perchance alleviate our dread of the natural threats posed by earthquakes, tsunamis, and volcanoes.

Acknowledgments

The authors wish to thank C. Hetzer, P. Caron, and S. McNamara for figures and analyses. Garces also extends his appreciation to M. Protti, M. Haggerty, and S. Schwartz for their contributions to the Arenal studies. D. Fee and R. Matoza provided very useful revisions to this chapter, for which we are grateful. Garces' contributions to this chapter were supported in part by the National Science Foundation (grant EAR-0609669).

Bibliography

Primary Literature

1. Alcoverro B, Le Pichon A (2005) Design and optimization of a noise reduction system for infrasonic measurements using elements with low acoustic impedance. *J Acoust Soc Am* 117:1717–27. doi:10.1121/1.1804966
2. Alcoverro B, Martysevich P, Starovoi Y (2005) Mechanical sensitivity of microbarometers MB2000 (DASE, France) and Chaparral 5 (USA) to vertical and horizontal ground motion. *Inframatics* 9:1–10 <http://www.inframatics.org/>. Accessed Mar 2005
3. Bhattacharyya, Bass JH, Drob D, Whitaker R, Revelle D, Sandoval T, Woodward R (2003) Description and Analysis of Infrasound and Seismic Signals Recorded from the Watusi High-explosive Experiment of September 28, 2002. SAIC technical report SAIC-03/2206
4. Bedard AJ (1971) Seismic response of infrasonic microphones. *J Res Nat Bur Stand* 75:41–45

5. Berninghausen WH (1969) Tsunamis and seismic seiches of Southeast Asia. *BSSA* 59:289–297
6. Bolt BA (1964) Seismic air waves from the great 1964 Alaskan earthquake. *Nature* 202:1095–1096
7. Bouchon M (1976) Teleseismic body wave radiation from a seismic source in a layered medium. *Geophys J Int* 47(3):515–530. doi:10.1111/j.1365-246X.1976.tb07099.x
8. Bouchon M (1981) A simple method to calculate Greens functions for elastic layered media. *Bull Seism Soc Am* 71:959–971
9. Bowman HS, Bedard AJ (1971) Observations of Infrasound and Subsonic Disturbances Related to Severe Weather. *Geophys J Int* 26(1–4):215–242. doi:10.1111/j.1365-246X.1971.tb03396.x
10. Brown D, Garcés M (2009) Ray Tracing in an Inhomogeneous Atmosphere with Winds, *Handbook on Signal Processing in Acoustics*. Springer (in press)
11. Chouet B (2003) Volcano Seismology. *Pageoph* 160:739–788
12. Cansi Y (1995) An automatic seismic event processing for detection and location: the PMCC method. *Geophys Res Lett* 22:1021–1024
13. Che I, Lee H, Jeon J, Kang T (2007) An analysis of the infrasound signal from the Miyagi-Oki earthquake in Japan on 16 August 2005. *Earth Planets Space* 59:e9–e12
14. Cook RK (1971) Infrasound radiated during the Montana earthquake of 1959 August 18. *Geophys J R Astr Soc* 26:191–198
15. Cook RK, Young JM (1962) Strange sounds in the atmosphere, Part II. *Sound* 1:25–33
16. Cox EF, Plagge HJ, Reed JW (1954) Meteorology Directs Where Blast Will Strike. *Bull Am Meteorol Soc* 35:95–103
17. Dessa JX, Virieux J, Lambotte S (2005) Infrasound modeling in a spherical heterogeneous atmosphere. *Geophys Res Lett* 32:L12808.1–5 doi:10.1029/2005GL022867
18. Donn WL, Posmentier ES (1964) Ground-coupled air waves from the great Alaskan earthquake. *J Geophys Res* 69:5357–5361
19. Drob DP, Picone MJ, Garces M (2003) Global morphology of infrasound partitioning. *J Geophys Res* 108(D21):4680. doi:10.1029/2002JD003307
20. Fee D, Garcés M (2007) Infrasonic tremor in the diffraction zone. *Geophys Res Lett* 34:L16826. doi:10.1029/2007GL030616
21. Garcés MA (2000) Theory of acoustic propagation in a multiphase stratified liquid flowing within an elastic-walled conduit of varying cross-sectional area. *J Volcanol Geotherm Res* 101:1–17
22. Garces M, Hetzer C (2003) Optimizing the Progressive Multi-Channel Correlation Detector for the Discrimination of Infrasonic Sources. In: *Proceedings of the 25th Seismic Research Review*, Tucson, 23–25 Sept 2003
23. Garcés MA, Hagerty MT, Schwartz SY (1998) Magma acoustics and time-varying melt properties at Arenal Volcano, Costa Rica. *Geophys Res Lett* 25:2293–2296
24. Garcés M, Hansen RA, Lindquist KG (1998) Traveltimes for infrasonic waves propagating in a stratified atmosphere. *Geophys J Int* 135:255–263
25. Garces M, Hetzer C, Lindquist K, Drob D (2002) Source Location Algorithm for Infrasonic Monitoring. 24th Annual DTRA/NSA Seismic Research Review, Ponte Vedra, 17–19 Sept 2002
26. Garces M, Harris A, Hetzer C, Johnson J, Rowland S, Marchetti E, Okubo P (2003) Infrasonic tremor observed at Kilauea Volcano, Hawaii. *Geophys Res Lett* 30:2023–2027 2003
27. Garces M et al (2004) Forensic studies of infrasound from massive hypersonic sources. *EOS* 85(43):433
28. Garcés M, Willis M, Hetzer C, Le Pichon A, Drob D (2004) On using ocean swells for continuous infrasonic measurements of winds and temperature in the lower, middle, and upper atmosphere. *Geophys Res Lett* 31:L19304
29. Garcés M, Willis M, Hetzer C (2004) The Hunt for Leaky Elevated Infrasonic Waveguides. 26th Seismic Research Review, Orlando
30. Garces M, Caron P, Hetzer C, Le Pichon A, Bass H, Drob D, Bhattacharyya J (2005) Deep infrasound from the Sumatra earthquake and tsunami. *EOS* 86(35):317–320
31. Garces M et al (2008) An acoustic fingerprint of stratospheric ash injection. *EOS* 89(40):377–378
32. Georges TM (1968) *Acoustic Gravity Waves in the Atmosphere*. Proceedings of the ESSA-ARPA Symposium. US Government Printing Office, Washington
33. Gossard EE, Hooke WH (1975) *Waves in the Atmosphere: Atmospheric Infrasonic and Gravity Waves – their Generation and Propagation*. Elsevier, London
34. Goerke VH, Young JM, Cook RK (1965) Infrasonic observations of the May 16, 1963, volcanic eruption on the Island of Bali. *J Geophys Res* 70:6017–6022
35. Hagerty M, Schwartz S, Garces M, Protti M (2000) Analysis of seismic and acoustic observations at Arenal Volcano, Costa Rica, 1995–1997. *J Volcanol Geotherm Res* 101:27–65
36. Ham FM, Park S (2002) A Robust Neural Network Classifier for Infrasound Events using Multiple Array. presented at WCCI-2002, (IJCNN-2002), Honolulu, 12–17 May, pp 2615–2619
37. Harris A, Ripepe M (2008) Synergy of multiple geophysical approaches to unravel explosive eruption conduit and source dynamics – A case study from Stromboli. *Chemie der Erde - Geochemistry* 67(1):1–35
38. Hedin AE, Biondi MA, Burnside RG, Hernandez G, Johnson RM, Killeen TL, Mazaudier C, Meriwether JW, Salah JE, Sica RJ, Smith RW, Spencer NW, Wickwar VB, Virdi TS (1996) Revised global model of upper thermospheric winds using satellite and ground-based observations. *J Geophys Res* 96:7657–7688
39. Hedlin M, Garces M, Bass H, Hayward C, Herrin G, Olson JV, Wilson C (2002) Listening to the Secret Sounds of Earth's Atmosphere. *EOS* 83:557, 564–565
40. Hedlin MAH, Alcoverro B, D'Spain G (2003) Evaluation of rosette infrasonic noise-reducing spatial filters. *J Acoust Soc Am* 114:1807–1820
41. Heil C, Urban M (1992) Sound fields radiated by arrayed multiple sound sources. paper presented at the 92nd Convention of the Audio Engineering Society. Preprint no 3269, Vienna, 24–27 March 1992
42. Hercz AR (1987) *Fundamentals of Sound Ranging*. published by Arthur R Hercz
43. Herrin E, Golden P, Negraru P, Andre W, Bass H, Garces M, Hedlin M, McKenna M, Norris D, Osborne D, Whitaker R (2006) *Infrasound Calibration Explosions from Rockets Launched at White Sands Missile Range*. 28th Seismic Research Review, Orlando, 19–21 Sept 2006
44. Kajiura K (1970) Tsunami source, energy and the directivity of wave radiation. *Bull Earthq Res Inst* 48:835–869
45. Kim TS, Hayward C, Stump B (2004) Local infrasound signals from the Tokachi-Oki earthquake. *Geophys Res Lett* 31:L20605 doi:10.1029/2004GL021178
46. Larson RJ, Craine LB, Thomas JE, Wilson CR (1971) Correlation of winds and geographic features with production of certain

- infrasonic signals in the atmosphere. *Geophys J R Astr Soc* 26:201–214
47. Le Pichon A, Garcés M, Blanc E, Barthélémy M, Drob DP (2002) Acoustic propagation and atmosphere characteristics derived from infrasonic waves generated by the Concorde. *J Acoust Soc Am* 111:629–641
 48. Le Pichon A, Guilbert J, Vega A, Garcés M, Brachet N (2002) Ground-coupled air waves and diffracted infrasound from the Arequipa earthquake of June 23, 2000. doi:10.1029/2002GL015052
 49. Le Pichon A, Guilbert J, Vallée M, Dessa JX, Ulziibat M (2003) Infrasonic imaging of the Kunlun Mountains for the great 2001 China earthquake. *Geophys Res Lett* 30(15):1814. doi:10.1029/2003GL017581
 50. Le Pichon A, Blanc E, Drob DP, Lambotte S, Dessa JX, Lardy M, Bani P, Vergnolle S (2004) Infrasound monitoring of volcanoes to probe high altitude winds. *J Geophys Res* 110:D13106 doi: 10.1029/2004JD005587
 51. Le Pichon A, Herry P, Mialle P, Vergoz J, Brachet N, Drob D, Garcés M, Ceranna L (2005) Infrasound associated with large Sumatra earthquakes and tsunami. *Geophys Res Lett* 32:L19802 doi:10.1029/2005GL023893
 52. Le Pichon A, Mialle P, Guilbert J, Vergoz J (2006) Multistation infrasonic observations of the Chilean earthquake of 2005 June 13. *Geophys J Int* 167(2):838–844. doi:10.1111/j.1365-246X.2006.03190.x
 53. Lin A, Fu B, Guo J, Zeng Q, Dang G, He W, Zhao Y (2002) Coseismic strike-slip and rupture length produced by the 2001 Ms 8.1 Central Kunlun earthquake. *Science* 296:2015–2017
 54. Matoza R, Hedlin M, Garcés M (2007) An infrasound array study of Mount St. Helens. *J Volcanol Geotherm Res* 160:249–262
 55. McKisic JM (1997) Infrasound and the infrasonic monitoring of atmospheric nuclear explosions. DOE document PL-TR-97-2123
 56. McGarr A, Vorhis RC (1968) Seismic seiches from the March 1964 Alaska earthquake. US Geological Survey Professional Paper 544-E, pp E1–E43, 1 sheet, scale 1:5,000,000
 57. Mikumo T (1968) Atmospheric pressure waves and tectonic deformation associated with the Alaskan earthquake of March 28, 1964. *J Geophys Res* 73:2009–2025
 58. Mikumo T, Shibutani T, Le Pichon A, Tsuyuki T, Watada S, Garcés M, Fee D, Morii W (2008) Low-Frequency Acoustic-Gravity Waves from Tectonic Deformation Associated with the 2004 Sumatra-Andaman Earthquake (Mw=9.2). *J Geophys Res* 113:B12402, doi:10.1029/2008JB005710
 59. Mutschlecner P, Whitaker R (1998) Infrasonic observations of earthquakes. Tech Rep LA-UR-98-2689, Los Alamos Laboratory, New Mexico
 60. Mutschlecner JP, Whitaker RW (2005) Infrasound from earthquakes. *J Geophys Res* 110:D01108. doi:10.1029/2004JD005067
 61. National Academy of Sciences (2002) Technical Issues Related to the Comprehensive Nuclear Test Ban Treaty. National Academy Press, Washington, International Standard Book Number 0-309-08506-3
 62. Olson JV, Wilson CR, Hansen R (2003) Infrasound associated with the 2002 Denali fault earthquake, Alaska. *Geophys Res Lett* 30:N0232195. doi:10.1029/2003GL018568
 63. Park J, Song T-RA, Tromp J, Okal E, Stein S, Roullet G, Clevede E, Laske G, Kanamori H, Davis P et al (2005) Earth's Free Oscillations Excited by the 26 December 2004 Sumatra-Andaman Earthquake. *Science* 308:1139–1144
 64. Sakai T, Yamasato H, Uehira K (1996) Infrasound accompanying C-type tremor at Sakurajima volcano. *Bull Volcanol Soc Japan* 41:181–185
 65. Shumway RH (2001) Detection and location capabilities of multiple infrasound arrays. *Proceedings of the 23rd NNSA Research Review: Worldwide Monitoring of Nuclear Explosions*, pp 160–167
 66. Sutherland L, Bass H (2004) Atmospheric absorption in the atmosphere up to 160 km. *J Acoust Soc Am* 115:1012–1032
 67. Sutherland L, Bass H (2006) Erratum: Atmospheric absorption in the atmosphere up to 160 km. *J Acoust Soc Am* 120:2985
 68. Tahira M, Nomura M, Sawada Y, Kamo K (1996) Infrasonic and acoustic-gravity waves generated by the Mount Pinatubo eruption of June 15, 1991. In: Newhall C, Punongbayan R (ed) *Fire and Mud - Eruption and Lahars of Mount Pinatubo*, Philippines. Univ. Washington Press, Seattle
 69. Takahashi Y, Koyama Y, Isei T (1994) In situ measured infrasound at Sapporo associated with an earthquake occurring offshore in southwest Hokkaido on July 12, 1993. *J Acoust Soc Jpn* 15:409–411
 70. Thatcher W (1990) Order and diversity in the modes of circum-Pacific earthquake recurrence. *J Geophys Res* 95:2609–2623
 71. Tsuda K, Steidl J, Archuleta R, Assimaki D (2006) Site-Response Estimation for the 2003 Miyagi-Oki Earthquake Sequence Considering Nonlinear Site. *Response Bull Seismol Soc Am* 96(4A):1474–1482. doi:10.1785/0120050160
 72. Vallée M, Bouchon M (2004) Imaging coseismic rupture in far field by slip patches. *Geophys J Int* 156:615–630
 73. Virieux J, Farra V (1991) Ray-tracing in 3D complex isotropic media: an analysis of the problem. *Geophysics* 56:2057–2069
 74. Virieux J, Garnier N, Blanc E, Dessa JX (2004) Paraxial ray-tracing for atmospheric wave propagation. *Geophys Res Lett* 31:L20106. doi:10.1029/2004GL020514
 75. Weber ME, Donn WL (1982) Ducted propagation of Concorde-generated shock waves. *J Acoust Soc Am* 71:340–347
 76. Willis M, Garcés M, Hetzer C, Businger S (2004) Infrasonic observations of open ocean swells in the Pacific: Deciphering the song of the sea. *Geophys Res Lett* 31:L19303
 77. Winchester S (2004) *Krakatoa, The Day the World Exploded*. Harper Perennial
 78. Young JM, Greene GE (1982) Anomalous infrasound generated by the Alaskan earthquake of 28 March 1964. *J Acoust Soc Am* 71:334–339
 79. Yamasato H (1998) Nature of infrasonic pulse accompanying low frequency earthquake at Unzen Volcano. *Japan Bull Volcanol Soc Japan* 43:1–13

Books and Reviews

- Bouchon M, Bouin M-P, Karabulut H, Toksöz MN, Dietrich M, Rosakis AJ (2001) How fast is the rupture during an Earthquake? New insights from the 1999 Turkey Earthquakes. *Geophys Res Lett* 28:2723–2726
- Donn WL, Balachandran NK (1981) Mount St. Helens eruption of 18 May 1980: Air waves and explosive yield. *Science* 213:539–541
- Garcés M, Iguchi M, Ishihara K, Morrissey M, Sudo Y, Tsutsui T (1999) Infrasonic precursors to a Vulcanian eruption at Sakurajima volcano, Japan. *Geophys Res Lett* 26:2537–2540

- Garcés MA, Hansen RA, McNutt SR, Eichelberger J (2000) Application of wave-theoretical seismoacoustic models to the interpretation of explosion and eruption tremor signals radiated by Pavlof volcano, Alaska. *J Geophys Res* 105:3039–3058
- Hagerty M, Schwartz SY, Protti M, Garcés M, Dixon T (1997) Observations at Costa Rican volcano offers clues to causes of eruptions. *EOS Trans Am Geophys Union* 78:565–571
- Harkrider DG (1964) Theoretical and observed acoustic-gravity waves from explosive sources in the atmosphere. *J Geophys Res* 69:5295–5321
- Johnson JB, Aster RC, Ruiz MC, Malone SD, McChesney PJ, Lees JM, Kyle PR (2003) Interpretation and utility of infrasonic records from erupting volcanoes. *J Volc Geotherm Res* 121:15–63
- Kamo K, Ishihara K, Tahira M (1994) Infrasonic and seismic detection of explosive eruptions at Sakurajima Volcano, Japan, and the PEGASAS-VE early warning system. In: Casadevall T (ed) *Proceedings, 1st International Symposium on Volcanic Ash and Aviation Safety*. 8–12 July 1991. US Geological Survey Bulletin 2047, pp 357–365
- Ripepe M, Poggi P, Braun T, Gordeev E (1996) Infrasonic waves and volcanic tremor at Stromboli. *Geophys Res Lett* 23:181–184
- Tahira M (1982) A study of the infrasonic wave in the atmosphere. II, Infrasonic waves generated by the explosions of the volcano Sakura-jima. *J Meteorol Soc Jpn* 60:896–907
- Tahira M, Ishihara K, Iguchi M (1988) Monitoring volcanic eruptions with infrasonic waves. In: *Proceedings of the Kagoshima International Conference on Volcanoes*, Tokyo, 19–23 July 1998. National Institute for Research Advancement, and Kagoshima, Kagoshima Prefectural Government, p 530–533

Inspection Games

RUDOLF AVENHAUS¹, MORTON J. CANTY²

¹ Armed Forces University Munich, Neubiberg, Germany

² Institute for Chemistry and Dynamics of the Geosphere, Forschungszentrum Jülich, Jülich, Germany

Article Outline

Glossary
 Definition
 Introduction
 Selected Inspection Models
 Future Directions
 Bibliography

Glossary

Noncooperative game An n -person noncooperative game in *normal* or *strategic form* is a list of actions, called *pure strategies*, for each of n players, together with a rule for specifying each player's payoff (utility) when every player has chosen a specific action. Each player seeks to maximize her own payoff.

Mixed strategy A mixed strategy for a player in a noncooperative game is a probability distribution over that player's pure strategies.

Extensive form The extensive form of a noncooperative game is a graphical representation which describes a succession of moves by different players, including chance moves, and which can handle quite intricate information patterns.

Zero-sum game A zero-sum game is a noncooperative game in which the payoffs of all players sum to zero for any specific combination of pure strategies.

Nash equilibrium A Nash equilibrium in a noncooperative game is a specification of strategies for all players with the property that no player has an incentive to deviate unilaterally from her specified strategy. A *solution* of a noncooperative game is a Nash equilibrium which is either unique, or which, for some reason, has been selected among alternative equilibria.

Saddle point A saddle point is a Nash equilibrium of a two-person zero-sum game. The *value* of the game is the (unique) equilibrium payoff to the first player.

Utility Utilities are sequences of numbers assigned to the outcomes of any strategy combination which mirror the order of preferences of each player and which fulfill the axioms of von Neumann and Morgenstern.

Deterrence In an inspection game, deterrence is said to be achieved by a Nash equilibrium in which the inspectee behaves legally, or in accordance with the agreed rule.

Inspector leadership Leadership in inspection games is a strategic concept by which, through persuasive announcement of her strategy, the inspector can achieve deterrence.

Verification Verification is the independent confirmation by an inspector of the information reported by an inspectee. It is used most commonly in the context of arms control and disarmament agreements.

Definition

Inspection games deal with the problem faced by an *inspector* who is required to control the compliance of an *inspectee* to some legal or otherwise formal undertaking. One of the best examples of an inspector, in the institutional sense, is the International Atomic Energy Agency (IAEA) which, under a United Nations mandate, verifies the activities of all States – inspectees – signatory to the Nuclear Weapons Non-proliferation Treaty. An inspection regime of this kind is a true conflict situation, even if the inspectee voluntarily submits to inspection (in multinational or bilateral treaties this is invariably the case), because the *raison d'être* of any control authority must be the assumption

that the inspected party has a real incentive to violate its commitments. The primary objective of the inspector is to *deter* the inspectee from illegal behavior or, barring this, to catch him out. It is thus natural that quantitative models of inspections should be non-cooperative games with at least two players, inspector and inspectee(s). This survey will be limited to just one inspectee, that is, we shall restrict ourselves to two-person non-cooperative inspection games.

Inspection games should be distinguished from related topics such as quality control or the prevention of random accidents, for which there are no adversaries that act strategically, or from search-and-destroy problems. The salient feature of an inspection game is that an inspector tries to prevent an inspectee from behaving illegally in terms of some commitment. The inspectee might, for example, decide not to violate, so that there is nothing to search for. In fact, deterrence is generally the inspector's highest priority. Nevertheless, a sharp distinction between, e. g., inspection games and quality control models cannot be made in all cases, as we will see in Subsect. "[Illegal Production](#)".

Introduction

Immediately after von Neumann and Morgenstern's pioneering book *Theory of Games and Economic Behavior* [48], Arms Control and Disarmament (ACD) inspections may have been analyzed game-theoretically as classified military research; this is not known for sure but may be inferred from papers published later. Non-classified work started vigorously in the early 1960s with analyses for the United States Arms Control and Disarmament Agency (ACDA). These dealt with very general ACD problems, and also with concrete problems of test ban treaty verification. In that context probably the first genuine inspection game in the open literature was the recursive game developed by Drescher [18]. Since it was seminal for later work it is presented in some detail in Subsect. "[Customs and Smugglers](#)".

A second phase of inspection game development started around 1968 in connection with the verification of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). There was no model for this verification system, therefore new principles and tools had to be developed and analyzed, see, e. g., [11,12,24]. Because of its importance for the further development of the whole discipline of inspection games, one of the major components of NPT verification measures, namely material accountancy, will be discussed in Subsect. "[Diversion of Nuclear Material](#)".

In Economics game-theoretic work on Accounting and Auditing was begun in the late 1960s. A first survey was given by Borch [13]. Since that time papers and books have been published regularly but on a limited scale along similar lines, placing emphasis on auditing practice, see, e. g., [14,15]; Wilks and Zimbelman [51] provided an updated review of theoretical and empirical research. In economic models known as *principal-agent problems*, in which inspections of economic transactions raise the question of their most efficient design, game-theoretic methods have been applied, early surveys having been given by Baiman [9], Kanodia [28] and Dye [19].

In the last decade, new models have been developed and analyzed again in the context of recent ACD verification developments, in particular the more stringent requirements on re-negotiated NPT verification measures [27]. Whereas under the previous regime purely technical aspects like size of the nuclear fuel cycle and accuracy of the measurement systems were considered, now qualitative features such as behavior and intentions of States had to be taken into account. This required the introduction of State-specific utilities, for first analyses of this kind see, e. g., [3,29]. Independently of concrete applications there has been an ongoing interest of mathematicians in the refinement and generalization of existing models, a few examples will be given below.

Inspections cause conflicts in many real world situations. In economics, there are services of many kinds the fulfillment or payment of which has to be verified. For example one is concerned with the central problem of principal-agent relationships, where the principal, e. g., an employer, delegates work or responsibility to the agent, the employee, and chooses a payment schedule that best exploits the agent's self-interests. The agent, of course, behaves so as to maximize her own utility given the fee schedule proposed by the principal.

Environmental agreements obviously give rise to inspection problems, but these have not yet received as much attention from modelers as one might have expected (and as they might deserve). To date most methodological analyses of inspection games have been performed in the context of arms control and disarmament.

There exist previous reviews of inspection games with objectives somewhat different to those of this survey. Avenhaus et al. [7] restrict discussion to arms control and disarmament, and emphasize the historical development. Avenhaus et al. [8] stress the methodological, and in particular the mathematical aspects. Here we undertake a new approach to organize the material, one which gives more credit to the diversity of the applications and the techniques necessary for their solution. We focus on selected

game theoretic inspection models which, together with their variants and generalizations, we hope will span the full range of the subject.

Selected Inspection Models

In the following, five inspection problems are chosen to illustrate applications of inspection games together with their analysis. They are complemented with discussion of some of their variants and generalizations. In the last illustration, the special role of the leadership concept in inspection models is emphasized.

Passenger Ticket Control

A commuter on the Munich subway is, consciously or unconsciously, involved in a non-cooperative game each time he boards a train. He can buy a valid ticket or travel “schwarz”, risking a fine if he is controlled (checked). In its edition of July 18th, 1996, the daily *Süddeutsche Zeitung* reported the complaint of the Munich City Treasurer to the effect that the deployment of ticket inspectors by the local transit authority (MMV) was not worthwhile, the collected fines paying for only about half the cost of the inspectors themselves.

From the game theorist's viewpoint there is obviously an optimum control intensity that would alleviate this problem: employing just a single inspector would encourage many violations and result in her collecting more than enough fines to pay for herself, but would clearly not be in the MVV's interests. Using an army of inspectors would ensure compliance, but, there now being no fines at all, would not finance the inspectors. The optimum must lie between these two extremes.

Solution The problem can be formulated as a two-person game in *normal form* involving the MVV as player 1 and the transit passenger as player 2 [1]. The pure strategies for the MVV are to control or not to control, whereas the passenger will decide whether or not to buy a valid ticket. Let f be the fare, $b > f$ the fine and $e < b$ the control costs per passenger, all in euros. The game's normal form (also called *bimatrix form*) is shown in Fig. 1.

In the figure, the pure strategies of player 1 (control/no control) are depicted as rows and those of player 2 (legal/illegal) as columns. The payoffs to player 1 for each pure strategy combination are shown in the lower left hand corners of the corresponding squares, those for player 2 in the upper right hand corners. (This simple formulation ignores MVV overhead costs and any material gain the passenger may have from his trip.)

		←	
1 ↙	2	legal q	illegal $1 - q$
	control p	$-f$ $f - e$	$-b$ $b - e$
	no control $1 - p$	$-f$ f	0 0
		→	
		↑	

Inspection Games, Figure 1

Normal form of the two-person game between MVV (player 1) and passenger (player 2). The arrows indicate the preference directions for the two players, the horizontal arrows for player 2, the vertical arrows for player 1

A solution of the game will be a *Nash equilibrium* [35], that is, a pair of strategies, called *equilibrium strategies*, with the property that neither player has an incentive to deviate unilaterally from his or her equilibrium strategy. Equivalently, the strategies are said to be *mutual best replies*. In the Figure the preference directions, i. e., the deviation incentives, are seen to be cyclical. This means that there can be no Nash equilibrium involving pure strategies. However the equilibrium concept can be generalized to involve *mixed strategies* which are probability distributions over the sets of pure strategies. In the present case, the MVV controls with some probability p and the passenger behaves legally with probability q . The expected payoffs to the two players are then given by

$$\begin{aligned} E_1(p, q) &= (f - e)pq + (b - e)p(1 - q) + f(1 - p)q \\ E_2(p, q) &= -fpq - bp(1 - q) - f(1 - p)q. \end{aligned} \quad (1)$$

If we designate the mixed equilibrium strategies by p^* and q^* and the equilibrium payoffs as $E_i^* = E_i(p^*, q^*)$, $i = 1, 2$, then the conditions for Nash equilibrium are

$$\begin{aligned} E_1^* &\geq E_1(p, q^*) \quad \text{for all } p \in [0, 1] \\ E_2^* &\geq E_2(p^*, q) \quad \text{for all } q \in [0, 1]. \end{aligned} \quad (2)$$

For this situation the equilibrium strategies can be determined simply by requiring that each protagonist be made

indifferent with regard to his or her own mixed strategy, see, e. g., Morris [34]. One obtains immediately

$$\begin{aligned} p^* &= \frac{f}{b}, & E_1^* &= f \left(1 - \frac{e}{b}\right), \\ q^* &= 1 - \frac{e}{b}, & E_2^* &= -f. \end{aligned} \quad (3)$$

At equilibrium the passenger behaves illegally with positive probability $1 - q^* = e/b$, that is, deterrence is not possible. We will return to this issue in Subsect. “Sharing Common Pool Resources”. Nevertheless on average he enjoys the same payoff as he would receive by paying his fare every time, namely $-f$.

Remarks The average control expenditure for the MVV is ep , whereas the mean profit from collection of fines is $bp(1 - q)$. The difference is $(e - b(1 - q))p$. If the passenger plays his equilibrium strategy as given by Eq. (3), then

$$(e - b(1 - q^*))p = 0 \quad (4)$$

for any control probability p . The control costs are thus exactly compensated by the collected fines. It might be mentioned that the actual figures for inspection probability and income per passenger for Munich approximately satisfy (3). We may speculate that the City Treasurer’s complaint arises from the fact that regular violators develop strategies to recognize and avoid inspectors. Other complications not taken into account in the model are variations in frequency, hour of day and dwelling time of passengers using the system.

There are many inspection problems which can be described with models equivalent or similar to the one presented here. Inspection of metered parking spaces provides another example. Control of the sharing of common pool resources, as discussed in the last inspection model, below, belongs to the same category.

Illegal Production

In treaties prohibiting the production, acquisition and/or proliferation of weapons of mass destruction, one is often concerned with the misuse of ostensibly legitimate production facilities. A commercial chemical plant, for example, may be used for production of forbidden precursors of chemical weapons, or a uranium enrichment facility may illegally enrich its product to weapons-grade U-235. The consequences of non-detection by an inspecting authority can be dire, and the *timeliness* of control procedures – the interval between the onset of plant misuse and its detection – may be especially important.

To illustrate, consider the following simple model. At the end of some reference time interval, for instance a calendar year or a production campaign, a major inspection takes place at a facility, one which would detect prior illegal production with certainty. Additionally, a single *interim inspection* is carried out, timed at the inspector’s discretion, which will likewise detect prior violation with certainty. The interim inspection is intended to enhance the timeliness of detection should illegal activity be underway. The inspector would like to know precisely when it should take place.

Solution This example entails solution of a *zero-sum game on the unit square*. The onset of illegal production and the time of the interim inspection are chosen on the interval strategically by the respective protagonists, plant operator and inspector. We take the payoff to the former as the time to detection of illegal production, and to the latter to be the negative of that quantity.

Representing the reference time by the interval $[0, 1]$, the operator’s so-called *payoff kernel* is

$$A(y, x) = \begin{cases} y - x & \text{for } x \leq y \\ 1 - x & \text{for } x \geq y, \end{cases} \quad (5)$$

here $x \in [0, 1]$ denotes a pure strategy for the operator, the onset of illegal production and similarly $y \in [0, 1]$ is a pure inspection strategy. Let the inspector’s and operator’s *mixed strategy distribution functions* be $G(y)$ and $F(x)$, respectively. $G(y)$ is the probability of an inspection taking place at time y or earlier, $F(x)$ the probability of illegal activity beginning at time x or earlier. The Nash equilibrium conditions, or, since we are dealing with a zero-sum game, the *saddle point criteria* determining the equilibrium mixed strategies G^* and F^* , are given by

$$E(G^*, x) \leq E(G^*, F^*) \leq E(y, F^*) \quad \text{for all } x, y \in [0, 1]. \quad (6)$$

Since the final inspection is certain, clearly the operator shouldn’t wait too long to act. Rather, in constructing his optimal probability distribution $F^*(x)$, he might plausibly choose x randomly on an interval $[0, b]$ with $b < 1$. Consequently the inspector will not act later than b either. Let us assume that she chooses her inspection time y according to the probability density function $g^*(y)$, where

$$\int_0^b g^*(y) dy = 1. \quad (7)$$

The expected payoff to the operator for diversion at time $x \in [0, b]$ is then

$$\begin{aligned} E(G^*, x) &= \int_0^x (1-x)g^*(y)dy + \int_x^b (y-x)g^*(y)dy \\ &= \int_0^x g^*(y)dy + \int_x^b yg^*(y)dy \\ &\quad - x \int_0^b g^*(y)dy. \end{aligned} \quad (8)$$

But if the operator randomizes across the interval $[0, b]$ as assumed, this payoff must be constant for all $x \in [0, b]$ and equal to the value of the game, i. e., the equilibrium payoff to the operator. If this were not the case for some x in the interval, the operator would not have included it in his mixed strategy F^* in the first place. Thus the derivative of Eq. (8) with respect to x must vanish. This gives immediately

$$g^*(y) = \frac{1}{1-y} \quad (9)$$

and from Eq. (7) $b = 1 - 1/e$. The expected payoff to the operator and value of the game is then easily seen to be $E(G^*, x) = 1/e$ for all $x \in [0, b]$.

Getting the operator's optimal strategy F^* is a bit more subtle because it requires a so-called *atom* at $x = 0$, that is to say, a finite probability of starting illegal production at precisely the beginning of the interval, as well as the probability density $f^*(x)$ on the remaining half-open interval $(0, b]$. In terms of the distribution function $F^*(x)$ that characterizes this mixed strategy, the atom is $F^*(0)$ and $f^*(x)$ is the derivative of $F^*(x)$ on $(0, b]$. The operator's payoff for some inspection time $y \in [0, b]$ is

$$\begin{aligned} E(y, F^*) &= yF^*(0) + \int_0^y (y-x)f^*(x)dx \\ &\quad + \int_y^b (1-x)f^*(x)dx \\ &= yF^*(0) + y \int_0^y f^*(x)dx + \int_y^b f^*(x)dx \\ &\quad - \int_0^b xf^*(x)dx \\ &= (y-1)F^*(y) + F^*(b) - \int_0^b xf^*(x)dx. \end{aligned} \quad (10)$$

Arguing as before, if the the inspector randomizes over the interval, this expression must be constant for all $y \in [0, b]$.

This is true if $(y-1)F^*(y)$ is independent of y . The requirement that $F^*(b) = 1$ then leads to

$$F^*(x) = \frac{1}{e} \cdot \frac{1}{1-x} \quad (11)$$

for $x \in [0, b]$ and $F^*(x) = 1$ for $x > b$. The atom is $F^*(0) = 1/e$, and the construction is complete. Verifying that F^* and G^* satisfy Eq. (6) is straightforward.

Remarks Owen [38] discussed the existence of Nash equilibria for continuous games on the unit square and methods for their solution. The above result was first obtained by Diamond [17], who gave a generalization to $k > 2$ inspections. Prior to Diamond's work, Derman [16] treated a somewhat similar minimax inspection problem.

Both models were motivated by reliability control problems: production units have to be inspected regularly and the earlier a failure is detected, the less costly it is for the production facility owner. Of course the production unit is not acting strategically, but in order to be on the safe side a minimax approach was chosen, which was then generalized by Diamond to give a saddle point solution. Thus we see that an approach which originally was not an inspection game according to our definition in Sect. "Definition" turned out to become one, with interesting applications such as the one discussed above.

Rothenstein and Zamir [42] extended Diamond's model with a single inspection to include errors of the first and second kind. Krieger [30] considered a time-discrete variant of the model. All variants require that both the operator and inspector commit themselves before the reference period begins. Thus if in the solution in Subsect. "Solution" the operator simply waits for the interim inspection and then violates, he will achieve an expected time to detection of

$$\int_0^b (1-x)f^*(x)dx = b = 1 - \frac{1}{e} > \frac{1}{e},$$

and the inspector's advantage will have evaporated. But of course $f^*(x)$ is not the inspector's equilibrium in such a sequential game. Prior commitment may be justified in some cases, but not in others. If there is no requirement for commitment, the operator may prefer to start his illegal action immediately, i. e., at the beginning of the reference period, or delay his decision until the first intermediate inspection. This situation has to be modeled as an *extensive form game*. Its time-continuous version, which also considered errors of the first and second kind, was studied by Avenhaus and Canty [4]. Surprisingly it turned out that an equilibrium strategy of the inspector is a pure strategy, contrary to the equilibrium strategies of the Diamond-type models.

Diversion of Nuclear Material

As already mentioned in the introduction, a large number of game theoretic models of inspection situations have been developed in the framework of IAEA verification activities. The basis of the IAEA inspection system is the verification of the continuing presence of fissile material in the peaceful nuclear fuel cycle of the State under consideration [26], therefore statistical measurement errors and, consequently, statistical decision theory must be taken into account.

Over a single accounting period, typically one year, we define the *material flow* as the measured net transfer of fissile material across the facility boundary, consisting of inputs (receipts of raw material) R and outputs (shipments of purified product and waste) S . If the physical inventory within the facility at the start of the period was I_0 , then the *book inventory* at the end of the accounting period is defined as

$$B = I_0 + R - S = I_0 + Y, \quad (12)$$

where $Y = R - S$ is the net material flow into the facility.

At the end of the period a new physical inventory I_1 is taken and compared with the book inventory,

$$Z = B - I_1 = I_0 + Y - I_1. \quad (13)$$

This expression, which is a random variable as a consequence of measurement error, defines the *material balance statistic*. If there are no unknown losses or diversions of material, its expectation value is

$$E(Z) = E(I_0) + E(Y) - E(I_1) = 0 \quad (14)$$

from conservation of mass. The quantity Z is commonly referred to as MUF, meaning *material unaccounted for*. Its reliable determination forms the basis for the inspector's conclusion with respect to non-diversion.

We shall focus upon a single nuclear facility and an inventory period of one year and pose the question: can the taking of *additional interim inventories* improve the detection sensitivity of the overall accountancy procedure?

Solution The additional inventories essentially define a series of shorter material balance periods, say n in all. At the beginning of the first balance period, the amount I_0 of material subject to control is measured in the facility. Then, during the i th period, $i = 1 \dots n$, some net measured amount Y_i of material enters the area. At the end of that period the amount of material, now I_i , is again measured. The quantity

$$Z_i = I_{i-1} + Y_i - I_i, \quad i = 1 \dots n,$$

is the material balance test statistic for the i th inventory period. Under the *null hypothesis* that no material was diverted, its expected value is, as before,

$$E_0(Z_i) = 0, \quad i = 1 \dots n.$$

The *alternative hypothesis* is that material is diverted from the balance area according to some specific pattern. Thus

$$E_1(Z_i) = \mu_i, \quad i = 1 \dots n, \quad \sum_{i=1}^n \mu_i = \mu > 0,$$

where the amount μ_i diverted in the i th period may be positive, negative or nil, while μ , the total amount of material missing, is hypothesized to be positive.

For the purpose of determining the best test procedure we now define a two-person zero-sum game, wherein the set of strategies of the inspector is the set of all possible test procedures $\{\delta_\alpha\}$, i. e., significance thresholds, for fixed false alarm probability α . The set of strategies of the operator is the set of diversion patterns $\mu = (\mu_1 \dots \mu_n)^T$, $\sum_i \mu_i = \mu$. The payoff to the inspector is the probability of detection $1 - \beta(\delta_\alpha, \mu)$, where $\beta(\delta_\alpha, \mu)$ is the *second kind error probability* (= non-detection probability). A solution of the game is any strategy pair (δ_α^*, μ^*) which satisfies the saddle point conditions

$$1 - \beta(\delta_\alpha^*, \mu) \geq 1 - \beta(\delta_\alpha^*, \mu^*) \geq 1 - \beta(\delta_\alpha, \mu^*) \quad \text{for any } \delta_\alpha, \mu. \quad (15)$$

With the aid of the Lemma of Neyman and Pearson, one of the most fundamental theorems in statistical decision theory, see, e. g., Rohatgi [41], we can derive the following solution first obtained by Avenhaus and Jaech [5]. Suppose that Σ is the covariance matrix of the multivariate normally distributed random vector $Z = (Z_1, Z_2, \dots, Z_n)^T$ and define $e = (1, 1 \dots 1)^T$. Then the equilibrium strategies are in fact given by

$$\mu^* = \frac{\mu}{e^T \cdot \Sigma \cdot e} \cdot \Sigma \cdot e, \quad (16)$$

and by the test δ_α^* characterized by the critical region

$$\{z | e^T \cdot z > k_\alpha\} \quad (17)$$

where k_α is determined by α . The value of the game, that is, the guaranteed probability of detection, is given by

$$1 - \beta(\delta_\alpha^*, \mu^*) = \Phi \left(\frac{\mu}{(e^T \cdot \Sigma \cdot e)^{\frac{1}{2}}} - U(1 - \alpha) \right), \quad (18)$$

where Φ is the normal distribution and U is its inverse.

We can demonstrate that this is the case as follows: Under the null hypothesis of legal behavior

$$1 - \alpha = \text{Prob}_0(\mathbf{Z}^T \cdot \mathbf{e} < k_\alpha) = \Phi\left(\frac{k_\alpha}{(\text{var}(\mathbf{Z}^T \cdot \mathbf{e}))^{\frac{1}{2}}}\right)$$

and therefore

$$\frac{k_\alpha}{(\text{var}(\mathbf{Z}^T \cdot \mathbf{e}))^{\frac{1}{2}}} = U(1 - \alpha).$$

Thus the left hand side of Eq. (15) is fulfilled as equality:

$$\begin{aligned} 1 - \beta(\delta_\alpha^*, \mu) &= \text{Prob}_1(\mathbf{Z}^T \cdot \mathbf{e} > k_\alpha) \\ &= 1 - \Phi\left(\frac{k_\alpha - E_1(\mathbf{Z}^T \cdot \mathbf{e})}{(\text{var}(\mathbf{Z}^T \cdot \mathbf{e}))^{\frac{1}{2}}}\right) \\ &= \Phi\left(\frac{\mu}{(\mathbf{e}^T \cdot \Sigma \cdot \mathbf{e})^{\frac{1}{2}}} - U(1 - \alpha)\right). \end{aligned}$$

As for the right hand side, the critical region which maximizes the detection probability for μ^* and for fixed α is, according to the Lemma of Neyman and Pearson, given by

$$\left\{z \mid \frac{f_1(z)}{f_0(z)} > k'_\alpha\right\} = \{z \mid \mu^{*T} \cdot \Sigma^{-1} \cdot z > k'_\alpha\},$$

where $f_i(z)$ are the joint density functions under hypothesis i , $i = 0, 1$. But from Eq. (16)

$$\mu^{*T} \cdot \Sigma^{-1} \cdot z \propto (\Sigma \cdot \mathbf{e})^T \cdot \Sigma^{-1} \cdot z = \mathbf{e}^T \cdot z.$$

Thus δ_α^* is indeed a best reply to μ^* and the right hand inequality (15) is fulfilled as well.

Remarks According to Eq. (17), the inspector's optimal test statistic is

$$\mathbf{e}^T \cdot \mathbf{Z} = \sum_{i=1}^n Z_i = I_0 + \sum_{i=1}^n Y_i - I_n,$$

which is just the overall material balance for the entire time period involved. All of the intermediate inventories I_i , $i = 1 \dots n - 1$, are *ignored*. This gives a definitive answer to the question as to whether additional inventory taking can improve the sensitivity of the material balancing system for detecting diversion. The answer is no.

Satisfying as it may be from a decision theoretic point of view, this result ignores the aspect of detection time. Waiting one year or one complete production campaign before evaluating the overall material balance may be too long to meet timeliness constraints. Therefore, under the name *near real-time material accountancy*, test procedures

have been discussed which indeed subdivide the year into several inventory periods (at the cost of reduced overall detection sensitivity, as was just explained). To date, it has not been possible to define or to solve satisfactorily a decision theoretic model which takes the critical time aspect into account.

The IAEA safeguards system is organized in such a way that the inspector compares the material balance data reported by the plant operators (via their national authorities) with her own findings and thereafter, if no discrepancies are found, closes the material balance with the help of the operator's reported data. Thus, along with material accountancy, *data verification* comprises the second foundation of the IAEA safeguards system. Due to the possibility that data may be intentionally falsified to make the material balance appear to be correct, data verification again poses game theoretic problems.

Two kinds of sampling procedures have to be considered in this context. In case of identifying items or checking of seals, so-called *attribute sampling* procedures are used in which only sampling errors have to be minimized. This leads on the one hand to stratified sampling solutions similar to those found in the context of accounting and auditing. One of these solutions became well-known, at least in expert circles, under the name IAEA formula, see e.g. [2]. On the other hand, in case of quantitative destructive and non-destructive verification measurements, statistical measurement errors can no longer be avoided, leading to consideration of *variable sampling* procedures. A decision problem arises in this instance, since discrepancies between reported and independently verified data can be caused either by measurement errors or by real and intentionally generated differences (data falsification). Stewart [46] was the first to propose the so-called *D-statistic* for use in IAEA data verification. For one class of data consisting of N items, n of which are verified, the *D-statistic* is the sum of the differences of reported data X_j and independently measured data Y_j , extrapolated to the whole class population, i. e.,

$$D_1 = \frac{N}{n} \sum_{j=1}^n (X_j - Y_j).$$

For K classes of data (for instance one class for each component of a closed material balance) the *D-statistic* is given by

$$D_K = \sum_{i=1}^K \frac{N_i}{n_i} \sum_{j=1}^{n_i} (X_{ij} - Y_{ij}).$$

These quantities then form the basis for the test procedure of the inspector, which goes along similar lines as outlined

before: Two hypotheses have to be formulated which permit the determination of significance thresholds for fixed false alarm probabilities and, from them, the associated detection probabilities.

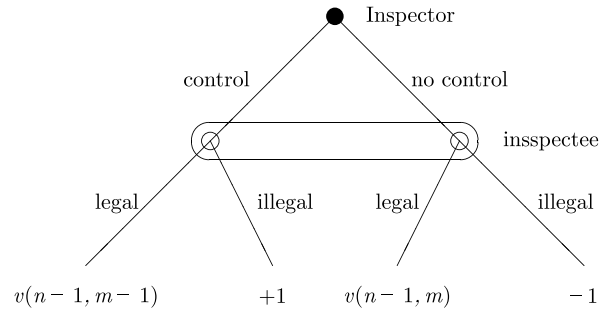
Later on [3] it was proven, again using the saddle point criterion and the Lemma of Neyman and Pearson, that the use of the D -statistic is optimal for a “reasonable” class of data falsification strategies, and it was shown how the sample sizes can be determined such that they maximize the overall probability of detecting a given total falsification for a total given inspection effort.

Customs and Smugglers

In the previous examples evidence of violation (illegal production, diversion) is assumed to persist: the illegal action can be detected after the fact. There are of course inspection problems where this kind of model is not appropriate. Probably the first genuine inspection game in the open literature was a recursive zero-sum game developed by Dresher [18] which treated the situation in which the violator can only be caught red-handed, that is, if the illegal act is actually in progress when an inspection occurs. It's not difficult to imagine real situations where this is the case, a much-discussed example being the customs-smuggler problem [47]. In its simplest form, a smuggler has n nocturnal opportunities to bring his goods safely across a channel. The customs office, equipped with a patrol boat, would very much like to apprehend him, but budget restrictions require that the boat can only patrol on $m < n$ nights. If a patrol coincides with a crossing attempt, the smuggler will be apprehended with certainty. Moreover the smuggler observes all patrols that take place. All that being the case, one can ask how customs should time its patrols.

Solution The game theoretic model developed by Dresher [18] at the RAND Corporation mentioned above fits this situation rather well. It illustrates nicely the special character of sequential games, and has an elegant recursive solution. We summarize it here, see von Stengel [50] for a more thorough discussion as well as some interesting variations on the same theme.

In Dresher's model there are n time periods, during each of which the inspector can decide whether or not to control the inspectee, using up one of a total of m inspections available to her if she does so. The inspectee knows at each stage the number of past inspections. He can violate at most once, but can also choose to behave legally. Detection occurs only if violation and inspection coincide in the same period. The conflicting interests of the two play-



Inspection Games, Figure 2
Dresher's game in reduced extensive form

ers are again modeled as a zero-sum game, that is, the inspectee's loss on detection is the negative of the inspector's gain. Legal action gives a payoff of nil to both players. The game is shown in reduced *extensive form*, i. e., as a decision tree, in Fig. 2.

The inspectee's *information set*, shown as an oval in the figure, encompasses both of her decision points. This is meant to imply that she doesn't know at which node she is situated when choosing her strategy.

The entries at the leaf nodes of the tree are the payoffs to the inspector. The value of the game prior to the first period is denoted $v(n, m)$. If the single violation occurs, the inspector achieves $+1$ if an inspection takes place, otherwise -1 . In the latter case, the game proceeds trivially with the inspectee behaving legally (he has already violated) and the inspector inspecting or not, as she chooses. If the inspectee behaves legally, the continuation of the game has, by definition, value $v(n-1, m-1)$ to the inspector if she decided to control in the first period, otherwise value $v(n-1, m)$. These values are the corresponding payoffs to the inspector after the first period, thus giving the recursive form of the game tree shown. The game terminates after detected violation or after the n periods, in the latter case with a payoff of either 0 (legal behavior) or -1 (illegal behavior) to the inspector.

The function $v(n, m)$ is subject to two boundary conditions. If there are no periods left, and no violation has occurred,

$$v(0, m) = 0, \quad m \geq 0. \quad (19)$$

If the inspector has no inspections left, then the inspectee is aware of this and can safely violate (and will do so, since his payoff is higher):

$$v(n, 0) = -1, \quad n > 0. \quad (20)$$

We shall now seek an equilibrium of the game in the domain of mixed strategies. Let $p(n, m) \in [0, 1]$ be the prob-

ability with which the inspector chooses to inspect in the first period. The equilibrium choice for $p(n, m)$ makes the inspectee indifferent to legal or illegal behavior, so that he receives the same payoff $-v(n, m)$ given by

$$\begin{aligned} v(n, m) &= p(n, m)v(n-1, m-1) \\ &\quad + (1-p(n, m))v(n-1, m) \quad (\text{legal}) \\ v(n, m) &= p(n, m)(+1) + (1-p(n, m))(-1) \quad (\text{illegal}). \end{aligned}$$

In a similar way the inspector chooses her probability $q(n, m)$ for violation at the first stage so as to make the inspectee indifferent as to control or no control, leading to

$$\begin{aligned} q(n, m)v(n-1, m-1) + (1-q(n, mZ))1 \\ = q(n, m)v(n-1, m) + (1-q(n, m))(-1). \end{aligned}$$

Solving these three equations for $p(n, m)$, $q(n, m)$ and $v(n, m)$ we obtain

$$p(n, m) = \frac{v(n-1, m) + 1}{v(n-1, m) + 2 - v(n-1, m-1)}, \quad m < n \quad (21)$$

$$q(n, m) = \frac{2}{v(n-1, m) + 2 - v(n-1, m-1)} \quad (22)$$

and

$$v(n, m) = \frac{v(n-1, m) + v(n-1, m-1)}{v(n-1, m) + 2 - v(n-1, m-1)}. \quad (23)$$

Equation (23) along with the two boundary conditions (19) and (20) determine $v(n, m)$ and therefore $p(n, m)$ and $q(n, m)$ uniquely. The explicit solution is given as follows: Define

$$t(n, m) = \sum_{i=1}^m \binom{n}{i}, \quad (24)$$

where $\binom{n}{i}$ denotes the binomial coefficient. For $0 < m < n$ the value of the game and the equilibrium strategies for both players are

$$v(n, m) = -\frac{\binom{n-1}{m}}{t(n, m)} \quad (25)$$

$$p(n, m) = \frac{t(n-1, m-1)}{t(n, m)} \quad (26)$$

$$q(m, n) = \frac{2}{\frac{\binom{n-2}{m}}{t(n-1, m)} - \frac{\binom{n-2}{m-1}}{t(n-1, m-1)} - 2}. \quad (27)$$

It is not quite trivial to prove that this solution satisfies its determinants. In fact one has to use the following recursive

relations for $t(n, m)$:

$$\begin{aligned} t(n-1, m) &= t(n-1, m-1) + \binom{n-1}{m} \\ t(n, m) &= t(n-1, m) + t(n-1, m-1). \end{aligned}$$

Remarks Since the value of the game is negative, the inspectee will behave illegally with positive probability. (If at some stage in an actual play, the inspector has as many inspections left as there are remaining opportunities, the inspectee will obviously behave legally). Dresher's model uses abstract payoffs – utilities in the sense of von Neumann and Morgenstern [48] – and therefore the zero-sum assumption becomes questionable. In some circumstances it can be argued that detected illegal action is, compared to legal action, *worse for both players*, inspector and inspectee. However, the non-zero-sum version of the game is not much more difficult to analyze and gives no additional structural insight.

Two technical remarks on Dresher's inspection game:

First of all, it works so well only because the concept of *behavioral strategies*, introduced by Kuhn [31], can be applied. Kuhn showed that in a game in extensive form with *perfect recall* (i. e., a game in which both players remember their previous moves) mixed strategies can be replaced by sequences of behavioral strategies. These are probabilities which define the actions of both players at all information sets of the game. In fact, $p(n, m)$ and $q(n, m)$ defined above are behavioral strategies.

Second, the recursive approach requires that at each decision node there follow a *subgame* (i. e. a branch of the tree) which by definition may not be connected by a joint information set to other subgames. This, however is *not* the case in Dresher's game: if the control doesn't take place, the inspector doesn't know whether the inspectee has already behaved illegally and thus whether a true recursive subgame still exists. Fortunately this doesn't matter, since if the inspectee did in fact behave illegally the game is de facto over. Whatever strategy the inspector chooses subsequently will not affect her payoff.

As indicated initially, variants and extensions of Dresher's original model probably represent the largest class of inspection games formulated and analyzed so far, and new variants are still being published, see, e. g., [25,39]. As also mentioned above, Dresher's zero sum assumption may be questioned – detected violation may still be worse for the inspector than no violation at all and, thus, negative for both players. Höpfinger [23] was the first to introduce generalized utilities and to determine the equilibria of the resulting recursive game. Von Stengel [50] considered the case that the inspectee in-

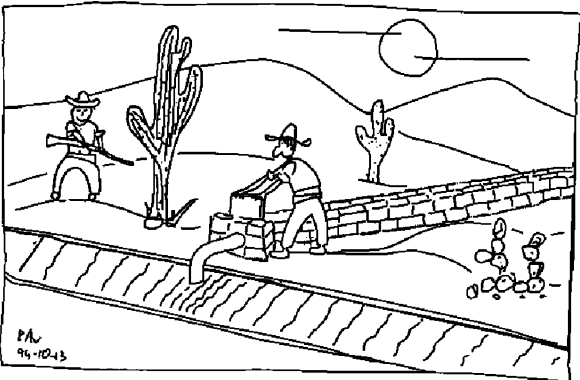
tends to violate more than once, however only once at a given stage. He was able to solve the recursive game for at most k violations on n stages, $k < n$, under the assumption that the inspector is fully informed after each stage whether or not there is a violation, even when no inspection takes place, see also [20,43]. A third category of extensions is associated with the customs-smuggler problem, see e. g. [10,21,22], where customs and the smuggler have more possibilities, e. g., customs has more than one boat, and the smugglers alternative routes. Finally, statistical errors of the first and second kind were considered, firstly by Maschler [33] and later by Rinderle [40]. However, due to the complexity of the models and the many additional assumptions which had to be made, explicit solutions have so far been found only in special cases .

Sharing Common Pool Resources

The sharing of so-called *common-pool resources* (see, e. g., [37]) involves an implicit inspection problem: once agreement has been reached, the parties concerned will want to make sure that the rules are followed. The following idealized common-pool resource problem illustrates this.

The Hatfields and the McCoy's (who else!) farm neighboring areas of irrigated land, the Hatfields having first access to the water. They have agreed on a fair sharing procedure, but there exists a certain amount of mutual distrust. (In fact, the two clans have not been on speaking terms for 40 years.) Broadly speaking, farmer Hatfield has two strategies: to take more water than his fair share, with the twofold incentive of better crops and "doing-in" the McCoy's, or to stand by the agreement. Farmer McCoy, who waits his turn for the water, has the option of controlling Hatfield at some penalty (danger to life and limb) or of simply doing nothing and taking what he gets. The strategic situation is illustrated in Fig. 3.

We may normalize the payoffs to zero for legal behavior and no control. Relative to this normalization, let e be the cost to McCoy of monitoring Hatfield's activities and a the loss he entails by not getting his fair share, but having the satisfaction of catching Hatfield out, $e > 0$, $a > 0$. Thus if Hatfield violates and McCoy monitors, McCoy's payoff is $-a$ relative to the normalization, while if he monitors but there is no violation his payoff is $-e$. Suppose, however, that McCoy's highest priority is to keep Hatfield honest. Then, necessarily, $a > e$. The worst outcome for McCoy is certainly undetected violation, with payoff $-c$, and so $c > a$. Hatfield's payoffs are simply $+d$ for undetected violation, and $-b$ for detected violation, $(b, d) > (0, 0)$. The game is depicted in bimatrix form in Fig. 4.



Inspection Games, Figure 3
The games irrigators play. (Reprinted from [3] with permission of Cambridge University Press)

		→	
1 \ 2		take share $1 - t$	take more t
	no control β	0 0	$+d$ $-c$
	control $1 - \beta$	0 $-e$	$-b$ $-a$
		←	

Inspection Games, Figure 4
Normal or bimatrix form of the irrigation game. Left lower numbers are the payoffs to player 1 (McCoy), upper right numbers the payoffs to player 2 (Hatfield), whereby $c > a > e$ and $(a, b, c, d, e) > (0, 0, 0, 0, 0)$. The horizontal arrows are incentive directions for player 2, the vertical arrows for player 1. The variables t and β define mixed strategies

Solution Just as in the passenger ticket control game, the preference arrows are cyclic. There is again a unique equilibrium in mixed strategies:

$$\beta^* = \frac{b}{b + d}, \quad t^* = \frac{e}{e + c - a}, \tag{28}$$

with payoffs to McCoy and Hatfield, respectively, given by

$$I_M(\beta^*, t^*) = -e \cdot \frac{c}{e + c - a}, \quad I_H(\beta^*, t^*) = 0. \tag{29}$$

Hatfield violates his commitment with probability $t^* > 0$ even though his payoff is the same as for behaving legally.

From a moralist's viewpoint this may not be particularly satisfactory, but Hatfield's equilibrium behavior is, given the circumstances, rational.

However it was postulated that McCoy's highest priority was to keep Hatfield honest, and there is in fact a way for him to do it. Suppose that McCoy takes the initiative and *announces credibly with what precise probability* he intends to monitor Hatfield's activities. This so-called *leadership game* can no longer be expressed as a bimatrix as in Fig. 4, because McCoy's set of pure strategies, that is, his choices of which monitoring probability $1 - \beta$ he will announce in advance, is infinite. Hatfield's set of strategies on the other hand consists of all functions which assign to each value of $1 - \beta$ the decision "take share" or "take more". The appropriate representation is the extensive form game shown in Fig. 5.

Due to its structure (formally, an extensive form game with *perfect information* in which all information sets are singletons) the game can be solved by *backward induction*. If this procedure leads to an equilibrium, then that equilibrium is said to be *subgame perfect*. A subgame perfect Nash equilibrium is one in which every subgame is also a Nash equilibrium.

The first step is to replace the outcome payoffs in Fig. 5 by their expected values, as shown in Fig. 6. The argument

then proceeds as follows: Hatfield, knowing the probability $1 - \beta$ of being controlled, decides

$$\begin{array}{ll} \text{take share} & \text{if } 0 > -b + (b + d)\beta \\ \text{indifferent} & \text{if } 0 = -b + (b + d)\beta \\ \text{take more} & \text{if } 0 < -b + (b + d)\beta, \end{array} \quad (30)$$

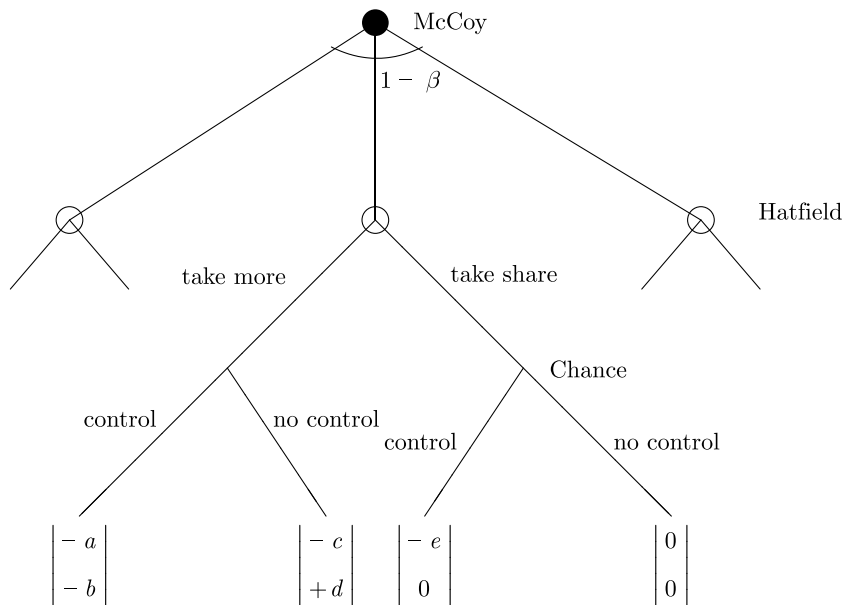
since this strategy will always maximize his expected pay-off. McCoy's equilibrium strategy will be shown below to be

$$\beta^* = \frac{b}{b + d}, \quad (31)$$

so (30) is equivalent to Hatfield's following decision:

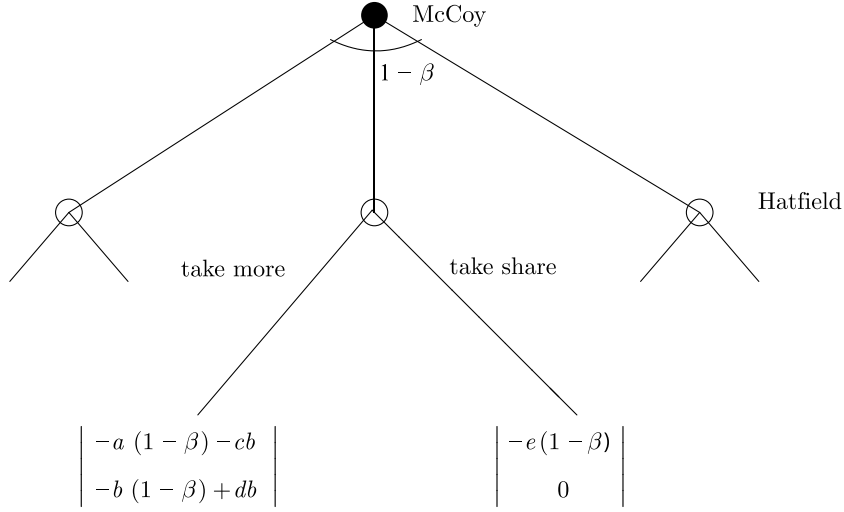
$$\begin{array}{ll} \text{take share} & \text{if } \beta < \beta^* \\ \text{indifferent} & \text{if } \beta = \beta^* \\ \text{take more} & \text{if } \beta > \beta^*. \end{array} \quad (32)$$

What is Hatfield's equilibrium strategy? In order to determine it, we must first define his set of strategies a little more carefully. A typical element of the set will be a recipe which tells him, for every conceivable announcement by McCoy of a value of the non-monitoring probability β , whether or not to take more than his share.



Inspection Games, Figure 5

Extensive form of the inspector leadership irrigation game. McCoy begins by choosing his monitoring probability $1 - \beta$ and announces it to Hatfield, the latter then deciding whether or not to take more than his share of water. Finally, chance decides if the monitoring actually takes place. The payoffs for each possible outcome are shown at the corresponding leaf nodes, McCoy above, Hatfield below



Inspection Games, Figure 6

Reduced extensive form of the inspector leadership irrigation game of Fig. 5

Recipe (30) is certainly one such, although it leaves an ambiguity for $\beta = \beta^*$. In that case Hatfield may decide to make his choice randomly, in other words to use a mixed strategy. In general, such a mixed strategy is given by a probability $t(\beta)$ for 'take more', and the complementary probability $1 - t(\beta)$ for 'take share', for all $\beta \in [0, 1]$. Hatfield's complete strategy set is therefore the set of all functions which map the unit interval onto itself, $\{t(\beta) \mid t: [0, 1] \mapsto [0, 1]\}$.

We now assert that Hatfield's equilibrium strategy t^* is in fact always a pure strategy, namely

$$t^*(\beta) = \begin{cases} 0 = \text{take share} & \text{for } \beta \leq \beta^* \\ 1 = \text{take more} & \text{for } \beta > \beta^* \end{cases}, \quad (33)$$

where β^* is given by (31). This is just the conclusion we reached in Eq. (32) by backward induction, except for the case $\beta = \beta^*$. We still have to show that $t^*(\beta^*) = 0$, that is, that Hatfield stays honest at equilibrium. To do this, we now have to consider McCoy's payoffs.

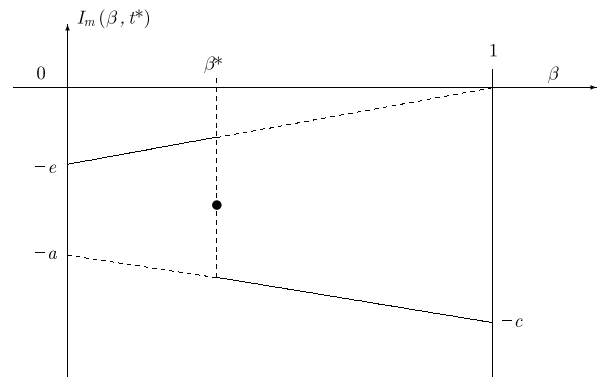
The expected payoff to McCoy, as a function of β , is given by

$$I_M(\beta, t^*) = \begin{cases} -e(1 - \beta) & \text{if } \beta < \beta^* \\ -e(1 - \beta^*)(1 - t(\beta^*)) \\ +(-c\beta^* - a(1 - \beta^*))t(\beta^*) & \text{if } \beta = \beta^* \\ -c\beta - a(1 - \beta) & \text{if } \beta > \beta^* \end{cases}. \quad (34)$$

It is plotted in Fig. 7. McCoy also wishes to maximize his expected payoff. For $\beta < \beta^*$, when Hatfield stays hon-

est, it is at least $-e$ and increases with increasing β . For $\beta > \beta^*$, when Hatfield takes more than his share, it is between $-c$ and $-a$ and certainly worse for McCoy. For $\beta = \beta^*$, McCoy's payoff is something intermediate, depending on Hatfield's behavior. The argument seems to be getting circular, but we are almost done.

McCoy's equilibrium strategy, if he has one at all, has to be β^* as given by (31): for $\beta > \beta^*$ McCoy would do better by choosing alternatively a small β to make Hatfield act honestly, and for $\beta < \beta^*$ he could always do a little better by choosing a larger β , closer to β^* . So the only maximum of his payoff curve, as seen in Fig. 7, is at $\beta = \beta^*$. However, as Eq. (34) shows, the maximum exists only if Hatfield's equilibrium strategy is such that $t^*(\beta^*) = 0$. The



Inspection Games, Figure 7

McCoy's payoff as a function of β according to (34). The ● indicates the equilibrium payoff of the simultaneous game, Eq. (28)

unique subgame perfect equilibrium strategies must therefore be (β^*, t^*) given by (31) and (33), with payoffs

$$I_M(\beta^*, t^*) = -c \cdot \frac{d}{b+d}, \quad I_H(\beta^*, t^*) = 0. \quad (35)$$

Thus McCoy's equilibrium strategy is the same as before, as is Hatfield's payoff, the decisive difference being that Hatfield does not take more than his share of the water.

Remarks The leadership concept was first introduced by von Stackelberg [49] in the context of economic theory, well before game theory became a scientific discipline. In game-theoretic terminology Schelling [44] probably was the first to formulate its importance:

A strategic move is one that influences the other person's choice in a manner favorable to one's self, by affecting the other person's expectations on how one's self will behave. One constrains the partner's choice by constraining one's own behavior. The object is to set up for one's self and communicate persuasively to the other player a mode of behavior (including conditional responses to the other's behavior) that leaves the other a simple maximization problem whose solution for him is the optimum for one's self, and to destroy the other's ability to do the same.

Simaan and Cruz [45] and Wölling [52] refined the concept and formulated conditions for the existence of Nash equilibria in leadership games. Maschler was the first to apply the leadership concept to inspection games [32,33]. Later on it was widely used in the analysis of IAEA verification procedures, in particular for variable sampling inspection problems, see Avenhaus et al. [6].

The importance of deterring the inspectee from illegal behavior, or more positively, of inducing him to behave legally, depends on the specific nature of the problem. For example in the ticket control problem of Subsect. "Passenger Ticket Control", maximization of intake – fares plus fines – is no doubt the highest priority of the transit authority, even though the inspector leadership concept would work here as well, at least in theory. In principal agent models the situation is similar. In the context of arms control and disarmament, on the other hand, deterrence is fundamental: the community of States party to such an agreement have a vital interest that all members adhere to its provisions. There exists a large literature on the subject; a comprehensive survey of the leadership concept in game theory in general has been given by Wölling [52].

Future Directions

We hope that, with our chosen examples, we have given a representative, although certainly not exhaustive, overview of the concepts and models making up the area of inspection games. At the same time we have tried to give some idea of its wide range of applications.

Of course we cannot predict future developments in the field. To be sure, there are still unsolved mathematical problems associated with present inspection models, in particular in arms control and disarmament. For example in Subsect. "Diversion of Nuclear Material" it was pointed out that near real time accountancy poses fundamental difficulties that have not yet been solved satisfactorily. Active research is proceeding and interesting results may be expected.

As mentioned at the outset, in the area of environmental control the number of published investigations is surprisingly small. With the growing awareness of the importance of international agreement on environmental protection the need for effective and efficient control mechanisms will become more and more apparent. Here we expect that the inspection game approach, especially as a means of structuring verification systems, can and will play a useful role. As Barry O'Neill concludes his examination of game theory in peace and war [36]:

... game theory clarifies international problems exactly because they are more complicated. [...] The contribution of game models is to sort out concepts and figure out what the game might be.

Bibliography

1. Avenhaus R (1997) Entscheidungstheoretische Analyse der Fahrgast-Kontrollen. *Der Nahverkehr* 9:27
2. Avenhaus R, Canty MJ (1989) Re-examination of the IAEA formula for stratified attribute sampling. *Proc 11th ESARDA Symposium, JRC, Ispra*, pp 351–356
3. Avenhaus R, Canty MJ (1996) *Compliance Quantified*. Cambridge University Press, Cambridge
4. Avenhaus R, Canty MJ (2005) Playing for time: a sequential inspection game. *Eur J Oper Res* 167(2):474–492
5. Avenhaus R, Jaech JL (1981) On subdividing material balances in time and/or space. *J Inst Nucl Mater Manag* IV(3):24–33
6. Avenhaus R, Okada A, Zamir S (1991) Inspector leadership with incomplete information. In: Selten R (ed) *Game equilibrium models*, vol IV. Springer, Heidelberg, pp 319–361
7. Avenhaus R, Canty MJ, Kilgour DM, von Stengel B, Zamir S (1996) Inspection games in arms control. *Eur J Oper Res* 90:383–394
8. Avenhaus R, von Stengel B, Zamir S (2002) Inspection games. In: Aumann R, Hart S (ed) *Handbook of game theory*. Elsevier, Amsterdam, pp 1947–1987

9. Baiman S (1982) Agency research in managerial accounting: a survey. *J Account Lit* 1:154–213
10. Baston VJ, Bostock FA (1991) A remark on the customs smuggler game. *Nav Res Logist* 41:287–293
11. Bierlein D (1968) Direkte Überwachungssysteme. *Oper Res Verfahren* 6:57–68
12. Bierlein D (1969) Auf Bilanzen und Inventuren basierenden Safeguards-Systeme. *Oper Res Verfahren* 6:36–43
13. Borch K (1990) Economics of insurance. North-Holland, Amsterdam
14. Cavasoglu H, Raghunathan S (2004) Configuration of detection software: a comparison of decision and game theory. *Decis Anal* 1:131–148
15. Cook J, Nadeau L, Thomas LC (1997) Does cooperation in auditing matter? a comparison of a non-cooperative and a cooperative game model of auditing. *Eur J Oper Res* 103:470–482
16. Derman C (1961) On minimax surveillance schedules. *Nav Res Logist Quarterly* 8:415–419
17. Diamond H (1982) Minimax policies for unobservable inspections. *Math Oper Res* 7(1):139–153
18. Drescher M (1962) A sampling inspection problem in arms control agreements: a game theoretical analysis. Memorandum RM-2972-ARPA. RAND Corporation, Santa Monica
19. Dye RA (1986) Optimal monitoring policies in agencies. *RAND J Econ* 17:339–350
20. Ferguson TS, Melolidakis C (1998) On the inspection game. *Nav Res Logist* 45:327–334
21. Garnaev AY (1991) A generalized inspection game. *Nav Res Logist* 28:171–188
22. Goutal P, Garnaev A, Garnaeva G (1997) On the infiltration game. *Int J Game Theory* 26(2):215–221
23. Höpfinger E (1971) A game-theoretic analysis of an inspection problem. University of Karlsruhe. (unpublished manuscript)
24. Höpfinger E (1974) Zuverlässige Inspektionsstrategien. *Z Wahrscheinlichkeitstheorie Verwandte Geb* 31:35–46
25. Hozaki R, Kuhdoh D, Komiya T (2006) An inspection game: taking account of fulfillment probabilities of players. *Nav Res Logist* 53:761–771
26. IAEA (1972) The structure and content of agreements between the agency and states required in connection with the treaty on the non-proliferation of nuclear weapons. IAEA, Vienna, INF/CIRC 153 (corrected)
27. IAEA (1997) Model protocol additional to the Agreement(s) between state(s) and the international atomic energy agency for the application of safeguards. IAEA, Vienna, INF/CIRC 140
28. Kanodia CS (1985) Stochastic and moral hazard. *J Account Res* 23:175–293
29. Kilgour DM (1992) Site selection for on-site inspection in arms control. *Arms Control* 13:439–462
30. Krieger T (2008) On the asymptotic behavior of a discrete time inspection game. *Math Model Anal* 13(1):37–46
31. Kuhn HW (1953) Extensive games and the problem of information. In: Kuhn HW, Tucker AW (eds) *Contributions to the theory of Games*, vol II. Princeton University Press, Princeton, pp 193–216
32. Maschler M (1966) A price leadership method for solving the inspector's non-constant-sum game. *Nav Res Logist* 13:11–33
33. Maschler M (1967) The inspector's non-constant-sum-game: its dependence on a system of detectors. *Nav Res Logist* 14:275–290
34. Morris P (1994) *Introduction to game theory*. Springer, New York
35. Nash JF (1951) Non-cooperative games. *Ann Math* 54:286–295
36. O'Neill B (1994) Game theory models of peace and war. In: Aumann R, Hart S (eds) *Handbook of game theory*. Elsevier, Amsterdam, pp 995–1053
37. Ostrom E, Gardner R, Walker J (1994) *Rules, games and common pool resources*. University of Michigan Press, Ann Arbor
38. Owen G (1968) *Game theory*. W. B. Saunders, Philadelphia
39. Pavlovic L (2002) More on the search for an infiltrator. *Nav Res Logist* 49:1–14
40. Rinderle K (1996) *Mehrstufige sequentielle Inspektionsspiele mit statistischen Fehlern erster und zweiter Art*. Kovac, Hamburg
41. Rohatgi VK (1976) *An introduction to probability theory and mathematical statistics*. Wiley, New York
42. Rothenstein D, Zamir S (2002) Imperfect inspection games over time. *Ann Oper Res* 109:175–192
43. Sakaguchi M (1994) A sequential game of multi-opportunity infiltration. *Math Jpn* 39:157–166
44. Schelling TC (1960) *The strategy of conflict*. Harvard University Press, Cambridge
45. Simaan M, Cruz JB (1973) On the Stackelberg strategy in nonzero-sum games. *J Optim Theory Appl* 11(5):533–555
46. Stewart KB (1971) A cost-effectiveness approach to inventory verification. *Proc of the IAEA Symposium on Safeguards Techniques*, vol II. International Atomic Energy Agency, Vienna, pp 387–409
47. Thomas MU, Nisgav Y (1976) An infiltration game with time-dependent payoff. *Nav Res Logist* 23:297–320
48. von Neumann J, Morgenstern O (1947) *Theory of games and economic behavior*. Princeton University Press, Princeton
49. von Stackelberg H (1934) *Marktform und Gleichgewicht*. Springer, Vienna
50. von Stengel B (1991) *Recursive inspection games*, Report No. S 9106. Computer Science Faculty, Armed Forces University Munich
51. Wilks TJ, Zimbelman MF (2004) Using game theory and strategic reasoning concepts to prevent and detect fraud. *Account Horiz* 18(3):173–184
52. Wölling A (2002) *Das Führerschaftsprinzip bei Inspektionsspielen*. Kovac, Hamburg

Intelligent Control

CLARENCE W. DE SILVA

Department of Mechanical Engineering,

University of British Columbia, Vancouver, Canada

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Soft Computing](#)

[Fuzzy Logic and Fuzzy Sets](#)

[Composition and Inference](#)

Fuzzy Control
 Future Directions
 Bibliography

Glossary

Adaptive system A system that can be automatically modified or reorganized to meet a set of performance specifications.

Artificial intelligence (AI) Computer-generated intelligence or characteristic (externally displayed) of a man-made system that behaves like a naturally intelligent system (biological system) to some degree.

Back-propagation This is a “learning by example” or “supervised learning” technique of a neural network. Using previously obtained input-output data (training data) of a system, a neural network is trained by assigning the training output data to the network output and then computing the corresponding signals in the previous layers of the network, sequentially. The error is computed and minimized in this manner.

Defuzzification The process of converting a fuzzy quantity into a crisp (non-fuzzy) quantity. One method, called the “centroid defuzzification” determines the centroid of the membership function of the fuzzy quantity and uses it as the crisp representation of the fuzzy quantity.

Evolutionary computing A non-analytical optimization technique that mimics biological evolution. Based on an optimization function (fitness function), solutions are selected by “mating” good solutions to create better solutions, and discarding poor solutions.

Expert systems Computer-based intelligent systems which take in data, match to a knowledge base, and generate inferences (solutions, advice, prescriptions, predictions, etc.) in a manner that somewhat mimics a human expert. Expert systems are developed by gathering human expertise in a specific domain and properly representing it in a computer system with a human interface.

Fuzzification The process of converting a crisp quantity into a fuzzy quantity. A membership function is determined to “fuzzify” the crisp quantity. For example, the crisp quantity may form the peak value or centroid of a membership function of sufficient width.

Fuzzy control A model-free control technique where control knowledge is represented by “if-then” statements that contain qualitative or “fuzzy” terms such as “small” and “fast” as present in human statements. System’s outputs are observed (or measured) and

“matched” with the control knowledge base to arrive at control actions.

Fuzzy logic A logic that is more generalized than binary crisp logic. Instead of the two states in binary logic, multiple states are possible, with a degree of overlap among states. Here, for example, the state of “warm” and the state of “hot” have some overlap, as is common with human perception.

Fuzzy set In the same manner that binary logic and Boolean sets go together, fuzzy logic and fuzzy sets are related. A fuzzy set has a non-crisp boundary. Elements that fall on the boundary have some level of presence within the set and a complementary level of presence outside the set.

Genetic algorithms A non-analytical optimization technique that mimics biological evolution. Based on an optimization function (fitness function), solutions are selected by “mating” good solutions to create better solutions, and discarding poor solutions.

Intelligent control A control approach that mimics a human who has expertise to generate a suitable control action for a particular system. A common approach uses a control knowledge base, and an inferencing mechanism which matches observed/measured information with the knowledge base to generate the control action (controller output or control input to the system).

Intelligent machine Machine with an artificial (computer-generated) brain, so that it can behave like a naturally intelligent biological system. It uses techniques of artificial intelligence (AI) for this purpose. Often, the term “machine” refers to a computer. More appropriately, the machine is a physical device that performs an engineering task, and its brain is a computer with AI software.

Knowledge-based system An artificially intelligent system that uses a knowledge base (KB) to represent expert knowledge in a particular application domain. A decision making method (inference engine) is used to “match” data or observed information with the KB to generate inferences.

Membership function A function that represents a fuzzy set. In this function, the membership of a quantity in a fuzzy set is represented by a numerical value between 0 and 1. Set elements that are clearly within the set are represented by a membership grade of 1 and elements that are clearly outside the set are represented by a membership grade of 0. Elements that lie on the set boundary have membership grades between 0 and 1, as determined by the degree of membership in the set.

Neural networks Massively parallel networks of artificial neurons that represent highly nonlinear systems or processes, without using analytical models. By adjusting the network parameters, the behavior of the network is made to resemble the actual system. It somewhat resembles the activity of a biological brain.

Probabilistic system A system that possesses a degree of randomness or uncertainty and uses methods that involve probability distribution functions to generate decisions, actions, or estimates.

Soft computing An approach that uses one or more techniques of fuzzy logic, neural networks, evolutionary computing, and probabilistic methods to perform numerical operations by somewhat mimicking biological systems.

Definition of the Subject

An intelligent controller may be interpreted as a computer-based controller that can somewhat “emulate” the reasoning procedures of a human expert in the specific area of control, to generate the necessary control actions. Here, techniques from the field of *artificial intelligence (AI)* are used for the purpose of acquiring and representing knowledge and for generating control decisions through an appropriate reasoning mechanism. With steady advances in the field of AI, especially pertaining to the development of practical *expert systems* or *knowledge systems*, there has been a considerable interest in using AI techniques for controlling complex processes. Complex engineering systems use intelligent control to cope with situations where conventional control techniques are not effective.

Intelligent control depends on efficient ways of representing and processing the control knowledge [5]. Specifically, a knowledge base has to be developed and a technique of reasoning and making “inferences” has to be available. Knowledge-based intelligent control relies on knowledge that is gained by intelligently observing, studying, or understanding the behavior of a plant, rather than explicitly modeling the plant, to arrive at the control action. In this context, it also heavily relies on the knowledge of experts in the domain, and also on various forms of general knowledge. Modeling of the plant is implicit here. Soft computing is an important branch of study in the area of intelligent and knowledge-based systems. It has effectively complemented conventional AI in the area of machine intelligence (computational intelligence). Fuzzy logic, probability theory, neural networks, and genetic algorithms are cooperatively used in soft computing for knowledge representation and for mimicking the reasoning and decision-

making processes of a human. Decision making with soft computing involves *approximate reasoning*, and is commonly used in intelligent control. This chapter presents the background theory, concepts, and development of intelligent control, with a particular emphasis on fuzzy logic control.

Introduction

Future generations of industrial machinery, plants, and decision support systems may be expected to carry out round-the-clock operation, with minimal human intervention, in manufacturing products or providing services. It will be necessary that these systems maintain consistency and repeatability of operation and cope with disturbances and unexpected variations within the system, its operating environment, and performance objectives. In essence, these systems should have the capability to accommodate rapid reconfiguration and adaptation. For example, a production machine should be able to quickly cope with variations ranging from design changes for an existing product to the introduction of an entirely new product line. This will call for tremendous flexibility and some level of autonomous operation in automated machines, which translate into a need for a higher degree of intelligence in the supporting devices. Smart systems will exhibit an increased presence and significance in a wide variety of engineering applications. Products with a “brain” are found, for example, in household appliances, consumer electronics, space technology, transportation systems, industrial processes, manufacturing systems, and services. There is clearly a need to incorporate a greater degree of intelligence and a higher level of autonomy into automated machines and systems. This will require the appropriate integration of such devices as sensors, actuators, and controllers, which themselves may have to be “intelligent” and furthermore, appropriately distributed throughout the system. Design, development, production, and operation of intelligent machines have been possible today through ongoing research and development in the field of intelligent systems and intelligent control [3,4,6,10,13].

Soft Computing

Soft computing is an important branch of study in the area of intelligent and knowledge-based systems. It has effectively complemented conventional AI in the area of machine intelligence (computational intelligence). Human reasoning is predominantly approximate, qualitative, and “soft.” Humans can effectively handle incomplete, imprecise, and fuzzy information in making intelligent decisions. Fuzzy logic, probability theory, neural networks,

and genetic algorithms are cooperatively used in soft computing for knowledge representation and for mimicking the reasoning and decision-making processes of a human [5,7]. Quite effective are the mixed or hybrid techniques, which synergistically exploit the advantages of two or more of these areas. Decision making with soft computing involves *approximate reasoning*. Now we will give an introduction to the subject of soft computing. Fuzzy logic and its use in intelligent control will be covered in detail in subsequent sections.

Fuzzy Logic

Fuzzy logic is useful in representing human knowledge in a specific domain of application and in reasoning with that knowledge to make useful inferences or actions [14]. The conventional binary (bivalent) logic is crisp and allows for only two states. This logic cannot handle fuzzy descriptors, examples of which are “fast” which is a *fuzzy quantifier* and “weak” which is a *fuzzy predicate*. They are generally qualitative, descriptive, and subjective and may contain some overlapping degree of a neighboring quantity, for example, some degree of “slowness” in the case of the fuzzy quantity “fast.” Fuzzy logic allows for a realistic extension of binary, crisp logic to qualitative, subjective, and approximate situations, which often exist in problems of intelligent machines where techniques of artificial intelligence are appropriate.

In fuzzy logic, the knowledge base is represented by if-then rules of fuzzy descriptors. Consider the general problem of approximate reasoning. In this case the knowledge base K is represented in an “approximate” form, for example, by a set of if-then rules with *antecedent* and *consequent* variables that are fuzzy descriptors. First, the data D are preprocessed according to

$$F_D = FP(D) \quad (1)$$

which, in a typical situation, corresponds to a data abstraction procedure called “fuzzification” and establishes the membership functions or membership grades that correspond to D . Then for a fuzzy knowledge base F_K , the fuzzy inference F_I is obtained through fuzzy-predicate approximate reasoning, as denoted by

$$F_I = F_K \circ F_D . \quad (2)$$

This uses a *composition* operator “ \circ ” for fuzzy matching of data (D) with the knowledge base (K), and making inferences (I) on that basis.

Fuzzy logic is commonly used in “intelligent” control of processes and machinery. In this case the inferences of

a fuzzy decision making system are the control inputs to the process. These inferences are arrived at by using the process responses as the inputs (context data) to the fuzzy decision-making system.

Neural Networks

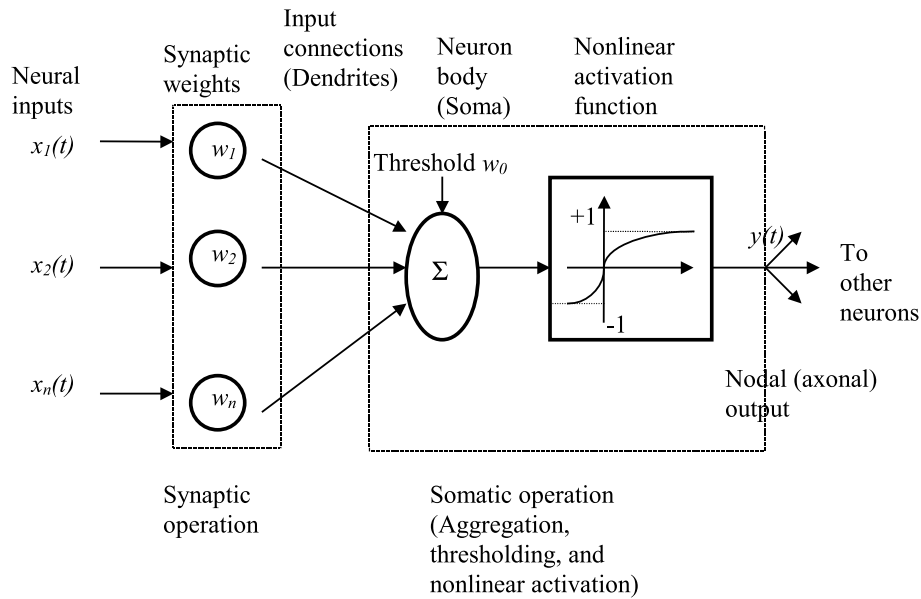
Artificial neural networks (NN) are massively connected networks of computational “neurons,” and represent parallel-distributed processing structures. The inspiration for NN has come from the biological architecture of neurons in human brain. A key characteristic of neural networks is their ability to approximate arbitrary nonlinear functions. Since machine intelligence involves a special class of highly nonlinear decision making, neural networks would be effective there. Furthermore, the process of approximation of a nonlinear function (i. e., system identification) by interacting with a system and employing data on its behavior, may be interpreted as “learning.” Through the use of neural networks, an intelligent system would be able to learn and perform high-level cognitive tasks. For example, an intelligent system would only need to be presented with a goal; it could achieve its objective through continuous interaction with its environment and evaluation of the responses by means of neural networks [1].

A neural network consists of a set of nodes, usually organized into layers, and connected through weight elements called synapses. At each node, the weighted inputs are summed (aggregated), thresholded, and subjected to an activation function in order to generate the output of that node. These operations are shown in Fig. 1. The analogy to the operations in a biological neuron is highlighted. Specifically, in a biological neuron, the dendrites receive information from other neurons. The soma (cell body) collects and combines this information, which is transmitted to other neurons using a channel (tubular structure) called axon. This biological analogy, apart from the abilities to learn by example, approximation of highly nonlinear functions, massive computing power, and memory, may be a root reason for inherent “intelligence” in a neural network. If the weighted sum of the inputs to a node (neuron) exceeds a threshold value w_0 , then the neuron is fired and an output $y(t)$ is generated according to

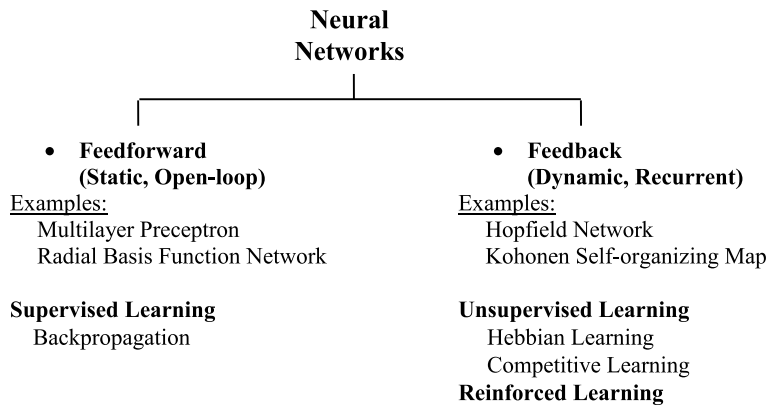
$$y(t) = \Psi \sum_{i=1}^n w_i x_i - w_0 , \quad (3)$$

where x_i are neuron inputs, w_i are the synaptic weights, and $\Psi[.]$ is the activation function.

As indicated in Fig. 2, there are two main classes of neural networks known as feedforward networks (or,



Intelligent Control, Figure 1
The operations at a node of a neural network

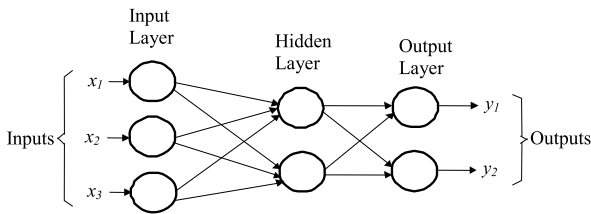


Intelligent Control, Figure 2
Classification of neural networks

static networks) and feedback networks (or, recurrent networks). In feedforward network, an example of which is a multilayer perceptron, the signal flow from a node to another node takes place in the forward direction only. There are no feedback paths. Figure 3 shows a multilayer perceptron consisting of an input layer, a hidden layer, and an output layer. Another example of feedforward network is the radial basis function network. Here there are only three layers. Furthermore, only the hidden layer uses nonlinear activation functions in its nodes. These functions are called radial basis functions, the Gaussian distribution function being a popular example. These functions form the basis for the capability of the NN to approximate

any nonlinear function. In a feedforward neural network, learning is achieved through example. This is known as supervised learning. Specifically, first a set of input-output data of the actual process is determined (e. g., by measurement). The input data are fed into the NN. The network output is compared with the desired output (experimental data) and the synaptic weights of the NN are adjusted using a gradient (steepest descent) algorithm until the desired output is achieved.

In a feedback NN, the outputs of one or more nodes (say, in the output layer) are fed back to one or more nodes in a previous layer (say hidden layer or input layer) or even to the same node. The feedback provides the capability of



Intelligent Control, Figure 3

A feedforward neural network (Multilayer perceptron)

“memory” to the network. An example is the Hopfield network. It consists of two layers: the input layer and the Hopfield layer. Each node in the input layer is directly connected to only one node in the Hopfield layer. The outputs of the network are fed back to the input nodes via a time delay (providing memory) and synaptic weights. Nodes in the Hopfield layer have nonlinear activation functions such as sigmoidal functions.

Feedback neural networks commonly use unsupervised learning algorithms. In these learning schemes, the synaptic weights are adjusted based on the input values to the network and the reactions of individual neurons, and not by comparing the network output to the desired output data. Unsupervised learning is called self-organization (or open-loop adaptation), because the output characteristics of the network are determined internally and locally by the network itself, without any data on desired outputs. This type of learning is particularly useful in pattern classification and grouping of data. Hebbian learning and competitive learning are examples of unsupervised learning algorithms. In the Hebbian learning algorithm, the weight between a neuron and an input is strengthened (increased) if the neuron is fired by the input. In competitive learning, weights are modified to enhance a node (neuron) having the largest output. An example is the Kohonen network, which uses a winner-takes-all approach. A Kohonen network has two layers, the input layer and the output layer (Kohonen layer). In the operation of the network, the node in the Kohonen Layer with weights that most closely resemble the current input, is assigned an output of 1 (the winner), and the outputs of all the remaining nodes are set to zero. In this manner, the input nodes organize themselves according to the pattern of the input data while the output nodes compete among themselves to be activated.

Genetic Algorithms

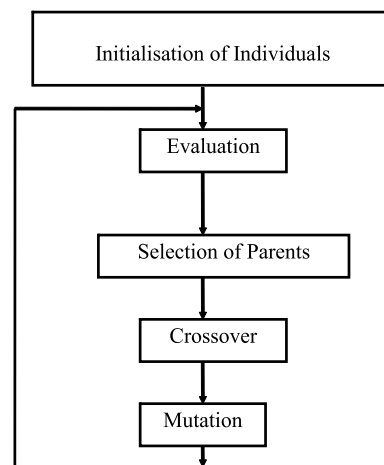
Genetic algorithms (GA) are derivative-free optimization techniques, which can evolve through procedures analogous to biological evolution. Genetic algorithms belong to the area of evolutionary computing. They represent an op-

timization approach where a search is made among a set of solutions (solution space) to “evolve” a solution algorithm, which will retain the “most fit” solutions, by using a procedure that is analogous to biological evolution through natural selection, crossover, and mutation. In the present context of intelligent machines, intellectual fitness rather than physical fitness is what is important for the evolutionary process. Evolutionary computing can play an important role in the development of an optimal and self-improving intelligent machine [5,7].

Evolutionary computing has the following characteristics:

1. It is based on multiple searching points or solution candidates (population based search).
2. It uses evolutionary operations such as selection, crossover and mutation.
3. It is based on probabilistic operations.

The basic operations of a genetic algorithm are indicated in Fig. 4. The algorithm works with a population of individuals, each representing a possible solution to a given problem. Each individual is assigned a fitness score according to how good its solution is to the problem. The highly fit (in an intellectual sense) individuals are given opportunities to reproduce by crossbreeding with other individuals in the population. This produces new individuals as offspring, who share some features taken from each parent. The least fit members of the population are less likely to get selected for reproduction and will eventually die out. An entirely new population of possible solutions is produced in this manner, by mating the best individuals (i. e., individuals with best solutions) from the current



Intelligent Control, Figure 4

The key operations of a genetic algorithm

generation. The new generation will contain a higher proportion of the characteristics possessed by the “fit” members of the previous generation. In this way, over many generations, desirable characteristics are spread throughout the population, while being mixed and exchanged with other desirable characteristics, in the process. By favoring the mating of the individuals who are more fit (i.e., who can provide better solutions), the most promising areas of the search space would be exploited. A GA determines the next set of searching points using the fitness values of the current searching points, which are widely distributed throughout the searching space. It uses the mutation operation to escape from a local minimum.

Two important activities of a GA are selection and reproduction. Selection is the operation which will choose parent solutions. New solution vectors in the next generation are calculated from them. Since it is expected that better parents generate better offspring, parent solution vectors that possess higher fitness values will have a higher probability of selection. There are several methods of selection. In the method of roulette wheel selection, the probability of winning is proportional to the area rate of a chosen number on a roulette wheel. In this manner, the selection procedure assigns a selection probability to individuals in proportion to their fitness values. In the elitist strategy, the best parents are copied into the next generation. This strategy prevents the best fitness value of the offspring generation from becoming worse than that in the present generation.

During the reproductive phase of a GA, individuals are selected from the population and recombined, producing offspring, which in turn will make up the next generation. Parents are selected randomly from the population using a scheme that favors the individuals who are more fit. After two parents are selected, their chromosomes are recombined using the mechanism of crossover and mutation. Crossover takes two individuals and cuts their chromosome strings at some randomly chosen position to produce two “head” segments and two “tail” segments. The tail segments are then swapped over to produce two new full-length chromosomes. Each of the two offspring will inherit some genes from each parent. This is known as a single-point crossover. Crossover is not usually applied to all pairs of individuals that are chosen for mating. A random choice is made, where the likelihood of the crossover being applied is typically between 0.6 and 1.0. Mutation is applied individually to each child, after crossover. It randomly alters each gene at a very low probability. Mutation provides a small degree of random search and helps ensure that every point in the search space has some probability of being examined.

Probabilistic Reasoning

Uncertainty and the associated concept of probability are linked to approximation. One can justify that probabilistic reasoning should be treated within the area of soft computing [5,7]. Probabilistic approximate reasoning may be viewed in an analogous manner to fuzzy-logic reasoning, considering uncertainty in place of fuzziness as the concept of approximation that is applicable. Probability distribution/density functions are employed in place of membership functions. The formula of knowledge-based decision making that corresponds to Eq. (2), in this case, depends on the specific type of probabilistic reasoning that is employed. The *Bayesian approach* is commonly used. This may be interpreted as a classification problem.

Suppose that an observation d is made, and it may belong to one of several classes c_i . The Bayes' relation states:

$$\max_i P(c_i|d) = \frac{P(d|c_i) \bullet P(c_i)}{P(d)}, \quad (4)$$

where

$P(c_i|d)$ = given that the observation is d , the probability that it belongs to class c_i (the a posteriori *conditional probability*)

$P(d|c_i)$ = given that the observation belongs to the class c_i , the probability that the observation is d (the *class conditional probability*)

$P(c_i)$ = the probability that a particular observation belongs to class c_i , without knowing the observation itself (the a priori *probability*)

$P(d)$ = the probability that the observation is d without any knowledge of the class.

In the Bayes' decision-making approach, for a given observation (data) d , the posteriori probabilities $P(c_i|d)$ are computed for all possible classes ($i = 1, 2, \dots, n$), using Eq. (4). The class that corresponds to the largest of these a posteriori probability values is chosen as the class of d , thereby solving the classification problem. The remaining $n - 1$ a posteriori probabilities represent the error in this decision.

Note the analogy between Eqs. (4) and (2). Specifically, $P(d)$ represents the “preprocessed” probabilistic data that correspond to the observation d . The knowledge base itself constitutes the two sets of probabilities:

1. $P(d|c_i)$ of occurrence of data d if the class (decision) is c_i , $i = 1, 2, \dots, n$
2. $P(c_i)$ of class (decision) c_i , $i = 1, 2, \dots, n$ without any knowledge of the observation (data) itself.

The knowledge-base matching is carried out, in this case, by computing the expression on the right side of Eq. (4)

Intelligent Control, Table 1

Techniques of computational intelligence

Technique	Characteristic	A Popular Analogy
Fuzzy Logic	Uses fuzzy rules and approximate reasoning	Human knowledge
Neural Networks	Network of massively connected nodes	Neuron structure in brain
Genetic Algorithms	Derivative-free optimization	Biological evolution
Probability	Incorporates uncertainty in predicting future events	Random action of a human
Conventional AI	Symbolic processing of information	Symbolic languages

for all possible i and then picking out the maximum value. Application areas of probabilistic decision making include forecasting, signal analysis and filtering, and parameter estimation and system identification.

Techniques of soft computing are powerful by themselves in achieving the goals of machine intelligence. Furthermore, they have a particular appeal in view of the biological analogies that exist. To summarize the biological analogies of fuzzy, neural, genetic, and probabilistic approaches: fuzzy techniques attempt to approximate human knowledge and the associated reasoning process; neural networks are a simplified representation of the neuron structure of a brain; genetic algorithms follow procedures that are crudely similar to the process of evolution in biological species; and probabilistic techniques can analyze random future action of a human. This is no accident because in machine intelligence it is the behavior of human intelligence that would be mimicked. Popular (e.g., biological) analogies and key characteristics of several techniques of machine intelligence are listed in Table 1. Conventional AI, which typically uses symbolic and descriptive representations and procedures for knowledge-based reasoning, is listed as well for completeness.

Fuzzy Logic and Fuzzy Sets

In the crisp Boolean logic, truth is represented by the state 1 and falsity is by the state 0. Boolean algebra and crisp logic have no provision for approximate reasoning. Fuzzy logic is an extension of crisp bivalent (two-state) logic in the sense that it provides a platform for handling approximate knowledge. Fuzzy logic is based on fuzzy set theory in a manner similar to how crisp bivalent logic is based on crisp set theory. Fuzzy logic provides an approximate yet practical means of representing knowledge regarding a system (e.g., describing the behavior of the system) that is too complex or ill-defined and not easy to tackle using precise mathematical means. The approach also provides a means of making inferences using that knowledge, which can be used in making correct decisions regarding the sys-

tem and for carrying out appropriate actions. In particular, human-originated knowledge can be effectively handled using fuzzy logic. As the complexity of a system increases, the ability to develop precise analytical models of the system diminishes until a threshold is reached beyond which analytical modeling becomes intractable. Under such circumstances, precise model-based decision making is not practical. Fuzzy knowledge-based decision making is particularly suitable then [9].

This section introduces fuzzy logic, which uses the concept of fuzzy sets. Applications of soft computing and particularly fuzzy knowledge-based systems rely on the representation and processing of knowledge using fuzzy logic. In particular, the *compositional rule of inference*, as presented in the subsequent section, is what is applied in decision-making with fuzzy logic. Some popular applications of fuzzy decision making and intelligent control are given next.

Process Temperature Control (OMRON): This is a proportional-integral-derivative (PID) controller integrated with a fuzzy controller. When temperature control is initiated, only the PID action is present. If the temperature overshoots the set point, the fuzzy controller takes over. This control system responds well to disturbance.

Air Conditioner (Mitsubishi): Conventional air conditioning systems use on-off controllers. Specifically when the temperature drops below a preset level, the unit is automatically turned off. When the temperature rises above a preset level, the unit is turned on. The former preset value is slightly lower than the latter preset value, providing a dead zone, so that high-frequency on-off cycling (chatter) is avoided. The thermostat in the system controls the on-off action. For example, “when the temperature rises to 25°C, turn on the unit, and when the temperature falls to 20°C, turn off the unit.” The Mitsubishi air conditioner is controlled by using fuzzy rules such as: “If the ambient air is getting warmer, turn the cooling power up a little; If the air is getting chilly, turn the power down moderately,” etc.

The operation becomes smoother as a result. This results in less wear and tear of the air conditioner, more consistent comfortable room temperatures, less noise, and increased efficiency (energy savings).

Vacuum Cleaner (Panasonic): Characteristics of the floor and the amount of dust are sensed by an infrared sensor, and the microprocessor selects the appropriate power by fuzzy control according to these characteristics. The floor characteristics include the type and nature (hardwood, cement, tile, carpet softness, carpet thickness, etc.). The changing pattern of the amount of dust passing through the infrared sensor is established as well. The microprocessor establishes the appropriate setting of the vacuum head and the power of the motor, using a fuzzy control scheme. Red and green lamps of the vacuum cleaner show the amount of dust left on the floor.

Automatic Transmission System (Nissan, Subaru, Mitsubishi): In a conventional automatic transmission system, electronic sensors measure the vehicle speed and throttle opening, and gears are shifted based on the predetermined values of these variables. According to Nissan, this type of system is incapable of uniformly providing satisfactory control performance to a driver because it typically provides only three different shift patterns. The fuzzy control transmission senses several variables including vehicle speed and acceleration, throttle opening, the rate of change of throttle opening, engine load, and driving style. Each sensed value is given a weight, and a fuzzy aggregate is calculated to decide whether to shift gears. This controller is said to be more flexible, smooth, and efficient, providing better performance. Also, an integrated system developed by Mitsubishi uses fuzzy logic for active control of the suspension system, four-wheel drive (traction), steering, and air conditioning.

Washing Machine (Matsushita, Hitachi): The control system senses both quality and quantity of dirt, load size, and fabric type, and adjusts the washing cycle and detergent amount accordingly. Clarity of water in the washing machine is measured by light sensors. At the start of the cycle, dirt from clothes would not have yet reached the water; so light will pass through it easily. The water becomes more discolored as the wash cycle proceeds, and less light will pass through. This information is analyzed and control decisions are made using fuzzy logic.

Camcorder (Panasonic, Sanyo, Fisher, Canon): The intelligent video camera determines the best focus and lighting, particularly when several objects are present in the picture. Also, it has a digital image stabilizer to

remove hand jitter. Fuzzy decision making is used in these actions. For example, the following scheme is used for image stabilization. The present image frame is compared with the previous frame from memory. A typically stationary object (e.g., house) is identified and its shift coordinates are computed. This shift is subtracted from the image to compensate for the hand jitter. A fuzzy algorithm provides a smooth control/compensation action.

Elevator Control (Fujitec, Toshiba): A fuzzy scheme evaluates passenger traffic and the elevator variables (load, speed, etc.) to determine car announcement and stopping time. This reduces waiting time and improves the efficiency and reliability of operation.

Handheld Computer (Sony): A fuzzy logic scheme reads hand-written input and interprets the characters for data entry.

Television (Sony): A fuzzy logic scheme uses sensed variables such as ambient lighting, time of day, and user profile, and adjusts such parameters as screen brightness, color, contrast, and sound.

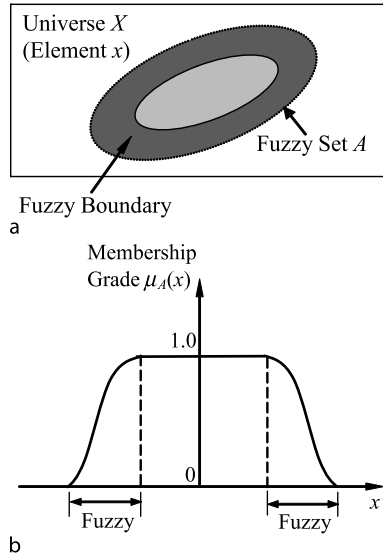
Antilock Braking System (Nissan): The system senses wheel speed, road conditions, and driving pattern, and the fuzzy ABS determines the braking action, with skid control.

Subway Train (Sendai): A fuzzy decision scheme is used by the subway trains in Sendai, Japan, to determine the speed and stopping routine. Ride comfort and safety are used as performance requirements.

Other applications of fuzzy logic include hot water heater (Matsushita), rice cooker (Hitachi), and cement kiln (Denmark). Also see [3,4,6,8,10,11,12,13].

The theory of fuzzy logic can be developed using the concepts of fuzzy sets similar to how the theory of crisp bivalent logic can be developed using the concepts of crisp sets. Specifically, there exists an isomorphism between sets and logic. In view of this, a good foundation in fuzzy sets is necessary to develop the theory of fuzzy logic. In this section, the mathematics of fuzzy sets is presented.

A fuzzy set is a set without clear or sharp (crisp) boundaries or without binary membership characteristics. Unlike an ordinary set where each object (or, element) either belongs or does not belong to the set, a partial membership in a fuzzy set is possible. In other words, there is a “softness” associated with the membership of elements in a fuzzy set. An example of a fuzzy set could be “the set of narrow streets in Vancouver.” There are streets that clearly belong to the above set, and others that cannot be considered as narrow. Since the concept of “narrow” is not precisely defined (for example, < 2 m), there will be a “gray”



Intelligent Control, Figure 5

a Venn diagram of a fuzzy set. **b** The membership function of a fuzzy set

zone in the associated set where the membership is not quite obvious. As another example, consider the variable “temperature.” It can take a fuzzy value (e. g., cold, cool, tepid, warm, hot). A *fuzzy value* such as “warm” is a *fuzzy-descriptor*. It may be represented by a fuzzy set because any temperature that is considered to represent “warm” belongs to this set and any other temperature does not belong to the set. Still, one cannot realistically identify a precise temperature interval (e. g., 25°C to 30°C), which is a crisp set, to represent warm temperatures.

Let X be a set that contains every set of interest in the context of a given class of problems. This is called the *universe of discourse* (or, simply universe), whose elements are denoted by x . A fuzzy set A in X may be represented by a Venn diagram as in Fig. 5a. Generally, the elements x are not numerical quantities. For analytical convenience, however, the elements x are assigned real numerical values.

Membership Function

A fuzzy set may be represented by a membership function. This function gives the grade (degree) of membership within the set, of any element of the universe of discourse. The membership function maps the elements of the universe on to numerical values in the interval $[0, 1]$. Specifically,

$$\mu_A(x): X \rightarrow [0, 1], \quad (5)$$

where $\mu_A(x)$ is the membership function of the fuzzy set A in the universe in X . Stated in another way, fuzzy set if A is a set of ordered pairs:

$$A = \{(x, \mu_A(x)); \quad x \in X, \quad \mu_A(x) \in [0, 1]\}. \quad (6)$$

The membership function $\mu_A(x)$ represents the grade of possibility that an element x belongs to the set A . It follows that a membership function is a *possibility function* and not a probability function. A membership function value of zero implies that the corresponding element is definitely not an element of the fuzzy set. A membership function value of unity means that the corresponding element is definitely an element of the fuzzy set. A grade of membership greater than 0 and less than 1 corresponds to a non-crisp (or fuzzy) membership, and the corresponding elements fall on the fuzzy boundary of the set. The closer the $\mu_A(x)$ is to 1 the more the x is considered to belong to A , and similarly the closer it is to 0 the less it is considered to belong to A . A typical membership function is shown in Fig. 5b.

Note: A crisp set is a special case of fuzzy set, where the membership function can take the two values 1 (membership) and 0 (non-membership) only. The membership function of a crisp set is given the special name *characteristic function*.

Fuzzy Logic Operations

It is well known that the “complement”, “union”, and “intersection” of crisp sets correspond to the logical operations NOT, OR, and AND, respectively, in the corresponding crisp, bivalent logic. Furthermore, it is known that, in the crisp bivalent logic, the union of a set with the complement of a second set represents an “implication” of the first set by the second set. Set inclusion (i. e., extracting a subset) is a special case of implication in which the two sets belong to the same universe. These operations (connectives) may be extended to fuzzy sets for corresponding use in fuzzy logic fuzzy reasoning. For fuzzy sets, the applicable connectives must be expressed in terms of the membership functions of the sets which are operated on. In view of the isomorphism between fuzzy sets and fuzzy logic, both the set operations and the logical connectives can be addressed together. Some basic operations that can be defined on fuzzy sets and the corresponding connectives of fuzzy logic are described next. Several methods are available to define the intersection and the union of fuzzy sets. The classical ones suggested by Zadeh are widely used, because of their simplicity and the analogy with crisp sets (binary logic).

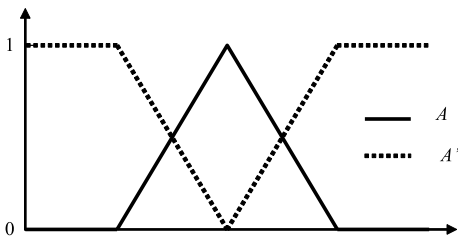
Complement (Negation, NOT) Consider a fuzzy set A in a universe X . Its complement A' is a fuzzy set whose membership function is given by

$$\mu_{A'}(x) = 1 - \mu_A(x) \quad \text{for all } x \in X. \quad (7)$$

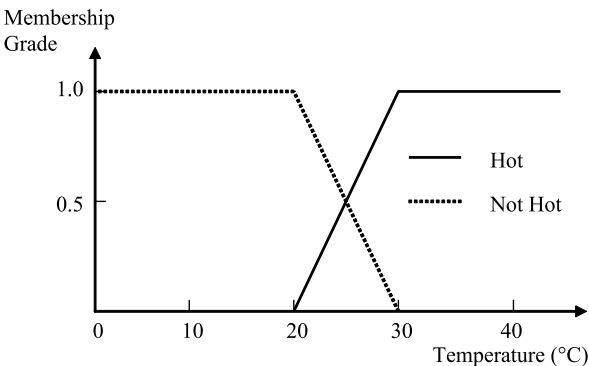
The complement in fuzzy sets corresponds to the negation (NOT) operation in fuzzy logic, just as in crisp logic, and is denoted by \bar{A} where A now is a fuzzy logic proposition (or a fuzzy state).

A graphic (membership function) representation of complement of a fuzzy set (or, negation of a fuzzy state) is given in Fig. 6.

Example As an example, consider the fuzzy set of “hot temperatures.” This may be represented by the membership function shown using a solid line in Fig. 7. This is the set containing all values of hot temperature in a specified universe. In fuzzy logic then, this membership function can represent the fuzzy logic state “hot” or the fuzzy statement, “the temperature is hot.” The complement of the fuzzy set is represented by the dotted line in Fig. 7. This is the set containing all temperature values that are not hot, in the given universe. As before, this membership function can also represent the fuzzy logic state “not hot” or the fuzzy-logic statement “the temperature is not hot.”



Intelligent Control, Figure 6
Fuzzy-set complement or fuzzy-logic NOT



Intelligent Control, Figure 7
An example of fuzzy-logic NOT

Furthermore, the solid line can represent a fuzzy temperature value and the dotted line can represent another (complementary) fuzzy temperature value. In this manner, various concepts of fuzzy sets and fuzzy logic can be quantified by the use of membership functions.

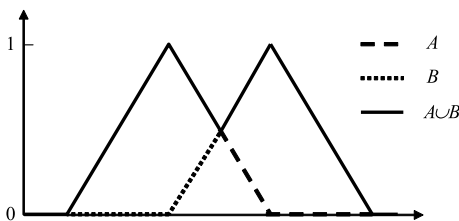
Union (Disjunction, OR) Consider two fuzzy sets A and B in the same universe X . Their union is a fuzzy set containing all the elements from both sets, in a “fuzzy” sense. This set operation is denoted by \cup . The membership function of the resulting set $A \cup B$ is given by

$$\mu_{A \cup B}(x) = \max[\mu_A(x), \mu_B(x)] \quad \forall x \in X. \quad (8)$$

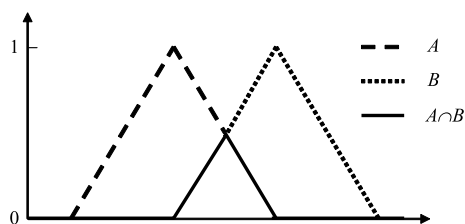
The union corresponds to a logical OR operation (called *Disjunction*), and is denoted by $A \vee B$, where A and B are fuzzy states or fuzzy propositions. The rationale for the use of *max* to represent fuzzy-set union is that, because element x may belong to one set or the other, the larger of the two membership grades should govern the outcome (union). Furthermore, this is consistent with the union of crisp sets. Similarly, the appropriateness of using *max* to represent fuzzy-logic operation “OR” should be clear. Specifically, since either of the two fuzzy states (or propositions) would be applicable, the larger of the corresponding two membership grades should be used to represent the outcome. A graphic (membership function) representation of union of two fuzzy sets (or, the logical combination OR of two fuzzy states in the same universe) is given in Fig. 8.

Even though set intersection is applicable to sets in a common universe, a logical “OR” may be applied for concepts in different universes. In particular, when the operands belong to different universes, orthogonal axes have to be used to represent them in a common membership function.

Intersection (Conjunction, AND) Again, consider two fuzzy set A and B in the same universe X . Their intersection is a fuzzy set containing all the elements that are common to both sets, in a “fuzzy” sense. This set operation is



Intelligent Control, Figure 8
Fuzzy-set union or fuzzy-logic OR



Intelligent Control, Figure 9

Fuzzy-set intersection or fuzzy-logic AND

denoted by \cap . The membership function of the resulting set $A \cap B$ is given by

$$\mu_{A \cap B}(x) = \min[\mu_A(x), \mu_B(x)] \quad \forall x \in X. \quad (9)$$

The union corresponds to a logical AND operation (called *Conjunction*), and is denoted by $A \wedge B$, where A and B are fuzzy states or fuzzy propositions. The rationale for the use of *min* to represent fuzzy-set intersection is that, because the element x must simultaneously belong to both sets, the smaller of the two membership grades should govern the outcome (intersection).

Furthermore, this is consistent with the intersection of crisp sets. Similarly, the appropriateness of using *min* to represent fuzzy-logic operation “AND” should be clear. Specifically, since both fuzzy states (or propositions) should be simultaneously present, the smaller of the corresponding two membership grades should be used to represent the outcome. A graphic (membership function) representation of intersection of two fuzzy sets (or, the logical combination AND of two fuzzy states in the same universe) is given in Fig. 9.

Implication (If-Then)

An if-then statement (a rule) is called an “implication”. In a knowledge-based system, the knowledge base is commonly represented using if-then rules. In particular, a knowledge base in fuzzy logic may be expressed by a set of linguistic rules of the if-then type, containing fuzzy terms. In fact a fuzzy rule is a *fuzzy relation*. A knowledge base containing several fuzzy rules is also a relation, which is formed by combining (aggregating) the individual rules according to how they are interconnected.

Consider a fuzzy set A defined in a universe X and a second fuzzy set B defined in another universe Y . The fuzzy implication “If A then B ,” is denoted by $A \rightarrow B$. Note that in this fuzzy rule, A represents some “fuzzy” situation, and is the *condition* or the *antecedent* of the rule. Similarly, B represents another fuzzy situation, and is the *action* or the *consequent* of the fuzzy rule. The fuzzy rule $A \rightarrow B$ is a *fuzzy relation*. Since the elements of A are de-

fined in X and the elements of B are defined in Y , the elements of $A \rightarrow B$ are defined in the *Cartesian product space* $X \times Y$. This is a two-dimensional space represented by two orthogonal axes (x -axis and y -axis), and gives the domain in which fuzzy rule (or fuzzy relation) is defined. Since A and B can be represented by membership functions, and additional orthogonal axis is needed to represent the membership grade.

Fuzzy implication may be defined (interpreted) in several ways. Two definitions of fuzzy implication are:

Method 1:

$$\mu_{A \rightarrow B}(x, y) = \min[\mu_A(x), \mu_B(y)] \quad \forall x \in X, \quad \forall y \in Y, \quad (10)$$

Method 2:

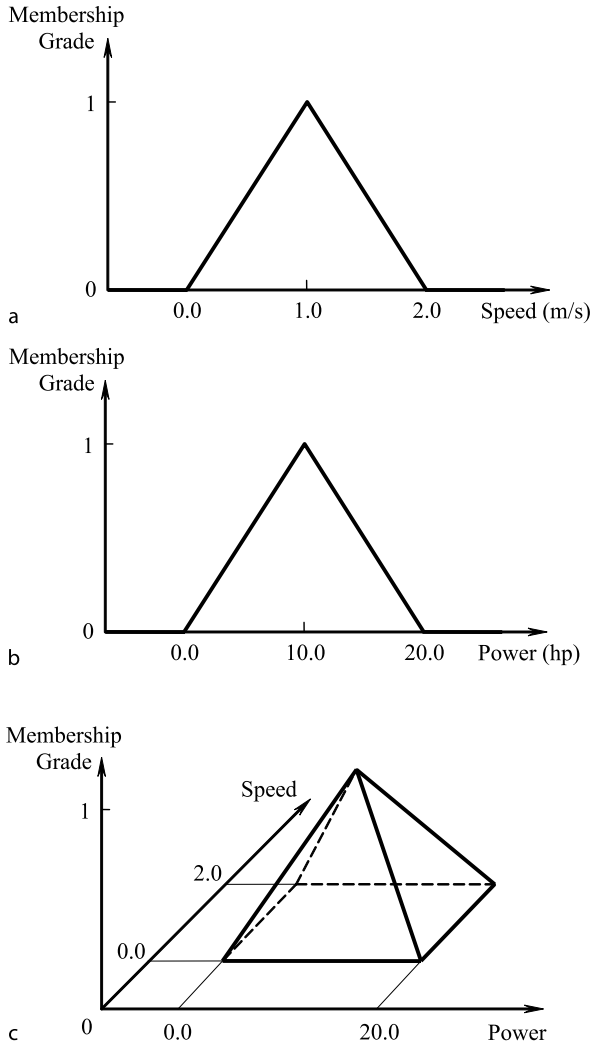
$$\mu_{A \rightarrow B}(x, y) = \min[1, \{1 - \mu_A(x) + \mu_B(y)\}] \quad \forall x \in X, \quad \forall y \in Y. \quad (11)$$

These two methods are approaches for obtaining the membership function of the particular fuzzy relation given by an if-then rule (implication). Note that the first method gives an expression that is symmetric with respect to A and B . This is not intuitively satisfying because “implication” is not a commutative operation (specifically, $A \rightarrow B$ does not necessarily satisfy $B \rightarrow A$). In practice, however, this method provides a good, robust result. The second method has an intuitive appeal because in crisp bivalent logic, $A \rightarrow B$ has the same truth table as $[(\text{NOT } A) \text{ OR } B]$ and hence are equivalent. Note that in Eq. (11), the membership function is upper-bounded to 1 using the *bounded sum* operation, as required (A membership grade cannot be greater than 1). The first method is more commonly used because it is simpler to use and often provides quite accurate results.

An example of fuzzy implication using the first method is shown in Fig. 10. Here the implication from a fuzzy set (a) to another fuzzy set (b) is given by (c).

Composition and Inference

If a fuzzy input is applied to a non-fuzzy (crisp) system, the output will still be fuzzy. This is known as the “extension principle,” and is illustrated by an example in Fig. 11. A fuzzy system may be represented by a set of fuzzy if-then rules, which may be aggregated into a single (multi-variable) membership function – a fuzzy relation. Application of a fuzzy input to a fuzzy relation in order to infer a fuzzy output is the basis of decision making in a fuzzy knowledge-based system. This idea is known as composition, and is illustrated in Fig. 12.



Intelligent Control, Figure 10

a Fuzzy set A; b Fuzzy set B; c Fuzzy implication $A \rightarrow B$

Decision making using fuzzy logic is known as fuzzy inference or *approximate reasoning*. The *compositional rule of inference* is utilized for this purpose.

Composition

Consider a fuzzy relation (fuzzy set) R in the r -dimensional subspace $X_1 \times X_2 \times \dots \times X_r$ and a second fuzzy relation (fuzzy set) S in the $(n - m + 1)$ -dimensional subspace $X_m \times X_{m+1} \times \dots \times X_n$ such that $m < r + 1$. Note that unlike the previous case of join, the two subspaces are never disjoint, and hence their intersection is never null (i.e., there is at least one common dimension in the subspaces R and S). The union of the two subspaces gives the overall n -dimensional space $X_1 \times X_2 \times \dots \times X_n$. The

composition of R and S is denoted by $R \circ S$ and is given by

$$R \circ S = \text{Proj}[\text{Join}(R, S); X_1, \dots, X_{m-1}, X_{r+1}, \dots, X_n]. \quad (12)$$

Specifically, we take the join (intersection) of the two sets R and S , and then project (max or supremum operation) the resulting fuzzy set (in the n -dimensional space) onto the subspace formed by the disjoint parts of the two subspaces in which the fuzzy sets R and S are defined. The membership function of the resulting fuzzy set is obtained from the membership functions of R and S , while noting that the operation *min* applies for *join* and the operation *supremum* applies for *projection*. Specifically,

$$\mu_{(R \circ S)} = \sup_{x_m, \dots, x_r} [\min(\mu_R, \mu_S)]. \quad (13)$$

This is called the *sup-min composition*.

Composition can be interpreted as a matching of the common (known) parts of two fuzzy sets, and then making an inference about the disjoint (unknown) parts according to the result of the matching process. Specially, the two sets are combined and matched (through *join*) over their common subspace. The result is then projected (through *supremum*) over to the inference subspace (non-overlapping subspace), giving a fuzzy set defined over the disjoint portions of the two subspaces. This process of matching is quite analogous to matching of data with the condition part of a rule in the conventional rule-based decision making. In this regard composition plays a crucial role in fuzzy inference (fuzzy decision making) and fuzzy knowledge based systems.

As a simple example, consider a fuzzy relation R defined in the $X \times Y$ space and another fuzzy relation S defined in the $Y \times Z$ space. The composition of these two fuzzy relations is given by

$$\mu_{R \circ S}(x, z) = \sup_{y \in Y} \{\min(\mu_R(x, y), \mu_S(y, z))\}, \quad (14)$$

which is defined in the $X \times Z$ space.

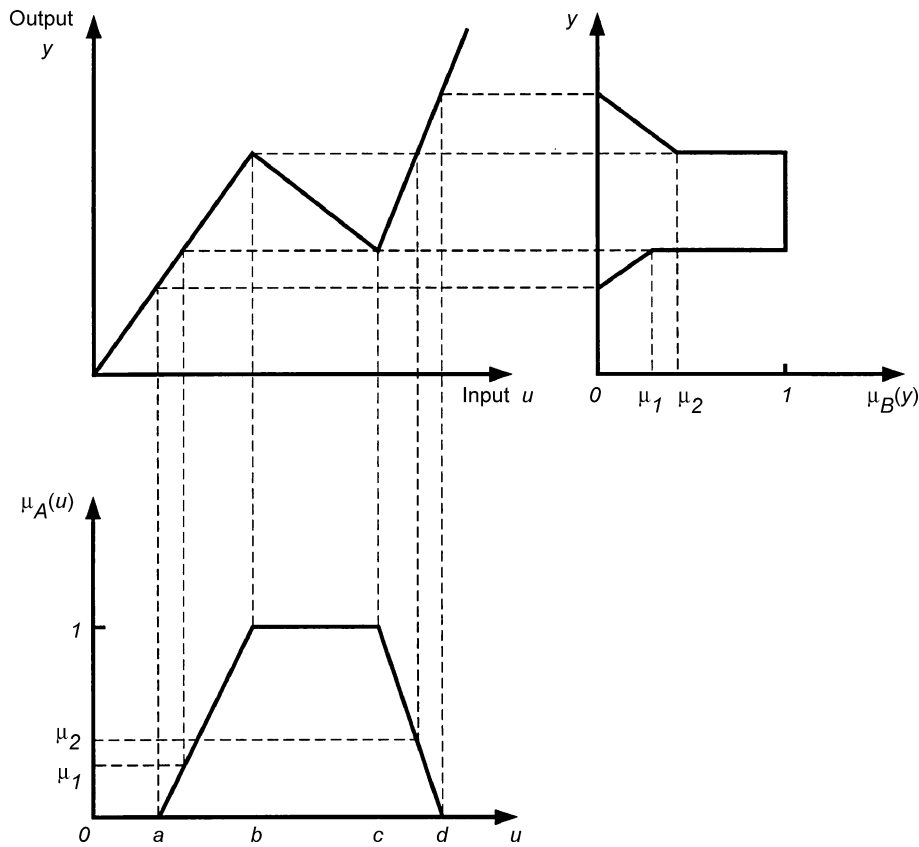
Sup-Product Composition In Eq. (13) if we use the “product” operation in place of “min” we get the *sup-product composition* given by

$$\mu_{(R \bullet S)} = \sup_{x_m, \dots, x_r} [\mu_R \times \mu_S]. \quad (15)$$

This composition is denoted by “ \bullet ” and is also known as the *sup-dot composition*.

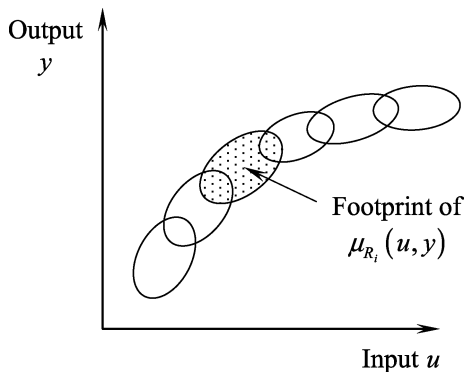
In summary, the composition operation involves:

1. Matching of data and a knowledge base, using the join (or, intersection) operation.



Intelligent Control, Figure 11

Fuzzy decision making using a crisp relation (extension principle)



Intelligent Control, Figure 12

Fuzzy decision making using a fuzzy relation (Composition)

Compositional Rule of Inference

In knowledge-based systems, knowledge is often expressed as rules of the form:

“IF condition Y_1 is y_1 **AND IF** condition Y_2 is y_2 **THEN** action C is c .”

In fuzzy knowledge-based systems (e. g., fuzzy control systems), rules of this type are linguistic statements of expert knowledge in which y_1 , y_2 , and c are fuzzy quantities (e. g., small negative, fast, large positive). These rules are fuzzy relations that employ the fuzzy implication (IF-THEN). The collective set of fuzzy relations forms the knowledge base of the fuzzy system. Let us denote the fuzzy relation formed by this collection of rules as the fuzzy set K . This relation is an aggregation of the individual rules, and may be represented by a multivariable membership function. In a fuzzy decision making process (e. g., in fuzzy logic control), the rulebase (knowledge base) K is first collectively matched with the available data (context). Next, an inference is made on another fuzzy variable that

2. Making an inference (decision) on the variables in the subspace of the knowledge base that is outside the subspace of data (i. e., nonoverlapping subspace), using the projection (sup) operation.

is represented in the knowledge base, on this basis. The matching and inference making are done using the composition operation, as discussed previously. The application of composition to make inferences in this manner is known as the *compositional rule of inference* (CRI).

For example, consider a control system. Usually the context would be the measured outputs Y of the process. The control action that drives the process is C . Typically, both these variables are crisp, but let us ignore this fact for the time being and assume them to be fuzzy, for general consideration. Suppose that the control knowledge base is denoted by R , a fuzzy relation. The method of obtaining the rule base R is analogous to model identification in conventional crisp control. Then, by applying the compositional rule of inference we get the fuzzy control action as:

$$\mu_C = \max_Y \min(\mu_Y, \mu_R). \quad (16)$$

Extensions to Fuzzy Decision Making

Thus far we have considered fuzzy rules of the form:

$$\text{IF } x \text{ is } A_i \text{ AND IF } y \text{ is } B_i \text{ THEN } z \text{ is } C_i \quad (17)$$

where, A_i , B_i , and C_i are fuzzy states governing the i th rule of the rulebase. In fact this is the Mamdani approach (Mamdani system or Mamdani model) named after the person who pioneered the application of this approach. Here, the knowledge base is represented as fuzzy protocols and represented by membership functions for A_i , B_i , and C_i , and the inference is obtained by applying the compositional rule of inference. The result is a fuzzy membership function, which typically has to be *defuzzified* for use in practical tasks.

Several variations to this conventional method are available. One such version is the *Sugeno model* (or, *Takagi-Sugeno-Kang model* or *TSK model*). Here, the knowledge base has fuzzy rules with crisp functions as the consequent, of the form

$$\text{IF } x \text{ is } A_i \text{ AND IF } y \text{ is } B_i \text{ THEN } c_i = f_i(x, y) \quad (18)$$

for Rule i , where, f_i is a crisp function of the condition variables (antecedent) x and y . Note that the condition part of this rule is the same as for the Mamdani model (17), where A_i and B_i are fuzzy sets whose membership functions are functions of x and y , respectively. The action part is a crisp function of the condition variables, however. The inference $\hat{c}(x, y)$ of the fuzzy knowledge-based system is obtained directly as a crisp function of the condition variables x and y , as follows:

For Rule i , a weighting parameter $w_i(x, y)$ is obtained corresponding to the condition membership functions, as for the Mamdani approach, by using either the “min” operation or the “product” operation. For example, using the “min” operation we form

$$w_i(x, y) = \min[\mu_{A_i}(x), \mu_{B_i}(y)]. \quad (19)$$

The crisp inference $\hat{c}(x, y)$ is determined as a weighted average of the individual rule inferences (crisp) $c_i = f_i(x, y)$ according to

$$\hat{c}(x, y) = \frac{\sum_{i=1}^r w_i c_i}{\sum_{i=1}^r w_i} = \frac{\sum_{i=1}^r w_i(x, y) f_i(x, y)}{\sum_{i=1}^r w_i(x, y)} \quad (20)$$

where, r is the total number of rules. For any data x and y , the knowledge-based action $\hat{c}(x, y)$ can be computed from (20), without requiring any defuzzification. The Sugeno model is particularly useful when the actions are described analytically through crisp functions, as in conventional crisp control, rather than linguistically. The TSK approach is commonly used in the applications of direct control and in simplified fuzzy models. The Mamdani approach, even though popular in low-level direct control, is particularly appropriate for knowledge representation and processing in expert systems and in high-level (hierarchical) control systems.

Fuzzy Control

Fuzzy Logic Control or simply “Fuzzy Control” belongs to the class of “Intelligent Control”, “Knowledge-Based Control”, or “Expert Control”. Fuzzy control uses knowledge-based decision making employing techniques of fuzzy logic, in determining the control actions. In this section, the basics of fuzzy logic control are presented.

There are many reasons for the practical deficiencies of conventional, crisp algorithmic control. Conventional control is typically based on “modeling” the plant that is to be controlled. If the model is not accurate the performance of the controller may not be acceptable. Furthermore, these conventional controllers rely on a complete set of data, including sensory information and parameter values, to produce control actions, and indeed, program instructions and data are conventionally combined into the same memory of the controller. If the model parameters and other necessary data are not completely known, say unexpectedly as a result of sensing problems, then appropriate estimates have to be made. Furthermore, if the available information is fuzzy, qualitative, or vague,

these “crisp” controllers that are based on such *incomplete information* will not usually provide satisfactory results. The environment with which the plant interacts may not be completely predictable either, and it is normally difficult for a crisp control algorithm to accurately respond to a condition that it did not anticipate and that it could not “understand”. Also, some conventional techniques of control assume that the plant is linear and time-invariant, an assumption that hardly holds in a large majority of practical problems. Conventional control techniques that are based on “crisp” algorithms or mathematical formulas can fail when the plant to be controlled is very complex, highly nonlinear, incompletely known, or contains large process delays. Often, a control strategy that is based on “soft knowledge” such as human “experience” in operating the plant, heuristics, commonsense, or expert opinion can provide the remedy to these problems.

Basics of Fuzzy Control

Fuzzy control uses the principles of fuzzy logic-based decision making to arrive at the control actions. The decision making approach is typically based on the *compositional rule of inference* (CRI), as presented before. In essence, some information (e.g., output measurements) from the system to be controlled is matched with a knowledge base of control for the particular system, using CRI. A fuzzy rule in the knowledge base of control is generally a “linguistic relation” of the form

$$\text{IF } A_i \text{ THEN IF } B_i \text{ THEN } C_i \quad (21)$$

where, A_i and B_i are fuzzy quantities representing process measurements (e.g., process error and change in error) and C_i is a fuzzy quantity representing a control signal (e.g., change in process input). What we have is a rule-base with a set of (n) rules:

$$\text{Rule 1: } A_1 \text{ and } B_1 \Rightarrow C_1$$

$$\text{Rule 2: } A_2 \text{ and } B_2 \Rightarrow C_2$$

$$\dots \dots \dots$$

$$\text{Rule } n: A_n \text{ and } B_n \Rightarrow C_n .$$

Because these fuzzy sets are related through IF-THEN implications and because an implication operation for two fuzzy sets can be interpreted as a “minimum operation” on the corresponding membership functions, the membership function of this fuzzy relation may be expressed as

$$\mu_{R_i}(a, b, c) = \min[\mu_{A_i}(a), \mu_{B_i}(b), \mu_{C_i}(c)] . \quad (22)$$

The individual rules in the rule-base are joined through ELSE connectives, which are OR connectives (“unions” of membership functions). Hence, the overall membership function for the complete rule-base (relation R) is obtained using the “maximum” operation on the membership functions of the individual rules; thus

$$\begin{aligned} \mu_R(a, b, c) &= \max_i \mu_{R_i}(a, b, c) \\ &= \max_i \min[\mu_{A_i}(a), \mu_{B_i}(b), \mu_{C_i}(c)] . \end{aligned} \quad (23)$$

In this manner the membership function of the entire rule-base can be determined (or, “identified” in the terminology of conventional control) using the membership functions of the response variables and control inputs. Note that a fuzzy knowledge base is a multivariable function – a multidimensional array (a three-variable function or a dimensional array in the case of Eq. (23)) of membership function values. This array corresponds to a fuzzy control algorithm in the sense of conventional control. The control rule-base may represent linguistic expressions of experience, expertise, or knowledge of the domain experts (control engineers, skilled operators, etc.). Alternatively, a control engineer may instruct an operator (or a control system) to carry out various process tasks in the usual manner; monitor and analyze the resulting data; and learn appropriate rules of control, say by using neural networks.

Once a fuzzy control knowledge base of the form given by Eq. (23) is obtained, we need a procedure to infer control actions using process measurements, during control. Specifically, suppose that fuzzy process measurements A' and B' are available. The corresponding control inference C' is obtained using the compositional rule of inference (i.e., inference using the composition relation). The applicable relation is

$$\mu_{C'}(c) = \sup_{a,b} \min[\mu_{A'}(a), \mu_{B'}(b), \mu_R(a, b, c)] . \quad (24)$$

Note that in fuzzy inference, the data fuzzy sets A' and B' are jointly matched with the knowledge-base fuzzy relation R . This is a “join” operation, which corresponds to an AND operation (an “intersection” of fuzzy sets), and hence the *min* operation applies for the membership functions. For a given value of control action c , the resulting fuzzy sets are then mapped (projected) from a three-dimensional space $X \times Y \times Z$ of knowledge onto a one-dimensional space Z of control actions. This mapping corresponds to a set of OR connectives, and hence the *sup* operation applies to the membership function values, as expressed in Eq. (24).

Actual process measurements are crisp. Hence, they have to be *fuzzified* in order to apply the compositional

rule of inference. This is conveniently done by reading the grade values of the membership functions of the measurement at the specific measurement values. Typically, the control action must be a crisp value as well. Hence, each control inference C' must be *defuzzified* so that it can be used to control the process. Several methods are available to accomplish defuzzification. In the *mean of maxima* method the control element corresponding to the maximum grade of membership is used as the control action. If there is more than one element with a maximum (peak) membership value, the mean of these values is used. In the *center of gravity* (or *centroid*) method the centroid of the membership function of control decision is used as the value of crisp control action. This weighted control action is known to provide a somewhat sluggish, yet more robust control.

There are several practical considerations of fuzzy control that were not addressed in the above discussion. Because representation of the rule-base R by an analytical function can be somewhat unrealistic and infeasible in general, it is customary to assume that the fuzzy sets involved in R have discrete and finite universes (or at least discrete and finite support sets). As a result, process response measurements must be quantized. Hence, at the outset, a decision must be made as to the element resolution (quantization error) of each universe. This resolution governs the cardinality of the sets and in turn the size of the multidimensional membership function array of a fuzzy rule-base. It follows that the required computational effort, memory and storage requirements, and accuracy are directly affected by the quantization resolution.

Because process measurements are crisp, one method of reducing the real-time computational overhead is to precompute a decision table relating quantized measurements to crisp control actions. The main disadvantage of this approach is that it does not allow for convenient modifications (e.g., rule changes and quantization resolution adjustments) during operation. Another practical consideration is the selection of a proper sampling period in view of the fact that process responses are generally analog signals. Factors such as process characteristics, required control bandwidth, and the processing time needed for one control cycle, must be taken into account in choosing a sampling period. Scaling or gain selection for various signals in a fuzzy logic control system is another important consideration. For reasons of processing efficiency, it is customary to scale the process variables and control signals in a fuzzy control algorithm. Furthermore, adjustable gains can be cascaded with these system variables so that they may serve as tuning parameters for the controller.

A proper tuning algorithm would be needed, however. A related consideration is real-time or on-line modification of a fuzzy rule-base. Specifically, rules may be added, deleted, or modified on the basis of some scheme of *learning and self-organization*. For example, using a model for the process and making assumptions such as input-output monotonicity, it is possible during control to trace and tag the rules in the rule-base that need attention. The control-decision table can be modified accordingly.

Steps of Fuzzy Logic Control

The main steps of fuzzy logic control, as described above, can be summarized in the following manner. A fuzzy knowledge base must be developed first (off-line), according to the following four steps:

1. Develop a set of linguistic control rules (protocols) that contain fuzzy variables as conditions (process outputs) and actions (control inputs to the process).
2. Obtain a set of membership functions for process output variables and control input variables.
3. Using the “fuzzy AND” operation (typically, *min*) and the “fuzzy implication” operation (typically, *min*) on each rule in Step 1, and using Step 2, obtain the multi-variable rule-base function R_i (a multi-dimensional array of membership values in the discrete case) for that rule.
4. Combine (aggregate) the relations R_i using the fuzzy connectives ELSE (fuzzy OR; typically, *max*) to obtain the overall fuzzy rule-base (relation) R (a multi-dimensional array in the discrete case).

Then, control actions may be determined in real time as follows:

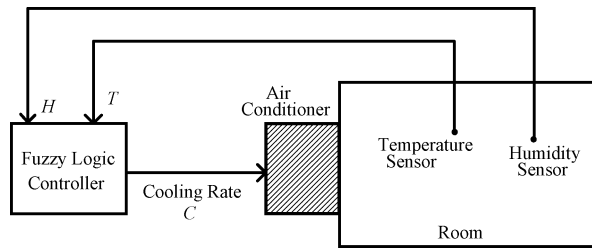
1. Fuzzify the measured (crisp) process variables (for example, a fuzzy singleton may be obtained by reading the grade value of the corresponding membership function at the measured value of the variable).
2. Match the fuzzy measurements obtained in Step 1 with the membership function (array in the discrete case) of the fuzzy rule-base (obtained in previous Step 4), using the compositional rule of inference.
3. Defuzzify the control inference obtained in Step 2 (for example, the mean of maxima method or the centroid method may be used here).

These steps reflect the formal procedure in FLC. There are several variations. For example, a much faster approach would be to develop a crisp decision table by combining the four steps of fuzzy knowledge base development and the first two steps of control, and using this table in

a table look-up mode to determine a crisp control action during operation. Also, hardware fuzzy processors (*fuzzy chips*) may be used to carry out the fuzzy inference at high speed. The rules, membership functions, and measured context data are generated as usual, through the use of a control “host” computer. The fuzzy processor is located in the same computer, which has appropriate interface (input/output) hardware and driver software. Regardless of all these, it is more convenient to apply the inference mechanism separately to each rule and then combine the result instead of applying it to the entire rule-base using the compositional rule of inference. This aspect is discussed next.

Fuzzy logic is commonly used in direct control of processes and machinery. In this case the inferences of a fuzzy decision making system form the control inputs to the process. These inferences are arrived at by using the process responses as the inputs (context data) to the fuzzy system. The structure of a direct fuzzy controller is shown in Fig. 13. Here, y represents the process output, u represents the control input, and R is the relation, which represents the fuzzy control knowledge base.

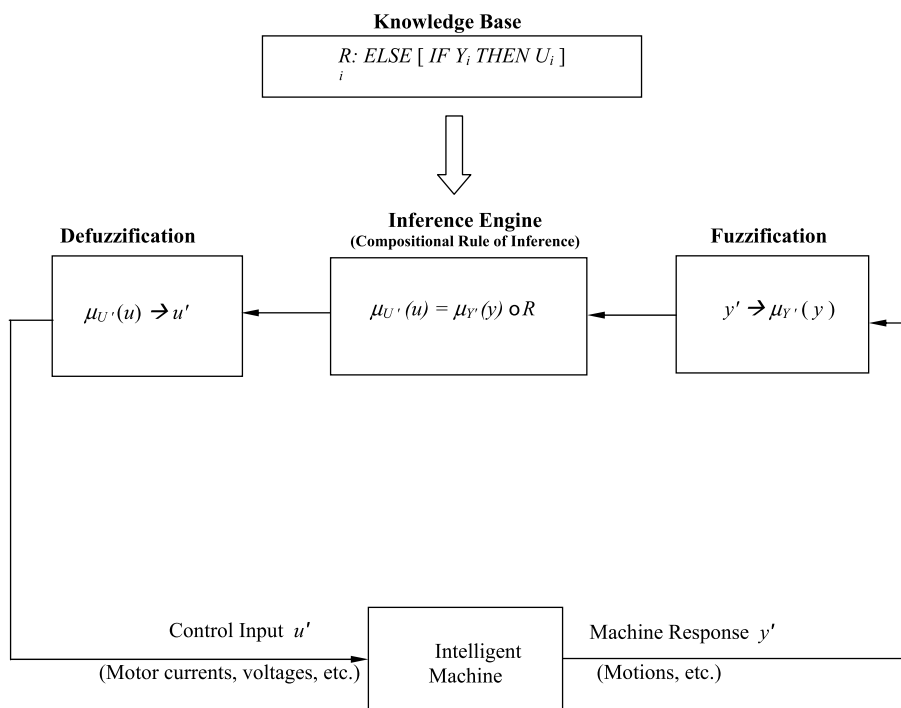
Example Consider the room comfort control system schematically shown in Fig. 14. The temperature (T) and



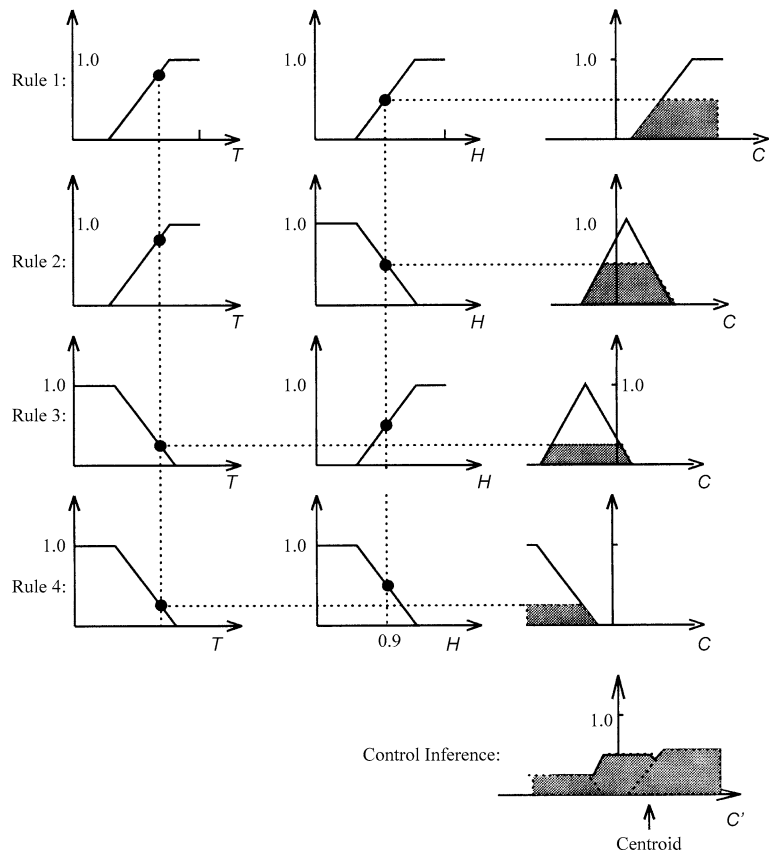
Intelligent Control, Figure 14
Comfort control system of a room

humidity (H) are the process variables that are measured. These sensor signals are provided to the fuzzy logic controller, which determines the cooling rate (C) that should be generated by the air conditioning unit. The objective is to maintain a particular comfort level inside the room.

A simplified fuzzy rule-base of the comfort controller is graphically presented in Fig. 15. The temperature level can assume one of two fuzzy states (HG, LW), which denote high and low, respectively, with the corresponding membership functions. Similarly, the humidity level can assume two other fuzzy states (HG, LW) with associated membership functions. Note that the membership functions of T are quite different from those of H , even though



Intelligent Control, Figure 13
Structure of a direct fuzzy controller



Intelligent Control, Figure 15
The fuzzy knowledge base of the comfort controller

the same nomenclature is used. There are four rules, as given in Fig. 15. The rule-base is:

- Rule 1: If T is HG and H is HG then C is PH
- Rule 2: else if T is HG and H is LW then C is PL
- Rule 3: else if T is LW and H is HG then C is NL
- Rule 4: else if T is LW and H is LW then C is NH and if

The nomenclature used for the fuzzy states is as follows:

Temperature (T)	Humidity (H)	Change in Cooling Rate (C)
HG = High	HG = High	PH = Positive High
LW = Low	LW = Low	PL = Positive Low
		NH = Negative High
		NL = Negative Low

Application of the compositional rule of inference is done here by using individual rule-based composition. Specifically, the measured information is composed with individ-

ual rules in the knowledge base and the results are aggregated to give the overall decision. For example, suppose that the room temperature is 30°C and the relative humidity is 0.9. Lines are drawn at these points, as shown in Fig. 15, to determine the corresponding membership grades for the fuzzy states in the four rules. In each rule the lower value of the two grades of process response variables is then used to clip (or modulate) the corresponding membership function of C (a *min* operation). The resulting “clipped” membership functions of C for all four rules are superimposed (a *max* operation) to obtain the control inference C' as shown. This result is a fuzzy set, and it must be defuzzified to obtain a crisp control action \hat{c} for changing the cooling rate. The centroid method may be used for defuzzification.

Fuzzy Control Surface

A fuzzy controller is a nonlinear controller. A well-defined problem of fuzzy control, with analytical membership functions and fuzzification and defuzzifica-

tion methods, and well-defined fuzzy logic operators, may be expressed as a nonlinear control surface through the application of the compositional rule of inference. The advantage then is that the generation of the control action becomes a simple and very fast step of reading the surface value (control action) for given values of crisp measurement (process variables). The main disadvantage is, the controller is fixed and cannot accommodate possible improvements to control rules and membership functions through successive learning and experience. Nevertheless, this approach to fuzzy control is quite popular. A useful software tool for developing fuzzy controllers is the MATLAB® Fuzzy Logic Toolbox.

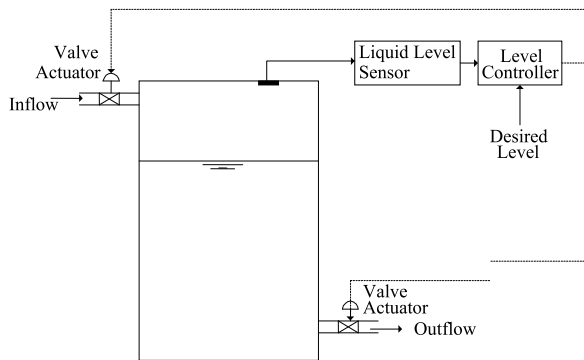
Example A schematic diagram of a simplified system for controlling the liquid level in a tank is shown in Fig. 16a. In the control system, the error (actually, correction) is given by

$$e = \text{Desired level} - \text{Actual level}.$$

The change in error is denoted by Δe . The control action is denoted by u , where $u > 0$ corresponds to opening the inflow valve and $u < 0$ corresponds to opening the outflow valve. A low-level direct fuzzy controller is used in this control system, with the control rule-base as given in Fig. 16b.

The membership functions for E , ΔE , and U are given in Fig. 16c. Note that the error measurements are limited to the interval $[-3a, 3a]$ and the Δ error measurements to $[-3b, 3b]$. The control actions are in the range $[-4c, 4c]$.

Following the usual steps of applying the compositional rule of inference for this fuzzy logic controller, we can develop a crisp control surface $u(e, \Delta e)$ for the system, expressed in the three-dimensional coordinate system $(e, \Delta e, u)$, which then can be used as a simple and fast controller. This method is described next.



Intelligent Control, Figure 16a
Liquid level control system

$E \backslash \Delta E$	NL	NS	ZO	PS	PL
NL	NL	NL	NM	NS	ZO
NS	NL	NM	NS	ZO	PS
ZO	NM	NS	ZO	PS	PM
PS	NS	ZO	PS	PM	PL
PL	ZO	PS	PM	PL	PL

Intelligent Control, Figure 16b
The control rule-base

The crisp control surface is developed by carrying out the rule-based inference for each point: $(e, \Delta e)$ in the measurement space $E \times \Delta E$, using individual rule-based inference. Specifically,

$$\mu_{U'}(u) = \max_{i,j} \min[\mu_{E_i}(e_o), \mu_{\Delta E_j}(\Delta e_o), \mu_{U_k}(u)], \quad (25)$$

where

- $\mu_{U'}$ control inference membership function
- e_o crisp context variable “error” defined in $[-3a, 3a]$
- Δe_o crisp context variable “change in error” defined in $[-3b, 3b]$
- E_i fuzzy states of “error”
- ΔE_j fuzzy states of “change in error”
- U_k fuzzy states of “control action”
- (i, j, k) possible combinations of fuzzy states of error, change in error, and control action, within the rule-base.

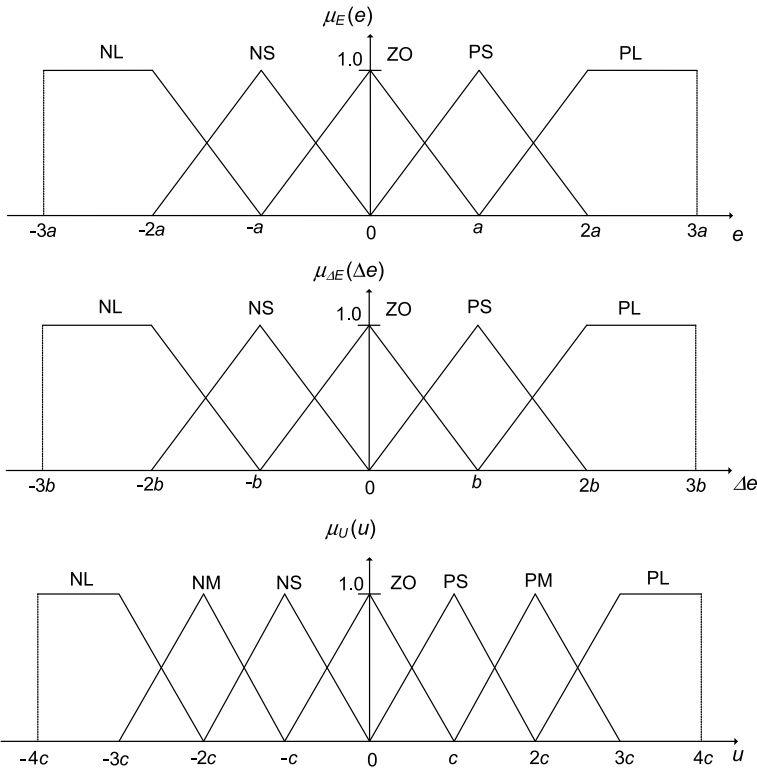
To find the crisp control inference (u') for a set of crisp context data $(e, \Delta e)$, the fuzzy inference $\mu_{U'}(u)$ is defuzzified using the center of gravity (centroid) method, which for the continuous case, is:

$$u' = \frac{\int_{u \in U} u \mu_{U'}(u) du}{\int_{u \in U} \mu_{U'}(u) du} \quad (26a)$$

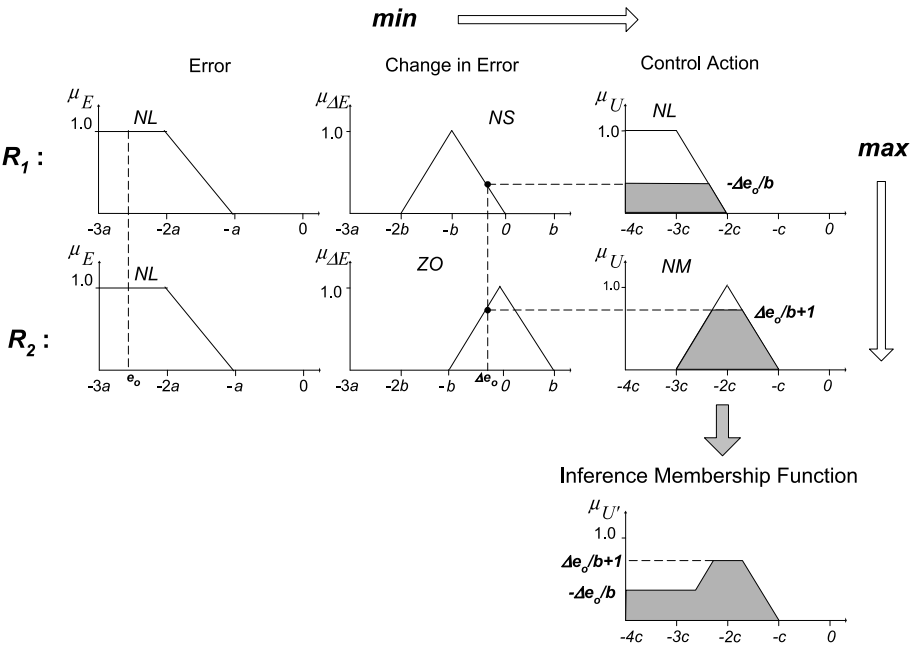
or, for the discrete case, it is:

$$u' = \frac{\sum_{u_i \in U} u_i \mu_{U'}(u_i)}{\sum_{u_i \in U} \mu_{U'}(u_i)}, \quad (26b)$$

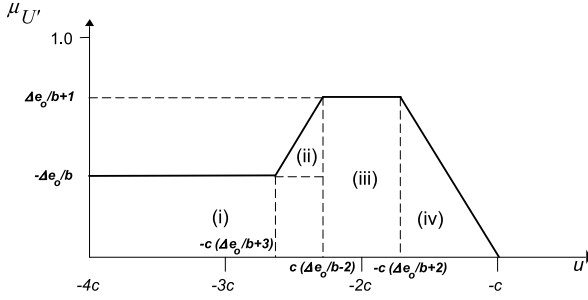
where $U = [-4c, 4c]$. Also, if the geometric shape of the inference is simple (e.g., piecewise linear), the centroid



Intelligent Control, Figure 16c
The membership functions of Error, Change in Error, and Control Action



Intelligent Control, Figure 17
Individual rule-based inference for $e_o[-3a, -2a]$ and $\Delta e_o[-b/2, 0]$



Intelligent Control, Figure 18

Sub-regions and critical points for calculation of the centroid

can be computed by the moment method; thus,

$$u' = \frac{\sum_{i=1}^n \text{area}_i m_i}{\sum_{i=1}^n \text{area}_i}, \quad (26c)$$

where

area_i = area of the i th sub-region

m_i = distance of the centroid of the i th sub-region,
on the control axis .

To demonstrate this procedure, consider a set of context data $(e_o, \Delta e_o)$, where e_o is in $[-3a, -2a]$ and Δe_o is in $[-b/2, 0]$. Then, from the membership functions and the

rule-base, it should be clear that only two rules are valid in this region, as given below:

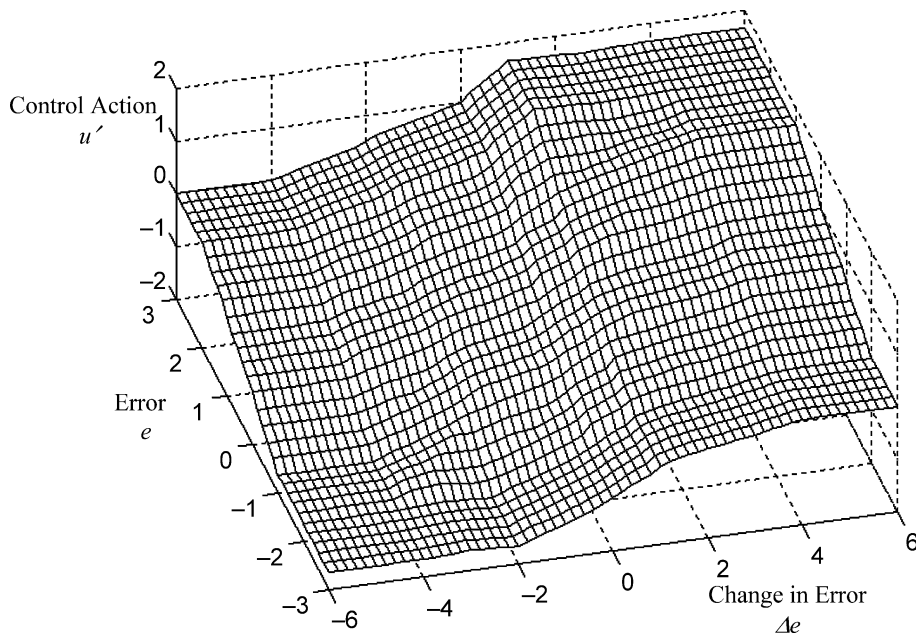
R_1 : if e is NL and Δe is NS then u is NL ,

R_2 : if e is NL and Δe is ZO then u is NM .

Since, in the range $[-3a, -2a]$, the membership grade of singleton fuzzification of e_o is always 1, the lower grade of the two context values is the one corresponding to the singleton fuzzification of Δe_o for both rules. Then, in applying the individual rule-based inference, the lower grade value of the two context variables is used to clip off the corresponding membership function of the control action variable U in each rule (this is a \min operation). The resulting membership functions of U for the two applicable rules are superimposed (this is a \max operation) to obtain the control inference U' , as shown in Fig. 17.

For defuzzification, we apply the moment method to find the centroid of the resulting membership function of control inference, as shown in Fig. 18. Note that the critical points in Fig. 17 and Fig. 18 (e. g., $-\Delta e_o/b$, $-c(\Delta e_o/b+3)$, etc.) are found from the corresponding membership functions.

From the moment method, we obtain the crisp control action as a function of e and Δe . The above procedure is repeatedly applied to all possible ranges of e $[-3a, 3a]$ and Δe $[-3b, 3b]$, to obtain the complete control surface.



Intelligent Control, Figure 19

Control surface with $a = 1$, $b = 2$, and $c = 0.5$

Also, the procedure can be implemented in a computer program to generate a control surface. A control surface with $a = 1$, $b = 2$, and $c = 0.5$ is shown in Fig. 19.

In the present example what we have applied is in fact the Mamdani approach (Mamdani system or Mamdani model). Sugeno model (or, Takagi–Sugeno–Kang model or TSK model) could have been used as well, thereby avoiding the defuzzification step.

Future Directions

Even though a considerable effort has gone into the development of machines that somewhat mimic humans in their actions, the present generation of intelligent machines do not claim to possess all the capabilities of human intelligence; for example, common sense, display of emotions, and inventiveness. Significant advances have been made, however, in machine implementation of characteristics of intelligence such as sensory perception, pattern recognition, knowledge acquisition and learning, inference from incomplete information, inference from qualitative or approximate information, ability to handle unfamiliar situations, adaptability to deal with new yet related situations, and inductive reasoning. Much research and development would be needed in these areas, pertaining to techniques, hardware, and software before a machine could reliably and consistently possess the level of intelligence of say, a dog.

For instance, consider a handwriting recognition system, which is a practical example of an intelligent system. The underlying problem cannot be solved through simple template matching, which does not require intelligence. Handwriting of the same person can vary temporally, due to various practical shortcomings such as missing characters, errors, nonrepeatability, physiological variations, sensory limitations, and noise. It should be clear from this observation that a handwriting recognition system has to deal with incomplete information and unfamiliar objects (characters), and should possess capabilities of learning, pattern recognition, and approximate reasoning, which will assist in carrying out intelligent functions of the system. Techniques of soft computing are able to challenge such needs of intelligent machines and intelligent control.

Increased attention will be paid to developments of intelligent controllers with inspiration from biological systems. In particular, hybrid systems that employ two or more soft computing approaches will become common. Various techniques of self-learning and distributed intelligent control will find an increased presence [12]. Embedded systems and integrated miniature systems with intelligent sensing, actuation, and control integrated and dis-

tributed within the same device will gain more prominence.

Bibliography

Primary Literature

1. Cao Y, de Silva CW (2006) Dynamic Modeling and Neural-Network Adaptive Control of a Deployable Manipulator System. *J Guid Control Dyn* 29(1):192–195
2. Cao Y, de Silva CW (2006) Supervised Switching Control of a Deployable Manipulator System. *Int J Control Intell Syst* 34(2):153–165
3. de Silva CW (1992) Research Laboratory for Fish Processing Automation. *Int J Robotics Comput-Integr Manuf* 9(1):49–60
4. de Silva CW (1997) Intelligent Control of Robotic Systems with Application in Industrial Processes. *Robotics Auton Syst* 21:221–237
5. de Silva CW (2003) The Role of Soft Computing in Intelligent Machines. *Philos Trans Royal Soc Ser A UK* 361(1809):1749–1780
6. de Silva CW, Wickramarachchi N (1997) An Innovative Machine for Automated Cutting of Fish. *IEEE/ASME Trans Mechatron* 2(2):86–98
7. Filippidis A, Jain LC, de Silva CW (1999) Intelligent Control Techniques. In: Jain LC, de Silva CW (eds) *Intelligent Adaptive Control*. CRC Press, Boca Raton
8. Goulet JF, de Silva CW, Modi VJ (2001) Hierarchical Control of a Space-Based Deployable Manipulator Using Fuzzy Logic. *AIAA J Guid Control Dyn* 24(2):395–405
9. Rahbari R, de Silva CW (2001) Comparison of Two Inference Methods for P-type Fuzzy Logic Control through Experimental Investigation Using a Hydraulic Manipulator. *Eng Appl Artif Intell* 14(6):763–784
10. Rahbari R, Leach BW, Dillon J, de Silva CW (2005) Expert System for an INS/DGPS Integrated Navigator Installed in a Bell 206 Helicopter. *Eng Appl Artif Intell* 18(3):353–361
11. Tang PL, Poo AN, de Silva CW (2001) Knowledge-Based Extension of Model-Referenced Adaptive Control with Application to an Industrial Process. *J Intell Fuzzy Syst* 10(3–4):159–183
12. Wang Y, de Silva CW (2008) A Machine Learning Approach to Multi-robot Coordination. *Eng Appl Artif Intell* 21(3):470–484
13. Yan GKC, de Silva CW, Wang GX (2001) Experimental Modeling and Intelligent Control of a Wood-Drying Kiln. *Int J Adapt Control Signal Process* 15:787–814
14. Zadeh LA (1984) Making Computers Think Like People. *IEEE Spectrum* 21(8):26–32

Books

- Anderson TW (1984) *An Introduction to Multivariate Statistical Analysis*. Wiley, New York
- Davis L (1991) *Handbook on Genetic Algorithms*. Van Nostrand–Rienhold, New York
- de Silva CW (1995) *Intelligent Control: Fuzzy Logic Applications*. CRC Press, Boca Raton
- de Silva CW (2000) *Intelligent Machines: Myths and Realities*. CRC Press, Boca Raton
- de Silva CW (2005) *Mechatronics: An Integrated Approach*. Taylor & Francis/CRC Press, Boca Raton

- Dubois D, Prade H (1980) *Fuzzy Sets and Systems: Theory and Applications*. Academic Press, Orlando
- Farrell JA, Polycarpou MM (2006) *Adaptive Approximation Based Control*. Wiley-Interscience, New York
- Forsyth R (1984) *Expert Systems*. Chapman & Hall, New York
- Gupta MM, Rao H (1994) *Neural Control Theory and Applications*. IEEE Press, Piscataway
- Hart A (1986) *Knowledge Acquisition for Expert Systems*. McGraw-Hill, New York
- Jain LC, de Silva CW (1999) *Intelligent Adaptive Control: Industrial Applications*. CRC Press, Boca Raton
- Jang JSR, Sun CTS, Mizutani E (1997) *Neuro-Fuzzy and Soft Computing*. Prentice Hall, Upper Saddle River
- Karray FO, de Silva CW (2004) *Soft Computing and Intelligent Systems Design – Theory, Tools, and Applications*. Addison Wesley, New York
- Klir GJ, Folger TA (1988) *Fuzzy Sets, Uncertainty, and Information*. Prentice Hall, Englewood Cliffs
- Pao YH (1989) *Adaptive Pattern Recognition and Neural Networks*. Addison-Wesley, Reading
- Passino KM (2005) *Biomimicry for Optimization, Control, and Automation*. Springer, London
- Ruano AE (2005) *Intelligent Control Systems Using Computational Intelligence Techniques*. The Institute of Electrical Engineers, Herts
- Vonk E, Jain LC, Johnson RP (1997) *Automatic Generation of Neural Network Architecture Using Evolutionary Computing*. World Scientific Publishing Co., Singapore

Intelligent Systems, Introduction to

JAMES HENDLER
Computer Science Department,
Rensselaer Polytechnic Institute, Troy, USA

The term “intelligent systems” has come to mean many different things in many different contexts and, like most things related to complex systems, it is hard to nail down a specific definition that is both rigorous enough to discriminate out those things which should not be included, but is loose enough to include those that are. As in defining terms like “agents” or “robots,” one is able to find overly inclusive definitions, such as “an autonomously acting entity” where a thermostat in the latter case, or hard disk controller in the former, would meet the definition. On the other hand, tighten up the definition and telerobotics or Google’s search bots no longer fit, despite being clearly related technologically. In the case of intelligent systems, too tight a definition of intelligence removes, say, the behaviors we see in a dog, which can seek out prey or be trained to beg for its dinner, but loosen the definition and we find ourselves talking about systems with the intelligence of a clam.

In this chapter, we are holding primarily to a tighter definition and starting to look at some of the kinds of behaviors that take us into areas traditionally associated with human intelligence. While several of the sections of this book deal with areas associated with, loosely defined, intelligent behaviors and others look specifically at aspects of Artificial Intelligence tied closely to data or a model, in this short section we pick up some of the missing pieces that help us complete the puzzle. The small number of papers in this section should not cause one to believe that there is little relation between intelligent systems and complexity, but rather that other sections of this encyclopedia necessarily included aspects of intelligence defined at some level – controlling against complexity demands it. The reader looking for other articles on Artificial Intelligence and complex systems will find them in many sections of this volume, particularly including those on:

- ▶ [Agent Based Modeling and Simulation](#),
- ▶ [Complexity and Non-linearity in Autonomous Robotics](#), [Introduction to](#),
- ▶ [Data-Mining and Knowledge Discovery](#), [Introduction to](#)
- ▶ [Data-Mining and Knowledge Discovery](#), [Neural Networks in](#)
- ▶ [Data-Mining and Knowledge Discovery: Case-Based Reasoning](#), [Nearest Neighbor and Rough Sets](#)
- ▶ [Soft Computing](#), [Introduction to](#), and
- ▶ [Nonsmooth Analysis in Systems and Control Theory](#)
- ▶ [Chronological Calculus in Systems and Control Theory](#)

However, despite the strong ties between these many sub-areas, there were some aspects of intelligent systems left out of these other topic areas, and this section is provided to cover these.

One such example is that of the use of mobile agents as the basis for intelligent cooperation among systems (see ▶ [Mobile Agents](#)). The primary need for mobility is in bandwidth-limited communications, and with the growth of modern computer networks, the area has gotten less attention of late. However, with the growing need for, and use of, sensor networks, wireless networks in noisy environments, deep-sea or space robotics, and other bandwidth-limited systems, agent-based modeling and simulation techniques can be used to model the networks, but not run on the networks themselves. Thus, an adaptive network that needs to do agent-to-agent communication for self-tuning will be seriously impacted in a bandwidth-

limited environment. Mobility can be used to help in such situations, moving small amounts of code to where the data it manipulates is stored, rather than moving the large amounts of data to where the computation could be run. Using such mobile agents bring both new capabilities, and new challenges, and as we try to increase the intelligence of systems operating in networks with disconnection, low bandwidth or high latency, such as many of the networks deployed in warfighting situations, agent mobility is regaining attention as an important area of research.

Another area that needs to be covered is that of the role of intelligent systems in the modeling and simulation area (see ► [Artificial Intelligence in Modeling and Simulation](#)). While agent-based modeling, as discussed elsewhere in this encyclopedia, is an important area, there are aspects of modeling it does not fully cover. One of these is the use of symbolic reasoning for use in validating models and simulation. Another is the actual modeling of the reasoning of other agents. For example, a baseball player recovering a ball that dropped in the outfield must be able to reason about what the base runners, who are trying to move up, and his team-mates, who are trying to hold them back, will most likely do. The player's skill in guessing the behaviors of these other agents, both competitive and cooperative, could be the difference between whether the game is won or lost.

As well as improving our capabilities in modeling and simulating systems, it is important to look at how, with the use of intelligent systems, we might better control the complex systems being modeled. Instead of trying to model a complex plant directly, in this section we look at work that takes a different approach. Instead we consider the knowledge-based control that results from observing, studying and understanding the behavior of a plant and/or the behavior of a human controlling it (see ► [Intelligent Control](#)). This area includes looking at soft-computing approaches to create an "approximate reasoning" solution that can be used for mimicking the control decisions that would be made by a human monitoring a plant, rather than for modeling the plant itself. Fuzzy logic, a particular branch of soft control has been successfully applied to the control of many complex systems, a number of which are described here. (The reader with a sense of humor may see a certain irony in many manufacturer's use of fuzzy logic to improve the picture on their camcorders and television sets, but let us leave that unexplored). Of particular import, this article outlines various ways in which the mathematical operations used in such control can be combined, allowing for the control of complex, non-linear systems that defy simpler control regimes.

Looking at very complex situations, a human operator, or at least a program simulating one, may want to look beyond soft computing and deal with the world at a higher level. One of the key abilities which separates the human from other primates and animals is our ability to learn over time to abstract away many details of the complex world in which we live, and to make plans for how to control it over time. Where planning itself can be a complex problem-solving task, learning how to abstract key aspects of situations and to apply plans to these is a critical need for dealing with complexity (see ► [Learning and Planning \(Intelligent Systems\)](#)). Exploring how we learn to plan is an area which has been gaining importance in the intelligent systems area as approaches which do not learn, but which apply brute force problem solving to larger and larger problems, are reaching the limits of their capabilities against the increasingly complex domains in which we wish to deploy our computational systems.

A recurring theme that arises in all of these attempts to provide intelligent behavior in evermore-complex systems environments is that of using a level of abstraction to reason not about some data or system itself, but about the meaning of the behaviors being observed. A critical aspect of performing such abstraction is the ability to represent a model of a system to the computer in a machine-readable way. The term "ontology" is used to describe this computer-based representation of the domain in which a system is trying to function.

Although ontologies have been around for a long time in AI, they have recently come back into their own in trying to help computer systems interact with one of the most complex human constructions in history – the millions of billions of dynamically changing bits of information that comprise the World Wide Web. While the Web has changed a great many aspects of our society, understanding its dynamics remains a major challenge (See Berners-Lee et al., *Creating a Science of the World Wide Web*, *Science*, 313(5788), August 2006, pp 769–771.) The use of ontologies, and other AI techniques, to help computers better process this massive information space is the *raison d'être* of the "Semantic Web," an overview of which is also presented in this section (see ► [Semantic Web](#)).

In short, this section, despite its brevity, picks up a number of themes that arise throughout this encyclopedia. These five articles will help the reader understand how many of the themes above are connected together via the use of technologies developed in artificial intelligence labs, allowing the creation of intelligent systems that provide a key tool in our arsenal for dealing with the complexity of the natural world and/or the complex human society that has evolved to let us live in it.

Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems

WOLFGANG TSCHACHER
University of Bern, Bern, Switzerland

Article Outline

Glossary
Definition of the Subject
Introduction
Intentionality and Representation
Synergetics
Discussion: Naturalization of Intentionality
Future Directions
Bibliography

Glossary

Cognition A general concept of psychology referring to all processes and structures of the mind. These comprise the processing of stimulus ‘input’ (i.e., perception) and the internal processing of represented information (e.g., memory functions, thinking, problem solving); the latter processes presuppose intentional features of the mind. Cognitive structures include knowledge, categories, memory, attitudes, and schemata, again intentional concepts.

Dynamical systems theory A system is any set of things (components, elements) that stand in relation to one another. If a rule or description exists that defines how the systems change over time (such as a differential equation or a mapping algorithm), the system is a dynamical system.

Intentionality A characterizing property of the mind. In contrast to physical systems, mental states have content, i.e. they ‘are about something’ in the sense that they contain a reference to an object, or the representation of an object. In addition to *aboutness*, intentionality demands a *functional* reference to the intentional object.

Mind-body problem The philosophical question as to if and how mind and brain/body interact. Analogously, the question if mental processes and physical processes are ontologically different.

Naturalization Explaining mental phenomena using concepts and models derived from the natural sciences. Naturalization efforts may be viewed as tools in order to develop viable simulation models of mental processes.

Definition of the Subject

In the philosophical and psychological tradition, intentionality is viewed as a characterizing property of mental (cognitive) acts. Mental acts have content, i.e. they ‘are about something’. This something is called the intentional object. Intentionality may take the form of a desired state (as in, “I wish it were Friday.”) or a goal (e.g., my plan for a weekend trip to the mountains). When viewing the constituents of the mind (the cognitive system) in this intentionalist manner, we stand in stark contrast to scientific descriptions of physical systems. These latter systems are material things, which are sufficiently described without reference to objects they would be about, or to states they might desire to realize. Therefore, are mental and physical systems qualitatively different with respect to intentionality? If yes, we are confronted with a dualist worldview entailing an aggravated mind-body problem. If no, a solution is demanded that can elucidate how mental phenomena may be explained avoiding intentional language or, conversely, how physical systems may show or mimic the features of intentionality. The project involved in the latter case may be named naturalization of intentionality.

Clarifying the problem of intentionality is important in several respects. First, psychology and other cognitive sciences are conceptually divided into two approaches, the phenomenological (first-person) approach and the behavioral and/or biological (third-person) approach. The conflation of first-person and third-person concepts may be viewed as a serious impediment to theorizing throughout psychology and cognitive neuroscience. Second, modern societies have a growing demand for machines and software that can function in ‘intelligent’ ways. Therefore, engineers of artificial knowledge-based systems need to know how intentionality can be implemented in physical information-processing machines. Third, the problem of intentionality is one of the foundational problems of the philosophy of mind and of consciousness research. Any, even if partial, solution to this problem that may be derived from a dynamical systems perspective is therefore welcome.

The phenomenology of intentional acts is well known; phenomenology yields the features of intentionality, which can indicate how closely a formal, mathematical, or physical model of intentionality approximates an understanding of the problem. These features are

- *Aboutness*, the intentional system’s state must be about something in the system’s environment
- *Functionality*, intentional states should be functional or instrumental with respect to what they are about

- *Mental-likeness*, in the sense that apart from being intentional, these model systems should have properties that resemble the properties of mental states. Especially, an interpreter (homunculus) who may account for missing links in the explanation of intentional mental states must not be allowed.

Introduction

If mind and matter are qualitatively different things, what is the nature of their difference? This is an elementary question of philosophy and scientific disciplines addressing the mind. This question and the potential answers to it, referred to as the mind–body problem, have a long history. What is the relationship between mind and body? If there is no essential difference between the two, or only a superficial difference, one may reach a monist understanding of the problem. The monists include idealists who posit that mind creates its world. A very distinct monist would be someone holding eliminative materialism, i. e. who would assume that the mind can be completely reduced to matter (e. g. [7]). If, on the other hand, a real difference between mind and body exists, we enter dualist notions (e. g. [41]). Dualism is confronted with intricate questions of how we may conceive of the interaction between mind and body. In the interest of the specific discussion of intentionality in this article, we will refrain from giving an account of philosophical theories of how mind and body may be associated in general.

Apart from presenting a core problem to the philosophy of mind, the mind–body problem is of interest to a wide range of contemporary scientific disciplines, especially psychology, neuroscience and computer science (artificial intelligence). In recent decades, a novel approach to cognitive science has appeared, combining dynamical systems theory with cognitive science. This dynamical approach to cognition addresses mind–body topics more or less explicitly. We will sketch this approach here because it provides the background for the ensuing treatment of intentionality.

The dynamical approach to cognition is founded on a number of studies and empirical paradigms. In various perceptual and behavioral tasks, researchers have observed a set of signatures of dynamical systems. These signatures are typically related to temporal patterns observed in the systems, especially asymptotic stability, and in many paradigmatic cases, multistability. Asymptotic stability means that a system's pattern of behavior is stable so that, in the face of an external disturbance, the system returns 'asymptotically' to this pattern. Multistability

means that several such patterns may coexist in the behavioral space of a single system.

Movement coordination has been at the center of several applications. Haken, Kelso and Bunz (1985) [22] provided a model of the coordination of two limbs (e. g. the hands, or the forefingers of both hands, of a person) using equations from dynamical systems theory. The rhythmic movement of the limbs generally becomes synchronized after a short time; these synchronous movement patterns are stable with respect to external inputs, i. e. they usually return to synchrony after externally induced disruptions of movement. Furthermore, they have been shown to undergo phase transitions depending on the values of the control parameter (in the Haken–Kelso–Bunz [22] system, the velocity of movements prescribed by a metronome). Characteristic phenomena were observed in the context of phase transitions such as hysteresis and critical fluctuations. Analogous findings were reported in animal locomotion. Here, especially multistability has attracted the attention of researchers because in certain regions of control parameter space two qualitatively different limb coordination patterns frequently occur. For instance, a horse may either gallop or trot at a certain velocity. In the seemingly unrelated field of visual perception, very similar signatures of dynamical systems were found in ambiguous stimuli ([18,29]). This research on perceptual organization may be viewed as a continuation of the tradition of Gestalt psychology ([27,45]). For example, perception of apparent motion (i. e. perception of motion in the absence of real motion of the stimuli, Wertheimer [50]) can be induced presenting a stimulus, e. g. a black disk, alternately in different positions of the visual field. Certain spatial configurations of the disks allow perception of two or more qualitatively different kinds of apparent motion although identical stimuli are presented. Hence this and related paradigms create perceptual multistability. Again, the signatures of self-organizing dynamical systems (as described by synergetics) can be found in the phase transitions between the different apparent motion perceptions.

The dynamical hypothesis in cognitive science ([46,49]) proposes that cognitive agents may be modeled as dynamical systems (instead of as physical symbol systems, Newell [36]). A common denominator in these dynamical approaches is to start from elementary perception–action cycles, an idea that is also implied in the concept of embodied cognition. Clark [10] elaborated three bridging assumptions by which the numerous empirical findings can be integrated in the theory of dynamical systems. First, the assumption of continuity refers to cognition as continuous with its developmental foundations [44]. Second, 'off-line' reasoning and thinking is

viewed as continuous with on-line motor control strategies. Therefore abstract cognition may be decoupled from the actual environment but may still be working on the same dynamical principles; thinking is accordingly understood as emulated sensorimotor loops of perception–action cycles. Third, due to the dynamical hypothesis, pattern is provided not by programs but is ‘soft-assembled’ by a continuing process of self-organization. This latter assumption was initially formulated in the framework of theories on complex systems in the natural sciences (especially Haken’s synergetics: Haken [17] and the theory of dissipative systems: [37]). The self-organization approach was successively introduced to cognitive science [21,26]. As already mentioned, prior to contemporary systems theory Gestalt psychology had developed a treatment of cognition and action which was very much akin to the dynamical systems approach, especially Köhler [27] and Lewin [30].

Cognitive science, especially its computational mainstream after behaviorist psychology turned cognitive in the 1960s, has had a tendency to start with ‘higher’ cognitive functions such as goals, beliefs and, in the context of goal-directed behavior, plans. This generated the symbol-grounding problem. The dynamical approach has avoided this problem arising in the computational framework [24] and has therefore proceeded in a bottom-up fashion instead. Higher cognition is assumed to emerge on the basis of elementary sensorimotor loops. Rather than focusing on symbol grounding, the dynamical view addresses symbol emergence that depends on control parameters. These control parameters comprise the ecological embedding of cognition, i. e. the environment of the cognitive agent. Therefore, the dynamical approach views cognition predominantly with reference to its embeddings, as embodied cognition [39] or situated cognition [16].

Intentionality and Representation

The dynamical view in cognitive science thus naturally leads to the concepts of embodied and situated cognition. Consequently, intentionality of mental acts can be discussed under these premises. Rather than searching, in a top-down fashion, for a fundament of experienced intentionality of the mind, the dynamical systems heuristic can be formulated differently: If the mind is conceptualized as arising from self-organization processes, is there a natural way by which emergent mental acts can be conceptualized as *being about something*?

Intentionality was introduced as a characterizing property of mental acts by Franz von Brentano, a philoso-

pher and early psychologist. In the late 19th century Brentano was professor at the University of Vienna where Edmund Husserl, Sigmund Freud, Carl Stumpf and other later protagonists of philosophy and psychology counted among his students. According to Brentano [4], mental phenomena are always directed towards an object (the intentional object). In other words, mental states contain within themselves something else (*‘intentionale Inexistenz’*, i. e. intentional existence within). No physical phenomenon has such intentional content, therefore according to Brentano intentionality constitutes the distinctive feature of the mind. Many concepts of contemporary cognitive psychology are in this sense intentional. The basic concepts *goal*, *wish*, *plan* and *intention* of volitional psychology [15,28] obviously have intentional content. The same applies to *achievement*, *valence* and *need* in motivation psychology [32]. *Affects* and *emotions* are also generally about something and thus intrinsically intentional concepts.

It should be noted, however, that intentionality may not provide a sufficient and necessary condition for a state or process to be mental. Not all mental states are intentional; some emotional states (e. g. moods such as a pervasive feeling of melancholy or of serenity) do not necessarily possess intentional content because they are not about something; yet they are undoubtedly experienced mental states. Furthermore, intentionality is likely not the only property that distinguishes mental from physical systems. Many current philosophers of mind suggest that in addition to intentional content, the phenomenal content of mental states must be considered [33]. This immediately leads to the topic of consciousness, which cannot be addressed here.

A concept very closely related to intentionality is *representation*. Representation plays a central role in cognitive psychology (as schema: Neisser [35]), in philosophy of mind (as language of thought: Fodor [11]) and in artificial intelligence (as physical symbol system: Newell [36]). In all of these fields representation of knowledge is a foundational concept, yet at the same time constitutes a core problem. If a cognitive agent is to have knowledge of its environment, the obvious idea is that there must be some kind of mental map or mental model of the environment ‘inside’ the agent. On the basis of information thus represented, the agent would then perform cognitive actions such as memory functions, manipulations for problem solving, and the like. One may note the close analogy of representation with the intentional object.

The naive, folk-psychological intuition of an inner map or depiction of the environment on which cognitive functions can be performed is however unsatisfac-

tory [2,9]. The map concept per se is not explanatory. The reason for this is simple: if the problem of an information-processing agent is to make sense of its physical environment, the solution to this problem will not be alleviated at all by representation alone; the agent's task of making sense of the *represented* environment is just as demanding. The theory of direct perception [13] has therefore proposed that information pick-up must occur right at the moment of perception, without any representational interlude. For analogous reasons, the 'storehouse metaphor' of memory has been rejected by researchers of cognitive science [14]. Memory is likely not a passive store out of which represented items, the memory engrams, can be retrieved at a later time, but a more active, constructive process. Eyewitness research (e.g., the so-called false memory syndrome, [31]) has emphasized how modifiable and adaptive the represented contents of memory actually are. Representation-as-mapping has therefore been criticized as merely providing a pseudo solution to a deeper problem, which is likely the very problem of naturalizing intentionality.

Synergetics

We have proposed that cognitive phenomena have attributes of dynamical systems, and that higher functions and more complicated contents of the mind are constructed bottom-up from simpler components by a process of self-organization. It is the goal of this line of argument to show that intentionality can be conceptualized using this framework. Since dynamical systems theory, as well as self-organization theory, are mathematical tools that have been developed for applications in physics, biology and other natural sciences, we may thereby approach the goal of a naturalization of intentionality. The final step prior to formulation of this naturalization proposal is to introduce self-organization theory.

We will rely largely on the interdisciplinary modeling approach of synergetics that deals with complex systems, i. e. systems composed of multiple components [17,18,19]. By way of interactions, these components can produce new qualitative features on macroscopic scales. Synergetics focuses on the emergence of these new qualities, while addressing the question of whether there are general principles governing the behavior of complex systems when such qualitative changes occur. In a large class of systems, it has been shown that they become accessible to unifying mathematical and conceptual approaches. A paradigmatic system is the Bénard system (e.g. [1]), which is comprised of a layer of fluid heated with temperature T_2 from below. The temperature at the upper surface of the fluid

is T_1 . Beyond a critical value of $\Delta T = T_2 - T_1$ extended coordinated motion of the components of the fluid system emerge. Compared to the erratic Brownian motion of the single components, these patterns are an example of the emergence of the new qualities focused on by synergetics.

Synergetics is based on observations that the behavior of many systems is strongly determined by environmental conditions. These conditions may be divided into constant (structural) constraints (e.g. the shape of solid walls that confine fluid systems such as the Bénard system), and into further environmental conditions that energize the systems (e.g. the heat source that drives the Bénard system). In the mathematical approach, these latter driving conditions are expressed by control parameters. In many cases, control parameters take the form of externally applied gradients, which are imposed upon the system from without, such as, for instance, the difference in temperature ΔT of the Bénard experiment. The general strategy in synergetics sets out from a state of a system that is already known under a certain control parameter value. When one or several control parameters are changed, this system can become unstable and show a tendency to leave its state to develop a new structure or behavior. The system in question is described by the states of its individual components, by means of a state vector q . The individual components in the Bénard system, for example, are the motions of single fluid molecules; components may also be, with respect to applications in psychology, the attributes of members within a social group or neurons in the brain.

Synergetics shows that the behavior of the system close to instability points is described and determined by few quantities, namely the order parameters. In the case of a single order parameter n of a complex system a typical equation reads

$$\frac{dn}{dt} = cn \quad (1)$$

where c is the 'effective' control parameter.

$$\begin{aligned} \text{For } c > 0, \quad n &\text{ increases exponentially,} \\ c < 0, \quad n &\text{ decreases exponentially,} \\ c = 0, \quad n &\text{ remains constant.} \end{aligned}$$

As was mentioned, c denotes the control parameter, a relevant parameter imposed on the system from outside, i. e. from the environment of the system. The generally few order parameters enslave, i. e. determine, the behavior of the many individual components. This implies an enormous information compression, because the description of the order parameters alone, rather than of each component, suffices. In the case of the Bénard system, description of

the coordinated motion yields a much more parsimonious description of system behavior than the description of all molecular movements in the Brownian motion state of the fluid. While they are being determined by the order parameters, it is the individual components that react on the order parameters and, by so doing, even generate the latter. The relationship between order parameters and components is, therefore, founded on *circular causality*, which can explain the generally avalanche-like onset of, and transition between, macroscopic states. In other words, this theory favors neither top-down nor bottom-up modeling but claims that both processes are entangled. Order parameters, after they have been generated in this fashion, quite often exhibit very simple behavior, for instance asymptotic stability.

Obviously, the system depicted in Fig. 1 is an open system. Self-organizing systems are invariably open systems in that they depend on control parameters. In terms of thermodynamics, they are non-equilibrium systems.

Let us however first focus on closed systems, in other words, systems in thermal equilibrium. Classic thermodynamics deals with closed systems throughout. The probability of all configurations of components within the (closed) system can be estimated. When dealing with a complex system consisting of a multitude of components, a great many possible realizations of the state vector q exist, namely the number of all combinations W of the states of components. Only a small fraction of these realizations are seen as regular, well-organized patterns. The vast majority of realizations, however, will represent a state of mixture. Should ordered patterns exist as an initial condition, it is far more probable that the temporally consecutive system states will be characterized by less or-

der. This is due to the statistical fact that the majority of possible consecutive states will be states with less order rather than states with the same, or even a higher degree, of order. Within a thermodynamics context, this touches on the concept of entropy S (disorder) in accordance with Boltzmann's statistical approach, in which S is directly related to the number of combinations W . The second law of thermodynamics states that any real closed system can only proceed in the direction of increasing entropy, hence following a maximum entropy principle. Hence, the spontaneous generation of order is highly improbable as indeed a spontaneous generation of disorder is to be expected. In other words, the emergence of pattern from a state of mixture requires explanation; the explanation is that the phenomenon of self-organization is driven by an external source, so that the premises of closed systems do not apply. Since the concept of entropy is defined only for equilibrium or close-to-equilibrium systems one may base the discussion of self-organizing systems on the concept of information [19,47].

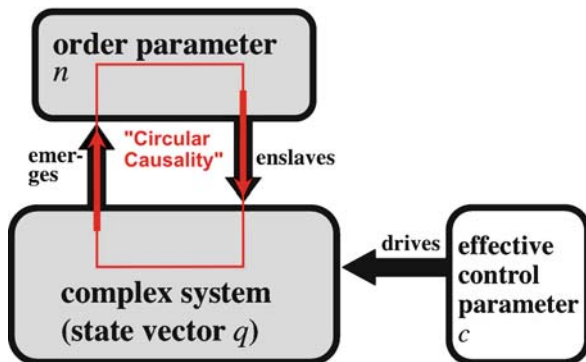
Some authors have applied the laws of thermodynamics in order to allow the study of self-organizing systems. The 'restated second law' of thermodynamics [42,43] addresses non-equilibrium systems, i. e. systems that are forced away from equilibrium by the application of gradients. The degree to which a system is moved away from equilibrium is measured by the gradients imposed on the system. As soon as such gradients dwell in the system's environment, the system will, as a consequence of the restated second law, "[...] utilize all avenues available to counter the applied gradients. As the applied gradients increase, so does the system's ability to oppose further movement from equilibrium" [42]. Schneider and Kay's restatement of the second law avoids some of the problems of defining entropy and entropy production by focusing on the destruction of gradients instead.

It should be kept in mind that this 'destruction of gradients' is only virtual (in analogy to the principle of virtual work in mechanics), because in open systems gradients are generally maintained by the environment. If, however, the self-organizing system and its *finite* environment act as a closed system, the gradient reduction becomes real. In other words, the effective control parameter c depends on the order parameter n ,

$$c = c_0 - \alpha n \quad (2)$$

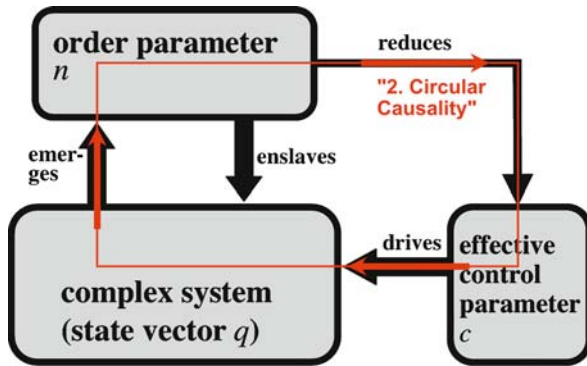
where c_0 is the control parameter prescribed from the outside and a constant. The effective control parameter (2) obeys the differential equation

$$\frac{dc}{dn} = -\frac{dV}{dc} \quad (3)$$



Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, Figure 1

Schematic illustration of circular causality as viewed in synergetics



Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems, Figure 2

Schematic illustration of the relationship between control parameter and order parameter ('second circular causality')

where

$$V = \alpha c \quad (4)$$

is a potential and thus the right-hand side of (3) a gradient.

Discussing the time evolution of n and c according to (1) and (2), we assume that, initially, $n = n(0)$ is close to zero and $c \approx c_0 > 0$.

Thus, according to (1), n increases exponentially. As a consequence, according to (2), c decreases, and the exponential increase of n slows down. This process goes on until $c = 0$ and n reaches a time-independent, i. e. steady, state. In practice, the transition to the new state is completed while the gradient has been reduced. In the Bénard example, the coordinated motion patterns have consumed the applied temperature difference and have reduced ΔT to 0. Then the motion patterns subside and steady state remains. This relationship between emergent pattern and control parameter is thus in line with the notion of gradient destruction. The reduction of c by n establishes a second kind of *circular causality*, which is schematically illustrated in Fig. 2.

Discussion: Naturalization of Intentionality

The steps above have put us in a position where the naturalization of intentionality comes within reach. We propose that this can be achieved on the basis of the properties of self-organizing complex systems. In this section we will therefore discuss to what extent the features of intentionality (listed in Sect. "Definition of the Subject") can be approximated by such systems.

Aboutness

Intentionality implies that a system state is about something else, namely the intentional object. In terms of cog-

nitive psychology, this process is called representation, by which a 'cognitive map' of the object is generated particularly during perceptual or memory processes.

To be capable of intentionality and representation, a minimum requirement is that the intentional system must be an open system. Many open systems can provide representations in the sense of mappings of environmental impacts. The silver particles of a light-sensitive surface of a photographic film can 'represent' the objects in front of the lens, however in a trivial, weak sense. As has been shown in the previous section, a self-organizing system can likewise 'represent' and thus generate the feature of being about something. In the latter system the order parameter is a component in the loop labeled 'second circular causality' (Fig. 2). Within this loop the complex system 'represents' an external object by the generation of an order parameter. The intentional object in this case is the external control parameter. The environmental condition described by control parameters is what self-organized patterns 'are about'.

The mechanisms of representation are clearly divergent in these two systems, as is the nature of the intentional objects. In the photographic system the mapping of the environmental objects onto the representing system is unidirectional, whereas in the self-organizing system there is continuous interaction between environmental objects and system. This circularity is illustrated in Fig. 2 and provides an important and desirable aspect of the kind of aboutness realized by such systems. Circularity guarantees that an intentional system is capable of exerting a retrograde effect back on what has been represented. This effect is generally a reduction of the gradient quantified by the control parameter.

Functionality

Intentionality must be functional in order to make representation explanatory, i. e. make representation more than just a mapping of environmental states onto the system's state.

The functionality of open systems has previously been approached from the angle of thermodynamics. Schneider and Sagan [43] pointed out that self-organizing systems maintain and increase their levels of organization by dissipating non-equilibrium gradients. If the gradient is to be kept constant, the demand on free energy (so-called exergy) that must be provided by an external source increases as the system becomes more organized. Alternatively, the efficiency of the system can be defined as the ratio of the change of work and the change of the gradient driving the system. This can be shown by findings

in simple physical systems that generate patterns; e. g. Bénard cells reduce the temperature gradient more efficiently as soon as they have generated ordered convection structures. Efficiency has thus increased in the self-organized convection regime in comparison to the linear conduction regime of the fluid. Analogous relationships are found in further self-organizing systems such as the laser when output power is plotted against input power [17]. In other words, pattern formation in these open systems is in the service of gradient reduction. The association of pattern formation with gradient reduction makes pattern formation functional.

In situations of multistability, several patterns are possible, so that each of these can be functional in reducing the gradients imposed on the system. These alternative patterns can be associated with different efficiencies. It is theoretically suggestive that the two circular loops pointed out in Fig. 1 and 2 interact and thereby create a darwinistic scenario in which a competition between microscopic modes q arises; the environmental forces c exert a selective impact on this competition of modes. From a mathematical modeling point of view it is not however clear if the optimal pattern is necessarily selected, nor under which circumstances the optimal pattern will be chosen. In some systems, such as quadruped movement coordination, it was empirically found that the specific pattern that provides the most efficient behavior (measured as the metabolic cost of transport of the animal) will be realized by the system [25]. The generalizability of such findings of optimality is not yet established.

Mental-Likeness

In this discussion we have so far found that the aboutness of intentionality and representation can be modeled by open systems. Only a subclass of open systems provides functional representations. Especially, certain self-organizing non-equilibrium systems appear to show the two features of aboutness and functionality.

It is obviously useful to reserve the predicate 'intentional' for mental systems alone. Thus, even though some physical self-organizing systems may show circular causality loops and thus stand the tests of aboutness and functionality, we would still categorize systems such as lasers, Bénard cells etc. as being not mental. We may say that such self-organizing physical systems behave 'as if they were intentional', i. e. they are proto-intentional systems. Let us finally investigate under which conditions self-organizing non-equilibrium systems may also show mental features and, with this, conclude the proposal of naturalizing intentionality. This final step is basically a discussion of the

validity of the statement that self-organizing systems can show intentionality.

- *Complexity reduction* is a core hallmark of mental processes. The ability of a system to simplify, group and coordinate environmental information is a necessary premise for any system to be mental. This property is addressed in the circular causality concept of synergetics illustrated in Fig. 1. With respect to information compression, self-organizing physical systems are mental-like.
- *Stability* together with related signatures of stability (e. g. hysteresis, critical fluctuations) is a further mental property, which is empirically well founded especially in the psychology of perception. Again, the emergent order parameters of self-organizing physical systems generally show this property.
- *Autonomy* is required of intentional mental systems, i. e. mental systems must be able to function in the absence of external agents. This ability addresses the 'homunculus-problem' that has already been introduced together with the feature of functionality. We may say that generally self-organizing systems do not require external supervision for producing order parameters (therefore, *self-organizing*). This does not rule out that several autonomous systems may be nested inside one another (cf. Minsky's society of mind [34], or the notion of subsumption architecture of Brooks' [5] robotic agents).
- What is the nature of the intentional content? Which aspect of the environment is being represented by self-organizing systems? Order parameters are intentional with respect to those environmental parameters that drive the system. In other words, the intentional content is generally connected to what energizes the system. This is a satisfactory model in all those instances where mental intentionality is of a motivational character, resembling a psychological 'drive' [12]. In the introduction we specified wishes, desires, affects, intentions, goals and the like as intentional. Such intentional states can be directly modeled by self-organizing systems that act to reduce the driving parameters (in Freudian terms, drive reduction). The self-organization model, however, is less applicable whenever intentionality is of a language-like, propositional type [11]. It is quite a different task to model, for instance, the intentionality inherent in the belief that "there is a unicorn grazing in the garden", using simple self-organizing systems. Therefore, the discussion of mental-likeness of self-organizing systems remains restricted to nonlinguistic intentionality.

- The only mental systems known to date, despite the efforts of several decades of artificial intelligence research, are *neural networks* inside biological organisms. Therefore, can the argumentation above be applied to brain dynamics and to pattern formation in neural networks? One may then associate the gradient of c with intentionality in a neurocognitive sense. Haken [20] has discussed synchronization, i. e. self-organization, of neural nets using various mathematical frameworks. Haken and Tschacher [23] have specifically addressed the reduction of the effective control parameter on the basis of the Wilson–Cowan equations describing cortical dynamics [51]. Haken and Tschacher suggested that the general findings on circular causality in the relationship between order parameter and control parameter can be readily applied to neural networks.

In conclusion, we have argued that a formulation of intentionality is feasible on the basis of the theory of nonlinear dynamical systems. When such systems are removed from thermodynamical equilibrium they acquire the capability of producing self-organized patterns. Pattern formation consequently puts the systems in a specific relationship to environmental parameters. Owing to the accompanying circular loops, these systems show aboutness, the defining property of intentionality, as well as functionality, which is essential for making the aboutness of intentional states explanatory. Finally, it can be shown that some of these self-organizing systems are mental-like because their behavior empirically shows signatures of mental phenomena beyond the features of intentionality. We therefore conclude that intentionality can be naturalized to a considerable extent using non-equilibrium complex systems.

Some caveats have turned up during this discussion, which may provide points of departure for future research. First, it is as yet unclear if the resulting self-organized patterns obey an optimization principle. In some empirically described systems showing multistability it could be shown that the more optimal pattern wins the competition among order parameters, but the generality of such findings is yet to be corroborated. Correspondingly, Haken and Tschacher [23] have discussed ‘second circular causality’ in the case of only one order parameter. A mathematical model comprising $m > 1$ order parameters would be a desirable next step. Second, we found that naturalization of intentional states is achievable when these states comprise motivation, goals, intentions, or drives; these may be viewed as basic intentional states related to behavioral strivings. At the moment, however, the intentionality problem appears intractable when these intentional states

are of a symbolic and propositional nature. The difficulties of modeling Fodor’s language-of-thought do indeed complicate the modeling of intentionality beyond its basic form (cf. Churchland [8], p. 304). At the present time, we may consider non-linguistic intentionality as the primary problem, which in fact seems accessible. The solution to this problem might support the clarification of symbolic, secondary intentionality in the future. This restriction to non-linguistic intentionality is likely related to a further limitation. Contrary to von Brentano’s supposition, mental phenomena may have to be viewed as intentional *as well as* phenomenal. Systems theory addresses solely the primary aspects of intentionality, yet not the ‘hard problem’ [6] of the phenomenal nature of intentionality. It does not lie within the scope of the present treatment to determine if and how the dynamical view can illuminate the hard question of consciousness.

Future Directions

The topic of intentionality has numerous implications beyond the philosophy of mind and theoretical psychology.

In psychiatry and in psychotherapy research, intentionality is a topic of considerable significance because intentional mental acts and states are often characteristically altered or disturbed during a mental disorder. Many such psychopathological conditions are found especially among the symptoms of schizophrenia, such as disorders of formal thought as well as of thought content, disorders of perception, and ego disorders. The symptoms are heterogeneous and manifold. A majority of these symptomatic alterations, however, involve changes in the cognitive coordination of the patients [40]. Recent schizophrenia research has shown that a considerable portion of the variance of psychotic symptoms can be explained by cognitive coordination measures [48]. While no generally accepted encompassing theory of schizophrenia exists, a theory of intentionality may have the potential to contribute to progress in psychopathology. It may help to link the phenomenology and neurobiology of schizophrenia and other psychiatric disorders by introducing a dynamical systems perspective.

Artificial intelligence is a completely different field, but is also confronted with intentionality as a core problem. In recent decades, the computational approach to machine intelligence has failed to a large extent; especially, no mental-like intelligence could be generated. Consequently the field has turned to the more basic tasks of designing autonomous agents and robots with rudimentary adaptivity in the real world [3,38]. In this framework, intelligence is expected to be closely associated with embodi-

ment (hence, embodied intelligence, mind–body co-evolution), rather than with symbol manipulation and the programming of symbol systems. The latter constituted the classic approach to artificial intelligence. One of the novel directions of development is using emergence principles for the design of intelligent agents or multi-agent systems; this is closely related to the view on intentionality proposed here. The engineering approach of embodied agents design and the dynamical systems proposal of intentionality have the potential of subserving each other in artificial intelligence research.

Bibliography

Primary Literature

- Bianciardi C, Ulgiati S (1998) Modelling entropy and exergy changes during a fluid self-organization process. *Ecol Modelling* 110:255–267
- Bickhard M, Terveen L (1995) Foundational issues in artificial intelligence and cognitive science. Holland, Amsterdam
- Braitenberg V (1986) Künstliche Wesen: Verhalten kybernetischer Vehikel. Vieweg, Braunschweig
- Brentano F von (1874) Psychologie vom empirischen Standpunkte. Duncker & Humblot, Leipzig
- Brooks R (1991) Intelligence without Representation. *Artificial Intell* 47:139–159
- Chalmers D (1996) The conscious mind. Oxford University Press, Oxford
- Churchland P (1986) Neurophilosophy: Toward a unified science of the mind-brain. MIT Press, Cambridge
- Churchland P (2002) Brain-Wise: Studies in neurophilosophy. MIT Press, Cambridge
- Clancey WJ (1993) Situated action. A neurophysiological interpretation response to Vera and Simon. *Cogn Sci* 17:87–116
- Clark A (1997) Being there: Putting brain, body, and world together again. MIT Press, Cambridge
- Fodor J (1975) The language of thought. Crowell, New York
- Freud S (1923) Das Ich und das Es [published in English (1949) The Ego and the Id. The Hogarth Press Ltd., London]
- Gibson JJ (1979) The ecological approach to visual perception. Houghton Mifflin, Boston
- Glenberg A (1997) What memory is for. *Behav Brain Sci* 20:1–19
- Gollwitzer PM, Bargh GA (eds) (1996) The psychology of action: Linking cognition and motivation to behavior. Guilford, New York
- Greeno J, Moore J (1993) Situativity and symbols: Response to Vera and Simon. *Cogn Sci* 17:49–59
- Haken H (1977) Synergetics – An introduction. Nonequilibrium phase-transitions and self-organization in physics, chemistry and biology. Springer, Berlin
- Haken H (1996) Principles of brain functioning: A synergetic approach to brain activity, behavior, and cognition. Springer, Berlin
- Haken H (2000) Information and self-organization: A macroscopic approach to complex systems. Springer, Berlin
- Haken H (2002) Brain dynamics. Synchronization and activity patterns in pulse-coupled neural nets with delays and noise. Springer, Berlin
- Haken H, Stadler M (eds) (1990) Synergetics of cognition. Springer, Berlin
- Haken H, Kelso J, Bunz H (1985) A theoretical model of phase transitions in human hand movements. *Biol Cybern* 51:347–356
- Haken H, Tschacher W (submitted) A theoretical model of intentionality based on self-organizing systems
- Harnad S (1990) The symbol grounding problem. *Physica D* 42:335–346
- Hoyt DT, Taylor CR (1981) Gait and the energetics of locomotion in horses. *Nature* 292:239–240
- Kelso JAS (1995) Dynamic Patterns: The self-organization of brain and behavior. MIT Press, Cambridge
- Köhler W (1920) Die physischen Gestalten in Ruhe und in stationärem Zustand. Vieweg, Braunschweig
- Kuhl J, Beckmann J (1994) Volition and personality. Action versus state orientation. Hogrefe and Huber Publishers, Göttingen
- Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. *Trends Cogn Sci* 3:254–264
- Lewin K (1936) Principles of topological psychology. McGraw-Hill, New York
- Loftus EL (2003) Our changeable memories: Legal and practical implications. *Nat Rev Neurosci* 4:231–234
- McClelland DC, Atkinson JW, Clark RA, Lowell EL (1953) The achievement motive. Van Nostrand, Princeton
- Metzinger T (2003) Being no one: The self-model theory of subjectivity. MIT Press, Cambridge
- Minsky M (1985) Society of mind. Simon & Schuster, New York
- Neisser U (1976) Cognition and reality. Principles and implications of cognitive psychology. Freeman, San Francisco
- Newell A (1980) Physical symbol systems. *Cogn Sci* 4:135–183
- Nicolis G, Prigogine I (1977) Self-organization in non-equilibrium systems. Wiley, New York
- Pfeifer R, Bongard JC (2006) How the body shapes the way we think. A new view of intelligence. MIT Press, Cambridge
- Pfeifer R, Scheier C (1999) Understanding intelligence. MIT Press, Cambridge
- Phillips WA, Silverstein SM (2003) Convergence of biological and psychological perspectives on cognitive coordination in schizophrenia. *Behav Brain Sci* 26:65–138
- Popper K, Eccles J (1977) The self and its brain. Springer, Berlin
- Schneider ED, Kay JJ (1994) Life as a manifestation of the second law of thermodynamics. *Math Comp Modelling* 19:25–48
- Schneider ED, Sagan D (2005) Into the cool. Energy flow, thermodynamics and life. University of Chicago Press, Chicago
- Thelen E, Smith LB (1994) A dynamic systems approach to the development of cognition and action. MIT Press, Cambridge
- Tschacher W (1997) Prozessgestalten [Processual gestalten]. Hogrefe, Göttingen
- Tschacher W, Dauwalder J-P (eds) (2003) The dynamical systems approach to cognition. World Scientific, Singapore
- Tschacher W, Haken H (2007) Intentionality in non-equilibrium systems? The functional aspects of self-organized pattern formation. *New Ideas in Psychol* 25:1–15
- Tschacher W, Schuler D, Junghan U (2006) Reduced perception of the motion-induced blindness illusion in schizophrenia. *Schizophrenia Res* 81:261–267
- van Gelder T (1998) The dynamical hypothesis in cognitive science. *Behav Brain Sci* 21:615–628
- Wertheimer M (1912) Experimentelle Studien über das Sehen

von Bewegungen [Experimental studies on the perception of motion]. *Z Psychol* 61:165–292

51. Wilson HR, Cowan JD (1972) Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys J* 12:1–24

Books and Reviews

- Beckermann A (2001) *Analytische Einführung in die Philosophie des Geistes*. de Gruyter, Berlin
- Carter R (2002) *Consciousness*. Weidenfeld & Nicolson, London
- Chaisson EJ (2001) *Cosmic evolution. The rise of complexity in nature*. Harvard University Press, Cambridge
- Ciampi L (1982) *Affektlogik*. Klett–Cotta, Stuttgart [published in English (1988): *The psyche and schizophrenia. The bond between affect and logic*. Harvard University Press, Cambridge]
- Dennett DC (1987) *The intentional stance*. MIT Press, Cambridge
- Dennett DC, Kinsbourne M (1992) Time and the observer: The where and when of consciousness in the brain. *Behav Brain Sci* 15:183–247
- Guastello S, Koopmans M, Pincus D (eds) (2008) *Chaos and complexity in psychology. Theory of nonlinear dynamical systems*. Cambridge University Press, Cambridge
- Kauffman S (1993) *The origins of order. Self-organization and selection in evolution*. Oxford University Press, New York
- Kruse P, Stadler M (eds) (1995) *Ambiguity in mind and nature*. Springer, Berlin
- Newell A, Simon HA (1972) *Human problem solving*. Prentice–Hall, Englewood Cliffs
- Port R, van Gelder TJ (eds) (1995) *Mind as motion: Explorations in the dynamics of cognition*. MIT Press, Cambridge
- Searle JR (1998) *The rediscovery of mind*. MIT Press, London
- Singer W, Gray CM (1995) Visual feature integration and the temporal correlation hypothesis. *Annu Rev Neurosci* 18:555–586
- Storch M, Cantieni B, Hühner G, Tschacher W (2006) *Embodiment*. Huber, Bern
- Swenson R, Turvey MT (1991) Thermodynamic reasons for perception–action cycles. *Ecol Psychol* 3:317–348
- Tschacher W, Dauwalder J-P (eds) (1999) *Dynamics, synergetics, autonomous agents. Nonlinear systems approaches to cognitive psychology and cognitive science*. World Scientific, Singapore
- Tschacher W, Scheier C (2003) *Der interaktionelle Ansatz in der Kognitionswissenschaft: Eine Positionsarbeit zu Konzepten, Methoden und Implikationen für die Psychologie* [The interactional approach to cognitive science: Concepts, methods, and implications for psychology]. *Zeitschrift für Psychologie* 211:2–16
- Varela F, Thompson E, Rosch E (eds) (1991) *The embodied mind*. MIT Press, Cambridge
- Varela FJ (1995) Resonant cell assemblies: A new approach to cognitive function and neuronal synchrony. *Biol Res* 28:81–95

Interaction Based Computing in Physics

FRANCO BAGNOLI

University of Florence, Florence, Italy

Article Outline

[Glossary](#)

[Definition](#)

[Introduction: Physics and Computers](#)

[From Trajectories to Statistics and Back](#)

[Artificial Worlds](#)

[Discussions and Conclusions](#)

[Bibliography](#)

Glossary

Nonlinear system A system composed by parts whose combined effects are different from the sum of the effects of each part.

Extended system A system composed by many parts connected by a network of interactions that may be regular (lattice) or irregular (graph).

Graph, lattice, tree A graph is set of nodes connected by links, oriented or not. If the graph is translationally invariant (it looks the same when changing nodes), it is called a (regular) lattice. A disordered lattice is a lattice with a fraction of removed links or nodes. An ordered set of nodes connected by links is called a path. A closed path not passing on the same links is a loop. A cluster is a set of connected nodes. A graph can be composed by one cluster (a connected graph) or more than one (a disconnected graph). A tree is a connected graph without loops.

Percolation The appearance of a “giant component” (a cluster that contains essentially all nodes or links) in a graph or a lattice, after adding or removing nodes or links. Below the percolation threshold the graph is partitioned into disconnected clusters, none of which contains a substantial fraction of nodes or links, in the limit of infinite number of nodes/links.

State of a system A complete characterization of a system at a given time, assigning or measuring the positions, velocities and other dynamical variables of all the elements (sometimes called a configuration). For completely discrete systems (cellular automata) of finite size, the state of the system is just a set of integer numbers, and therefore the state space is numerable.

Trajectory A sequence of states of a system, labeled with the time, i. e., a path in the state space.

Probability distribution The probability of finding a system in a given state, for all the possible states.

Mean field An approximate technique for computing the value of the observables of an extended system, neglecting correlations among parts. If necessary, the dynamics is first approximated by a stochastic process. In its simpler version, the probability of a state of the sys-

tem is approximated by the product of the probability of each component, neglecting correlations. Since the state of two components that depend on a common “ancestor” (that interact with a common node) is in general not uncorrelated, and this situation corresponds to an interaction graph with loops, the simplest mean field approximation consists in replacing the graph or the lattice of interactions with a tree.

Monte-Carlo A method for producing stochastic trajectories in the state space designed in such a way that the time-averaged probability distribution is the desired one.

Critical phenomenon A condition for which an extended system is correlated over extremely long distances.

Definition

Physics investigation is based on building models of reality: in order for a phenomenon to be *understood*, we need to represent it in our minds using a limited amount of symbols. However, it is a common experience that, even using simple “building blocks” one usually obtains systems whose behavior is quite complex. In this case one needs to develop new languages and new *phenomenological* models in order to manage this “complexity”.

Computers have changed the way a physical model is studied. Computers may be used to *calculate* the properties of a very complicated model representing a real system, or to investigate *experimentally* what are the essential ingredients of a complex phenomenon. In order to carry out these explorations, several basic models have been developed, which are now used as building blocks for performing simulations and designing algorithms in many fields, from chemistry to engineering, from natural sciences to psychology. Rather than being derived from some fundamental law of physics, these blocks constitute *artificial worlds* still to be completely explored.

In this article we shall first present a pathway from Newton’s laws to cellular automata and agent-based simulations, showing (some) computational approaches in classical physics. Then, we shall present some examples of *artificial worlds* in physics.

Introduction: Physics and Computers

Some sixty years ago, shortly after the end of Second World War, computers become available to scientists. Computers were used during the last years of the war for performing computations about the atomic bomb [30,41].

Up to then, the only available tool, except experiments, was paper and pencil. Starting with Newton and Leibnitz, humans discovered that continuous mathematics (i. e., dif-

ferential and integral calculus) allowed to derive many consequences of a given hypothesis just by the manipulation of symbols. It seemed natural to express all quantities (e. g., time, space, mass) as continuous variables. Notice however that the idea of a continuous number is not at all “natural”: one has to learn how to deal with it, while (small) integer numbers can be used and manipulated (added, subtracted) by illiterate humans and also by many animals. A point which is worth to be stressed is that any computation refers to a model of certain aspects of reality considered most important, while others are assumed to be not important

Unfortunately most of human-accessible explorations in physics are limited to almost-linear systems, or systems whose effective number of variables is quite small. On the other hand, most of naturally occurring phenomena can be “successfully” modeled only using nonlinear elements. Therefore, most of pre-computer physics is essentially linear physics, although astronomers (like other scientists) used to integrate numerically, by hand, the nonlinear equations of gravitation, in order to compute the trajectories of planets. This computation, however, was so cumbersome that no “playing” with trajectories was possible.

While analog computers have been used for integrating differential equations, the much more flexible digital computers are deterministic discrete systems. The way of working of a (serial) computer is that of a very fast automaton, that manipulates data following a program.

In order to use computers as fast calculators, scientists ported and adapted existing numerical algorithms, and developed new ones. This implied the development of techniques able to *approximate* the computations of continuous mathematics using computer algebra. However, numbers in computers are not exactly the same as human numbers, in particular they have finite (and varying) precision.

This intrinsic precision limit has deep consequences in the simulations of nonlinear systems, in particular of chaotic ones. Indeed, chaos was “numerically discovered” by Lorenz [38] after the observation that a simple approximation, a number that was retyped with fewer decimals, caused a macroscopic change in the trajectory under study.

With all their limits, computers can be fruitfully used *just* to speed-up computations that *could* eventually be performed by humans. However, since the increase in velocity is of several order of magnitude, it becomes possible to include more and more details into the basic model of the phenomenon under investigation, well beyond what would be possible with an army of “human computers”. The idea of exploiting the *brute power* of fast comput-

ers has originated a fruitful line of investigation in *numerical physics* especially in the field of chemistry, biological molecules, structure of matter. The power of computers has allowed for instance to include quantum mechanical effects in the computation of the structure of biomolecules [16], and although these techniques may be targeted as “brute force”, the algorithms developed are actually quite sophisticated.

However, a completely different usage of computers is possible: instead of exploiting them for performing computations on models that already proved to approximate the reality, one can use computers as “experimental” apparatus to investigate the patterns of *theoretical* models, generally non-linear. This is what Lorenz did after having found the first example of chaos in computational physics. He started simplifying his equations in order to enucleate the minimal ingredients of what would be called the *butterfly effect*.

Much earlier than Lorenz, Fermi, Pasta and Ulam (and the programmer Tsingou [20]) used one of the very first available computers to investigate the basis of statistical mechanics: how energy distributes among the oscillation modes of a chain of nonlinear oscillators [25].

Also in this case the model is simplified at its maximum, in order to put into evidence what are the fundamental ingredients of the observed pattern, and also to use all the available power of computers to increase the precision, the duration and the size of the simulation.

This simplification is even more fruitful in the study of systems with many degrees of freedom, that we may denote generically as *extended systems*. We humans are not prepared to manipulate more than a few symbols at once. So, unless there is a way of grouping together many parts (using averages, like for instance when considering the pressure of a gas as an average over extremely many particle collisions), we are in difficulties in *understanding* such systems. They may nevertheless be studied performing “experiments” on computers. Again, the idea is that of simplifying at most the original model, in order to isolate the fundamental ingredients of the observed behavior. It is therefore natural to explore systems whose *physics* is different from the usual one. These *artificial worlds* are preferably formulated in discrete terms, more suitable to be implemented in computers (see Sect. “Artificial worlds”).

This line of investigation is of growing interest today: since modern computers can easily simulate thousands or millions of *elementary automata* (often called *agents*), it may be possible to design *artificial worlds* in which *artificial people* behave similarly to real humans. The rules of these worlds are not obtained from the “basic laws” of the

real one, since no computer can at present simulate the behavior of all the elements of even a small portion of matter. These rules are designed so to behave similarly to the system under investigation, and to be easily implemented in digital terms. There are two main motivations (or hopes): to be able to understand real complex dynamics by studying simplified models, and to be so lucky to discover that a finely-tuned model is able to reproduce (or forecast) the behavior of its real counterpart.

This is so promising that many scientists are performing experiments on these artificial worlds, in order to extract their principal characteristics, to be subsequently analyzed possibly using paper and pencil!

In the following we shall try to elucidate some aspects of the interplay between computer and physics. In Sect. “From Trajectories to Statistics and Back”, we shall illustrate possible logic pathways (in classical mechanics) leading from Newton’s equations to research fields that use computers as investigative tool, like agent-based investigations of human societies. In Sect. “Artificial worlds”, we shall try to present succinctly some example of “artificial worlds” that are still active research topics in theoretical physics.

From Trajectories to Statistics and Back

The outline of this Section is the following. Physics is often denoted the most “fundamental” science, and one may think that, given powerful enough computers, one should be able to reconstruct any experimental situation simply implementing the fundamental laws of physics. I would like to show that any investigation is based on models, requiring approximations and additional assumptions, and that any change of scale implies a change of model. However, an important suggestion from physics is that similar models can be used to interpret different situations, and therefore the associated computational techniques can be “reused” in different contexts. We shall follow this line from Newton’s equations to agent-based modeling.

Let us assume that a working model of the reality can be built using a set of dynamical equations, for instance those of classical mechanics. We shall consider the model of a system formed by many particles, like a solid or a fluid. The state of the resulting system can be represented as a point in a high-dimensional space, since it is given by all coordinates and velocities of all particles. The evolution of the system is a trajectory in such space. Clearly, the visualization and the investigation of such a problem is challenging, even using powerful computers.

Moreover, even if we are able to compute one or many trajectories (in order to have an idea of fluctuations), this

does not imply that we have *understood* the problem. Let us consider for instance the meteorology: one is interested in the probability of rain, or in the expected wind velocity, not in forecasting the trajectories of all molecules of air. Similarly in psychology, one is interested in the expected behavior of an individual, not in computing the activity of all neurons in his brain.

Since physics is the oldest discipline that has been “quantified” into equations, it may be illuminating to follow some of the paths followed by researchers to “reduce” the complexity of a high-dimensional problem to something more manageable, or at least simpler to be simulated on a computer.

In particular, we shall see that many approaches consist in “projecting” the original space onto a limited number of dimensions, corresponding to the observables that vary in a slow and smooth way, and assuming that the rest of the dynamics is approximated by “noise”¹. Since the resulting system is stochastic, one is interested in computing

¹The noise can be so small, compared to the macroscopic observables that it can be neglected. In such cases, one has a deterministic, low-dimensional dynamical system, like for instance the usual models for rigid bodies, planets, etc.

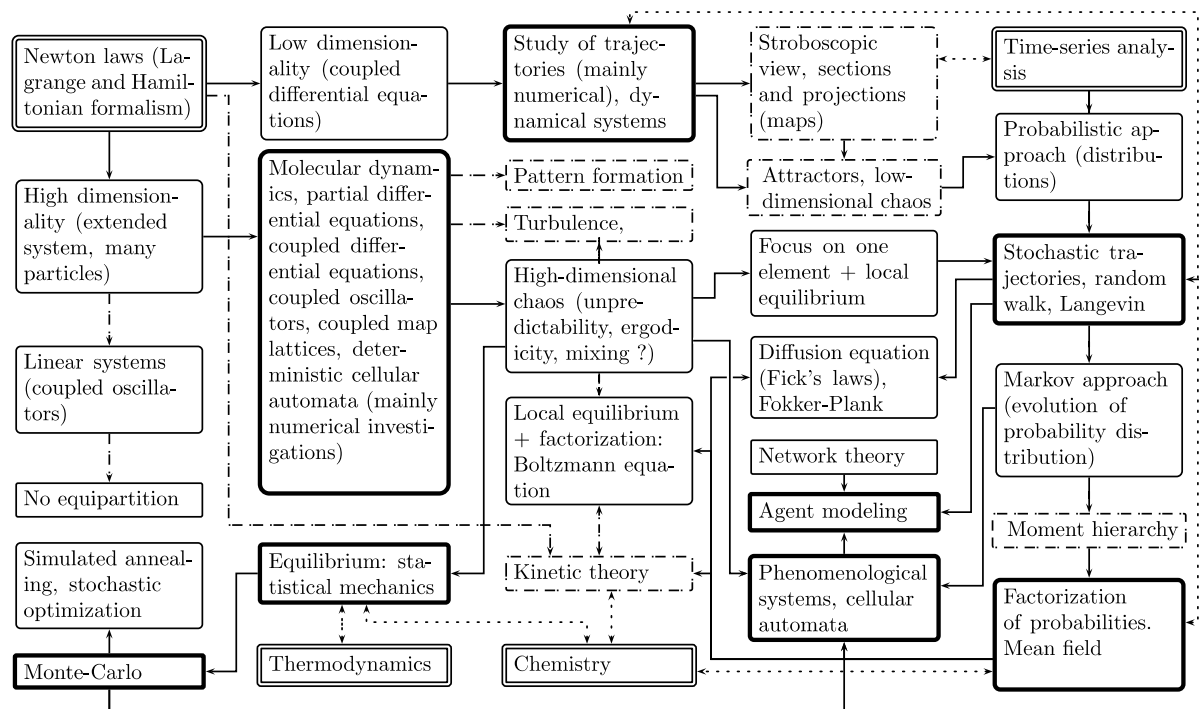
the average values of observables, over the probability distribution of the projected system.

However, the computation of the probability distribution may be hard, and so one seeks to find a way of producing “artificial” trajectories, in the projected space, designed in such a way that their probability distribution is the desired one. So doing, the problem reduces to the computation of the time-averaged values of “slow” observables. For the rest of this section, please make reference to Fig. 1.

Newton Laws

The success of Newton in describing the motion of one body, subjected to a static field force (say: gravitational motion of one planet, oscillation of a body attached to a spring, motion of the pendulum, etc.) clearly proved the validity of his approach, and also the validity of using simple models for dealing with natural phenomena. Indeed, the representation of a body as a point mass, the idea of massless springs and strings, the concept of force fields are all mathematical idealizations of reality.

The natural generalization of this approach is carried out in the XVIII century by Lagrange, Hamilton and many others. It brings to the mathematisation of mechanics and



Interaction Based Computing in Physics, Figure 1

Graphical illustration of the logic path followed in this introduction. Boxes with double frame are “starting points”, dashed boxes are topic that are not covered by discussion, boxes with darker frames mark topics that are investigated more in details

the derivation of *rational mechanics*. The resulting “standard” (or historical) way of modeling physical systems is that of using differential equations, i. e., a continuous description of time, space and other dynamical quantities.

From an abstract point of view, one is faced with two different options: either concentrate on systems described by a few equations (low-dimensional systems), or try to describe systems formed by many components.

Low Dimensionality

Historically, the most important problem of Newton’s times was that of three bodies interacting via gravitational attraction (the Sun, the Earth and the Moon). By approximating planets with point masses, one gets a small number of coupled differential equations. This reduction of dimensionality is an example of a “scale separation”: the variables that describe the motion of the planets vary slowly and smoothly in time. Other variables, for instance those that describe the oscillations of a molecule on the surface of a planet, can be approximated by a noise term so small that can be safely neglected. This approximation can also be seen as a “mean field” approach, for which one assumes that variables behave not too differently from their average. Using these techniques, one can develop models of many systems that result in the same mathematical scheme: a few coupled equations. The resulting equations may clearly have a structure quite different from that resulting from Newtonian dynamics (technically, Hamiltonian systems).

However, the reduction of the number of variables does not guarantee the simplicity of the resulting model. The problem of three gravitational bodies cannot be split into smaller pieces, and the computation of an accurate trajectory requires a computer. In general, a *nonlinear* system in a space with three or more dimensions is chaotic. This implies that there it may “react” to a small perturbation of parameters or initial conditions with large variations of its trajectory. This *sensibility* to variation implies the impossibility of predicting its behavior for long times, unless one is content with a probabilistic description.

High Dimensionality

In many cases, the “projection” operation results in a system still composed by many parts. For instance, models of nonequilibrium fluids neglect to consider the movement of the individual molecules, but still one has to deal with the values of the pressure, density and velocity in all points. In these cases one is left with a high-dimensional problem. Assuming that the “noise” of the projected dimensions can be neglected, one can either write down a large number of

coupled equation (e. g., in modeling the vibration of a crystal), or use a continuous approach and describe the system using partial differential equations (e. g., the model of a fluid).

Linear Systems

In general, high and low-dimensional approaches can be systematically developed (with paper and pencil) only in the linear approximation. Let us illustrate this point for the case of coupled differential equation: if the system is linear one can write the equations using matrices and vectors. One can in principle find a (linear) transformation of variables that make the system diagonal, i. e., that reduces the problem to a set of *uncoupled* equations. At this point, one is left with (many) one-dimensional independent problems. Clearly, there are mathematical difficulties, but the path is clear. A similar approach (for instance using Fourier transforms) can be used also for dealing with partial differential equations.

The variables that result from such operations are called normal modes, because they behave independently one from the other (i. e., they correspond to orthogonal or normal directions in the state space). For instance, the linear model of a vibrating string (with fixed ends) predicts that any pattern can be described as a superposition of “modes”, which are the standing oscillations with zero, one, two, ... nodes (the harmonics).

However, linear systems behave in a somewhat strange way, from the point of view of thermal physics. Let us consider for instance the system composed by two *uncoupled* oscillators. It is clear that if we excite one oscillator with any amount of energy, it will remain confined to that subsystem. With normal modes, the effect is the same: any amount of energy communicated to a normal mode remains confined to that mode, if the system is completely linear. In other words, the system never forgets its initial condition.

On the contrary, the long-time behavior of “normal” systems does not depend strongly on the initial conditions. One example is the *timbre*, or “sound color” of an object. It is given by the simultaneous oscillations on many frequencies, but in general an object emits its “characteristic” sound regardless of how exactly is perturbed. This would not be true for linear systems.

Since the *distribution* of energy to all available “modes” is one of the assumptions of equilibrium statistical mechanics, which allows us to “understand” the usual behavior of matter, we arrived at an unpleasant situation: linear systems, which are so “easy” to be studied, cannot be used to ground statistical physics on mechanics.

Molecular Dynamics

Nowadays, we have computers at our disposal, and therefore we can simulate systems composed by many parts with complex interactions. One is generally interested in computing *macroscopic* quantities. These are defined as *averages* of some function of the microscopic variables (positions, velocities, accelerations, etc.) of the system. A measurement on a system implies an average, over a finite interval of time and over a large number of elementary components (say: atoms, molecules, etc.) of some quantity that depends on the microscopic state of that portion of the body.

Chaos and Probabilities

It was realized by Poincaré (with paper and pencil) and Lorenz (with one of the very first computers) that also very few (three) coupled differential equations with nonlinear interactions may give origin to complex (chaotic) behavior. In a chaotic system, a small uncertainty amplifies exponentially in time, making forecasting difficult. However, chaos may also be simple: the equations describing the trajectory of dice are almost surely chaotic, but in this case the chaos is so strong that the tiniest perturbation or uncertainty in the initial conditions will cause in a very small amount of time a complete variation of the trajectory. Our experience says that the process is well approximated by a probabilistic description. Therefore, chaos is one possible way of introducing probability in dynamics.

Chaotic behavior may be obtained in simpler models, called maps, that evolve in discrete time steps. As May showed [40], a simple map with a quadratic non-linearity (logistic map) may be chaotic. One can also model a system using coupled maps instead of a system of coupled differential equations [34]. And indeed, when a continuous system is simulated on a computer, it is always represented as an array of coupled maps.

Discretization

There is a progression of discretization from partial differential equations, coupled differential equations, coupled map lattices: from systems that are continuous in space, time and in the dynamical variables to systems that are discrete in time and space, and continuous only in the dynamical variables. The further logic step is that of studying completely discrete systems, called *cellular automata*.

Cellular automata show a wide variety of different phenomenologies. They can be considered mathematical tools, or used to model reality. In many cases, the resulting phenomenological models follows probabilistic rules,

but it is also possible to use cellular automata as “building blocks”. For instance, is possible to simulate the behavior of a hydrodynamical system by means of a completely discrete model, called cellular automata lattice gas [26,29].

Statistics

The investigation of chaotic extended systems proceed generally using a statistical approach. The idea is the following: any system contains a certain degree of non-linearity, that couples otherwise independent normal modes. Therefore, (one hopes that) the initial condition is not too important for the asymptotic regime. If moreover one assumes that the motion is so chaotic that any trajectory spans the available space in a “characteristic” way (again, not depending on the initial conditions), we can use statistics to derive the “characteristic” probability distribution: the probability of finding the system in a given portion of the available space is proportional to the time that the system spends in that region. See also the paragraph on equilibrium.

Random Walks

Another approach is that of focusing on a small part of a system, for instance a single particle. The rest of the system is approximated by “noise”. This method was applied, for instance, by Einstein in the development of the simplest theory of Brownian motion, the random walk [31]. In random walks, each steps of the target particle is independent on previous steps, due to collisions with the rest of particles. Collisions, moreover, are supposed to be uncorrelated. A more sophisticated approximation consists in keeping some aspects of motion, for instance the influence of inertia or of external forces, still approximating the rest of the world by noise (which may contain a certain degree of correlation). This is known as the Langevin approach, which includes the random walk as the simplest case. Langevin equations are stochastic differential equations.

The essence of this method relies in the assumption that the behavior of the various parts of the systems is uncorrelated. This assumption is vital also for other types of approximations, that will be illustrated in the following. Notice that in the statistical mechanics approach, this assumption is not required.

In the Langevin formulation, by averaging over many “independent” realizations of the process (which in general is not the same of averaging over many particles that “simultaneously” move, due for instance to excluded volumes) one obtains the evolution equation of the probability of finding a particle in a given portion of space. This

is the Kolmogorov integro-differential equation, that in many cases can be simplified, giving a differential (Fokker-Planck) equation. The diffusion equation is just the simplest case [27,56].

It is worth noticing that a similar formulation may be developed for quantum systems: the Feynman path-integral approach is essentially a Langevin formulation, and the Schrödinger equation is the corresponding Fokker-Planck equation.

Random walks and stochastic differential equations find many applications in economics, mainly in stock market simulations. In these cases, one is not interested in the average behavior of the market, but rather in computing non-linear quantities over trajectories (time-series of good values, fluctuations, etc.).

Time-Series Data Analyses

In practice, a model is never derived “ab initio”, by projecting the dynamics of all the microscopic components onto a limited number of dimensions, but is constructed heuristically from observations of the behavior of a real system.

It is therefore crucial to investigate how observations are made, i. e., the analysis of a series of time measurements. In particular, a good exercise is that of simulating a dynamical or stochastic system, analyzing the resulting time-series data of a given observable, and see if one is able to reconstruct from it the relationships or the equations ruling the time evolution.

Let us consider the experimental study of a chaotic, low-dimensional system. The measurements on this system give a time series of values, that we assume discrete (which is actually the case considering experimental errors). Therefore, the output of our experiment is a series of symbols or numbers, a time-series. Let us assume that the system is stationary, i. e., that the sequence is statistically homogeneous in time. If the system is not extremely chaotic, symbols in the sequence are correlated, and one can derive the probability of observing single symbols, couples of symbols, triples of symbols and so on. There is a hierarchy in these probabilities, since the knowledge of the distribution of triples allows the computation of the distribution of couples, and so on.

It can be shown that the knowledge of the probability distribution of the infinite sequence is equivalent to the complete knowledge of the dynamics. However, this would correspond to performing an infinite number of experiments, for all possible initial conditions.

The usual investigation scheme assumes that correlations vanish beyond a certain distance, which is equivalent to assume that the probability of observing sequences

longer than that distance factorize. Therefore, one tries to model the evolution of the system by a probabilistic dynamics of symbols. See Sect. “[Probabilistic Cellular Automata](#)”. Time-series data analysis can therefore be considered as the main experimental motivation in developing probabilistic discrete models. This can be done heuristically comparing results with observations *a posteriori*, or trying to extract the rules directly from data, like in the Markov approach.

Markov Approximation

The Markov approach, either continuous or discrete, also assumes that the memory of the system vanishes after a certain time interval, i. e., that the correlations in time series decay exponentially. In discrete terms, one tries to describe the process under study as an automata, with given transition probabilities. The main problem is: given a sequence of symbols, what is the simplest automata (hidden Markov chains [48]) that can generate that sequence with maximum “predictability”, i. e., with transition probabilities that are nearest to zero or one? Again, it is possible to derive a completely deterministic automata, but in general it has a number of nodes equivalent to the length of the time-series, so it is not generalizable and has no predictability (see also Sect. “[Probabilistic Cellular Automata](#)”). On the contrary, an automata with a very small number of nodes will have typically intermediate transition probabilities, so predictability is again low (essentially equivalent to random extraction). Therefore, the good model is the result of an optimization problem, that can be studied using, for instance, Monte-Carlo techniques.

Mean-Field

Finally, from the probabilities one can compute averages of observables, fluctuations and other quantities called “moments” of the distribution. Actually, the knowledge of all moments is equivalent to the knowledge of the whole distribution. Therefore, another approach is that of relating moments at different times or different locations, truncating the recurrences at a certain level. The roughest approximation is that of truncating the relations at the level of averages, i. e., the mean field approach. It appears so natural that it is often used without realizing the implications of the approximations. For instances, chemical equations are essentially mean-field approximations of a complex phenomena.

Boltzmann Equation

Another similar approach is that of dividing an extended system into zones, and assume that the behavior of the sys-

tem in each zone is well described by a probability distribution. By disregarding correlations with other zones, one obtains the Boltzmann equation, with which many transport phenomena may be studied well beyond elementary kinetic theory. The Boltzmann equation can also be obtained from the truncation of a hierarchy of equations (BBGKY hierarchy) relating multi-particle probability distributions. Therefore, the Boltzmann equations is similar in spirit to a mean-field analysis.

Equilibrium

One of the biggest success of the stochastic approach is of course *equilibrium statistical mechanics*. The main ingredient of this approach is that of minimum information, which, in other words, corresponds to the assumption: *what is not known is not harmful*. By supposing that at equilibrium the probability distribution of the systems maximizes the information entropy (corresponding to a minimum of information on the system) one is capable of deriving the probability distribution itself and therefore the expected values of observables (ensemble averages, see Sect. “Ising”). In this way, using an explicit model, one is capable to compute the value of the parameters that appear in thermodynamics. Were it possible to show that the maximum entropy state is actually the state originated by the dynamics of a mechanical (or quantum) system, one could ground thermodynamics on mechanics. This is a long-investigated subject, dating back to Boltzmann, which is however not yet clarified. The main drawback in the derivations is about ergodicity. Roughly speaking, a system is called *ergodic* if the infinite-time average of an observable over a trajectory coincides with its average over a snapshot of infinitely many replicas. For a system with fixed energy and no other conserved quantities, a sufficient condition is that a generic trajectory passes “near” all points of the accessible phase space. However, most systems whose behavior is “experimentally” well approximated by statistical mechanics are not ergodic. Moreover, another ingredient, the capability of *forgetting* quickly the information about initial conditions appears to be required, otherwise trajectories are strongly correlated and averages over different trajectories cannot be “mixed” together. This capability is strongly connected to the chaoticity or *unpredictability* of extended systems, but unfortunately these ingredients makes analytic approaches quite hard.

An alternative approach, due to Jaynes [33], is much more pragmatic. In essence, it says: let design a model with the ingredients that one thinks are important, and assume that what is not in the model does not affect its

statistical properties. Compute the distribution that maximizes the entropy with the given constraints. Then, compare the results (averages of observables) with experiments (possibly, numerical ones). If they agree, one has captured the essence of the problem, otherwise one have to include some other ingredient and repeat the procedure. Clearly, this approach is much more general than the “dynamical” one, not considering trajectory or making assumptions about the energy, which is simply seen as a constraint. But physicists would be much more satisfied by a “microscopic” derivation of statistical physics.

In spite of this lack of strong basis, the statistical mechanics approach is quite powerful, especially for systems that can be reduced to the case of *almost* independent elements. In this situation, the system (the *partition function*) factorizes, and many computations may be performed by hand. Notice however that this behavior is in strong contrast to that of truly linear systems: the “almost” attribute indicates that actually the elements interact, and therefore share the same “temperature”.

Monte-Carlo

The Monte-Carlo technique was invented for computing, with the aid of a computer, thermal averages of observables of physical systems at equilibrium. Since then, this term is often used to denote the technique of computing the average values of observables of a stochastic system by computing the time-average values over “artificial” trajectories.

In equilibrium statistical physics, one is faced by the problem of computing averages of observables over the probability distribution of the system, and since the phase space is very high-dimensional, this is in general not an easy task: one cannot simply draw *random* configurations, because in general they are so different from those “typical” of the given value of the temperature, that their statistical weight is marginal. And one does not want to revert to the original, still-more-highly-dimensional dynamical system, which typically requires powerful computers just to be followed for tiny time intervals.

First of all, one can divide (*separate*) the model into almost independent subsystems, that, due to the small residual interactions (the “almost” independency), are at the same temperature. In the typical example of a gas, the velocity components appear into the formula of energy as additive terms, i.e., they do not interact with themselves or with other variables. Therefore, they can be studied separately giving the *Maxwell distribution* of velocities. The positions of molecules, however, are linked by the potential energy (except in the case of an ideal gas), and so the

hard part of the computation is that of generating configurations. Secondly, statistical mechanics guarantees that the asymptotic probability distribution does not depend on the details of dynamics. Therefore, one is free to look for the fastest dynamics still compatible with constraints. The Monte-Carlo computation is just a set of recipes for generating such trajectories. In many problems, this approach allows to reduce the (computational) complexity of the problem of several orders of magnitude, allowing to generate “artificial” trajectories that span statistically significant configuration with small computational effort. In parallel with the generation of the trajectory, one can compute the value of several observables, and perform statistical analysis on them, in particular the computation of time averages and fluctuations.

By extension, the same terms “Monte-Carlo” is used for the technique of generating sequences of states (trajectories) given the transition probabilities, and computing averages of observables on trajectories, instead of over the probability distribution.

Stochastic Optimization

One of the most interesting applications of Monte-Carlo simulations concerns stochastic optimization via *simulated annealing*. The idea is that of exploiting an analogy between the status of a system (and its energy) and the coding of a particular procedure with corresponding cost function. The goal is that of finding the *best* solution, i. e., the global minimum of the energy given the constraints. “Easy” systems have a smooth energy landscape, shaped like a funnel, so that usual techniques like that of always choosing the displacements that locally lowers the energy (gradient descent) are successful. However, when the energy landscape is corrugated, there are many local minima where algorithms like gradient descent tend to get trapped. Methods from statistical mechanics (Monte-Carlo), on the contrary, are targeted to generating trajectories that quickly explore the “relevant” parts of the state space, i. e., those that correspond to the largest contributions to the probability distribution, that depends on the temperature, an “external” or control parameter. If the temperature is high, the probability distribution is broad and the generated trajectory does not “see” the minima of energy that are below the temperature, i. e., it can jump over and off the local minima.

By lowering the temperature, the probability distribution of system obeying statistical mechanics concentrates around minima of energy, and the Monte-Carlo trajectory does the same. The energy (or cost) function of not-extremely-complex problems is shaped in such a way that

the global optimum belongs to a broad valley, so that this lowering of the temperature increases the probability of finding it.

Therefore, a sufficiently slow annealing should furnish the desired global minimum. Moreover, it is possible to “convert” constraints into energy terms. For many problems, it is difficult to generate configurations that satisfy the constraints. Let us think for instance to the problem of generating a school timetable, keeping into consideration that lessons should not last more than three consecutive hours, that a teacher or students cannot stay in two classes at the same time, that a teacher is not available on Monday, another prefers the first hours, etc. One can formulate a Monte-carlo algorithm that generates free configurations, but then weights them with a factor that depends on how many constraints are violated. At high temperature, constraints do not forbid the exploration of the state space, and therefore to try “creative” solutions. At low temperature, constraints become important. At zero temperature, the solutions with lower energy are those that satisfy all constraints, if possible, or at least the largest part of them.

In recent years, physics have dealt with extremely complex problems (e. g., spin glasses [22,42]), in which the energy landscape is extremely rough. Special techniques, based on a non-monotonous “walk” on temperature have been developed (simulated tempering [39]).

Critical Phenomena

One of the most investigated topics of statistical mechanics concerns phase transitions. This is a fascinating subject: in the vicinity of a continuous phase transitions correlation lengths diverge, and the system behave collectively, in a way which is largely independent of the details of the model. This universal behavior allows the use of extremely simplified models, that therefore can be massively simulated.

The *philosophy* of statistical mechanics may be *exported* to nonequilibrium systems: systems with absorbing states (that correspond to infinitely negative energy), driven systems (live living ones), strongly frustrated systems (that actually never reach equilibrium), etc. In general, one defines these systems in terms of transition probabilities, not in term of energy. Therefore, one cannot invoke a maximum entropy principle, and the results are less general.

However, many systems exhibit behavior reminiscent of equilibrium systems, and the same language can be used: phase transitions, correlations, susceptibilities,...

These characteristics, common to many different models, are sometimes referred as *emergent features*.

One of the most famous problems in this field is percolation: the formations of giant clusters in systems described by a local stochastic aggregation dynamics. This “basic” model has been used to describe an incredibly large range of phenomena [53].

Equilibrium and nonequilibrium phase transitions occur for a well-defined value of a control parameter. However, in nature one often observes phenomena whose characteristics resemble that of a system near a phase transition, a critical dynamics, without any *fine-tuned* parameter. For such system the term self-organized criticality has been coined [9], and they are the subject of active researches.

Networks

A recent “extension” of statistical physics is the theory of networks. Networks in physics are often regular, like the lattice of a crystal, or only slightly randomized. Characteristics of these networks are the fixed (or slightly dispersed around the mean) number of connections per node, the high probability of having connected neighbors (number of “triangles”), the large time needed to cross the network. The opposite of a regular network is a random graph, which, for the same number of connections, exhibits low number of triangles and short crossing time. The statistical and dynamical properties of systems whose connection are regular or random are generally quite different.

Watts and Strogatz [58] argued that *social networks* are never completely regular. They showed that the simple random rewiring of a small number of links in a regular network may induce the small world effect: local properties, like the number of triangles, are not affected, but large-distance ones, like the crossing time, quickly became similar to that of random graphs. Also the statistical and dynamical properties of models defined over a rewired networks are generally similar to those correlated to random graphs.

After this finding, many social networks were studied, and they revealed a yet different structure: instead of having a well-defined connectivity, many of them present a few highly-connected “hubs”, and a lot of poorly-connected “leaves”. The distribution of connectivity is often shaped as a power-law (or similar [43]), without a well-defined mean (scale-free networks [1]). Many of phenomenological models are presently re-examined in order to investigate their behavior over such networks. Moreover, scale-free networks cannot be “laid down”, they need to be “grown” following a procedure, similar in this to frac-

tals. It is natural therefore to include such a procedure in the model, so that not only they evolve “over” the networks, but also evolve “the” network [13].

Agents

Many of the described tools are used in the so-called agent-based modeling. The idea is that of exploiting the powerful capabilities of present computers to simulate directly a large number of *agents* that interact among them. Traditional investigations of *complex systems*, like crowds, flocks, traffic, urban models, and so on, have been performed using *homogeneous* representation: partial differential equations (i. e., mean-field), Markov equations, cellular automata, etc. In such an approach, it is supposed that each agent type is present in many identical copies, and therefore they are simulated as “macrovariables” (cellular automata), or aggregated like independent random walkers in the diffusion equation. But live elements (cells, organisms) do not behave in such a way: they are often individually unique, carry information about their own past history, and so on. With computers, we are now in the position of simulating large assemblies of individuals, possibly geographically located, like for instance humans in an urban simulation.

One of the advantages of such approach is that of offering the possibility of measuring quantities that are inaccessible to field researchers, and also to play with different scenarios. The disadvantages are the proliferation of parameters, that are often beyond experimental confirmation.

Artificial Worlds

A somewhat alternative approach to that of “traditional” computational physics is that of studying an artificial model, build with little or no direct connection with reality, trying to include only those aspect that are considered relevant. The goal is to be able to find the *simplest* system still able to exhibit the relevant features of the phenomena under investigation. The resulting models, though not directly applicable to the interpretation of experiments, may serve as interpretative tools in many different situations. For instance, the Ising model was developed in the context of the modelization of magnetic systems, but is has been applied to opinion formation, social simulations, etc.

Ising

The Ising (better: Lenz-Ising) model is probably one of the most known models in statistical physics. Its history [44] is particularly illuminating in this context, even if it took

place well before the advent of computers in physics. It is also a model for which the Monte-Carlo and simulated annealing techniques are readily applied.

Let us first illustrate schematically the model. I shall present the traditional version, with the terminology that arises from the physics of magnetic systems. However, it is an interesting exercise to reformulate it in the context, for instance, of opinion formation. Let simply replace “spin up/down” with “opinion A/B”, “magnetization” with “average opinion”, “coupling” with “exchange of ideas”, “external magnetic field” with “propaganda”, and so on.

The Ising model is defined on a lattice, that can be in one, two or more dimensions, or even on a disordered graph. We shall locate a cell with an index i , corresponding to the set of spatial coordinates for a regular lattice or a label for a graph. The dynamical variable x_i for each cell is just a binary digit, traditionally named “spin” that takes the values ± 1 . We shall indicate the whole configuration as \mathbf{x} . Therefore, a lattice with N cells has 2^N distinct configurations. Each configuration \mathbf{x} has an associated energy $E(\mathbf{x}) = -\sum_i (H + h_i)x_i$, where H represents the external magnetic field and h_i is a local magnetic field, generated by neighboring spins, $h_i = \sum_j J_{ij}x_j$. The coupling J_{ij} for the original Lenz-Ising model is one if i and j are nearest neighbors, and zero otherwise.

The maximum-entropy principle [33] gives the probability distribution

$$P(\mathbf{x}) = \frac{1}{Z} \exp\left(-\frac{E(\mathbf{x})}{T}\right)$$

from which averages can be computed. The parameter T is the temperature, and Z , the “partition function” is the normalization factor of the distribution.

The quantity $E(\mathbf{x})$ can be thought as a “landscape”, with low-energy configurations corresponding to valleys and high-energy ones to peaks. The distribution $P(\mathbf{x})$ can be interpreted as the density of a gas, each “particle” corresponding to a possible realization (a replica) of the system. This gas concentrates in the valleys for low temperatures, and diffuses if the temperature is increased. The temperature is related to the average level of the gas.

In the absence of the local field ($J = 0$), the energy is minimized if each x_i is *aligned* (same sign) with H . This ordering is counteracted by thermal noise. In this case it is quite easy to obtain the average magnetization per spin (order parameter)

$$\langle x \rangle = \tanh\left(\frac{H}{T}\right),$$

which is a plausible behavior for a paramagnet. A ferromagnet however presents hysteresis, i. e., it may maintain

for long times (metastability) a pre-existing magnetization opposed to the external magnetic field.

With coupling turned on ($J > 0$), it may happen that the local field is strong enough to “resist” H , i. e., a compact patch of spins oriented against H may be stable, even if the energy could be lowered by flipping all them, because the flip of a single spin would rise the energy (actually, this flip may happen, but is statistically re-absorbed in short times). The fate of the patch is governed by boundaries. A spin on a boundary of a patch feels a weaker local field, since some of its neighbors are oriented in the opposite. Straight boundaries in two or more dimensions separate spins that “know” the phase they belong to, since most of their neighbors are in that phase, the spins on the edges may flip more freely. Stripes that span the whole lattice are rather stable objects, and may *resist* an opposite external field since spins that occasionally flip are surrounded by spins belonging to the opposite phase, and therefore feel a strong local field that pushes them towards the phase opposed to the external field.

In one dimension with finite-range coupling, a single spin flip is able to create a “stripe” (perpendicularly to the lattice dimension), and therefore can destabilize the ordered phase. This is the main reason for the absence of phase transitions in one dimension, unless the coupling extends on very large distances or some coupling is infinite (see the part on directed percolation, Sect. “[Probabilistic Cellular Automata](#)”).

This model was proposed in the early 1920s by Lenz to Ising for his PhD dissertation as a simple model of a ferromagnet. Ising studied it in one dimension, found that it shows no phase transition and concluded (erroneously) that the same happened in higher dimensions. Most of contemporaries rejected the model since it was not based on Heisenberg’s quantum mechanical model of ferromagnetic interactions. It was only in the forties that it started gaining popularity as a model of cooperative phenomena, a prototype of order-disorder transitions. Finally, in 1940, Onsager [46] provided the exact solution of the two-dimensional Lenz-Ising model in zero external field. It was the first (and for many years the only) model exhibiting a non-trivial second-order transition whose behavior could be exactly computed.

Second-order transitions have interested physicists for almost all the past century. In the vicinity of such transitions, the elements (say, spins) of the system are correlated up to very large distances. For instance, in the Lenz-Ising model (with coupling and more than one dimension), the high-temperature phase is disordered, and the low-temperature phase is almost completely ordered. In both these phases the connected two-points correlation

function

$$G_c(r) = \langle x_i x_{i+r} \rangle - \langle x_i \rangle^2$$

decreases exponentially, $G_c(r) \simeq \exp(-r/\xi)$, with $r = |r|$. The length ξ is a measure of the typical size of patch of spins pointing in the same direction.

Near the critical temperature T_c , the correlation length ξ diverges like $\xi(T - T_c) \simeq (T - T_c)^{-\nu}$ ($\nu = 1$ for $d = 2$ and $\nu \simeq 0.627$ for $d = 3$, where d is the dimensionality of the system). In such case the correlation function is described by a power law

$$G_c(r) \simeq \frac{1}{r^{d-2+\eta}},$$

with $\eta = 1/4$ for $d = 1$ and $\eta \simeq 0.024$ for $d = 3$. This phase transition is an example of a *critical phenomenon* [12], ν and η are examples of *critical exponents*.

The divergence of the correlation length indicates that there is no characteristic scale (ξ) and therefore fluctuations of all sizes appear. In this case, the details of the interactions are not so important, so that many different models behave in the same way, for what concerns for instance the critical exponents. Therefore, models can be grouped into *universality classes*, whose details are essentially given by “robust” characteristics like the dimensionality of space and of the order parameter, the symmetries, etc.

The power-law behavior of the correlation function also indicates that if we perform a *rescaling* of the system, it would appear the same or, conversely, that one is unable to estimate the *distance* of a pattern by comparing the “typical size” of particulars. This scale invariance is typical of many natural phenomena, from clouds (whose height and size are hard to be estimated), trees and other plant elements, lungs, brain, etc.

Many examples of power laws and collective behavior can be found in natural sciences [52]. Differently from what happens in the Lenz-Ising model, in these cases there is no parameter (like the temperature) that has to be fine-tuned, so one speaks of *self-organized criticality* [9].

Since the Lenz-Ising model is so simple, exhibits a critical phase and can be exactly solved (in some cases), it has become the playground for a variety of modifications and applications to various fields. Clearly, most of modifications do not allow analytical treatment and have to be investigated numerically. The Monte-Carlo method allows to add a temporal dimension to a statistical model [35], i.e., to transform stochastic integrals into averages over fictitious trajectories. Needless to say, the Ising-Lenz model is the standard test for every Monte-Carlo beginner, and most of techniques for accelerating

the convergence of averages has been developed with this model in mind [55].

Near a second-order phase transitions, a physical system exhibits *critical slowing down*, i.e., it reacts to external perturbations with an extremely slow dynamics, with a convergence time that increases with the system size. One can extend the definition of the correlation function including the time dimension: in the critical phase also the temporal correlation length diverges (as a power law). This happens also for the Lenz-Ising model using the Monte-Carlo dynamics, unless very special techniques are used [55]. Therefore, the dynamical version of the Lenz-Ising model can be used also to investigate relaxational dynamics and how this is influenced by the characteristics of the energy landscape. In particular, if the coupling J_{ij} changes sign randomly for every couple of sites (or the field H has random sign for each site), the energy landscape becomes extremely rugged. When spins flip in order to align to the local field, they may invert the field felt by neighboring ones. This *frustration* effect is believed to be the basic mechanism of the large viscosity and memory effects of *glassy substances* [22,42].

The rough energy landscape of glassy systems is also challenging for optimization methods, like *simulated annealing* [36] and its “improved” cousin, *simulated tempering* [39]. Again, the Lenz-Ising model is the natural playground for these algorithms.

The dynamical Lenz-Ising model can be formulated such that each spin is updated in parallel [10] (with the precaution of dividing cells into sublattices, in order to keep the neighborhood of each cell fixed during updates). In this way, it is essentially a probabilistic cellular automata, as illustrated in Sect. “[Probabilistic Cellular Automata](#)”.

Cellular Automata

In the same period in which traditional computation was developed, in the early fifties, John Von Neumann was interested in the logic basis of life and in particular in self-reproduction, and since the analysis of a self-reproduction automata following the rules of real physics was too difficult, he designed a playground (a cellular automaton) with just enough “physical rules” in order to make its analysis possible. It was however just a theoretical exercise, the automaton was so huge that up to now it has not yet completely implemented [57].

The idea of cellular automata is quite simple: take a lattice (or a graph) and put on each cell an automaton (all automata are equal). Each automaton exhibits its “state” (which is one out of a small number) and is programmed

to react (change state) according to the state of neighbors and its present one (the *evolution rule*). All automata update their state synchronously.

Cellular automata share many similarities with the parallel version of the Lenz-Ising model. Differently from that, their dynamics is not derived from an energy, but is defined in terms of the transition rules. These rules may be deterministic or probabilistic. In the first case (illustrated in this section), cellular automata are *fully discrete, extended dynamical systems*. Probabilistic cellular automata are illustrated in Sect. “Probabilistic Cellular Automata”. The temporal evolution of deterministic cellular automata can be computed *exactly* (regardless of any approximation) on a standard computer.

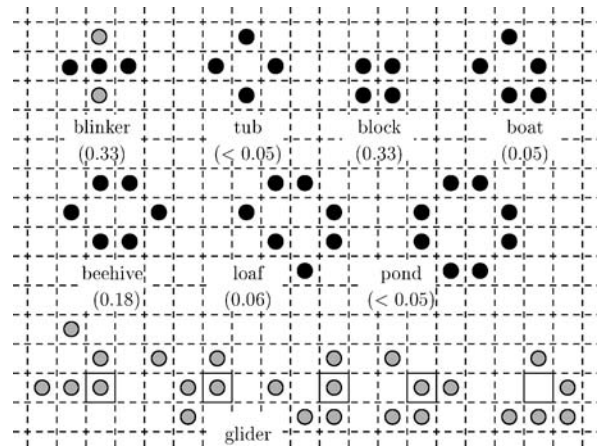
Let us illustrate the simplest case, *elementary cellular automata*, in Wolfram’s jargon [61]. The lattice is here one dimensional, so to identify an automaton it is sufficient to give one coordinate, say i , with $i = 1, \dots, N$. The state of the automaton on cell i at time t is represented by a single variable, $x_i(t)$ that can take only two values, “dead/live”, or “inactive/active” or 0/1. The time is also discrete, so $t = 1, 2, \dots$

The parallel evolution of each automaton is given by the rule

$$x_i(t+1) = f(x_{i-1}(t), x_i(t), x_{i+1}(t)).$$

Since $x_i = 0, 1$, there are only eight possible combinations of the triple $\{x_{i-1}(t), x_i(t), x_{i+1}(t)\}$, from $\{0, 0, 0\}$ to $\{1, 1, 1\}$. For each of them, $f(x_{i-1}(t), x_i(t), x_{i+1}(t))$ is either zero or one, so the function f can be simply coded as a vector of eight bits, each position labeled by a different configuration of inputs. Therefore, there are only $2^8 = 256$ different elementary cellular automata, that have been studied carefully (see for instance Ref. [61]).

In spite of their simplicity, elementary cellular automata exhibit a large variety of behaviors. In the language of dynamical systems they can be “visually” classified [61] as fixed points (class-1), limit cycles (class-2) and “chaotic” oscillations (class-3). A fourth class, “complex” CA, is present with larger neighborhood or in higher dimensions. A classical example is the *Game of Life* [11]. This two-dimensional cellular automaton is based on a simple rule. A cell may be either empty (0) or “alive” (1). A living cell survives if, among its 8 nearest neighbors, there are two or three alive cells, otherwise it dies and disappears. Generation is implemented through a rule for empty cells: they may become “alive” if surrounded by exactly three living cells. In spite of the simplicity of the rule, this automaton generates complex and long-living patterns, some of them illustrated in Fig. 2.



Interaction Based Computing in Physics, Figure 2

Some of the most common “animals” in the game of life, with the probability of encountering them in an asymptotic configuration [4]

Complex CA have large transients, during which interesting structures may emerge. They finally relax into class-1 automata. It has been conjectured that they are able of computation, i. e., that one can “design” an universal computer using these CA as building blocks, as has been proved to be possible with the Game of Life. Another hypothesis, again confirmed by the Game of Life, is that these automata are “near the edge” of self-organizing complexity. One can slightly “randomize” the Game of Life, allowing sometimes an exception to the rule. Let us introduce a parameter p , that measures this randomness, with the assumption that $p = 0$ is the true “life”. Well, it was shown [45] that the resulting model exhibits a second-order phase transition for a value of p very near zero.

Deterministic cellular automata have been investigated as prototypes of discrete dynamical systems, in particular for what concerns the definition of chaos. Visually, one is tempted to use this word also to denote the irregular behavior of “class 3” rules. However, the usual definition of chaos involves the sensitivity to an infinitesimally small perturbation: following the time dynamics of two initially close configurations one can observe an amplification of their distance. If the initial distance (δ_0) is infinitesimal, then the distance grows exponentially for some time ($\delta(t) \simeq \delta_0 \exp(\lambda t)$), after which it tends to saturate, since the trajectories are generally bounded inside an attractor, or due to the dimensions of the accessible space. The exponent λ depends on the initial configuration, and if this behavior is observed for different portions of the trajectory, it fluctuates: a trajectory spends some time in regions of high chaoticity, after which may pass through “quiet” zones. If

one “renormalizes” periodically this distance, considering one system as the “master” and the other as a measuring device, one can accumulate good statistics, and define a Lyapunov exponent λ , that gives indications about the chaoticity of the trajectory, through a limiting procedure.

The accuracy of computation poses some problems. Since in a computer numbers are always approximate, one cannot follow “one” trajectory. The small approximations accumulates exponentially, and the computer time series actually *jumps* among neighboring trajectories. Since the Lyapunov exponent is generally not so sensible to a change of precision in computation, one can assume that the chaotic regions are rather compact and uniform, so that in general one associates a Lyapunov exponent to a system, not to an individual trajectory. Nevertheless, this definition cannot apply to completely discrete systems like cellular automata.

In any case, chaoticity is related to unpredictability. As first observed by Lorenz, and following the definition of Lyapunov exponent, the precision of an observation over a chaotic system is related to the average time for which predictions are possible. Like in weather forecasts, in order to increase the time span of a prediction one has to increase the precision of the initial measurement. In extended system, this also implies to extend the measurements over a larger area. One can also consider a “synchronization” approach. Take two replicas of a systems and let them evolve starting from different initial configurations. With a frequency and a strength that depends on a parameter q , one of these replica is “pushed” towards the other one, so to reduce their distance. Suppose that $q = 0$ is the case of no push and $q = 1$ is the case of extremal push, for which the two systems synchronize in a very short time. There should be a critical value q_c that separates these two behaviors (actually, the scenario may be more complex, with many phases [4]). In the vicinity of q_c the distance between the two replicas is small, and the distance δ grows exponentially. The critical value q_c is such that the exponential growth compensates the shrinking factor, and is therefore related to the Lyapunov exponent λ .

Finite-size cellular automata always follow periodic trajectories. Let us consider for instance a Boolean automata, of N cells. The number of possible different states is 2^N and due to determinism, once that a state has been visited twice the automata has entered a limit cycle (or a fixed point). One may have limit cycles with large basins of transient configurations (configurations that do not belong to the cycle). Many scenarios are possible. The set of different configurations may be divided in many basins, of small size (small transient) and small period, like in class 1 and 2 automata. Or one may have large basins, with long

transients that lead to short cycles, like in class-4 automata. Finally, one may have one or very few large basins, with long cycles that include most of configurations belonging to the basin (small transients). This is the case of class-3 automata. For them, the typical period of a limit cycle grows exponentially, like the total number of configurations, with the system size, so that for moderately large system it is almost impossible to observe a whole cycle in a finite time. Another common characteristic of class-3 automata is that the configurations quickly decorrelates (in the sense of the correlation function) along a trajectory. If one takes into consideration as starting points two configurations that are the same but for a local difference, one observes that this difference amplifies and diffuses in class-3 automata, shrinks or remains limited in class-1 and class-2, and have an erratic transient behavior in class-4, followed by the fate of class-1 and 2. Therefore, if one considers the possibility of not knowing exactly the initial configuration of an automata, unpredictability grows with time also for such discrete systems. Actually, also (partially) continuous systems like coupled maps may exhibit this kind of behavior [4,17,19,47]. Along this line, it is possible to define an equivalent of the Lyapunov exponents for CA [7]. The synchronization procedure can be applied also to cellular automata, and it correlates well with the Lyapunov exponents [5].

An “industrial” application of cellular automata is their use for modeling gases. The *hydrodynamical equations*, like the Navier-Stokes ones, simply reflect the conservation of mass, momentum and energy (i. e., rotational, translational and time invariance) for the microscopic collision rules among particles. Since the modeling of a gas via molecular dynamics is rather cumbersome, some years ago it was proposed [26,29] to simplify drastically the microscopic dynamics using particles that may travel only along certain directions with some discrete velocities and jumping in discrete time only among nodes of a lattice. Indeed, a cellular automaton. It has been shown that their macroscopic dynamics is described by usual hydrodynamics laws (with some odd features related to the underlining lattice and finiteness of velocities) [51,60].

The hope was that these Lattice Gas Cellular Automata (LGCA) could be simulated so efficiently in hardware to make possible the investigation of turbulence, or, in other words, that they could constitute the Ising model of hydrodynamics. While they are indeed useful to investigate certain properties of gases (for instance, chemical reactions [37], or the relationship between chaoticity and equilibrium [6]), they resulted too noisy and too viscous to be useful for the investigation of turbulence. Viscosity is related to the transport of momentum in a direction perpen-

dicular to the momentum itself. If the collision rule does not “spread” quickly the particles, the viscosity is large. In LGCA there are many limitations to collisions, so that in order to lower viscosity one has to consider averages over large patches, thus lowering the efficiency of the method.

However, LGCA inspired a very interesting approximation. Let us consider a large assembly of replicas of the same system, each one starting from a different initial configuration, all compatible with the same macroscopic initial conditions. The macroscopic behavior after a certain time would be the *average* over the status of all these replicas. If one assumes a form of local equilibrium, i. e., applies the mean-field approximation for a given site, one may try to obtain the dynamics of the *average* distribution of particles, which in principle is the same of “exchanging” particles that happen to stay on the same node among replicas.

It is possible to express the dynamics of the average distribution in a simple form: it is the Lattice Boltzmann Equation (LBE) [18,54,60]. The method retains many properties of LGCA like the possibility of considering irregular and varying boundaries, and may be simulated in a very efficient way with parallel machines [54]. Differently from LGCA, there are numerical stability problems to be overcome.

Probabilistic Cellular Automata

In deterministic automata, given a local configuration, the future state of a cell is univocally determined. But let us consider the case of measuring experimentally some pattern and trying to analyze it in terms of cellular automata. In time-series analysis, it is common to perform averages over spatial patches and temporal intervals and to discretize the resulting value. For instance, this is the natural result of using a camera to record the temporal evolution of an extended system, for instance the turbulent and laminar regions of a fluid. The resulting pattern symbolically represents the dynamics of the original system, and if it is possible to extract a “rule” out of this pattern, it would be extremely interesting for the construction of a model. In general, however, one observes that sometimes a local configuration is followed by a symbol, and sometime the same local configuration is followed by another one. One should conclude that the neighborhood (the local configuration) does not univocally determines the following symbol.

One can extend the “range” of the rule, adding more neighbors farther in space and time [48]. By doing so, the “conflicts” generally reduce, but at the price of increasing the complexity of the rule. At the extremum, one could have an automaton with infinite “memory” in time and space, that perfectly reproduces the observed patterns but

with almost none predictive power, since it is extremely unlucky that the same *huge* local configuration is encountered again.

So, one may prefer to limit the neighborhood to some finite extension, and accept that the rule sometimes “outputs” a symbol and sometimes another one. One defines a local transition probability $\tau(x_i(t+1)|X_i(t))$ of obtaining a certain symbol x_i at time $t+1$ given a local configuration X_i at time t . Deterministic cellular automata correspond to the case $\tau = 0, 1$. The parallel version of the Lenz-Ising model can be re-interpreted as a probabilistic cellular automaton.

Starting from the local transition probabilities, one can build up the transition probability $T(x|y)$ of obtaining a configuration x given a configuration y ($T(x|y)$). $T(x|y)$ is given by the product of the local transition probabilities τ [2]. One can read the configurations x and y as indexes, so that T can be considered as a matrix. The normalization of probability corresponds to the constraint $\sum_x T(x|y) = 1, \forall y$.

Denoting with $P(x, t)$ the probability of observing a given configuration x at time t , and with $P(t)$ the whole distribution at time t , we have for the evolution of the distribution

$$P(t+1) = TP(t),$$

with the usual rules for the product of matrices and vectors. Therefore, the transition matrix T defines a Markov process, and the asymptotic state of the system is given by the eigenvalues of T . The largest eigenvalue is always 1, due to the normalization of the probability distribution, and the corresponding eigenvector is the asymptotic distribution. The theory of Markov processes says that if T is finite and irreducible, i. e., it cannot be rewritten (renumbering rows and columns) as blocks of noninteracting subspaces, like

$$T = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix},$$

then the second eigenvalue is strictly less than one and the asymptotic state is unique. In this case the second eigenvalue determines the convergence time to the asymptotic state. For finite systems, often the matrix T is irreducible. However, in the limit of infinite size, the largest eigenvalue may become degenerate, and therefore there are more than one asymptotic state. This is the equivalent of a phase transition for Markov processes.

For the parallel Lenz-Ising model, the elements of the matrix T are given by the product of local transition rules of the Monte-Carlo dynamics. They depend on the choice

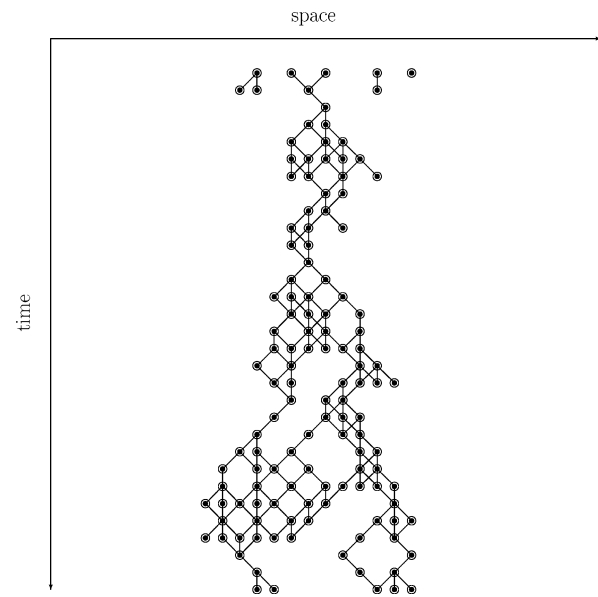
of the algorithm, but essentially have the form of exponentials of the difference in energy, divided by the temperature. Although a definitive proof is still missing, it is plausible that matrices with all elements different from zero correspond to some equilibrium model, whose transition rules can be derived from an energy function [28].

Since probabilistic cellular automata are defined in terms of the transition probabilities, one is free to investigate models that go beyond equilibrium. For instance, if some transition probability takes the value zero or one, in the language of equilibrium system this would correspond to some coupling (like the J of the Lenz-Ising model) that become infinite. This case is not so uncommon in modeling. The inverse of a transition probability corresponds to the average waiting time for the transition to occur in a continuous-time model (one may think to chemical reactions). Some transitions may have a waiting time so long with respect to the observation interval, to be practically irreversible. Therefore, probabilistic cellular automata (alongside other approaches like for instance annihilating random walks) allow the exploration of out-of-equilibrium phenomena.

One such phenomena is directed percolation, i.e., a percolation process with a special direction (time) along which links can only be crossed one-way [14]. Let think for instance to the spreading of an infection in a one-dimensional lattice, with immediate recovery (SIS model). An ill individual can infect one or both of his two neighbors but returns to the susceptible state after one step. The paths of infection (see Fig. 3) can wander in the space directions, but are directed in the time directions.

The parallel version of a directed percolation process can be mapped onto probabilistic cellular automata. The simplest case, in one spatial dimension and with just two neighbors, is called the Domany-Kinzel model [21]. It is even more general than the usual directed percolation, allowing “non-linear” interactions among sites in the neighborhood (e.g., two wet sites may have less probability of percolating than one alone).

These processes are interesting because there is an *absorbing state* [3,32], which is the dry state for the wetting phenomenon and the healthy state for the spreading of epidemics. Once the system has entered this absorbing state, it cannot exit, since the spontaneous appearing of a wet site or of an ill individual is forbidden. For finite systems, the theory of Markov chains says that the system will “encounter”, sooner or later, this state, that therefore corresponds to the unique asymptotic state. For infinitely extended systems, a phase transition can occur, for which wet sites percolate for all “times”, the epidemics become endemic, and so on.



Interaction Based Computing in Physics, Figure 3
An example of a directed percolation cluster

Again these are examples of *critical phenomena*, with exponents different from the equilibrium case.

Cellular Automata and Agent-Based Simulations

Cellular automata can be useful in modeling phenomena that can be described in lattice terms. However, many phenomena requires “moving particles”. Examples may be chemical reactions, ecological simulations, social models. When particles are required to obey hydrodynamics constraints, i.e., to collide conserving mass, momentum and energy, one can use lattice gas cellular automata or the lattice Boltzmann equation. However, in general one is interested in modeling just a macroscopic scale, assuming that what happens at lower level is just “noise”. According with the complexity (and “intelligence”) assigned to particles, one can develop models based on the concept of walkers, that move more or less randomly. From the simulation point of view, walkers are not very different from the graphs succinctly described above. In this case the identifier i is just a label, that allows to access walker’s data, among which there are the coordinates of the walker (that may be continuous), its status and so on.

In order to let walker interact, one is interested in finding efficiently all walkers that are nearer than a given distance from the one under investigation. This is the same problem one is faced with when developing codes for molecular dynamics simulations [49]: scanning all walkers

in order to compute their mutual distance is a procedure that grows quadratically with the number of walkers. One “trick” is that of dividing the space in cells, i. e., to define an associated lattice. Each cell contains a list of all walkers that are located inside it. In this way, one can directly access all walkers that are in the same or neighboring cells of the one under investigation.

Moreover, one can exploit the presence of the lattice to implement on it a “cellular automaton”, that may interact with walkers, in order to simulate the evolution of “fields”. Just for example, the simulation of a herd of herbivores that move according to the exhaustion of the grass, and the parallel growing of the vegetation can be modeled associating the “grass” to a cellular automaton, and the herbivores to walkers. This simulation scheme is quite flexible, allowing to implement random and deterministic displacements of moving objects or agents, continuous or discrete evolution of “cellular objects” and their interactions. Many simulation tools and games are based on this scheme [50,59,62], and they generally allow the contemporary visualization of a graphical representation of the state of the system. They are valuable didactic tool, and may be used to “experiment” with these artificial worlds. As often the case, the flexibility is paid in terms of efficiency and speed of simulation.

Discussions and Conclusions

Complex systems, like for instance human societies, cannot be simulated starting from “physical laws”. This is sometimes considered a weakness of the whole idea of studying such systems from a quantitative point of view. We have tried to show that actually even the “hardest” discipline, physics, always deals with models, that have finally to be simulated on computers making various assumptions and approximations. Theoretical physics is accustomed since a long time to “extreme simplifications” of models, hoping to enucleate the “fundamental ingredients” of a complex behavior. This approach has proved to be quite rewarding for our understanding of nature.

In recent years, physics has been seen studying many fields not traditionally associated to physics: molecular biology, ecology, evolution theory, neurosciences, psychology, sociology, linguistics, and so on. Actually, the word “physics” may refer either to the classical subjects of study (mainly atomic and subatomic phenomena, structure of matter, cosmological and astronomical topics), or to the “spirit” of the investigation, that may apply to almost any discipline. This spirit is essentially that of building simplified quantitative models, composed by many elements, and study them with theoretical instruments (most of

times, applying some form of mean-field treatment) and with computer simulations.

This approach has been fruitful in chemistry and molecular biology and nowadays many physical journals have sections devoted to “multidisciplinary studies”. The interesting thing is that not only have physicists brought some “mathematics” into fields that are traditionally more “qualitative” (which often corresponds to “linear” modeling, plus noise), but physicists have also discovered many interesting questions to be investigated, and new models to be studied. One example is given by the current investigations about the structure of social networks, that were “discovered” by physicists in the nontraditional field of social studies.

Another contribution of physicists to this “new way” of performing investigations, is the use of networked computers. Since a long time, physicists have used computers for performing computations, storing data and diffusing information using the Internet. Actually, the concept of what is now the World Wide Web was born at CERN, as a method for sharing information among laboratories [15]. The high-energy physics experiments require a lot of simulations and data processing, and physicists (among others) developed protocols to distribute this load on a grid of networked computers. Nowadays, an European project aims to “open” grid computing to other sciences [24].

It is expected that this combination of quantitative modeling and grid computing will stimulate innovative studies in many fields. Here is a small sparse list of possible candidates:

- Theory of evolution, especially for what concerns evolutionary medicine.
- Social epidemiology, coevolution of diseases and human populations, interplay between sociology and epidemics.
- Molecular biology and drug design, again driven by medical applications.
- Psychology and neural sciences, it is expected that the “black box” of traditional psychology and psychiatry will be replaced by explicit models based on brain studies.
- Industrial and material design.
- Earth sciences, especially meteorology, vulcanology, seismology.
- Archaeology: simulation of ancient societies, reconstruction of historical and pre-historical climates.

Nowadays, the term *cellular automata* has enlarged its meaning, including any system whose elements do not move (in opposition to “agent-based modeling”). There-

fore, we now have cellular automata on non-regular lattices, non-homogeneous, with probabilistic dynamics (see Sect. “[Probabilistic Cellular Automata](#)”), etc. [23]. They are therefore considered more as a “philosophy” of modeling rather than a single tool. In some sense, cellular automata (and agent-based) modeling is opposed to the spirit of describing a phenomena using differential equations (or partial differential equations). One of the reasons is that the language of “automata” and “agents” is simpler and requires less training than that of differential equations. Another reason is that at the very end, any “reasonable” problem has to be investigated using computers, and while the implementation using discrete elements is straightforward (even if careful planning may speed-up dramatically the simulation), the computation of partially differential equations is an art in itself.

However, the final success of this approach is related to the availability of high-quality experimental data that allow to discriminate among the almost infinite number of models that can be built.

Bibliography

Primary Literature

- Albert R, Barabasi AL (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Bagnoli F (2000) Cellular automata. In: Bagnoli F, Ruffo S (eds) *Dynamical Modeling in Biotechnologies*. World Scientific, Singapore, p 1
- Bagnoli F, Boccara N, Rechtman R (2001) Nature of phase transitions in a probabilistic cellular automaton with two absorbing states. *Phys Rev E* 63(4):046 116
- Bagnoli F, Cecconi F (2001) Synchronization of non-chaotic dynamical systems. *Phys Lett A* 282(1–2):9–17
- Bagnoli F, Rechtman R (1999) Synchronization and maximum Lyapunov exponents of cellular automata. *Phys Rev E* 59(2):R1307–R1310
- Bagnoli F, Rechtman R (2007) Entropy and chaos in a discrete hydrodynamical system. *arXiv:cond-mat/0702074*
- Bagnoli F, Rechtman R, Ruffo S (1992) Damage spreading and Lyapunov exponents in cellular automata. *Phys Lett A* 172:34
- Bagnoli F, Rechtman R, Ruffo S (1994) Some facts of life. *Physica A* 171:249
- Bak P, Tang C, Wiesenfeld K (1987) Self-organizing criticality: An explanation of $1/f$ noise. *Phys Rev A* 38:364–374
- Barkema GT, MacFarland T (1994) Parallel simulation of the Ising model. *Phys Rev E* 50(2):1623–1628
- Berlekamp E, Conway J, Guy R (1982) *What is Life?*, vol 2: Games in Particular. Academic Press, London, chap 25
- Binney J, Dowrick N, Fisher A, Newman MEJ (1993) *The Theory of Critical Phenomena*. Oxford Science Publications. Clarendon Press, Oxford
- Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Phys Rep* 424(4–5):175–308
- Broadbent S, Hammersley J (1957) Percolation processes I. Crystals and mazes. *Proc Camb Philos Soc* 53:629–641
- Cailliau R (1995) A short history of the web. http://www.netvalley.com/archives/mirrors/robert_cailliau_speech.htm Accessed 25 May 2008
- Car R, Parrinello M (1985) Unified approach for molecular dynamics and density-functional theory. *Phys Rev Lett* 55(22):2471–2474
- Cecconi F, Livi R, Politi A (1998) Fuzzy transition region in a one-dimensional coupled-stable-map lattice. *Phys Rev E* 57(3):2703–2712
- Chopard B, Luthi P, Masselot A, Dupuis A (2002) Cellular automata and lattice Boltzmann techniques: An approach to model and simulate complex systems. *Adv Complex Syst* 5(2):103–246
- Crutchfield J, Kaneko K (1988) Are attractors relevant to turbulence? *Phys Rev Lett* 60(26):2715–2718
- Daxois T, Peyrard M, Ruffo S (2005) The Fermi–Pasta–Ulam ‘numerical experiment’: History and pedagogical perspectives. *Eur J Phys* 26:S3–S11
- Domany E, Kinzel W (1984) Equivalence of cellular automata to Ising models and directed percolation. *Phys Rev Lett* 53(4):311–314
- Dotsenko V (1994) *An Introduction to the Theory of Spin Glasses and Neural Networks*. World Scientific, Singapore
- El Yacouby S, Chopard B, Bandini S (eds) (2006) *Cellular Automata*. Lecture Notes in Computer Science, vol 4173. Springer, Berlin
- Enabling grids for e-science. <http://www.eu-egee.org/> Accessed 25 May 2008
- Fermi E, Pasta J, Ulam S (1955) Los Alamos report la-1940. In: Segré E (ed) *Collected papers of Enrico Fermi*. University of Chicago Press, Chicago
- Frisch U, Hasslacher B, Pomeau Y (1986) Lattice-gas automata for the Navier-Stokes equation. *Phys Rev Lett* 56(14):1505–1508
- Gardiner CW (1994) *Handbook of Stochastic Methods for Physics, Chemistry, and the Natural Sciences*. Springer series in synergetics, vol 13. Springer, Berlin
- Georges A, le Doussal P (1989) From equilibrium spin models to probabilistic cellular automata. *J Stat Phys* 54(3–4):1011–1064
- Hardy J, Pomeau Y, de Pazzis O (1973) Time evolution of a two-dimensional classical lattice system. *Phys Rev Lett* 31(5):276–279
- Harlow H, Metropolis N (1983) Computing & computers - weapons simulation leads to the computer era. *Los Alamos Science* 4(7):132
- Haw M (2005) Einstein’s random walk. *Physics World* 18:19–22
- Hinrichsen H (1997) Stochastic lattice models with several absorbing states. *Phys Rev E* 55(1):219–226
- Jaynes E (1957) Information theory and statistical mechanics. *Phys Rev* 106(4):620–630
- Kaneko K (1985) Spatiotemporal intermittency in coupled map lattices. *Progr Theor Phys* 74(5):1033–1044
- Kawasaki K (1972) Kinetics of Ising model. In: Domb CM, Green MS (eds) *Phase Transitions and Critical Phenomena*, vol 2. Academic Press, New York, p 443
- Kirkpatrick S, Gelatt Jr CG, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
- Lawniczak A, Dab D, Kapral R, Boon JP (1991) Reactive lattice gas automata. *Phys D* 47(1–2):132–158

38. Lorenz E (1996) History of chaos: http://library.thinkquest.org/3120/old_htdocs.1/text/fraz1.txt Accessed 25 May 2008
39. Marinari E, Parisi G (1992) Simulated tempering: A new Monte Carlo scheme. *Europhys Lett* 19:451–458
40. May R (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
41. Metropolis N, Hewlett J, Rota GC (eds) (1980) *A History of Computing in the Twentieth Century*. Academic Press, New York
42. Mezard M, Parisi G, Virasoro MA (1987) *Spin Glass Theory and Beyond*. World Scientific Lecture Notes in Physics, vol 9. World Scientific, Singapore
43. Newman ME (2005) Power laws, pareto distributions and zipf's law. *Contemp Phys* 46:323–351
44. Niss M (2005) History of the Lenz–Ising model 1920–1950: From ferromagnetic to cooperative phenomena. *Arch Hist Exact Sci* 59(3):267–318
45. Nordfalk J, Alstrøm P (1996) Phase transitions near the “game of life”. *Phys Rev E* 54(2):R1025–R1028
46. Onsager L (1944) Crystal statistics. i. a two-dimensional model with an order-disorder transition. *Phys Rev* 65:117–149
47. Politi A, Livi R, Oppo GL, Kapral R (1993) Unpredictable behaviour of stable systems. *Europhys Lett* 22(8):571–576
48. Rabiner L (1989) A tutorial on hidden markov models and selected applications in speech recognition. *Proc IEEE* 77(2):257–286
49. Rapaport DC (2004) *The Art of Molecular Dynamics Simulation*. Cambridge University Press, Cambridge
50. Repast – recursive porous agent simulation toolkit. <http://repast.sourceforge.net/> Accessed 25 May 2008
51. Rothman DH, Zaleski S (2004) *Lattice-Gas Cellular Automata*. Monographs and Texts in Statistical Physics. Collection Alea-Saclay, Paris
52. Sornette D (2006) *Critical Phenomena in Natural Sciences*. Springer Series in Synergetics. Springer, Berlin
53. Stauffer D, Aharony A (1994) *Introduction To Percolation Theory*. Taylor Francis, London
54. Succi S (2001) *The Lattice Boltzmann Equation For Fluid Dynamics and Beyond*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford
55. Swendsen R, Wang JS (1987) Nonuniversal critical dynamics in Monte Carlo simulations. *Phys Rev Lett* 58(2):86–88
56. van Kampen NG (1992) *Stochastic Processes in Physics and Chemistry*. North-Holland, Amsterdam
57. Von Neumann universal constructor. http://en.wikipedia.org/wiki/Von_Neumann_Universal_Constructor Accessed 25 May 2008
58. Watts D, Strogatz SH (1998) Collective dynamics of ‘small-world’ networks. *Nature* 393:440–441
59. Wilensky U (1999) Netlogo. <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston; Accessed 25 May 2008
60. Wolf-Gladrow D (2004) *Lattice-Gas Cellular Automata and Lattice Boltzmann Models: An Introduction*. Lecture Notes in Mathematics, vol 1725. Springer, Berlin
61. Wolfram S (1983) Statistical mechanics of cellular automata. *Rev Mod Phys* 55:601–644
62. Wright W (1989) Simcity. <http://simcitysocieties.ea.com/> Accessed 25 May 2008

Books and Reviews

- Bungartz H-J, Mundani R-P, Frank AC (2005) *Bubbles, jaws, moose tests, and more: The wonderful world of numerical simulation*, Springer VideoMATH. Springer, Berlin (DVD)
- Gould H, Tobochnik J, Christian W (2007) *An Introduction to Computer Simulation Methods: Applications to Physical Systems*. Addison-Wesley, New York
- Landau RH (2005) *A First Course in Scientific Computing: Symbolic, Graphic, and Numeric Modeling Using Maple, Java, Mathematica, and Fortran90*. Princeton University Press, Princeton
- Chopard B, Droz M (2005) *Cellular automata modeling of physical systems*. In: Collection Alea-Saclay: Monographs and Texts in Statistical Physics. Cambridge University Press, Cambridge
- Boccara N (2004) *Modeling complex systems*. In: Graduate Texts in Contemporary Physics. Springer, Berlin
- Deisboeck S, Kresh JY (eds) (2006) *Complex systems science in biomedicine*. In: Topics in Biomedical Engineering. Springer, New York
- Resnick M (1994) *Turtles, Termites, and Traffic Jams. Explorations in Massively Parallel Microworlds*. In: Complex Adaptive Systems. MIT Press, Cambridge
- Shalizi C, Cosma's Home Page <http://www.cscs.umich.edu/~crshalizi/>
- Open Source Physics <http://www.opensourcephysics.org/about/index.html>

Intermittency and Localization

GUR YAARI^{1,2}, DIETRICH STAUFFER^{2,3},
SORIN SOLOMON^{1,2}

¹ Institute for Scientific Interchange, Turin, Italy

² Racah Institute of Physics, Hebrew University, Jerusalem, Israel

³ Institute for Theoretical Physics, Cologne University, Köln, Germany

Article Outline

Glossary

Definition of the Subject

Introduction

Logistic Systems: From Malthus Until Today

Minimal Extensions to the Classical Logistic System

Applying These Models to Real-Life Systems

Future Directions

Acknowledgments

Bibliography

Glossary

Auto-catalysis In systems that go through several reactions, the reaction is called **autocatalytic** if the reaction product is itself the catalyst for that reaction.

Exponential growth An autocatalytic reaction is usually described with a simple linear, first order differential equation. The solution for it is an exponential increasing/decreasing function.

Logistic growth If one adds a saturation term (of power two) to the linear first order differential equation which describes exponential growth, the resulting solution saturates instead of ever-lasting grow. The solution to this system is described with a **logistic curve** and the system is said to follow a **logistic growth**.

Reaction-diffusion systems Reaction-diffusion systems are mathematical models that describe how the concentration of one or more substances distributed in space changes under the influence of two processes: local reactions in which the substances are converted into each other, and diffusion which causes the substances to spread out in space.

Definition of the Subject

In this paper, we show how simple logistic growth that was studied intensively during the last 200 years in many domains of science could be extended in a rather simple way. The resulting extended model has, among other features, two very important ones: Intermittency and Localization. These features were observed repeatedly along the history of science in an enormous number of real-life systems in economics, sociology, biology, ecology and more. We suggest by this a unified theoretical umbrella that might serve in a surprising way many scientific disciplines who share similar observed patterns.

Introduction

A well known joke that many physicists like to tell during their talks in order to demonstrate the strength of simplifying the problem one has in hand is: “First, let us consider a spherical cow...”. Although, no one really believes in spherical cows – the power of simplification is well accepted and appreciated by the physics community, or as Albert Einstein put it, very accurately: “Everything should be made as simple as possible, but not simpler”.

There are many more such mantras like: “Keep it simple, stupid”, “Kill your darlings” and “Less is more”. As these lines of thought were adopted so strongly by physicist for so much time, the statement of P.W. Anderson that “More is different” made such a revolution in Science. In the paper that has this title, Anderson pushed the new scientific (inter-) discipline, now known as complexity. By introducing these new ideas, Anderson paved the way for many physicists carrying with them heavy weapons from

traditional physics to start thinking and attacking many problems from a variety of scientific disciplines.

A lot of criticism about such physicists that try to cross the borders of their discipline is about *over-simplifying* real-life problems in order to be able to solve the resulting models with the tools they already have. Due to that, it is important to emphasize here that by working inside the framework of complexity one tries not to lose the minimal theoretical ingredients of the problem that are sufficient to produce the complex observed outcome. Rather than do this, one tries to study to the best of one’s ability, the simplest *possible* model.

A common question that arises in the social sciences is: *Why are improbable things so frequent?* Fine-tuned irreducibly complex systems have generically a low probability to appear and highly integrated \simeq arranged systems are usually artificial (often man-made) and untypical. Yet many complex systems are found lately to be self-organized. More precisely, the amount of non-generic, fine tuned and highly integrated systems is much larger in nature from what would be reasonably expected from generic stochastic estimations. It often happens that even though the range of parameters necessary for some non-trivial collective phenomenon to emerge is very narrow (or even an isolated single point out of an continuum infinite range), the phenomenon does actually take place in nature. This leads to collective objects whose properties are not explainable by the generic dynamics of their components. The explanation of the generic emergence of systems which are non-generic from the multi-agent point of view seems to be related to self-catalyzing dynamics.

As suggested by the examples above, the frequency with which we encounter non-generic situations in self-catalyzing systems is not so surprising. Consider a space of all possible systems obtainable from certain chemical and physical parts. Even if a macroscopic number of those systems are not auto-catalytic and only a very small number happen to be auto-catalytic after enough time, one of the auto-catalytic systems will eventually arise. Once this happens, the auto-catalytic system will start multiplying leading to a final (or far-future) situation in which those auto-catalytic – a priory very improbable systems – are over-represented compared with their natural probability of occurrence. Basically, this is how life spread all over Earth.

In this paper, we show how simple logistic growth that was studied intensively during the last 200 years in many domains of science could be extended in a rather simple way and with these extensions is capable to produce a collection of behaviors widely observed in an enormous number of real-life systems in economics, sociology, biology,

ecology and more. For other reviews in this direction we recommend on [4,18,38].

The paper will start with a historical overview of the use of logistic-like systems in science since its introduction by Malthus in 1798 until today. The next section will present a view of the minimal, though sufficient, extensions to the classical logistic system that are able to bring this theoretical framework, closer to reality, but auto-catalysis yet still solvable analytically in many regions of the parameter's space. Then, we will show some of the successes we had in applying this framework to real-life systems. We will finish the paper by a short fantasy trying to describe a dream about the possible usages of this powerful theoretical framework in the so called soft sciences in the future.

Logistic Systems: From Malthus Until Today

Auto-Catalysis

One of the key concepts underlying the emergence of complex macroscopic features is auto-catalysis. We therefore give at this point a provisory definition of it: auto-catalysis = self-perpetuation, \simeq reproduction, \simeq multiplication. As opposed to the usual stochastic systems in which the microscopic dynamics changes typically the individual microscopic quantities by additive steps (e.g. a molecule receiving or releasing a quantum of energy), the auto-catalytic microscopic dynamics involve multiplicative changes (e.g. the market worth of a company changes by a factor (index) after each elementary transaction). Such auto-catalytic microscopic rules are widespread in chemistry (under the name of auto-catalysis), biology (reproduction \simeq multiplication, species perpetuation), social sciences (profit, returns, rate of growth).

The autocatalytic essence of the growth processes was formally expressed as early as 1798 by T.R. Malthus [24] who wrote a differential equation for describing the dynamics of a population of proliferating individuals:

$$\frac{dW(t)}{dt} = a \cdot W(t). \quad (1)$$

The growth rate of the population W is proportional to W itself and parametrized by a relative growth (/proliferation) rate a . The Malthus equation can be reinterpreted to represent a very wide range of phenomena in various fields: behavior adoption in sociology, proliferation in biology, capital returns in economics, or proselytizing in politics. The (exponential) solution $\sim e^{(a \cdot t)}$ of this equation influenced much of the subsequent ideas in various fields and in particular it roused the first worries about the sustainability of growth. Malthus himself expressed

great concern of the humanitarian catastrophe that unlimited population growth may lead to. However, Verhulst [39] introduced (in 1838) a nonlinear interaction term $-b \cdot W^2$ (that may represent (confrontation over) limited resources in biology, competition in economics, limited constituency in politics and finite population in sociology)

$$\frac{dW(t)}{dt} = a \cdot W(t) - b \cdot W^2(t). \quad (2)$$

By including this term, rather than increasing indefinitely, the solution saturates at a constant asymptotic value $W \rightarrow \frac{a}{b}$. For the following two centuries, this logistic dynamics was considered by the leading scientists as a crucial element in various fields from biology (Volterra [40]) to the everyday world of politics and economics (Lord May [26]).

Real-Life Examples

The A(utocatalysis)-Bomb The first and the most dramatic example of the macroscopic explosive power of the multi-agent auto-catalytic systems is the nuclear (atom) bomb. The simple microscopic interaction underlying it is that the U235 nucleus, when hit by a neutron splits into a few energetic fragments including neutrons:



On the basis of (autocatalysis equation 1) even without knowing what is a neutron or a U235 nucleus, it is clear that a macroscopic chain reaction may develop: if there are other U235 nuclei in the neighborhood, the neutrons resulting from the first (autocatalysis equation 1) may hit some of them and produce similar new reactions. Those reactions will produce more neutrons that will hit more U235 that will produce more neutrons.

The result will be a chain (or rather branching tree) of reactions in which the neutrons resulting from one generation of fission events induce a new generation of fission events by hitting new U235 nuclei. This chain reaction will go on until eventually, the entire available U235 population (of typically some 10^{26} nuclei) is exhausted and their corresponding energy is emitted: the atomic explosion. The crucial feature in the equation above, which we call auto-catalysis, is that by inputting one neutron n in the reaction one obtains two (or more) neutrons ($n + n$). The theoretical possibility of iterating it and have an exponentially increasing macroscopic number of reactions was explained in a letter from Einstein to President Roosevelt. But only the later attack on Pearl Harbor lead to the initiation of the Manhattan project and the eventual construction of the A-bomb.

It is not by chance that the basic multi-agent method (the Monte Carlo simulation algorithm used until this very day in physics applications) was invented by people (Metropolis, Rosenbluth, Rosenbluth, Teller, Teller) involved in nuclear weapons research: the multi-agent method is the best fit method to compute realistically the macroscopic effects originating in microscopic interactions!

The B-Bomb: Autocatalysis and Localization in Immunology In no field is the auto-catalysis and localization more critical than in the emergence of living organisms functions out of the elementary interactions of cells and enzymes. From the very beginning of an embryo development the problem is how to create a controlled chain reaction such that each cell (starting with the initial egg) divides into similar cells, yet spatio-temporal structures (systems and organs) emerge. Let us consider the immune system as an example. The study of the Immune System for the past half century has succeeded in characterizing the key: cells, molecules, and genes. As always in complex systems, the mere knowledge of the microscopic world is not sufficient (and, on the other hand, some details of the micros are not necessary). Understanding comes from the identification of the relevant microscopic interactions and the construction of a multi-agent simulation with which to demonstrate in detail how the complex behavior of the immune system emerges. Indeed, the immune system provides an outstanding example of the emergence of unexpectedly complex behavior from a relatively limited number of simple components interacting according to known simple rules. By simulating their interactions in computer experiments that parallel real immunology experiments, one can check and validate the various mechanisms for the emergence of collective functions in the immune system. (e.g. recognition and destruction of various threatening antigens, the oscillations characteristic to rheumatoid arthritis, the localization of diabetes 1 to pancreatic islets etc). This would allow one to design further experiments, to predict their outcome and to control the mechanisms responsible for various auto-immune diseases and their treatment.

The Tulip Bomb The tulip mania is one of the most celebrated and dramatic economic bubbles in history. It involved the rise of the tulip bulb prices in 1637 to the level of average house prices. In the same year, after an increase by a factor of 20 within a month, the market collapsed back within the next 3 months. After loosing a fortune in a similar event (triggered by the South Sea Co.) in 1720 at the

London Stock, Sir Isaac Newton was quoted to say, "I can calculate the motions of the heavenly bodies, but not the madness of people."

It might seem over-ambitious to try where Newton has failed but let us not forget that we are 300 years later, have big computers and had plenty of additional opportunities to contemplate the madness of people. One finds that global macroscopic (and often catastrophic) economic phenomena are generated by reasonably simple buy and sell microscopic operations. Much attention was paid lately to the sales dynamics of marketable products. Large amounts of data has been collected describing the propagation and extent of sales of new products, yet only lately one started to study the implications of the autocatalytic multi-agent reaction-diffusion formalism in describing the underlying microscopic process [12,37,41,42].

Extensions of the Classical Logistic System

One of the great early successes of the logistic dynamics was its application to the spread of malaria in humans and mosquitos. Sir Ronald Ross was awarded the Nobel prize [30] for this work. His ideas were expressed by Lotka [19] in terms of a coupled system of two equations generalizing (2):

$$\begin{aligned} dw_1(t)/dt &= a_1 \cdot w_1(t) + a_{12} \cdot w_2(t) - a_{112} \cdot w_1(t) \cdot w_2(t) \\ dw_2(t)/dt &= a_2 \cdot w_2(t) + a_{21} \cdot w_1(t) - a_{212} \cdot w_1(t) \cdot w_2(t). \end{aligned} \quad (4)$$

Lotka has studied numerically this system in order to predict the ratios between the infected mosquitoes and the infected humans and the stability of the system. Vito Volterra advocated independently the use of equations in biology and social sciences [40] and re-deduced the logistic curve by reducing the Verhulst equation (2) to a variational principle that maximized a function that he named quantity of life [19]. Later, R.A. Fisher [11] extended of (2) to spatial distributed systems and expressed it in terms of partial differential equations:

$$\frac{\partial W(\vec{x}, t)}{\partial t} = a \cdot W(\vec{x}, t) - b \cdot W^2(\vec{x}, t) + D \cdot \nabla^2 W(\vec{x}, t). \quad (5)$$

He applied this to the spread of a mutant superior gene within a population and showed that as opposed to usual diffusion, the propagation consists of a sharp frontier (Fisher wave) that advances with constant speed (rather

then proportional to \sqrt{t} as in usual diffusion). Following its formulation, the mathematical study of (5) was taken over by mathematicians [17] and lead eventually a large number of physics studies (especially on the anomalous and fractal properties of the interface [1,6,14,15]).

A crucial step was then taken by Eigen [9] and Eigen and Schuster [10] who generalized the Lotka system of two equations for two populations to an arbitrary number of equations \simeq populations. They used the new system in the study of the Darwinian selection and evolution in prebiotic environments. More precisely, they considered quasi-species of auto-catalytic (self reproducing RNA sequences) molecules which can undergo mutations. Each sequence i self-replicates at a rate a_i and undergoes mutations to other sequences j at rates a_{ij} . The resulting system of equations is:

$$\begin{aligned} dW_i(t)/dt = & a_i W_i(t) + \sum_{j=1}^N a_{ij} W_j(t) \\ & - \sum_{j=1}^N a_{ji} W_i(t) - b(\overrightarrow{W(t)}, t) W_i(t). \quad (6) \end{aligned}$$

The arbitrary function $b(\overrightarrow{W(t)}, t)$ represents generically the interaction with the environment (in the specific case of ref [10] the result of replenishing and stirring the container continuously).

The extension of the logistic framework to social sciences was strongly advanced by Elliot Montroll who based a book on social dynamics on the principle that almost all the social phenomena, except in their relatively brief abnormal times obey the logistic growth [27].

An analogy that was often exploited in economics was the ecology-market metaphor (e.g. [31]) which was advanced in parallel with the more mechanical physics analogies. The connection to the logistic framework was strengthened by the evolutionary economics metaphor (e.g. [8,16,28]). This lead to the extension of (6) to economics with the a_i 's representing capital \simeq GDP growth rates and the a_{ij} 's representing trade, social security, mutual help or other mechanisms of wealth transfer (e.g. taxes \simeq subsidies). More recently [25] the logistic dynamics was applied to the dynamics of the equities i within a personal portfolio. Then $a_i(t)$'s are interpreted as the rate of growth of the equity i (at time t) and a_{ij} as the periodic redistribution of capital between the equities by the owner of the portfolio (in order to optimize it). Stochastic generalizations of the logistic \simeq Lotka-Volterra equations were studied also in a large body of mathematical literature (e.g. [17]), and in order to get meaningful results out

of the model, one has to introduce the noise in a proper way that will stand for its effect in real-life systems.

The Danger of Being Mean – Simple Examples

In this subsection we argue why microscopic (i.e. agent-based) studies are needed and why simplification in the style of mean field theories can be seriously wrong.

If we deal with a small biological population, then due to random accidents it may die out completely and irreversibly. For example, poachers may kill the two surviving males of a small elephant herd which is isolated from other elephants. It does not help the herd if one shows that *on average* there is enough food and space for two adult males, two adult females, and several calves. For larger populations usually such extreme fluctuations are less probable, and the time until it happens may increase exponentially with the population size.

Also, a hurricane may sink a ship even if averaged over the whole Atlantic Ocean the absolute value of the wind speed and wave height are moderate. In a marriage a husband is supposed to be faithful to his wife and should not average his efforts to become a father over 10^9 women; at least that's what wives often demand.

A less trivial example is demography. If you want to know how many people of retirement age are there for every thousand people of working age, usually one takes into account mortalities, birth rates, and migration. Let us assume, however, that one group of the population has a higher birth rate than the rest and that this difference is given on to the following generations, either genetically or culturally. Then, if everything else is the same, the group with the higher birth rate will finally dominate in the population, and using the *average* birth rate is not correct. (Of course, if the difference is small and we want to extrapolate over less than a century, then the average birth rate is still a good approximation.) One could remedy this error by simulating the two populations together; but then there could be other inherited traits which are demographically relevant, and thus with more and finer subdivisions we finally end up with agent-based demography [5], dealing with each individual.

This explains the conceptual gap between sciences: in conditions in which only a few exceptional individuals dominate, it is impossible to explain the behavior of the collective by plausible arguments about the typical or most probable individual. In fact, in the emergence of nuclei from nucleons, molecules from atoms, DNA from simple molecules, humans from apes, there are always the atypical cases (with accidentally exceptional advantageous properties) that carry the day. This effect seems to embrace the

emergence of complex collective objects in a very wide range of disciplines from bacteria to economic enterprises, from emergence of life and Darwinism to globalization and sustainability.

In the following section we will bring examples [33] where these effects lead to strong localization, such that the mean-field approximations give qualitatively wrong results, like predicting extinction where survival is possible. The approximations do not become good if only the population is large.

In conclusion, generic logistic ideas hinted by (2) arose for the last century in an extremely wide-ranging set of applications. For each discipline, subject and system, the variables of the model had to be interpreted in terms of the empirical observables and adapted to the relevant range of parameters and initial conditions. Once the parameters are specified, the generic framework (6) (plus noise) becomes a well defined model for a specific system. Then, one can derive from it precise predictions and confront them with the data.

Minimal Extensions to the Classical Logistic System

Here, we show how by restricting the parameter's regime of the generic framework (6) (plus noise), one ends up with a model that has a very strong prediction's power [2,3,18,29,34,35,36].

Case 1: The Generalized Lotka–Volterra System

If one considers a uniform interaction in (6), the resulting equation can be written as:

$$\begin{aligned} dW_i(t)/dt \\ = a_i \cdot W_i(t) + \alpha \cdot W(t) - b(\overrightarrow{W(t)}, t) \cdot W_i(t) \end{aligned} \quad (7)$$

where $W(t)$ is the average value of the W_i 's; then it was shown [18,29,34,35,36] that:

- The system has a steady state for the normalized quantity $X_i(t) \equiv \frac{W_i(t)}{W(t)}$.
- The steady state distribution of the X_i could be calculated analytically and the resulting distribution has the following form: $P(X) = e^{-2\alpha/XD} \cdot X^{-2-2\alpha/D}$ where D is the variance of the distribution from which the growth rates (a_i 's) is drawn out of.
- The fluctuations of the average ($W(t)$) have a wide distribution with a power-law tail that is closely connected with the value of the steady state distribution $(-2 - 2\alpha/D)$.

Obviously, as there is no explicit space in this system, one cannot see localization effects. However, intermittency is

very clear here: The fluctuations of the average value are enormous but changing around a fixed value. The possible interpretation of such a model are very diverse:

Income Distribution: $W_i(t)$ can represent the annual income of each individual in the society – then, the $W\alpha$ term is connected to social benefits one gets from the being part of the society, such as social security, charity and minimum wage. The a_i 's stand for the relative change between this year and the previous one. $b(\overrightarrow{W(t)}, t) \cdot W_i(t)$ then represents the overall trend of the market – periods of depression and of external investments.

Stock Market: $W_i(t)$ can represent the value of a specific stock in the stock market (at the closing time of the market for example) – then, the $W\alpha$ term is connected to correlations among the different stocks in the market. The a_i 's stand again for the relative change between the value today and the previous one. $b(\overrightarrow{W(t)}, t) \cdot W_i(t)$ represents the overall trend of the market – periods of depression and of external investments.

Population Dynamics: $W_i(t)$ can represent the number of individuals from a specific species in animals or of a specific nation in humans – then, the $W\alpha$ term is connected to immigration or mutations connecting the different populations. $b(\overrightarrow{W(t)}, t) \cdot W_i(t)$ represents the conditions for breeding.

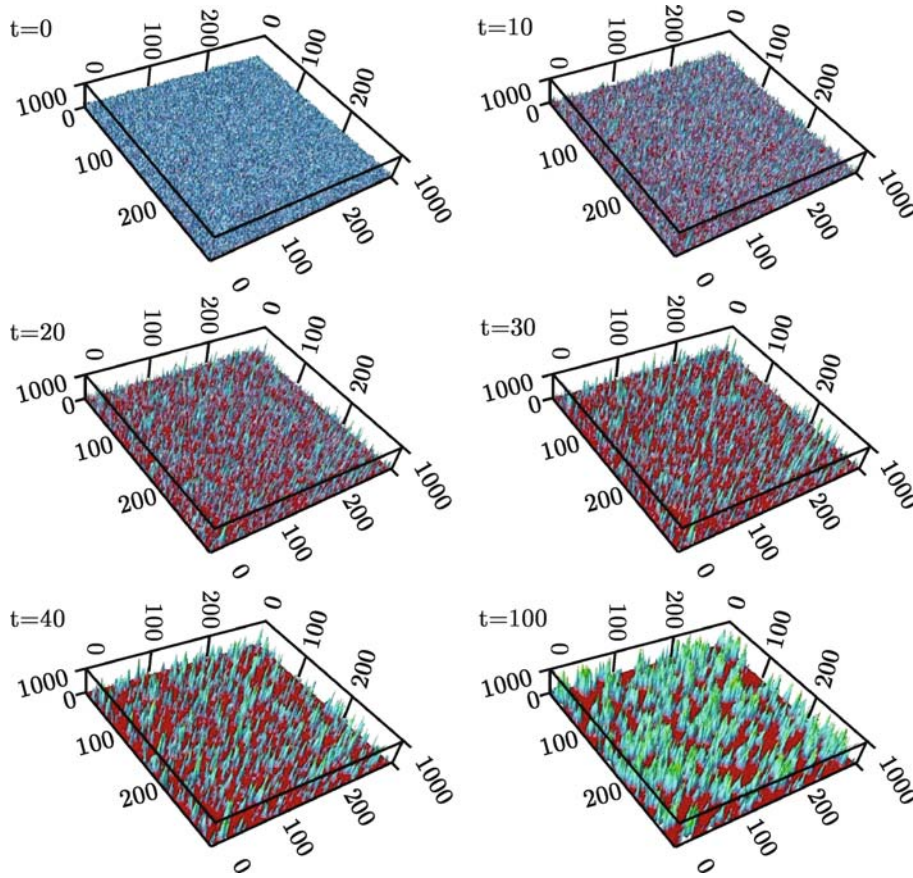
There are many more possible interpretations but the point is clear. For each interpretation one can argue that the uniform choice of the interaction matrix is unrealistic – of course it might be true, but as it turns out lately, the power-law prediction is very robust and can stand many different choices of this matrix.

Case 2: The AB Model

The AB Model [20,22,23,32,33] is actually a reaction-diffusion system which has two types of agents: A and B . It is a discrete system, both in space and in the fields it describes (A and B in any spatial point are natural numbers, never negative) and as such needs to be described with a set of rate equations. Then the agents may go through the following possible processes with the corresponding rates:

Diffusion: at each time step, with probabilities $D_a/2d$ and $D_b/2d$, respectively, an A or B moves to a nearest neighbor site on a d -dimensional lattice.

Reaction: at each time step, with probabilities μ and $\lambda \cdot N_A$, a single B dies or gives birth to a new B , respectively, where N_A is the number of A 's in the same location.



Intermittency and Localization, Figure 1

Snapshots of the AB Model in two dimensions [Log scale] in these 6 snapshots the time evolution of the AB Model is demonstrated: The two dimension lattice is set in $t = 0$ to be in a random distribution of the B s drawn from a Poisson distribution. The parameters are set in a way that if one looks at the mean-field approximation one will guess that the system needs to go to extinction in a relatively short time. However, due to the discreteness of the catalysts and the reactants, the B s adapt themselves to the rich (in food) areas and the famous island structure is formed

Naively, this system can be mapped into two partial differential equations:

$$\frac{dB(x, t)}{dt} = D_b \cdot \nabla^2 B(x, t) + (\lambda \cdot A(x, t) - \mu) \cdot B(x, t) \quad (8)$$

$$\frac{dA(x, t)}{dt} = D_a \cdot \nabla^2 A(x, t) \quad (9)$$

It is tempting to say that we can solve Eq. (9) to get:

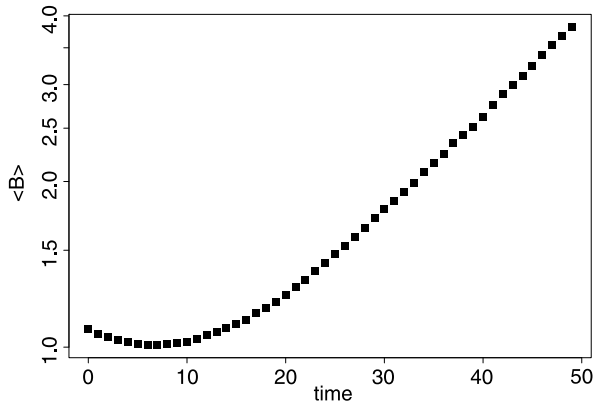
$$A(x, t) \longrightarrow n_A \quad (10)$$

in long times and then to plug it into Eq. (8) to say that depending on the parameter $m \equiv (n_A \cdot \lambda - \mu)$ the total number of B s will either increase exponentially (if $m > 0$) or decrease exponentially (if $m < 0$). It turns out that this mean-field treatment is totally wrong and as was shown in [33] in low enough dimensions ($d \leq 2$) the B 's will

asymptotically increase exponentially no matter what the rest of the parameters are! The intuitive explanation for this surprising result is that the B 's somehow adapt themselves to be localized around regions with good conditions (large number of A_i). One can see a typical snapshot of this system in Fig. 1. Another prediction [23] of this model is the intermittency of the total number of B 's even when one adds a saturation term similar to the second term in Eq. (2). Yet one more prediction of this model is the J-shape in the total number of B 's: i. e. initial decline followed by lasting exponential growth, Fig. 2.

Applying These Models to Real-Life Systems

As mentioned in Subsect. "Real-Life Examples", many real-life systems have characteristics that can be explained with the AB Model or the GLV:



Intermittency and Localization, Figure 2

The time evolution of the average B number in the AB Model in two dimensions For the same parameters as Fig. 1 if one plots the average B number as a function of time, the above picture is revealed: a J-shape. It turns out that this shape is a ubiquitous feature for the GDP evolution of many countries that went through a major shock in their economies like the collapse of the Soviet bloc for example

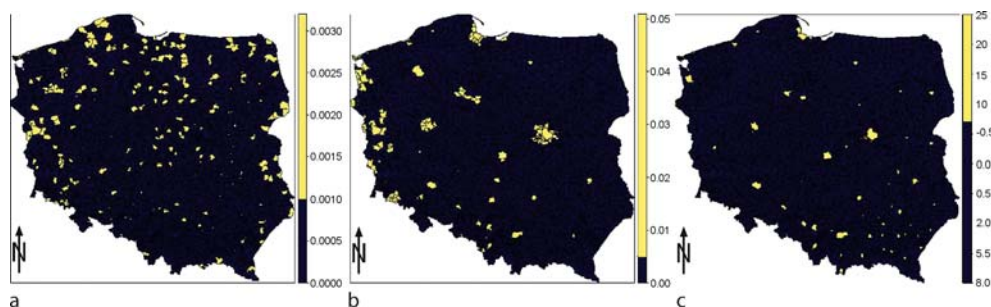
In the immune system, it was shown in [21] that the body's B cells tend to grow in places in the genome space where they are needed (depending on the diseases existing in the system).

On the Internet, it was shown in [13] that one can use the theoretical understanding of the model in order to plan strategies that will improve the way the Internet works.

In global economics it was shown in [23] that the global economic system can be mapped into a modification of the AB Model. In [18] the power law's distribution of the income and of the returns in the stock market were

compared to give extraordinary fit with reality. One more paper was just written [7] on how could decision makers gain knowledge on the economic system they are in charge of under the light of this model.

In this section we show in some more details one specific example of how one can take the predictions resulting from these types of models and apply them to real life systems: The system we will discuss is the Polish economy following the collapse of the Soviet bloc. We chose to present the system with the aid of Eq. (6). In the present application, the index i of the equations in the system (6) ranges from 1 to 2945 and labels the economy of each of the 2945 counties composing Poland. Each equation represents the evolution of the economic activity W_i of the county i . More precisely, W_i is the number of enterprises per capita in the county i . The a_i s represent endogenous growth rate of the country i and vary from county to county depending on local factors such as social capital, availability of natural resources or infrastructure. In fact the data indicate that the most important factor affecting the economic growth is the education level in the county. This dependence of a purely economical quantity on a social quantity is of great methodological importance and emphasizes in a dramatic way the importance of interdisciplinary studies (in this case economics, social science and physics). A recent work [43] led to a list of nine specific predictions resulting from the model. The data confirmed in a clear way the model predictions: Following the liberalization, the counties behaved in divergent ways: while most of the counties' economies plunged by factors of two, a few counties tripled their economic activity. This in turn lead to a quick increase in inequality between the counties.



Intermittency and Localization, Figure 3

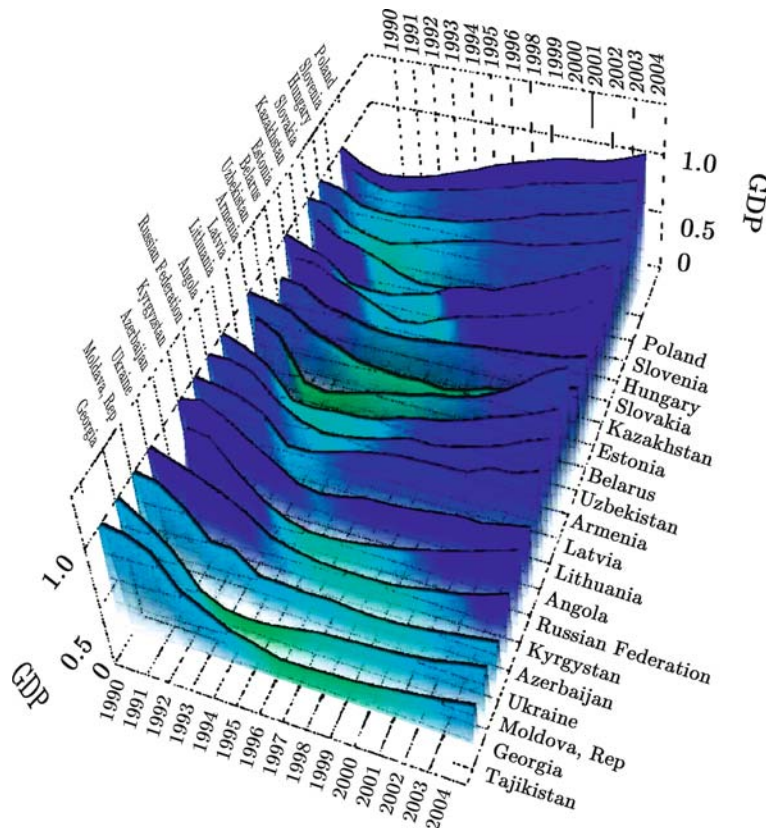
The influence of Education on Economic Activity before and after liberalization a The number of enterprises per capita in each county in 1989. b The number of enterprises per capita in each county in 1998. c The years of education per capita in various counties in 1988. a maps the number of the enterprises per capita in the year preceding the economic transition. This initial distribution does not display any spatial pattern: is very close to a uniform random (Poisson) distribution (similar to Fig. 1 at $t = 0$). b After the liberalization there is a clear spatial pattern: the economic activity is concentrated around the singular growth centers which are strongly correlated with the education levels *before* the transition (c). In the language of the AB Model, the A s – represent the education level, while the B s represent the economic activity (enterprises per capita)

During the preceding (socialist) regime, all counties were allocated roughly the same amount of economic activity by the central government. Thus the counties with high post-liberalization growth rate represented initially a negligible part of the country GDP and could not avert the fast global decay. However, within a couple of years, following their dramatic growth, the fast developing counties became the economic force, driving up the GDP. Moreover, their influence expanded to the neighboring regions until, eventually the entire country reached an uniform growth rate. This is not to say that the economic activity per capita equalized. Quite contrary, in the asymptotic regime in which the growth of the weak regions was due to the diffusion of economic activity from the fast developing regions, the very wide differences in GDP per capita persisted and in fact increased. One can see in Fig. 3 the

spatial structure of the system, in $t = 0$ (year 1989, before the liberalization) and in 1994 and compare it with the social conditions (education level) that catalyzed the economic growth. The localization effect is very clear. In Fig. 4 one sees how the generic prediction of the J-shape resulting from the AB Model is present in all of the formerly communist countries!

Future Directions

In his science fiction novel "Foundation" (1951), Isaac Asimov was playing with the idea of having a reliable predictions of the human society under the new scientific discipline he invented and called psychohistory. In this novel he was dealing mainly with the fact that unlike other scientific disciplines – here, due to the fact that human beings



Intermittency and Localization, Figure 4

The J-curve economic recovery after the liberalization of the Soviet Block The GDPs of the Eastern European countries experienced strong decay immediately after the economic liberalization. This was generally followed by a growth period. The resulting pattern resembles the letter J which explains the name J-curve. While the magnitude of the initial decay and the time and rate of recovery varies among the various countries, the J-shape is universal. The marked departure (exponential growth) of the Polish economic activity from an exponential decay extrapolated curve indicates that the classical global logistic framework cannot explain the observed pattern. The AB Model however does! (as can be seen in Fig. 2)

are involved, they are able to read the predictions and by that changing them... Without entering this discussion, we feel that Asimov succeeded to put his finger on a very crucial point: unlike physics for example, there is no such tool that people can rely upon when trying to predict the possible future outcome of today's deeds. Many scientists that come from the natural sciences are very suspicious towards their colleagues from the soft sciences because of this reason. On the other hand the soft scientists are claiming that the problems that they deal with are far too complex to put into solvable equations. We do understand the positions of both sides, but feel that the time has come to try and close the cultural gap between the two. The methods of the accurate sciences have been improved dramatically since the availability of computer power, on the other hand the social sciences are able today (also due to computers and Internet) to measure many social indexes on a very wide scale and for many years. What is needed now is first to make efforts to quantify the observations that social scientists agree upon, and by that to create a set of so called stylized facts. After having such list of qualitative and quantitative motifs that science agree they are present in reality – the road to having models that could be validated or invalidated by comparing their theoretical predictions to reality is closer than ever. We do believe that what was described in this paper are first steps towards Asimov's fantasy. Maybe this line of research will help us understand a little better the complex nature of human society.

Acknowledgments

The present research was partially supported by the STREPs CO3 and DAPHNet of EC FP6, and by GIACS (General Integration of the Applications of Complexity in Science).

Bibliography

Primary Literature

1. Ben-Avraham D, Havlin S (2000) *Diffusion and reactions in fractals and disordered systems*. Cambridge University Press, Cambridge
2. Biham O, Huang ZF, Malcai O, Solomon S (2001) Long-time fluctuations in a dynamical model of stock market indices. *Phys Rev E* 64(2):026101
3. Biham O, Malcai O, Levy M, Solomon S (1998) Generic emergence of power law distributions and lévy-stable intermittent fluctuations in discrete logistic systems. *Phys Rev E* 58:1352–1358
4. Billari F, Fent T, Prskawetz A, Scheffran J (2006) Agent-based computational modelling. *Physica*, Heidelberg
5. Bońkowska K, Szymczak S, Cebrat S (2006) Microscopic modeling the demographic changes. *Int J Mod Phys C* 17:1477–1484
6. Cardy JL, Tauber U (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. *Phys Rev Lett* 77:4780
7. Challet D, Yaari G, Solomon S (2008) The therapy to shock-therapy: optimal dynamical policies for transition economies. *Eur Phys J B*
8. Ebeling W, Feistel R (1982) *Physik der Selbstorganisation und Evolution*. Akademie, Berlin
9. Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* 58:465–523
10. Eigen M, Schuster P (1979) *The Hypercycle*. Springer, Berlin
11. Fisher RA (1937) The wave of advance of advantageous genes. *Ann Eugen* 7:355–369
12. Goldenberg J, Libai B, Solomon S, Jan N, Stauffer D (2000) Marketing percolation. *Phys Stat Mech Appl* 284:335–347
13. Goldenberg J, Shavit Y, Shir E, Solomon S (2005) Distributive immunization of networks against viruses using the 'honey-pot' architecture. *Nat Phys* 1:184–188
14. Grassberger P (1982) On phase transitions in schlogl's second model. *Z Phys B Condens Matter* 47:365–374
15. Janssen HK (1981) On the nonequilibrium phase transition in reaction-diffusion systems with an absorbing stationary state. *Z Phys B Condens Matter* 42:151
16. Jiménez-Montano MA, Ebeling W (1980) A stochastic evolutionary model of technological change. *Collect Phenom* 3:107–114
17. Kesten H (1980) *Random Processes in Random Environments*. Springer, Heidelberg
18. Levy M, Levy H, Solomon S (2000) *Microscopic simulation of financial markets: From investor behavior to market phenomena*. Academic Press, New York
19. Lotka AJ (1923) Contribution to the analysis of malaria epidemiology. *Am J Hyg* 3:1–121
20. Louzoun Y, Shnerb NM, Solomon S (2007) Microscopic noise, adaptation and survival in hostile environments. *Eur Phys J B* 56:141–148
21. Louzoun Y, Solomon S, Atlan H, Cohen IR (2001) The emergence of spatial complexity in the immune system ELSEVIER. *Physica A* 297:242–252
22. Louzoun Y, Solomon S, Atlan H, Cohen IR (2003) Proliferation and competition in discrete biological systems. *Bull Math Biol* 65(3):375–396
23. Louzoun Y, Solomon S, Goldenberg J, Mazursky D (2003) World-size global markets lead to economic instability. *Artif Life* 9(4):357–370
24. Malthus TR (1798) *An Essay on the Principle of Population*. Printed for Johnson J by Bensley T. Reprinted with an introduction by Antony Frew. Harmondsworth, Middx: Penguin Books
25. Marsili M, Maslov S, Zhang YC (1998) Dynamical optimization theory of a diversified portfolio. *Physica A* 253:403–418
26. May RM (1976) Simple mathematical models with very complicated dynamics. *Nature* 261:459–467
27. Montroll EW (1978) Social dynamics and the quantifying of social forces. *Proc Natl Acad Sci USA* 75(10):4633–4637
28. Nelson RR, Winter SG (1982) *The theory of economic development*. Harvard Univ Press, Cambridge
29. Richmond P, Solomon S (2001) Power laws are disguised Boltzmann Laws. *Int J Mod Phys C* 12(3):333–343
30. Ross R (1911) *The prevention of malaria*. John Murray, London
31. Schumpeter J (1934) *The theory of economic development*. Harvard Univ Press, Cambridge

32. Shnerb NM, Bettelheim E, Louzoun Y, Agam O, Solomon S (2001) Adaptation of autocatalytic fluctuations to diffusive noise. *Phys Rev E* 63:21103–21108
33. Shnerb NM, Louzoun Y, Bettelheim E, Solomon S (2000) The importance of being discrete: Life always wins on the surface. *Proc Natl Acad Sci USA* 97(19):10322–10324
34. Solomon S (2000) Generalized Lotka Volterra (GLV) models of stock markets. In: Ballot G, Weisbuch G (eds) *Applications of simulation to social sciences*. Hermes Science Publications, Paris, pp 301–322
35. Solomon S, Richmond P (2001) Power laws of wealth, market order volumes and market returns. *Physica A* 299(1):188–197
36. Solomon S, Richmond P (2001) Stability of Pareto–Zipf law in non-stationary economies. *Computing in Economics and Finance 2001* 11, Society for Computational Economics, Apr available at <http://ideas.repec.org/p/sce/scecf1/11.html>
37. Solomon S, Weisbuch G, de Arcangelis L, Jan N, Stauffer D (2000) Social percolation models. *Physica A* 277(1–2):239–247
38. Stauffer D, de Oliveira SM, de Oliveira PMC (2006) *Biology, Sociology, Geology by Computational Physicists*. Elsevier, Amsterdam
39. Verhulst PF (1838) Notice sur la loi que la population suit dans son accroissement. *Corresp Math Phys* 10:113–121
40. Volterra V (1931) *Variations and Fluctuations of the Number of individuals in animal Species living together*. Mc Graw Hill, NY
41. Weisbuch G, Solomon S, Stauffer D (2001) Economics with heterogeneous interacting agents. In: *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, p 43
42. Yaari G, Deissenberg C, Solomon S (2006) Advertising, negative word-of-mouth and product acceptance. *Eur J Econ Soc Syst* 19(2):257–268
43. Yaari G, Nowak A, Rakocy K, Solomon S (2008) Microscopic study reveals the singular origins of growth. *Eur Phys J B* 62(4):505–513

Books and Reviews

For general books and reviews see [4,18], and [38].

Internet Topology

YIHUA HE, GEORGOS SIGANOS, MICHALIS FALOUTSOS
University of California, Riverside, USA

Article Outline

Glossary

Definition of the Subject

Introduction

Data Sources and Their Limitations

Power-Laws of the Internet

Topology Generating Models and Tools

Conceptual Models for the Internet Topology

The Complete Internet Topology

Conclusion and Future Directions

Bibliography

Glossary

Autonomous system (AS) An autonomous system is a connected group of one or more IP prefixes, run by one or more network operators, which has a single and clearly defined routing policy. A unique AS number (or ASN) is allocated to each AS for identification purpose in inter-domain routing among ASes. For example, an organization, such as an ISP or a university, is an example of an AS. Some organizations may have more than one AS and thus have more than one AS number.

BGP (border gateway protocol) The Border Gateway Protocol (BGP) version 4 is the *de facto* routing protocol used in the Internet to exchange reachability information among ASes and interconnect them. The current BGP is version 4.

Degree The degree of an AS (or a node) is the number of neighbors of this AS (or node).

CCDF The complementary cumulative distribution function (CCDF) of a degree is the percentage of nodes that have a degree greater than the degree of interest.

Degree rank The degree rank of a node is its index in a list of degrees, ranked in decreasing order.

Eigenvalue Let A be an $N \times N$ matrix. If there is a vector $X \in \mathcal{R}^N \neq 0$ such that $AX = \lambda X$ for some scalar λ , then λ is called the eigenvalue of A with corresponding eigenvector X .

Definition of the Subject

Internet topology is the structure by which hosts, routers or autonomous systems (ASes) are connected to each other. The majority of existing Internet topology research focuses on the AS-level. There are three reasons for this. First, AS-level Internet topology is at the highest granularity of the Internet; other levels of Internet topology partially depend on AS-level topology. Second, the AS-level topology is relatively easy to obtain; other levels of topology are sometimes regarded as private information and are harder to get. Third, AS-level topology is not directly engineered by humans; instead, it is the aggregate result of technological and economical forces and, therefore, its origin and evolution attract considerable interest from investigators.

Research on Internet topology is driven by the explosive growth of the Internet, which has been accompanied by a wide range of inter-networking problems related to routing, resource reservation and administration. The study of algorithms and policies to address such problems often requires topological information and models. In 1999, Faloutsos et al. [27] discovered that the seemingly

random Internet topology does follow some rules: it follows power-law distributions. This finding revitalized the research on Internet topology and generated significant follow-up work.

Introduction

The Internet can be decomposed into connected subnetworks that are under separate administrative authorities, as shown in Fig. 1. These subnetworks are called *domains* or *autonomous systems* (ASes). The Internet community develops and employs different routing protocols inside an AS and between ASes. An intra-domain protocol, such as RIP, IS-IS, or OSPF, is runs within an AS, while an inter-domain protocol, such as BGP, runs between ASes. This way, the topology of the Internet can be studied at two different granularities. At the **router level**, each router is represented by a node [56], and a direct connection (either inter-domain or intra-domain) between any pair of routers is represented by an edge. At the AS level, each AS is represented by a single node [31] and each edge is an inter-domain interconnection. The study of the topology at each level is equally important.

This article focuses on AS-level Internet topology. Note that, at this level, only one edge exists between two nodes, although in practice there may be multiple connections between two ASes. This is a limitation of the nature of the data that are currently available.

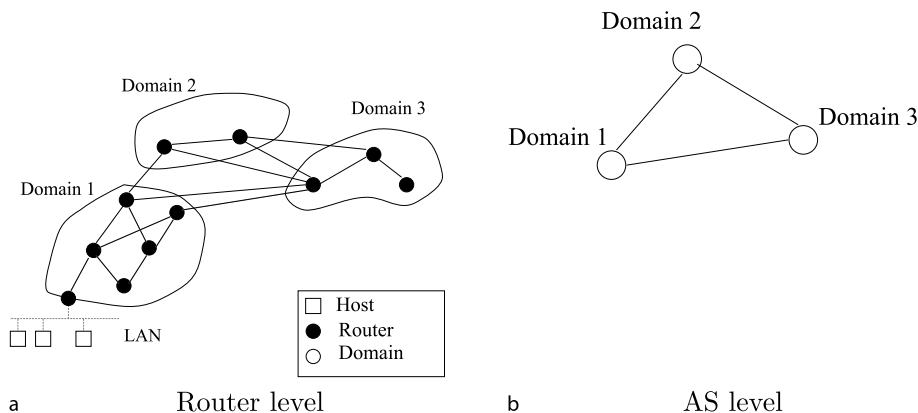
There are multiple benefits from understanding the topology of the Internet. For example, we want to be able to answer questions such as the following: “What does the Internet look like?”, “Are there any topological properties that don’t change in time?”, “How will it look a year from now?”, “How can I generate Internet-like graphs for my

simulations?”. Modeling Internet topology is an important open problem despite the attention it has attracted recently. Paxson and Floyd consider this problem as a major reason why we don’t know how to simulate the Internet [28]. An accurate topological model can have significant impact on network research. First, we can design more efficient protocols that take advantage of its topological properties. Second, we can create more accurate artificial models for simulation purposes. And third, we can derive estimates for topological parameters that are useful for the analysis of protocols and for speculation regarding the Internet topology in the future.

The rest of this article is structured as follows: We first will review how and where Internet topology information is collected in Sect. “[Data Sources and Their Limitations](#)”. We also compare the pros and cons of different data sources and discuss why they have such different properties. In Sect. “[Power-laws of the Internet](#)”, we will review the power-laws of the Internet, which constitute one of the most important discoveries of Internet topology. The power-laws lead to significant follow-up research in modeling the Internet topology, and we will discuss these models in Sects. “[Topology Generating Models and Tools](#)” and “[Conceptual Models for the Internet Topology](#)”. Then, we present the challenges and techniques in discovering the complete Internet topology in Sect. “[The Complete Internet Topology](#)”. Finally, we discuss future research directions of Internet topology in Sect. “[Conclusion and Future Directions](#)”.

Data Sources and Their Limitations

We first describe the data sources for collecting AS-level Internet topology. All of these sources and methods have



Internet Topology, Figure 1
The structure of Internet at two levels

Internet Topology, Table 1

An excerpt of an entry in typical Cisco “sh ip bgp” format BGP table dumps from BGP collector “route-views.oregon-ix.net” on May 1, 2007

Network	Next Hop	(Other Fields)	Path
2.0.0.0/24	157.130.10.233	(...)	701 1299 34211 41856 41856

their shortcomings, which will be discussed at the end of this section.

BGP Routing Tables

Border Gateway Protocol (BGP) routing table dumps are probably the most widely used resource that provides topological information. Typically, these routing table dumps are obtained from special BGP collectors, each of which connects with one or more Internet backbone routers in different ASes with special agreements. These BGP collectors do not advertise any prefixes (that is, IP blocks) to the Internet, while they are configured to receive all routes that the Internet backbone routers advertise to them. Therefore, these collectors are passive, and they have no effect on the global Internet. Periodically, each collector dumps its full routing table to Internet archives, which are usually publicly available for download.

Table 1 shows an entry of a typical Cisco “sh ip bgp” format BGP table dump from BGP collector “route-views.oregon-ix.net” on May 1, 2007. This entry indicates that, the destination network 2.0.0.0/24 can be reached via AS path “701 1299 34211 41856”. Therefore, the instance of Internet topology should include four ASes (AS701, AS1299, AS34211 and AS41856) and three links (AS701–AS1299, AS1299–AS34211 and AS34211–AS41856). Typically, a BGP collector’s routing table has more than a hundred thousands such entries from each peer AS; the total number of entries often exceeds several million. The number of ASes or routers of each BGP collector varies from a few to approximately one hundred. There are two well-maintained BGP routing table collector agents: Oregon Routeviews [54] and RIPE RIS [61]. Each of these agents maintains a number of BGP collectors around the world.

Besides routing tables, a BGP collector may also periodically dump routing updates received from its peers. Routing updates have a similar format to routing tables. A BGP update message displays the current route to a prefix, and therefore, a collection of BGP updates is able to reveal the dynamics of BGP routing.

Traceroute

Traceroute [36] is a tool that discovers the route that IP datagrams follow from one host to another. Traceroute takes advantage of the fact that each router decrements

the TTL (Time To Live) field by 1 for each IP packet that passes through it, and each router must discard any IP packet with TTL = 0 and send an ICMP “time-exceeded” error message back to the sender of the original IP packet. The original purpose of the action is to prevent IP packets from circulating the network forever. Traceroute operates by sending IP packets to a destination with small but increasing TTL values. These packets expire at the routers along the path to the destination. Since each router along the path sends an ICMP “time-exceeded” message back to the traceroute source, the identities of these routers (or more precisely, the outgoing IP interfaces of those routers) can be discovered. Although there is no guarantees that two consecutive IP packets will traverse the same route to the same destination, most often they do.

There are several large scale measurement projects using traceroute-like probes. Skitter [66] is a part of CAIDA’s [70] topology measurement project. CAIDA [70] maintains a set of (about 20) active monitors distributed around the globe. Each monitor uses a modified version of traceroute to probe a large set of IP addresses which cover nearly the whole IP address space. Rocketfuel [67] is a topology discovery project from the University of Washington. Rocketfuel uses a larger number of traceroute sources (a few hundred) from public traceroute servers as their sources. Therefore, Rocketfuel has a significantly higher number of vantage points than CAIDA. However, due to restrictions of the public traceroute servers, the rate of traceroute probing is limited in Rocketfuel. As a result, Rocketfuel is better at probing specific ISP networks rather than the whole Internet. Another promising project is NetDimes [62] from Tel Aviv University. To increase the number of vantage points, NetDimes distributes a large number of probing agents (tens of thousands) to global Internet users on a volunteer basis. These agents perform traceroutes according to the NetDimes center controls. Since the agents are mostly volunteers, coordination is still difficult when attempting to probe the Internet topology from anywhere at any given time.

All traceroute probes reflect only router-level topology. In order to obtain AS level topology, the probed IP addresses must be mapped to the ASes to which they belong. The conventional way to map an IP address to its AS number is by looking up, in the BGP routing tables, the IP block with the longest prefix match. For example, if

there is an IP address 2.0.0.18 and the longest prefix that it matches in the routing table is 2.0.0.0/24 (as shown in Table 1), then the announcing AS, which is the AS at the end of the “Path” field (AS41856 in Table 1), is the AS to which the IP 2.0.0.18 should be mapped. However, the accuracy of this method may be low in certain situations, as reported in the literature [48,49].

Internet Routing Registry (IRR)

The need for cooperation between autonomous systems is fulfilled today by the Internet Routing Registry (IRR) [54]. ASes use the Routing Policy Specification Language (RPSL) [61,66] to describe their routing policy, and router configuration files can be produced from it. At present, 55 registries exist, which form a global database from which to obtain a view of the global routing policy. Some of these registries are regional, such as RIPE or APNIC; other registries describe the policies of an autonomous system and its customers. The main uses of the IRR registries are to provide an easy way for consistent configuration of filters, and a way to facilitate the debugging of Internet routing problems. From the registered routing export and import policies, Internet topology can be extracted from IRR. For example, in Table 2, an excerpt of *aut-num* record for AS3303 in the IRR is shown. From the registered import and export policy in this excerpt, the Internet topology should include three ASes (AS3303, AS701 and AS1239), and two edges (AS3303–AS701 and AS3303–AS1239).

Data Source Comparison

BGP table dumps, especially the one from the Oregon Routeview project, are the most widely used source for Internet topology studies. An advantage of the BGP routing tables is that their link information is considered reliable.

Internet Topology, Table 2

An excerpt of the IRR in plain text format for AS3303

aut-num:	AS3303
as-name:	SWISSCOM
descr:	Swisscom Solutions Ltd
descr:	IP-Plus Internet Backbone
...	...
import:	from AS701 action pref=700; accept ANY
export:	to AS701 announce AS -SWCMGLOBAL
import:	from AS1239 action pref=700; accept ANY
export:	to AS1239 announce AS -SWCMGLOBAL
...	...

If an AS link appears in a BGP routing table dump, it is almost certain that the link exists. However, a limited number of vantage points makes it hard to discover a complete view of the AS-level topology. A single BGP routing table has the union of “shortest” or, more accurately, “preferred” paths with respect to this point of observation. As a result, such a collection will not see edges that are not on the preferred path for this point of observation. Several theoretical and experimental efforts explore the limitations of such measurements [1,21]. Furthermore, the incompleteness is statistically biased based on the type of the links: peer-to-peer links are more likely to be missing from BGP routing tables than provider-customer links, due to the selective exporting rules of BGP. Here, a provider-customer link means the two ASes incident to the edge have a provider-customer relationship, that is, one AS pays the other AS for traffic transit service. A peer-to-peer edge means the two ASes incident to the edge have a peer-to-peer relationship, that is, these two ASes have a mutual agreement that they carry traffic for each other with no or little fee. The classification of AS relationships can be inferred fairly accurately from the BGP routing tables by a number of algorithms [23,25,30,77]. The majority of the edges (approximately 80%) are provider-customer edges and most of the rest are peer-to-peer edges [33]. Typically, a peer-to-peer link can be seen only in a BGP routing table of the two peering ASes or their customers. Thus, given a peer-to-peer edge, unless a BGP collector peers with a customer of either AS incident to the edge, the edge can not be detected from the table dumps of the BGP collector. A recent work [18] discusses this limitation in depth. Thus, apart from being incomplete, the measured graph may not fairly represent the different types of links. Furthermore, BGP table dumps are likely to miss alternative and back-up paths. By definition, a router advertises only the best path to each destination, namely an IP prefix. Therefore, the back-up paths will not show up unless the primary link breaks. To address the problem, a recent effort suggests the need for actively probing backup links [19].

Previous studies [24,80] using BGP updates as a source of topological information show that by collecting BGP updates over a period of time, more AS links are visible. This is because as the topology changes, BGP updates provide transient and ephemeral route information. However, if the window of observation is long, an advertised link may cease to exist [80] by the time that we construct a topology snapshot. In other words, BGP updates may provide a superimposition of a number of different snapshots that existed at some point in time. Recently, Oliveira et al. [53] explicitly distinguished this commonly overlooked “liveness problem”. Note that BGP updates are col-

lected at the same vantage points as the BGP tables in most collection sites. Naturally, topologies derived from BGP updates share the same statistical bias per link type as topologies derived from BGP routing tables: peer-to-peer links are advertised only to the peering ASes and their customers. This further limits the additional information that BGP updates can currently provide. On the other hand, over a long period of observation, BGP updates could be useful in revealing ephemeral backup links which are not visible in the Internet at large, unless the primary link breaks down. At the same time, we could obtain erroneous BGP updates which may introduce fictitious links. To tell ephemeral and erroneous links apart, we need highly targeted probes. Recently, active BGP probing [19] has been proposed for identifying backup AS links. This is a promising approach that could complement other measurement work and provide the needed capability for discovering more AS links.

By using traceroute, one can explore IP paths and then translate the IP addresses to AS numbers, thus obtaining AS paths. Similarly to BGP tables, the traceroute path information is considered reliable, since it represents the path that the packets actually traverse. On the other hand, a traceroute server explores the routing paths from its location towards the rest of the world, thus the collected data has the same limitations as BGP data in terms of completeness and link bias. One additional challenge with the traceroute data is the mapping of an IP path to an AS path, as we previously mentioned. The problem is far from trivial, and it has been the focus of several recent efforts [48,49].

The Internet Routing Registry (IRR) [35] is the union of a growing number of world-wide routing policy databases that use the Routing Policy Specification Language (RPSL). In principle, each AS should register routes to all its neighbors (that reflect the AS links between the AS and its neighbors) with this registry. IRR information is manually maintained and there is no stringent requirement for updating it. Therefore, without any processing, AS links derived from the IRR are prone to human error and could be outdated or incomplete. However, up-to-date IRR entries provide a wealth of information that could not be obtained from any other source. A recent effort [63] shows that, with careful processing of the data, we can extract a non-trivial amount of correct and useful information.

Power-Laws of the Internet

The power-laws for Internet topology were first observed by Faloutsos et al. [27], and later summarized in [64]. In

those two papers, the authors showed that Internet topology at the AS level can be described efficiently with power-laws. The elegance and simplicity of the power-laws provide a novel perspective into the seemingly chaotic Internet structure.

Power-laws are expressions of the form $y \propto x^a$, where a is a constant, x and y are the measures of interest, and \propto stands for “proportional to”. Pareto was among the first to introduce power-laws in 1896 [57]. He used power-laws to describe distribution of income where there are a few very rich people, but most people have a low income. Another classical law, the Zipf law [81], was introduced in 1949, for the frequencies of English words and the populations of cities. More recently, power-laws have been observed in communication networks. Power-laws have been observed in traffic [20,43,59]. In addition, the topology of the World Wide Web [5,42] can be described by power-laws. Furthermore, power-laws describe the topology of peer-to-peer networks [40] and properties of multi-cast trees [16,60,71,76].

The initial work on power-laws [27] has generated significant follow-up work. In fact, [27] is one of the top five most cited computer science papers published in 1999 [17]. Various researchers have verified the power-law observations with different datasets [32,38,45]. In addition, significant work has been devoted to understanding the origin [50], and generating power-law topologies [8,13,37,39,50,51,55,69,78].

For the Internet topology, three power-laws have been identified: the rank power-law, the degree power-law and the eigen power-law.

Rank power-law

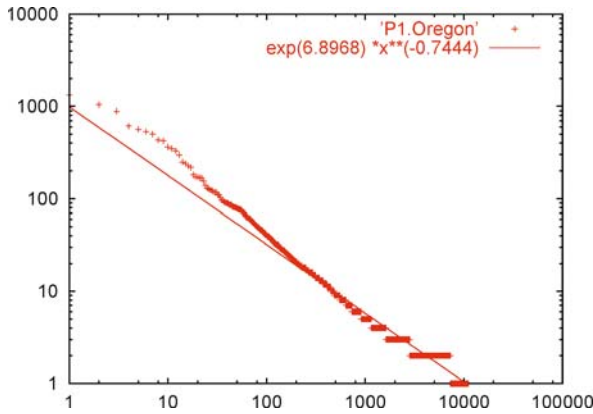
We empirically observe the following property for the Internet.

Power-law 1 (rank exponent) Given a graph, the degree d_v of a node v is proportional to the rank of the node r_v to the power of a constant \mathcal{R} :

$$d_v \propto r_v^{\mathcal{R}}$$

Definition 1 Let us sort the nodes of a graph in decreasing order of degree. We define the rank exponent \mathcal{R} to be the slope of the plot, on a log-log scale, of the degrees of the nodes versus the rank of the nodes.

Figure 2 shows the (r_v, d_v) pairs on a log-log scale after the nodes in an Internet topology are sorted in decreasing order of degree d_v . The measured data are obtained from the Oregon Routeviews [54] collector and are represented by



Internet Topology, Figure 2

Log-log plot of the degree d_v versus the rank r_v in the sequence of decreasing degree

points, while the solid line represents the least-squares approximation. A striking observation is that the plots are approximated well by linear regression. The correlation coefficient is over 0.97 in this case. The authors of [64] also inspected more than 1000 Internet topology instances over a six year span (1997 and 2003), and they found that for every instance of the inter-domain topology, the correlation coefficient was always higher than 0.97. This linearity is unlikely to be a coincidence.

Intuitively, power-law 1 correlates the degrees of the nodes and their rank and reflects a principle of the way domains connect. Such a relationship can be used to calculate the number of edges as a function of the number of nodes for an exponent of a given rank. In fact, in a graph where power-law 1 holds, it can be shown [27] that the number of edges E of the graph can be estimated as a function of the number of nodes N and the rank exponent R as follows:

$$E = \frac{N}{2(R+1)} \left(1 - \frac{1}{N^{R+1}} \right)$$

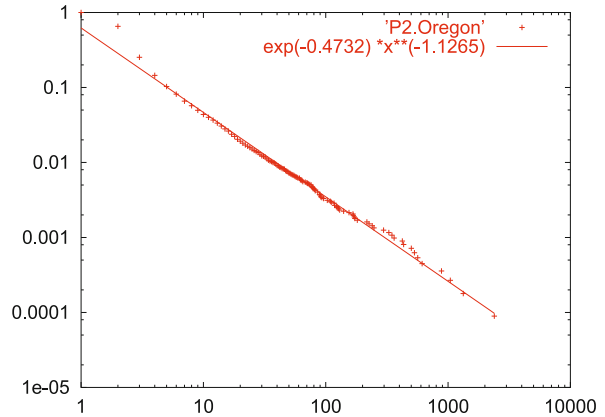
For additional discussion on estimates using this formula, see [27].

Degree Power-Law

Power-law 2 (degree exponent) Given a graph, the CCDF D_d of a degree d is proportional to the degree to the power of a constant \mathcal{D} :

$$D_d \propto d^{\mathcal{D}}.$$

Definition 2 We define the degree exponent \mathcal{D} to be the slope of the plot of the cumulative distribution of the degrees versus the degrees in log-log scale.



Internet Topology, Figure 3

The log-log plot of D_d versus the degree for the Oregon topologies

In Fig. 3 D_d is plotted versus the degree d in log-log scale. The major observation is that the plot is linear. The correlation coefficient is more than 0.996 for data obtained from Oregon Routeviews [54]. The authors in [64] found that the degree power-law holds for all the instances they inspected from 1997 to 2003, with correlation coefficient higher than 0.99.

The intuition behind this power-law is that the distribution of the degree of Internet nodes is not arbitrary. The qualitative observation is that lower degrees are more frequent. The power-law manages to quantify this observation by a single number, the degree exponent. This way, the realism of a graph can be tested with a simple numerical comparison. If a graph does not follow power-law 2, or if its degree exponent is considerably different from real exponents, it probably does not represent a realistic topology.

The exponents of rank and degree power-laws have been shown to be related [2,15]. More specifically, in a perfect power-law distribution, the exponent of the rank power-law is equal to the multiplicative inverse of the exponent of the degree power-law. However, in reality, the two exponents do not have such a perfect relationship. The discrepancy could be attributed to measurement imperfections and inaccuracies. In that regard, both the rank and the degree power-laws characterize the degree distribution from different angles, and it is useful to report both exponents when characterizing a topology.

Eigen Power-Law

Eigenvalues of a graph are the eigenvalues of its adjacency matrix. We observe the following property for the Internet graph.

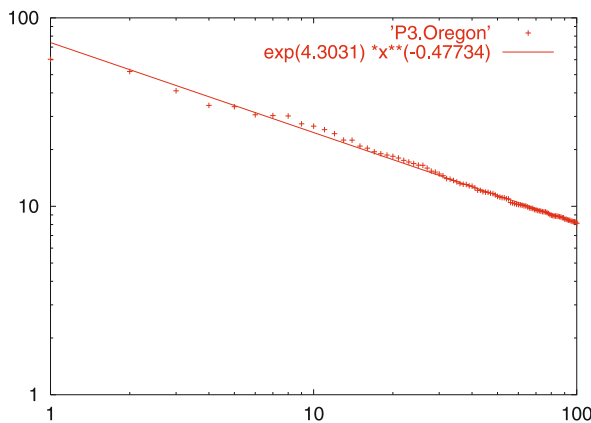
Power-law 3 (eigen exponent) Given a graph, the eigenvalues λ_i are proportional to the order i to the power of a constant \mathcal{E} :

$$\lambda_i \propto i^{\mathcal{E}}.$$

Definition 3 We define the eigen exponent \mathcal{E} to be the slope of the plot of the sorted eigenvalues versus their order in log-log scale.

In Fig. 4, the eigenvalues are plotted versus their order, in decreasing sequence, in log-log scale. The eigenvalues are shown as points in the figure, and the solid lines are approximations using a least-squares fit. Similar observations with equally high correlation coefficients were observed for all instances obtained between 1997 and 2003 [64]. The plot is practically linear with a correlation coefficient of 0.996, which constitutes an empirical power-law of the Internet topology.

Eigenvalues are fundamental graph metrics. There is a rich literature that proves that the eigenvalues of a graph are closely related to many basic topological properties such as the diameter, the number of edges, the number of spanning trees, the number of connected components, and the number of walks of a certain length between vertices, as shown in [22]. Interestingly, Mihail et al. [52] show that there is a surprising relationship between the eigen exponent and the degree exponent: the eigen exponent is approximately half of the degree exponent. In practice, the exponents adequately obey the mathematical relationship, although the match is, naturally, not perfect. All of the above suggest that the eigenvalues intimately relate to topological properties of graphs.



Internet Topology, Figure 4
The eigenvalues plot for the Oregon topologies

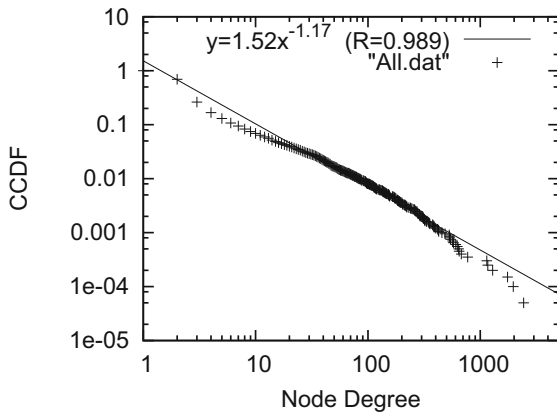
The Doubts and the Settlement

There has been a long debate on whether the degree distribution of the Internet at the AS level can be accurately characterized by power-laws [11,13]. The major concern is that by adding new edges discovered from sources other than Oregon Routeviews [54], the degree distribution of the Internet topology deviates from a perfect power-law. Interestingly, most of these new edges are of the peer-to-peer type which is underrepresented in the initial topologies as we explained earlier.

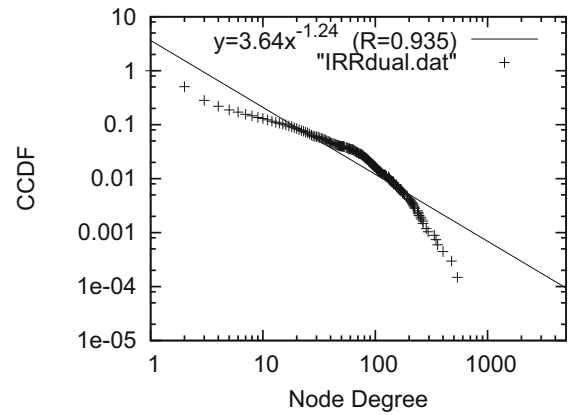
There are at least two reasons for this debate. First, this debate is partly due to the absence of a definitive statistical test. In Fig. 5a, the CCDF of node degrees is plotted for an Internet topology instance obtained from multiple resources, including Routeviews [54] and verified edges from IRR [35]. The distribution is highly skewed, and the correlation coefficient of a least square errors fitting is 98.9%. However, one could still use different statistical metrics and argue against the accuracy of the approximation [13]. Second, the answer could vary depending on which source we think is more complete and accurate, and the purpose or the required level of statistical confidence of a study. In Fig. 5b, the CCDF is plotted for an Internet topology instance obtained from the IRR after being filtered by the Nemecis tool [63]. The correlation coefficient is only 93.5%.

A recent paper [33] proposes a reconciliatory divide-and-conquer approach to explain and settle the debate. The authors propose to separately model the degree distribution according to the types of the edges: provider-customer and peer-to-peer. In Fig. 5, an indicative set of degree distribution plots are shown. The graphs obtained from multiple sources (Oregon Routeviews and traceroute-verified IRR links) are plotted in the left column (Fig. 5a,c,e), and the topology obtained from Nemecis-filtered IRR are plotted in the right column (Fig. 5b,d,f). The distributions for the whole graph are shown in the top row, only the provider-customer edges in the middle row, and only the peer-to-peer edges in the bottom row. The power-law approximation in the first two rows of plots and the Weibull approximation in the bottom row of plots are shown.

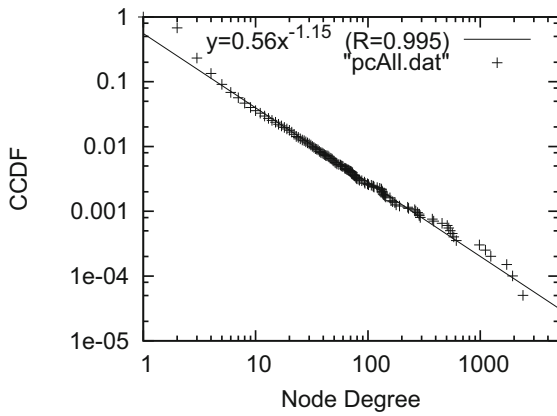
The following two properties can be observed from Fig. 5: (1) The provider-customer-only degree distribution can be accurately approximated by a power-law. The correlation coefficient is 99.5% or higher in the plots of Fig. 5c,d. Note that, although the combined degree distribution of the topology in the IRR does not follow a power law as shown in Fig. 5b, its provider-customer subgraph follows a strict power law in Fig. 5d. (2) The peer-to-peer-



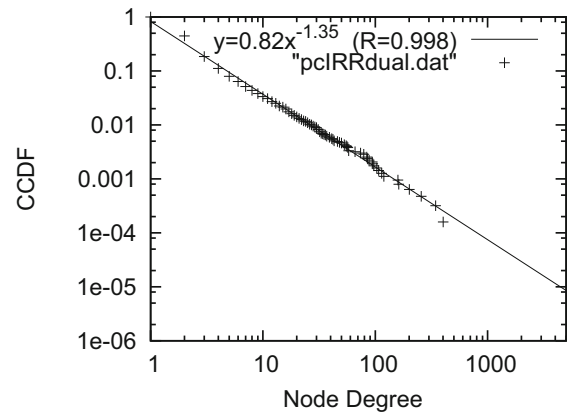
a Oregon + verified IRR



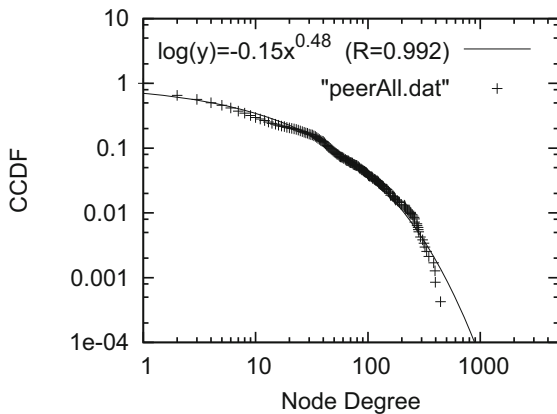
b Nemecis-filtered IRR



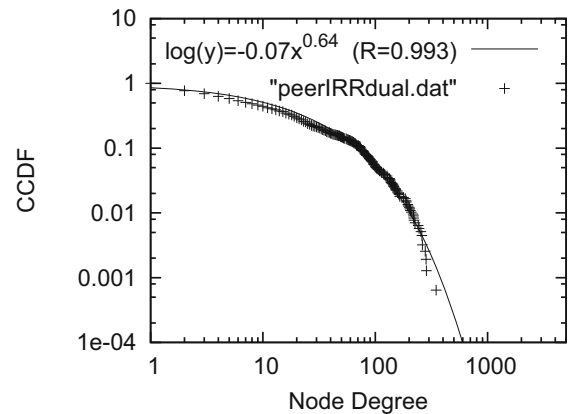
c Provider-customer links from a



d Provider-customer links from b



e Peer-to-peer links from a



f Peer-to-peer links from b

Internet Topology, Figure 5

The degree distributions of *Oregon* + verified IRR (left) and *Nemecis*-filtered IRR (right) in the *top row*, their provider-customer degree distributions in the *middle row*, and their peer-to-peer degree distributions in the *bottom row*

only degree distribution can be accurately approximated by a Weibull distribution [75]. The correlation coefficient is 99.2% or higher in the plots of Fig. 5e,f. It is natural to ask why the two distributions differ. The following could be one explanation: Power-laws are related to the rich-get-richer behavior; low degree nodes “want” to connect to high degree nodes. For provider-customer edges, this makes sense: an AS wants to connect to a high-degree provider, since that provider would likely provide shorter paths to other ASes. This is less obviously true for peer-to-peer edges. If AS1 becomes a peer of AS2, AS1 does not benefit from the other peer-to-peer edges of AS2 due to routing policies [72]: an AS normally will not carry traffic from one of its peers to its other peers. Therefore, high peer-to-peer degree does not make a node more attractive as a peer-to-peer neighbor. The validity of this explanation is still under investigation [33].

Topology Generating Models and Tools

So far, we have tried to model the learned topology as a snapshot. In this section, we present models and tools that attempt to generate Internet-like topologies.

Early Models

The simplest model is probably the **pure random** model. In this model, a set of nodes is distributed in a plane, and an edge is added between each pair of the nodes with a fixed probability p . Although this model does not explicitly attempt to reflect any structure of real networks, it is attractive for its simplicity: it is controlled by a single parameter.

The **Waxman** [73] model, on the other hand, adds edges with a probability that is a function of the distance between the nodes. Specifically, this probability for an edge between node u and node v is given by

$$P(u, v) = \alpha e^{-d/(\beta L)}, \quad (1)$$

where $0 < \alpha, \beta \leq 1$, d is the Euclidean distance from u to v , and L is the maximum distance between any two nodes. There are several variations of the Waxman model [26, 74, 79].

The **transit-stub** [9] method tries to impose a more Internet-oriented hierarchical structure as follows: In this model, each routing domain in the Internet can be classified as either a *stub* domain or a *transit* domain. A domain is a stub domain if the path connecting any two nodes u or v is in that domain; transit domains do not have this restriction – in other words, transit domains carry traffic not only for its own, but also for other domains as well. In

more detail, a connected random graph is first generated (by using, for example, the Waxman method described above). Each node in that graph represents an entire transit domain. Each transit domain node is expanded to form another connected random graph representing the backbone topology of that transit domain. Next, for each node in each transit domain, a number of connected random graphs are generated, representing stub domains that are attached to that transit node. Finally, some extra connectivity is added, in the form of “back-door” links between pairs of nodes, where a pair consists of a node from a transit domain and another from a stub domain, or one node from each of two different stub domains. By having nodes of different types, it is possible to generate large sparsely-connected Internet-like topologies with typically low node degrees. **GT-ITM** (Georgia Tech Internetwork Topology Models) is a tool that uses this transit-stub model.

The problem of these early models is that they do not generate power-law distributions. Medina et al. [51] tested the generated topologies from Waxman and Transit-Stub, and found both exhibit a weak presence or lack of the power-laws.

Pure Power-Law Models

Since the discovery of power-laws by Faloutsos et al. [27], the main focus of generating an Internet-like topology has shifted to matching the power-law exhibited in the Internet.

Palmer et al. [55] proposed the **PLOD** (power law out-degree) model. In this model, a degree credit is first assigned to each node in a graph with a given number of nodes. The degree distribution complies with the appropriate power-laws. Then, an edge placement loop is executed: it randomly picks two nodes and assigns an edge if they are not connected and each node still has remaining degree credit. After an edge is assigned, the degree credit of the nodes incident to the edge is deducted accordingly. The loop continues until there are no more pairs of nodes that fulfill the condition.

The concept of **PLRG** (Power Law Random Graph) was proposed by Aiello et al. [3] in the year 2000, and this model is also sometimes called **Model A**. In this model, a random graph is produced with a power-law degree distribution depending on two parameters that roughly delineate the size and density but are natural and convenient for describing a power law degree sequence. The power-law random graph model $P(\alpha, \beta)$ is described as follows: Let y be the number of nodes with degree x . $P(\alpha, \beta)$ assigns uniform probability to all graphs with $y = e^\alpha / x^\beta$, where α is the intercept and β is the (negative) slope when the de-

gree sequence is plotted on a log–log scale. After the degree distribution is defined, a set, L , which contains $\deg(v)$ distinct copies of each node v , will be formed. Then a random matching of the elements of L is chosen. For two nodes u and v , the number of edges joining u and v is equal to the number of edges in the matching of L joining copies of u to copies of v . The graph formed in the end is the PLRG.

These generators do not attempt to answer how a graph comes to have a power law degree sequence. Interestingly, these method seem to be able to match many other (but not all) topology properties of the real Internet [65].

Dynamic Growth Models

In contrast to the pure power-law models, dynamic growth models try to generate the Internet topology graph by simulating the growth of the Internet.

Barabasi and Albert [6] proposed a generic model (the **BA Model**) for network growth that relies on:

1. **Incremental growth:** The network expands continuously by the addition of new nodes.
2. **Preferential attachment:** A new node attaches preferentially to nodes that are already well connected.

In more detail, the network begins with a small number (m_0) of connected nodes. New nodes are added to the network one at a time. The probability $p(v)$ that a new node is connected to an existing node v is determined as the following:

$$p(v) = d_v / \sum_j d_j, \quad (2)$$

where d_v is the degree of node v and $\sum_j d_j$ is the sum of degrees of all existing nodes. In BA model, heavily linked nodes tend to quickly accumulate even more links, while nodes with only a few links are unlikely to be chosen as the destination for a new link. The new nodes have a “preference” to attach themselves to the already heavily linked nodes. This is so called “rich-get-richer” phenomenon.

The **AB model** [4] extends the BA model by adding a third operation called “rewiring”. The rewiring operation consists of choosing m links randomly and rewiring each end of them according to the same preference rule used in the BA model.

Bu et al. [8] found that the graphs generated by PLRG, BA and AB models have different characteristic values from the real Internet graph in terms of path length and clustering coefficient. They proposed **GLP** (Generalized

Linear Preference) [8], in which the probability p is

$$p(v) = (d_v - \beta) / \sum_j (d_j - \beta), \quad (3)$$

where $\beta \in (-\infty, 0)$ is a tunable parameter. The smaller the value of β , the less preference given to high degree nodes.

All these dynamic growth models produce graphs with power-law distribution. However, it is still difficult for these models to capture every topological property of the Internet. Authors in [65] show that even GLP does not follow some hierarchical properties of the Internet.

Sampling

All aforementioned models attempt to grow a graph, an approach which we call “constructive”. These methods depend on the principles of construction and the choice of parameter values. Furthermore, they often focus on matching a certain number of topology properties, while failing to match some others. At the same time, one can observe that we need small realistic graphs for simulation purposes and we do have several large measured topologies. To create small realistic graphs, Krishnamurthy et al. [41] proposed a “reductive” approach: instead of trying to construct a graph, they try to “sample” real topologies to produce a smaller graph. The idea is that the original properties, either well-known or unnoticed, can be kept during the process of reduction.

In more detail, they propose several reduction methods: **DRV** (Deletion of Random Vertex): Remove random vertexes, each with the same probability. **DRE** (Deletion of Random Edge): Remove random edges, each with the same probability. **DRVE** (Deletion of Random Vertex or Edge): Select a vertex uniformly at random, and then delete an edge chosen uniformly at random from the edges incident on this vertex. **DHYB- w** (Hybrid of DRVE/DRE): In this method, DRVE is executed with probability w and DRE is executed with probability $1 - w$, where $w \in [0, 1]$. This method was motivated by the study showing that sometimes DRVE and DRE had opposite performances with respect to different metrics.

The topologies sampled by both DRV and DRE methods are mathematically proved to follow power-law degree distribution. By comparing experimental data, the authors in [41] concluded that DHYB-0.8 is the best reduction method, and it also compares favorably to graph generation methods proposed previously in the literature. These sampling methods are successful to reduce a topology down to 30% of the original size. Beyond that the statistical confidence is found to be low.

Topology Generation Tools

Several publicly available topology generation tools exist.

BRITE (Boston university Representative Internet Topology gEnerator) [7] is a universal topology generator. It implements a single generation model that has several degrees of freedom with respect to how the nodes are placed in the plane and the properties of the interconnection method to be used. With different parameter settings, BRITE can generate either a Waxman model or a BA model.

Inet [34] is an AS-level Internet topology generator. Inet aims at reproducing the connectivity properties of Internet topologies as power-laws with additional improvements. It initially assigns node degrees from a power-law distribution and then proceeds to interconnect them using various rules. The current version, Inet-3.0, improves on previous versions by creating topologies with more accurate degree distributions and minimum vertex covers as compared to Internet topologies. Inet-3.0's topologies still do not well represent the Internet in terms of maximum clique size and clustering coefficient. These related problems stress a need for a better understanding of Internet connectivity and will be addressed in future work.

Conceptual Models for the Internet Topology

The Internet topology is large, complex and constantly changing. Even with the introduction of power-laws, which appear as a necessary though not sufficient condition for a topology to be realistic, a conceptual model of the topology [45,56,64] is still hard to obtain. Although the Internet is widely believed to be hierarchical by construction, it is too interconnected for an obvious hierarchy [69]. Several efforts to visualize the router-level topology have been made [14,66], however they can not be recreated manually and they do not provide a memorable model.

One goal here is to develop an effective conceptual model: a model that can be easily drawn by hand, while at the same time, is able to capture significant macroscopic properties. The jellyfish [65] and medusa [10] models are two conceptual models proposed for the inter-domain Internet topology.

The Jellyfish Model

The jellyfish model classifies ASes into different hierarchical layers. The highest layer is called the **Core**, which can be constructed in the following way: First, we sort all ASes in non-increasing degree order. The highest degree node is selected as the first member of the Core. Then, each AS is examined in the previous order; a node is added to the Core only if it forms a clique with the nodes already in the Core. In other words, the new node must connect to all the nodes already in the Core. The procedure stops when no more nodes can be added. The constructed Core is a clique of high-degree ASes, but not necessarily the maximal clique of the graph. The Core is a starting point to construct a jellyfish topology, and the ASes in the Core are probably the most important ASes in the Internet, since they have high degrees. The rest of the nodes are defined according to their proximity to the Core. The first **layer** is defined as all the ASes adjacent to the Core. Similarly, the second layer is defined as the non-labeled neighbors of the first layer. By repeating this procedure, six layers can be identified from the instances of Internet AS-level topology if the Core is counted as layer zero. Table 3 shows the number and percentage of ASes with each layer for three Internet topology instances at different times.

The jellyfish model also separately studies the one-degree ASes. First, the one-degree nodes are not useful in terms of connectivity to the rest of the network. Second, one-degree nodes are a large percentage of the network, and it is important to clarify and isolate their role. In fact,

Internet Topology, Table 3
Distribution of nodes in layers for three Internet instances

	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
Layer No	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Layer-0	8	0.23	14	0.176	13	0.08
Layer-1	1354	44.90	3659	46.25	7330	46.27
Layer-2	1202	39.866	3090	39.05	7116	45.51
Layer-3	396	13.134	1052	13.29	1078	6.89
Layer-4	43	1.425	86	10.87	96	0.61
Layer-5	12	0.398	10	0.12	1	0.0063

Internet Topology, Table 4

Distribution of nodes in shell and hang classes

Layer ID	Instance					
	Int-11-1997		Int-06-2000		Int-07-2003	
	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	8	0.23	14	0.176	13	0.08
Hang-0	465	15.42	798	10.08	1174	7.5
Shell-1	889	29.49	2861	36.16	6156	39.37
Hang-1	623	20.66	1266	16	2821	18.04
Shell-2	579	19.2	1824	23.05	4295	27.47
Hang-2	299	9.92	662	8.36	808	5.16
Shell-3	97	3.22	390	4.92	270	1.72
Hang-3	41	1.36	74	0.93	84	0.53
Shell-4	2	0.66	12	0.15	12	0.07
Hang-4	12	0.4	10	0.12	1	0.006

35%–45% of the ASes in the Internet are one-degree. In the jellyfish model, each layer is separated into two classes: a) the multiple-degree or **shell** nodes, and b) the one-degree or **hang** nodes. The one-degree nodes hanging from k th shell are referred as the k th hang class. For example, shell-0 is the Core, and its one-degree neighbors are denoted as hang-0, while the rest of the neighbors constitute shell-1. Naturally, the number of ASes in the layers, shells and hangs have the following relationship:

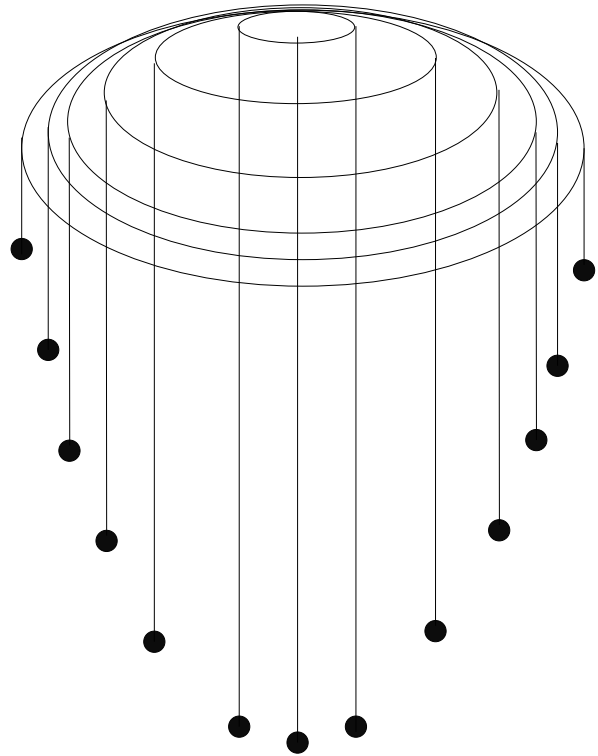
$$\text{Layer}_k = \text{Shell}_k + \text{Hang}_{k-1}$$

Table 4 shows the size of each group of nodes in the classification.

The conceptual the jellyfish model is described by the layer-shell-hang classification. The Core is the center of the head of the jellyfish surrounded by shells of nodes. Figure 6 shows a graphical illustration of this model. The hang nodes form the tentacles of the jellyfish. The length of the tentacle represents the concentration of one-degree neighbors for each shell.

Besides being a conceptual model, the jellyfish captures and represents concisely some fundamental properties of the Internet topology.

1. Core: The topology has a Core of highly connected important nodes, which is represented by the center of the jellyfish cap.
2. Center-heavy: Approximately 80% of the ASes are layer-1, layer-2 and layer-3 (see Table 3).
3. Node distance: Distances between ASes are small; maximum distance less than 11 hops, and 80% of the ASes are within 5 hops.
4. Edge types: Approximately 70% of the edges are between different node layers. The rest are horizontal to the hierarchy providing connectivity between nodes of the same layer.



Internet Topology, Figure 6

The internet topology as a jellyfish

5. One-degree nodes: There is a non-trivial percentage (35–45%) of one-degree nodes.

Note that the jellyfish structure seems to hold over time. We observe a 5–10% (in terms of total number of nodes) change of the number of nodes in each layer over 6 years, during which the size of Internet increased dramatically (see Table 4). In the future, with more years of observation,

one would attempt to derive trends in the evolution of the jellyfish.

The Jellyfish Model Tells the Difference

As shown in the previous section, the Internet topology fits the jellyfish profile. However, not every graph can be modeled as a jellyfish. For example, if a tree with N nodes were to be modeled into a jellyfish, the number of shells would be proportional to $O(\log N)$. This does not fit into the jellyfish profile of the Internet, where the number of shells is constant at five despite the rapid growth of the number of nodes. This provides an opportunity to use the jellyfish model as a criterion of the realism of Internet-like graphs.

Siganos et al. [65] use the jellyfish model to test the GLP methodology proposed in [8], and the PLRG approach proposed in [3]. The GLP approach depends on a preferential model. On the other hand, the PLRG generator is based on an interesting theoretical model for scale-free graphs, and takes the degree distribution as a given. In [8], these two generators were compared and it was concluded that the best generator was the GLP. PLRG was shown to fail in capturing properties such as the characteristic path length and the clustering coefficient. However, by using the jellyfish model, Sigamos et al. [65] were able to show that GLP does not capture the macro structure by using the jellyfish. Incidentally, PLRG seems to pass the test, although it fails with regard to other properties.

In more detail, two graphs are generated by the GLP model and the PLRG model, respectively. Both of them have similar number of nodes to an Internet topology instance in June, 2000. In Table 5, these graphs are decomposed using the jellyfish model. These results clearly show

that the graph generated using the GLP methodology is qualitatively different than the Internet graph. First, the Core of the network is much bigger in GLP (21) compared to the Internet (14). Second, the number of hanging nodes (degree one) in GLP far exceeds the number of shell nodes. The ratio is approximately 70% hanging nodes to 30% shell nodes. In the case of the Internet, this ratio is the opposite. Third, the GLP topology has only up to five layers, with the fifth layer having only three members, while the real Internet has six layers. On the other hand, PLRG seems to maintain similar structure according to the jellyfish model. The only differences between PLRG and the Internet are that the clique is smaller, having only 11 nodes, and that there is a slightly smaller shell-1 and a bigger shell-2.

The Medusa Model

One problem of the jellyfish model is that the identities of the Jellyfish Core are not particularly robust when the completeness of the Internet topology is uncertain. Carmi et al. [10] found that by adding or deleting an edge, the ASes in the Jellyfish Core could change up to 25%, mostly affecting some European ASes. To address the problem, they proposed a model called *medusa*. The medusa model depends on an informative functional decomposition of the Internet ASes called k -pruning, which proceeds as follows:

First, each AS with only one neighbor is removed. The link to that neighbor along with the node is removed as well. As this pruning proceeds, further nodes with one neighbor (or fewer) may be created. They will be removed until there is no longer a one-degree AS in the remaining graph. ASes removed in this way make up what is called 1-shell. The remaining graph is called 2-core. Sec-

Internet Topology, Table 5
Distribution of nodes in shell and hang classes

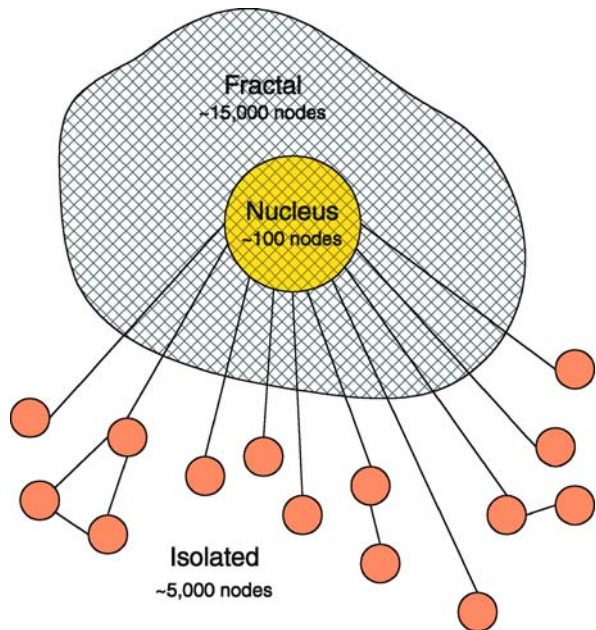
Layer ID	Instance					
	GLP		Int-06-2000		PLRG	
	Nodes	% of Nodes	Nodes	% of Nodes	Nodes	% of Nodes
Core/Shell-0	21	0.2	14	0.176	11	0.13
Hang-0	1885	23.82	798	10.08	565	7.1
Shell-1	1672	21.13	2861	36.16	2346	29.6
Hang-1	3371	42.6	1266	16	1298	16.4
Shell-2	688	8.7	1824	23.05	2305	29.13
Hang-2	221	2.79	662	8.36	525	6.6
Shell-3	3	0.037	390	4.92	325	4.1
Hang-3	3	0.037	74	0.93	125	1.5
Shell-4	0	0	12	0.15	41	0.51
Hang-4	0	0	10	0.12	23	0.29

ond, the pruning process is repeated and is characterized by an index k . For example, when $k = 2$, all nodes with two neighbors will be removed from the 2-core, and the removed nodes in this step is called 2-shell. The process continues, eliminating any nodes reduced to a degree of two (or fewer) by this pruning, until all nodes remaining have three or more neighbors. The remaining graph is called 3-core. The process is repeated to identify the 3-shell and 4-core, and so on. The process stops when no further nodes remain. The last nonempty k -core provides a very robust and natural definition of the heart or nucleus of any communications network. Finally, the k -crust is defined as the union of the nodes in the 1- through k -shells, and the links that join them. Thus $k - 1$ crust is the complement of the k -core.

For small values of k , the k -crust consists of many small connected components or clusters. For sufficiently large k , the largest connected cluster of a k -crust consists of a significant fraction of the whole k -crust, while no smaller cluster contains more than a few nodes. The change occurs at a well-defined threshold value of k . There is a significant fraction of the nodes within each large- k crust which is not part of its largest cluster, and remains isolated. Thus, the AS graph can be decomposed into three distinct components as shown in Fig. 7:

1. The nucleus (the innermost k -core),
2. The giant connected component of the last crust, in which only the nucleus is left out,
3. The isolated components of the last crust, nodes forming many small clusters. These connect to the connected component of the last crust only through the nucleus.

These three classes of nodes are quite different in their functional roles within the Internet. The nucleus plays a critical role in BGP routing, since its nodes lie on a large fraction of the paths that connect different ASes. It allows redundancy in path construction, which gives immunity to multiple points of failure. The connected component of the large- k crusts could be an effective substrate on which to develop additional routing capacity for messages that do not need to circle the globe. Finally, the isolated nodes and isolated groups of nodes in the last crust essentially leave all routing up to the nodes in the nucleus of the network. Because all their message traffic passes through the nucleus, even when the destination is relatively close by, they may be contributing unnecessary load to the most heavily used portions of the Internet. The relative size of this component could be a key indicator of the evolution of the topography of the Internet.



Internet Topology, Figure 7
The Internet topology as a medusa

This model can be visualized as Fig. 7. The Core of the medusa includes the most important nodes that are found in the Core and the first ring of the jellyfish's mantle. The jellyfish has relatively few rings around its Core, while the medusa's mantle is more extended and differentiated. The tendrils hanging from the jellyfish (leaf nodes) descend mostly from the Core, but also from all the other rings, while all the tendrils of the medusa are, by construction, attached to its nucleus.

The Complete Internet Topology

The accuracy of a topological model is important for effectively simulating, analyzing, and designing future protocols [28]. With an accurate Internet AS-level topology, first, one can design and analyze new inter-domain routing protocols, such as HLP [68], that take advantage of the properties of the Internet AS-level topology. Second, one can create more accurate models for simulation purposes [44]. Third, one can analyze phenomena such as the spread of viruses [29,58] more accurately. In addition, the current initiatives of rethinking and redesigning the Internet and its operation from scratch would also benefit from such an accurate Internet topology.

Toward Finding the Complete Internet Topology

Developing an accurate representation of the Internet topology at the AS level remains as a challenge despite the

recent flurry of studies [11,12,18,19,24,47,62,80]. One of the major problems is that, although the majority of ASes are represented completely in a snapshot, the edges among the ASes are not. As seen in an earlier section, each source has its own advantages, but each of them also provides an incomplete, sometimes inaccurate view of the Internet AS topology.

Recently, He et al. [33] present a systematic framework for extracting and synthesizing the AS level topology information from different sources. Instead of simply taking the union of all resources, a careful synthesis and cross-validation is performed. In addition to the sources mentioned above, they also utilize information gathered from IXPs (Internet Exchange Points), which have not received attention in terms of Internet topology discovery, although they play a major role in Internet connectivity.

He et al. identify and validate several properties of the missing AS links: (1) most of the missing AS edges are of the peer-to-peer type, (1) many of the missing AS edges from BGP tables appear in IRR, and (3) most of the missing peer-to-peer AS edges are incident at IXPs

Their work consists of four steps.

First, BGP routing tables are compared. They consider the AS edges derived from multiple BGP routing table dumps [80], and compare them to the Routeview data (OBD). The question to answer is “What is the information that the new BGP tables bring?”. Table 6 lists a portion of the collection of BGP table dumps that were collected in May, 2005. One observation, here, is that about 80% of the missing links that do not appear in a single table dump (OBD) but appear in a collection of table dumps (BD) are peer-to-peer type. For example, among 8702 edges in BD

but not in OBD, 7183 of them are classified as peer-to-peer type.

Second, He et al. systematically analyze the IRR data and identify topological information that seems trustworthy using the Nemecis tool [63]. They follow a conservative approach, given that the IRR may contain some out-dated and/or erroneous information. They do not accept new edges from the IRR, even after the first processing, unless they are confirmed by traceroutes using public traceroute servers. Overall, they find that the IRR is a good source of missing links. For example, they discover that more than 80% of the new edges found in the extra tables already exist in the IRR [35]. On the other hand, the IRR still has significantly more edges.

Third, He et al. provide a state-of-the-art methodology for identifying the ASes which participate at Internet Exchange Points (IXPs). An IXP is a relatively low cost solution by which an AS can peer with many other peers who are also participants at the same IXP. The exhaustive identification of IXP participants has received limited attention. Most previous work focuses on identifying the existence of IXPs. The finding here is that many of the ASes incident to the peer-to-peer edges missing from the different data sets are IXP participants. Note that even if two ASes peer at the same IXP, that does not necessarily mean there is an AS edge between these two ASes, because this totally depends on peering agreement between these two ASes. Therefore, in order to test whether or not these missing edges are indeed at the IXPs, they proceed to the next step.

Fourth, He et al. use their traceroute tool, RETRO, to verify potential edges from the IRR and IXPs. RETRO is

Internet Topology, Table 6
A collection of BGP table dumps

Route collector or Router server name	# of Nodes	# of Edges	# of edges with type inferred			edges not in OBD	edges not in OBD w/ type		
			total	p-p	p-c		total	p-p	p-c
route -views (OBD)	19843	42643	42570	5551	36766	0	0	0	0
route -views2	19837	41274	41230	4464	36514	1029	1028	835	191
route -views.eqix	19650	34889	34876	1027	33640	674	674	530	143
route -views.linx	19655	37259	37246	3246	33765	2511	2511	2188	319
route -views.isc	19753	36152	36139	1915	34004	784	783	663	118
rrc00.ripe	19770	36479	36465	1641	34605	655	654	543	111
rrc01.ripe	19640	34193	34180	1121	32855	617	617	512	105
rrc03.ripe	19737	39147	39129	3850	35042	3233	3228	2609	616
rrc05.ripe	19765	32676	32659	1122	31324	1095	1091	658	432
rrc07.ripe	19618	32811	31797	1219	30394	804	803	724	79
rrc12.ripe	19628	33841	33827	2024	31606	1611	1610	1417	193
Total (BD)	19950	51345	51259	12734	38265	8702	8689	7183	1499

a tool that collects public traceroute server configurations, sends out traceroute requests, and collects traceroute results dynamically. They confirm the existence of many potential edges identified in the previous steps. The results show that more than 94% of the RETRO-verified AS edges in the IRR indeed go through IXPs. They even discover edges that were not previously seen in either the BGP table dumps or IRR. In total, 300% more peer-to-peer links than those in the conventional BGP table dumps from Routeviews have been validated.

Towards Finding Internet Backup Links

One limitation of the previous method is that it ultimately depends on traceroute to verify the existence of a suspected edge. It is plausible that the suspected edge is a primary link, which means it exists most of the time. If a suspected edge is a backup, and it does not show unless some other links break down, it is unlikely to be seen by traceroute.

Recently, active BGP probing [19] has been proposed for identifying backup AS links. The main idea is to inject false AS path loops for an unused IP block. Since AS path loops are prohibited in inter-domain routing, BGP routers are forced to switch to backup links for this unused IP block. These links can be observed from any route collector, such as Routeviews or RIPE/RIS. This probing technique does not affect normal Internet routing because every change is restricted to the unused IP block.

In more detail, the principle of active BGP probing is the following: An active probing AS announces one of its prefixes with AS-paths including a number of other ASes. These ASes, due to loop detection, will not use or propagate the announcement. Then, if there is any alternative path available, it will show up. To avoid influencing AS-path length, the prohibited ASes are placed in an AS-set at the end of the path. For example, to stop its announcement from being propagated by ASes 1, 2, and 3, an AS (say AS12654) might announce one of its prefixes with an AS-path of [12654 {1,2,3}]. This allows AS 12654 to discover who propagates its announcements, find backup paths, and deduce the policies of other ASes with respect to its prefixes. By properly selecting the “prohibited” AS sets, one may be able to discover all backup links visible to the probing AS.

Note that, this is a promising method for discovering backup links, but so far it has been limited by the small number of probing ASes. This method and the method of discovering missing peer-to-peer links [33] are complementary to each other.

Conclusion and Future Directions

In this paper, we have surveyed current achievement in the research of Internet topology. In summary, power-law distributions constitute an undeniable property of the Internet topology, that has been consistent from 1997 to present, a period in which the size of Internet has grown eight-fold in terms of number of ASes. We examined different data sources for Internet topology research and provide some insight into their pros and cons. A number of important topology generating models and tools, as well as conceptual models of the Internet, were also studied. Finally we introduced current progress on finding the complete Internet topology.

Similar to any other science field, such as Physics, one of the ultimate goals of research is to identify the invariant laws from among seemingly uncontrolled appearance and phenomena. In terms of Internet topology, the power-law appears to be one such law emerging from the ever-changing and growing Internet. Obviously, there are more to discover. For example, a recent work [46] examines the correlation among node degrees and finds that by replicating node degree correlation, one can essentially reproduce the original topology. Another work [12] argues that one should take into account AS relationships to study the evolution of the Internet topology. The implications of these discoveries remain to be extended.

Bibliography

Primary Literature

1. Achlioptas D, Clauset A, Kempe D, Moore C (2005) On the bias of traceroute sampling, or power-law degree distributions in regular graphs. In: STOC 2005
2. Adamic LA (2000) Zipf, power-laws, and Pareto – a ranking tutorial. <http://www.parc.xerox.com/iea/>
3. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proceedings of the ACM symposium on theory of computing
4. Albert R, Barabasi A (2000) Topology of complex networks: local events and universality. *Phys Rev* 85:24
5. Albert R, Jeong H, Barabasi AL (1999) Diameter of the world wide web. *Nature* 401(6749):130
6. Barabasi A, Albert R (1999) Emergence of scaling in random networks. *Science* 286(5439):509
7. Brite. <http://www.cs.bu.edu/brite/>
8. Bu T, Towsley D (2002) On distinguishing between internet power law topology generators. In: Proceedings of the 21st Annual Joint Conference of the IEEE Computer and Communications Societies (Infocom IEEE 2002). vol 2. IEEE, pp 638–647
9. Calvert K, Doar M, Zegura EW (1997) Modeling internet topology. *Communication IEEE Magazine* 35(6):160–163
10. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of internet topology using k-shell decomposition. *Proc Natl Acad Sci USA* 104:11150–11154

11. Chang H, Govindan R, Jamin S, Shenker S, Willinger W (2004) Towards capturing representative AS-level internet topologies. *Comput Netw* 44(6):737–755
12. Chang H, Jamin S, Willinger W (2006) To peer or not to peer: modeling the evolution of the internet's topology AS. In: *Infocom IEEE* 2006
13. Chen Q, Chang H, Govindan R, Jamin S, Shenker S, Willinger W (2002) The Origin of power laws in internet topologies revisited. In: *Infocom IEEE* 2002
14. Cheswick B, Burch H (1998) Internet mapping project. *Wired Magazine* 6(12):216–217. See <http://cm.bell-labs.com/cm/cs/who/ches/map/index.html>
15. Chou H (2000) A note on power-laws of Internet topology, e-print cs. NI/0012019, Harvard University digital library for physics and astronomy
16. Chuang J, Sirbu M (1998) Pricing multicast communications: a cost based approach. In: *Proc. INET'98*
17. Citeseer. <http://citeseer.ist.psu.edu/articles1999.html>
18. Cohen R, Raz D (2006) The internet dark matter – on the missing links in the connectivity AS map. In: *Infocom IEEE* 2006
19. Colitti L, Di G Battista, Patrignani M, Pissonia M, Rimondini M (2006) Investigating prefix propagation through active BGP probing. In: *IEEEISCC* 2006
20. Crovella M, Bestavros A (1996) Self-similarity in world wide web traffic, evidence and possible causes. In: *SIGMETRICS*, pp 160–169
21. Crovella M, Lakhina A, Byers JW, Matta I (2003) Sampling biases in ip topology measurements. In: *Infocom IEEE* 2003
22. Cvetković DM, Boob M, Sachs H (1979) *Spectra of Graphs*. Academic press, New York
23. di Battista G, Erlebach T, Hall A, Patrignani M, Pizzonia M, Schank T (2007) Computing the types of the relationships between autonomous systems. *IEEE/ACM Trans Netw* 15(2): 267–280
24. Dimitropoulos X, Krioukov D, Riley G (2005) Revisiting internet AS-level topology Discovery. In: *PAM*
25. Dimitropoulos X, Krioukov D, Fomenkov M, Huffaker B, Hyun Y, Claffy KC, Riley G (2007) AS relationships: inference and validation. *ACM Sigcomm Comput Commun Rev (CCR)* 37(1): 29–40
26. Doar M, Leslie I (1993) How bad is naive multicast routing? In: *Proc. IEEEINFOCOM*, pp 82–89
27. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the Internet topology. In: *ACMSigcomm*, pp 251–262
28. Floyd S, Paxson V (2001) Difficulties in simulating the Internet. *IEEE Trans Netw* 9(4):392–403
29. Ganesh A, Massoulié L, Towsley D (2005) The effect of network topology on the spread of epidemics. In: *Infocom IEEE* 2005
30. Gao L (2000) On inferring autonomous system relationships in the internet. In: *Global IEEE Internet*
31. Govindan R, Reddy A (1997) An analysis of internet inter-domain topology and route stability. In: *Proc. IEEEINFOCOM*, Kobe, Japan, 7–11 April 1997
32. Govindan R, Tangmunarunkit H (2000) Heuristics for internet map discovery. In: *Proc. IEEEINFOCOM*, Tel Aviv, Israel, March 2000
33. He Y, Siganos G, Faloutsos M, Krishnamurthy S (2007) A systematic framework for unearthing the missing links: measurements and impact. In: *USENIXNSDI*
34. Inet topology generator. <http://topology.eecs.umich.edu/inet/>
35. Internet routing registry. <http://www.irr.net>
36. Jacobson V (1995) Traceroute. Internet measurement tool. Available at <ftp://ftp.ee.lbl.gov/traceroute.tar.gz>
37. Jaiswal S, Rosenberg A, Towsley D (2004) Comparing the structure of power law graphs and the internet AS graph. In: *ICNP* 2004
38. Jamin S, Jin C, Jin Y, Raz D, Shavitt Y, Zhang L (2000) On the placement of Internet instrumentation. In: *Proc. IEEEINFOCOM*, Tel Aviv, Israel, March 2000
39. Jin C, Chen Q, Jamin S (2000) Inet: Internet topology generator. Technical Report UMCSE-TR-433–00, Michigan
40. Jovanovic M (2001) Modeling large-scale peer-to-peer networks and a case study of gnutella. Master thesis, University of Cincinnati
41. Krishnamurthy V, Faloutsos M, Chrobak M, Cui J, Lao L, Percus AG (2007) Sampling large internet topologies for simulation purposes. *Comput Netw* 51(15):4284–4302
42. Kumar R, Raghavan P, Rajagopalan S, Sivakumar D, ATomkins, Upfal E (2000) The web as a graph. In: *Symposium ACM on Principles of Database Systems*
43. Leland WE, Taqqu MS, Willinger W, Wilson DV (1994) On the self-similar nature of ethernet traffic. *IEEE Trans Netw* 2(1):1–15 (earlier version in *Sigcomm'93*, pp 183–193)
44. Maennel O, Feldmann A (2002) Realistic BGP traffic for test labs. In: *Sigcomm ACM*
45. Magoni D, Pansiot JJ (2001) Analysis of the autonomous system network topology. *ACM Sigcomm Comput Commun Rev (CCR)* 31(3):26–37
46. Mahadevan P, Krioukov D, Fall K, Vahdat A (2006) Systematic topology analysis and generation using degree correlations. In: *Sigcomm ACM*
47. Mahadevan P, Krioukov D, Fomenkov M, Huffaker B, Dimitropoulos X, Claffy KC, Vahdat A (2006) The internet AS-level topology: three data sources and one definitive metric. *ACM Sigcomm Comput Commun Rev (CCR)* 36(1):17–26
48. Mao Z, Rexford J, Wang J, Katz R (2003) Towards an accurate AS-level traceroute tool. In: *Sigcomm ACM* 2003
49. Mao Z, Johnson D, Rexford J, Wang J, Katz R (2004) Scalable and accurate identification of AS-Level forwarding paths. In: *Infocom IEEE* 2004
50. Medina A, Matta I, Byers J (2000) On the origin of powerlaws in Internet topologies. *ACM Sigcomm Comput Commun Rev CCR* 30(2):18–34
51. Medina A, Lakhina A, Matta I, Byers J (2001) Brite: an approach to universal topology generation. In: *MASCOTS* 2001
52. Mihail M, Papadimitriou CH (2002) On the eigenvalue power law. In: *Proceedings of the 6th International Workshop on Randomization and Approximation Techniques*, pp 252–262
53. Oliveira R, Zhang B, Zhang L (2007) Observing the evolution of internet as topology. In: *Sigcomm ACM* 2007
54. Oregon Routeview Project. <http://www.routeviews.org>
55. Palmer CR, Stefan JG (2000) Generating network topologies that obey powerlaws. In: *Proceedings of the Global Internet Symposium, GLOBECOM* 2000
56. Pansiot J-J, Grad D (1998) On routes and multicast trees in the Internet. *ACM Sigcomm Comput Commun Rev* 28(1):41–50
57. Pareto V (1896) *Cours d'economie politique*. Dronz, Geneva
58. Park K, Lee H (2001) On the effectiveness of route-based packet

filtering for distributed DoS attack prevention in power-law Internets. In: Sigcomm ACM 2001

59. Paxson V, Floyd S (1995) Wide-area traffic: The failure of Poisson modeling. *IEEE Trans Netw* 3(3):226–244 (earlier version in *Sigcomm'94*, pp 257–268)
60. Philips G, Shenker S, Tangmunarunkit H (1999) Scaling of multicast trees: Comments on the chuang-sirbu scaling law. In: *Sigcomm ACM* 1999
61. Ripe Route Information Service. <http://www.ripe.net/ris>
62. Shavitt Y, Shir E DIMES: let the internet measure itself. *ACM Sigcomm Comput Commun Rev (CCR)* 35(5):71–74
63. Siganos G, Faloutsos M (2004) Analyzing BGP policies: methodology and tool. In: *Infocom IEEE* 2004
64. Siganos G, Faloutsos M, Faloutsos P, Faloutsos C (2003) Power-laws of the internet topology. *IEEE Trans Netw* 1(4):514–524
65. Siganos G, Tauro S, Faloutsos M (2006) Jellyfish: a conceptual model for the as internet topology. *J Commun Netw* 8(3):339–350
66. Skitter. <http://www.caida.org/tools/measurement/skitter/>
67. Spring N, Mahajan R, Wetherall D, Anderson T (2004) Measuring ISP topologies with rocketfuel. *IEEE Trans Netw* 12(1):2–16
68. Subramanian L, Caesar M, Ee CT, Handley M, Mao M, Shenker S, Stoica I (2005) HLP: next a-generation interdomain routing protocol. In: *Sigcomm ACM* 2005
69. Tangmunarunkit H, Govindan R, Jamin S, Shenker S, Willinger W (2002) Network topology generators: degree based vs. structural. In: *Sigcomm ACM* 2002
70. The Cooperative Association for Internet Data Analysis. <http://www.caida.org>
71. van Mieghem P, Hooghiemstra G, van der Hofstad R (2001) On the efficiency of multicast. *IEEE Trans Netw* 9(6):719–732
72. Wang F, Gao L (2003) Inferring and characterizing internet routing policies. In: *ACMIMW* 2003
73. Waxman BM (1988) Routing of multipoint connections. *IEEE J Sel Areas Commun* 6(9):1617–1622
74. Wei L, Estrin D (1994) The trade-offs of multicast trees and algorithms. In: *Proceedings of the International Conference on Computer Communications and Networks*
75. Weibull Distribution (2003) NIST/SEMATECH e-handbook of statistical methods. <http://www.itl.nist.gov/div898/handbook/>
76. Wong T, Katz R (2000) An analysis of multicast forwarding state scalability. In: *Proceedings of the International Conference on Network Protocols*
77. Xia J, Gao L (2004) On the evaluation of as relationship inferences. In: *Globecom IEEE* 2004
78. Yook SH, Jeong H, Barabasi A (2002) Modeling the internet's large-scale topology. *Proc Natl Acad Sci USA* 99(21):13382–13386
79. Zegura EW, Calvert KL, Donahoo MJ (1997) A quantitative comparison of graph-based models for internetworks. *IEEE Trans Netw* 5(6):770–783
80. Zhang B, Liu R, Massey D, Zhang L (2005) Collecting the internet AS-level topology. *ACM Sigcomm Comput Commun Rev (CCR)* 2005
81. Zipf GK (1949) Human behavior and principle of least effort: An introduction to human ecology. Addison Wesley, Cambridge

Books and Reviews

- Newman M, Barabasi A-L, Watts DJ (2006) *The structure and dynamics of networks*. Princeton University Press, Princeton

Invasion Percolation

MARK KNACKSTEDT¹, LINCOLN PATERSON²

¹ Department of Applied Maths, RSPHysSE,
Australian National University, Canberra, Australia

² CSIRO Petroleum, Clayton, Australia

Article Outline

Glossary

Definition of the Subject

Introduction

Classical Definitions, Algorithms and Results

Modifications to Invasion Percolation

Effect of Pore Scale Structure on IP

Effect of Correlated Heterogeneity on IP

Future Directions

Bibliography

Glossary

B	Bond number
D_b	Backbone
D_f	Fractal dimension
D_{\min}	Fractal dimension of minimum path
fBm	Fractional Brownian motion
fLm	Fractional Lévy motion
H	Hurst exponent
IP	Invasion percolation
NTIP	Non-trapping invasion percolation
OP	Ordinary percolation
p_c	Ordinary percolation threshold
S_r	Residual saturation
TIP	IP with trapping
ξ_w	Correlation length
ν	Percolation correlation length exponent
ϕ	Porosity
g	Field gradient
Z	Coordination number
σ	Standard deviation

Terms

Defender Fluid initially within pore space.

Invader Second fluid injected to displace defending fluid (defender).

Drainage Displacement of a wetting fluid by a non-wetting fluid.

Definition of the Subject

Invasion percolation is a simple dynamic process describing the slow displacement of one fluid by another in

a porous material. This is a common phenomena with many important applications; these include the penetration of nonaqueous polluting liquids into soil, the penetration of air into porous media such as soil, concrete, wood and ceramic powder during drying and the displacement of water from soil and rocks by gases generated by buried waste. A final important example, which is the primary focus of this review, is the accumulation during initial migration and the subsequent recovery and production of hydrocarbon reservoirs. Experiments on idealized systems have shown that the simple invasion percolation model provides a very realistic description of the slow fluid-fluid displacement processes associated with these important applications. Experiments and simulations of invasion percolation have been extended to consider the role of gravity, wettability and the complex nature of real porous materials on the dynamic immiscible displacement processes.

Introduction

Invasion percolation is a dynamic process that was proposed to describe the slow immiscible displacement of one fluid by another in a porous medium. Many porous media can be represented a network of pores (sites) connected via throats (bonds) [9,33,71]. In this representation all the pores and throats are initially filled with *defending* fluid. When a second fluid is injected very slowly into the porous medium such that the capillary forces dominate the viscous forces, the dynamics is controlled by the size of the local pore or throat. In a drainage process (the displacement of a wetting fluid by a nonwetting fluid), capillary forces are strongest at the narrowest places in the medium; in the pore throats. A drainage process is therefore represented as a series of discrete jumps across throats in which the non-wetting fluid displaces the wetting fluid via the *largest throat* (offering the least resistance to displacement). This is the version of the model originally considered by Chandler et al. [9], equivalent to bond invasion percolation. Wilkinson and Willemsen [71], who were the first to use the term *invasion percolation*, considered the process of imbibition (a non-wetting fluid being displaced by a wetting fluid) at a constant but infinitesimal (capillary dominated) flow rate. In this scenario the capillary forces are again strong in the throats, so the wetting fluid invades the throat quickly, but slows when entering the larger pores. This motion can be described by a series of discrete jumps in which at each time step the wetting fluid advances through the *smallest available pore*. This is site invasion percolation. In the absence of trapping, the bond version can be reduced to the site version through bond

to site transformations. However, when trapping is introduced where regions of defending fluid are incompressible and cannot escape, the transformation becomes a correlated site-bond problem [71].

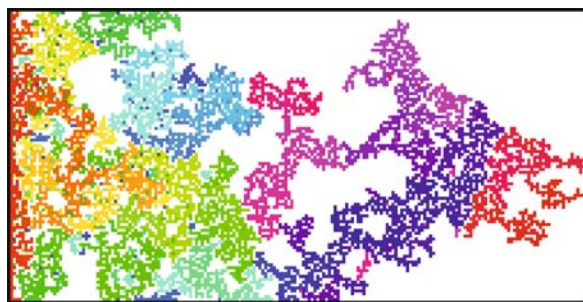
Most implementations of invasion percolation have been on a regular lattice. Sites or bonds representing pores and throats are usually assigned uniformly distributed random numbers as it is the sequencing of the events that is important, independent of the distribution. For drainage, invasion percolation is generally considered to be a good model. Although Wilkinson and Willemsen [71] used site invasion percolation for imbibition, later work by Blunt and Scher [7] has this as a special case where different alternative modes of wetting invasion are possible.

Classical Definitions, Algorithms and Results

Invasion vs. Ordinary Percolation

There are distinct differences between invasion percolation (IP) and ordinary percolation (OP). Invasion percolation starts with a well defined interface (inlet) and displaces the defending phase in a systematic way until spanning the system. In this way the concepts of history and the sequence of invading pores are naturally built into the model. The cluster generated in IP always spans a region between the injection face and the outlet face; there is no analogue to the percolation occupation probability p and there is only a single invasion cluster (no finite disconnected clusters). Figure 1 shows the spanning cluster at breakthrough of the invading phase for invasion percolation in two dimensions.

The implementation of trapping in the defending fluid introduces more differences. Monte-Carlo simulations of invasion percolation with trapping (TIP) in two dimensions [71] found that the fractal dimension of the sample



Invasion Percolation, Figure 1

Invasion percolation on 2D lattice. The invader (*colored*) enters from sites on the *left hand edge* and the defender exits from the *right hand edge*. Different colors indicate sites added within different time intervals

spanning cluster was $D = 1.82$, significantly smaller than $D = 1.896$ for OP. Simulation of non-trapping IP (NTIP) gave similar values to OP. In three dimensions no significant difference in the IP and OP models was originally observed. The difference in the 2D values was assumed to be associated with trapping and the effect of trapping thought to be negligible in 3D. The question of the universality class of IP was not conclusively established at this time.

There was extensive experimental evidence in support of the IP model of two phase flow in porous media. Lenormand and Zarcone [35] performed air drainage of oil in a transparent two-dimensional etched network of pores. Analysis of the fractal dimension of the invading phase gave $D = 1.82$ consistent with 2D simulations of TIP. In a further range of drainage experiments with a variety of wetting and non-wetting fluids including oil, air, water and different sucrose solutions Lenormand [34] showed that the results were completely consistent with an IP description of the phenomena. Stokes et al. [60] considered displacement patterns in a cell packed with unconsolidated glass beads. They found that under drainage conditions the resultant fluid displacement patterns were consistent with IP. Chen and Wada [10] used a technique of index matching fluids to visualize the fluid distributions of a quartz bead pack; again the fluid distribution patterns are reminiscent of IP. While the main application of IP has been to the description of the evolution of the interface between two immiscible fluids, IP also has applications to other problems including the characterization of optimal paths and domain walls in strongly disordered media [11,51] minimum spanning trees [12] and the simulation of the Ising model at the critical temperature [17]. Moreover, IP is one of the simplest parameter-free models which exhibits self-organized critically [58].

Methods and Algorithms

An elegant and fast algorithm for invasion percolation can be found in the book by Richard Gaylord and Paul Wellin [20], together with a detailed explanation. This algorithm, written in Mathematica, is included for convenience in Algorithms 1 and 2. As listed, this code starts on a square lattice with a single seed at 0,0 and maintains a list of potential invasion sites on the boundary as the cluster grows. It is straightforward to modify this algorithm to include boundaries and a different initial cluster.

Trapping IP results were for many years limited to small lattice sizes due to the time needed to search for the trapped regions at each time step. In the conventional algorithms the search for the trapped regions was done after every invasion event using a Hoshen–Kopelman algo-

rithm [24,59], which traverses the whole lattice, labels all the connected regions, and then only those sites (bonds) that are connected to the outlet face are considered as potential invasion sites (bonds). A second sweep of the lattice is then done to determine which of the potential sites is to be invaded in the next time step. Thus each invasion event required $O(N^2)$ calculations and limited TIP simulations to small lattice sizes.

Sheppard et al. [53] developed more efficient algorithms for generating TIP simulations. They noted, firstly, after each invasion event only a small local change is made in the interface; implementing the global Hoshen–Kopelman search is unnecessary. Secondly, it is wasteful to traverse the entire lattice at each time step to find the most favorable site (bond) on the interface since the interface is largely static. The first problem is tackled by searching the neighbors of each newly invaded site (bond) to check for trapping. This is ruled out in almost all instances. If trapping is possible, then several simultaneous breadth first “forest-fire” searches are used to update the cluster labeling as necessary [4]. This restricts the changes to the most local region possible. Since each site (bond) can be invaded or trapped at most once during an invasion, this part of the algorithm scales as $O(N)$. The second problem (identifying the sites for invasion) was solved by storing the sites (bonds) on the fluid-fluid interface in a list, sorted according to the capillary pressure threshold (or size) needed to invade them. This list is implemented via a balanced binary search tree, so that insertion and deletion operations on the list can be performed in $\log(n)$ time, where n is the list size. The sites (bonds) that are designated as trapped using the procedures described above are removed from the invasion list. Each site (bond) is added and removed from the interface list at most once, limiting the cost of this part of the algorithm to $O[N \log(n)]$. Thus, the execution time for N sites (bonds) is dominated (for large N) by list manipulation and scales at most as $O[N \log(N)]$. This allowed multiple simulations of TIP at scales of 4000^2 in 2D and 500^3 in 3D.

Universality Class of TIP

TIP describes waterflooding processes in secondary oil recovery. If differences in the topology of the transport pathways for the bond based TIP (drainage) and site-based TIP (imbibition) exist, this has a profound effect on the conductivity of the invading phase at the breakthrough point where a sample spanning cluster first forms (water breaks through during recovery). Differences have been noted in experimental measurements on rocks under different wettability conditions. Probing this question required simu-

lating TIP processes on very large lattices. This allows one to obtain precise estimates for the fractal dimensions of the sample spanning cluster, backbone and minimal path.

In two dimensions it was found that these fractal dimensions are non-universal and vary with the coordination number Z of the lattices [53,54]. The fractal dimension of the sample spanning cluster (SSC) of lattices with low Z exhibited the standard value of $D_f = 1.82$, the fractal dimension crosses over to the value given by OP for large Z (> 6). Values for the triangular lattice ($Z = 6$) seemed to give an intermediate value close to the prediction for OP. The same trends were seen with the value of the backbone and minimal path dimensions (Table 1). These results showed that the scaling properties of TIP in 2D are lattice dependent and hence non-universal.

It was initially thought that site- and bond-based IP were identical; results on large lattices showed distinct differences in the scaling properties of site-based and bond-based TIP. Results in 3D showed that site and bond TIP were in two distinct universality classes [54]. Site TIP had the same scaling behavior as NTIP and OP. A second universality class was observed for bond TIP.

Overall the results show that while ordinary percolation, a static process, is described by a unique universality class, TIP, a dynamic phenomena does not possess a unique universality class. The difference in the topology of site-based and bond-based TIP can assist in the interpretation of experimental measurements of waterflooded rock under different wettability conditions; for an oil wet rock the conductivity of the water channels can be sev-

eral orders of magnitude smaller at breakthrough than for a comparable water wet rock; this difference is consistent with the different topology of the flow paths for site vs bond based TIP.

Trapping Thresholds for TIP

As the TIP cluster is a fractal at breakthrough, the initially spanning cluster has essentially zero density. However a percolation threshold can be defined for TIP; this occurs beyond the percolation point, and is associated with the disconnection of the defending phase. This second percolation threshold is reached when the defender phase no longer percolates through the system and consists only of isolated clusters. At this point the TIP process ends. In two dimensions this second threshold corresponds to the invading phase breaking through. In three dimensions the trapping does not occur until the invader has occupied a significant fraction of the pore space. In a porous rock with two immiscible fluids this second threshold is associated with the residual saturation of the defending phase; no more defending fluid can be displaced from the sample without increasing the flow rate and introducing viscous forces into the displacement process. The value of the threshold is therefore of important practical interest. Numerical estimates of the trapping threshold for TIP on a cubic lattice were given by [71] as 0.66. Percolation and trapping thresholds have been related to the mean coordination number of the lattice [23]. Simulations of ordinary percolation on a range of simple lattices has led to the approximate relationships that $p_c^{\text{bond}} = 1.5/Z$ and $p_c^{\text{site}} = 2/Z$ [52]. Galam and Mauger [19] expanded on this idea and proposed a universal formula which in 3D is given by $p_c = p_0(Z - 1)^{-a}$ where p_0 and a are constants which depend on the type of percolation considered. Percolation threshold for OP and OP with trapping followed this relationship [47]. In a set of simulations on lattices of varying coordination number it was shown that the formula of [19] matches the TIP threshold of the lattices with $Z > 6$. However values of the threshold for coordination numbers $Z < 5$ [50] diverged from the prediction of [19]. The results are significant because of the application to the prediction of residual phase saturations in multiphase flow through porous media.

Modifications to Invasion Percolation

The invasion percolation model was originally developed to model two-phase displacements in porous media, with obvious application to oil recovery. However, even the early researchers in the field were aware that to be of practical significance and have utility the invasion percolation

Invasion Percolation, Table 1

The most accurate estimates of various fractal dimensions for IP in 2D and 3D, and their comparison with those of random percolation (OP) [52]

Model	D_f	D_{\min}	D_b
2D			
NTIP	1.8959	1.1293	1.6422
Site TIP ($Z < 6$)	1.825	1.203	1.217
Site TIP ($Z = 6$)	1.890	1.132	1.616
Site TIP ($Z > 6$)	1.895	1.136	1.642
Bond TIP ($Z < 6$)	1.822	1.214	1.214
Bond TIP ($Z = 6$)	1.823	1.215	1.215
Bond TIP ($Z > 6$)	1.895	1.221	1.221
OP	1.895	1.1307	1.6432
3D			
Site NTIP	2.524	1.3697	1.868
Site TIP	2.524	1.3697	1.861
Bond TIP	2.524	1.458	1.458
OP	2.524	1.374	1.87

model would need to be extended to include gravity, viscosity, and other effects.

Introduction of Gravity

Gravity was first introduced into the invasion percolation algorithm by Wilkinson [69] through the application of a simple linear weighting on the invasion thresholds in the direction of buoyancy.

An important application of invasion percolation with buoyancy is to the secondary migration of oil. Secondary migration is the slow process occurring over geological timescales where oil migrates from the source rocks where it is formed into structural or stratigraphic traps. It is in these traps where oil is found and produced, so a knowledge of secondary migration can be a key input into oil exploration. Secondary migration is a very slow process dominated by capillary and buoyancy forces, so invasion percolation is well suited to modeling this process as viscous forces can be safely ignored.

In a series of papers, researchers at the University of Oslo have explored invasion percolation as a model for secondary migration [42,43,64,66,67]. In particular they have studied the structure formed when the non-wetting fluid disintegrated into fragments [41]. They performed extensive two- and three-dimensional computer simulations and found that with a destabilizing external field on invasion percolation that displacement patterns were dominated by the growth of a single branch. This branch could be described in terms of a connected string of blobs of size ξ_w , which form a directed random walk along the direction of the field. On length scales smaller than ξ_w , the displacement patterns had the structure of invasion percolation clusters without a destabilizing field. They found that dependence of the correlation length ξ_w on the magnitude of the field gradient g is given by $\xi_w \sim |g|^{-\nu/(\nu+1)}$ (where ν is the ordinary percolation correlation length exponent) in accord with the theoretical arguments of Wilkinson [69,70].

A fundamental difference between invasion percolation and ordinary percolation is that invasion percolation generates a spanning cluster for lesser numbers of invaded sites, and this fraction of invaded sites tends toward zero as the network size increases. This is consistent with field observations; it can be very hard to detect secondary migration pathways that are only a thin filament occupying only an infinitesimal proportion of the exploration volume. Experiments conducted by Hirsch and Thompson on sandstone samples of different sizes found saturations consistent with these predictions from invasion percolation. Heterogeneity does however create local pools along

the migration path, so there are occasions where parts of the path can be detected.

Pore-network simulations and concepts from invasion percolation in a gradient have been used to study the effect of gravity on the critical gas saturation in a porous medium. Critical gas saturation denotes the volume fraction of the gas phase at the onset of bulk gas flow during the depressurization of a supersaturated liquid in a porous medium. Tsimpanogiannis and Yortsos [63] found that the critical gas saturation approaches two plateau values at low and high Bond numbers B . In the intermediate region it scales as a power law of B , which for a 2D lattice is $B^{-0.91}$.

Introduction of Viscous Forces

Invasion percolation corresponds to very slow displacements where viscous forces can be ignored. However, commercial extraction of oil is often at rates where viscous forces cannot be ignored hence there is incentive to expand the model to incorporate viscous forces. Wilkinson [69] was the first to suggest how percolation may be extended to include viscous forces through the use of a mean field description of the fluid flow. His ideas were developed further by Xu et al. [72] in the limits of high and low viscosity ratios. For small viscosity ratios, they determined that displacement could be modeled by a form of gradient percolation in a stabilizing gradient, involving a particular percolation probability profile. In the opposite case, the displacement can be described by gradient percolation in a destabilizing gradient and this leads to capillary-viscous fingering.

Introduction of Wetting

Invasion percolation was originally formulated with very simple concepts of solid surfaces with uniformly wetting and non-wetting fluids. Wetting refers to the fluid that preferentially contacts the solid surface. Experiments in sedimentary rocks over time necessarily led to the development of models involving additional mechanism such as snap-off and film flow, extending basic invasion and ordinary percolation models [7]. Further work on wettability showed that many rocks exhibit mixed-wettability, with the contact angle between the fluids varying with the diverse rock mineralogy [6]. This has led to more complicated network flow models that increasingly depart from the simple original invasion percolation model [8].

Effect of Pore Scale Structure on IP

In this section we consider the application of IP to the description of multiphase fluid flow in porous media at the

pore scale. In order to produce realistic descriptions of the multiphase flow behavior of real porous materials requires one to obtain an accurate description of the morphology of the porous material. In this section we consider the importance of sizes and shapes of pores and throats and the presence of pore to throat correlations to IP. Equally important are the connection patterns of the pores and throats or the topology of the real porous network. Network topology is defined by parameters which include mean connectivity, coordination number distributions and numbers of isolated clusters. In this section we describe the topological and geometric properties of a range of porous materials obtained from 3D images and compare them to classical lattices. Calculation of IP properties illustrates the importance of accurately reproducing the topology of real porous samples.

Measurement of Pore Scale Topology and Geometry

Direct Analysis of Pore Geometry and Topology Most studies of IP in 3D before 1995 were limited to regular cubic lattices. In recent years there has been mounting evidence that real rock topologies exhibit much lower coordination numbers. Ioannidis et al. [25] measured the average coordination number \bar{Z} from serial sections of a sandstone core and found $\bar{Z} = 4.1$. Bakke and Øren [5,45] developed a process-based reconstruction procedure which incorporates grain size distribution and other petrographical data obtained from 2D thin sections to build network models that are analogues of real sandstones. The average coordination number \bar{Z} of the resultant pore networks was significantly less than $Z = 6$. Recently, direct measurement of a 3D pore structure has become more readily available via synchrotron X-ray computed microtomography (micro-CT) [15,16,57], conventional micro X-ray CT [2] and laser confocal scanning microscopy [18]. Coupled with skeletonization algorithms [5,36,45,62] one can extract microstructural parameters for the direct input into network models. Lindquist et al. [37] originally made measurement of the pore coordination number in equivalent network models derived from a suite of Fontainebleau sandstone samples with porosity varying from 7.5% to 22%. The average coordination number varied from $\bar{Z} = 3.37$ at $\phi = 7.5\%$ to $\bar{Z} = 3.75$ at $\phi = 22\%$. Moreover, the coordination number of the pores within each sample exhibited a broad distribution; the majority of pores were 3-connected, however some pores with $Z > 20$ were observed.

Extensive topological analysis of rock core material [55] from a range of geological settings (sandstones, carbonates) has shown that the mean coordination num-

ber Z is usually $Z < 6$. In Fig. 2a–c we show a slice of a sand pack, the partitioning of the subvolumes into over 300,000 individual pores and throats and the resultant pore network structure. The coordination number or the pore space is found to be $Z = 5.4$. The full coordination number distribution is shown in Fig. 2d. Some pores exhibit coordination numbers > 20 . From the partitioning one can also directly define the pore size, throat or constriction size, shapes and tortuosity of the pore structure in 3D. Examples of distributions of these properties are given in Fig. 2e–f.

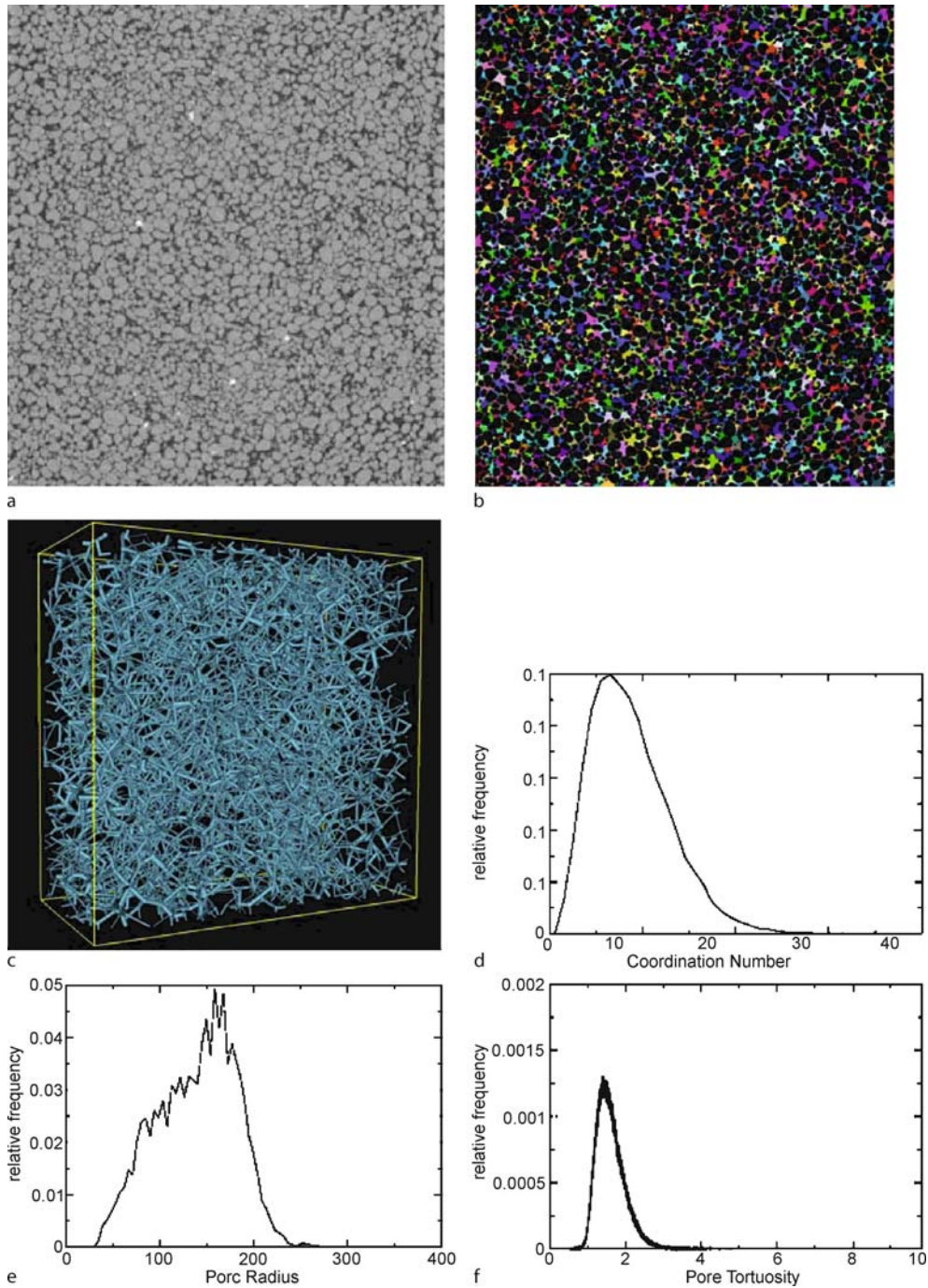
In Fig. 3 and Table 2 we illustrate and quantify the pore structure and resultant pore topologies of three other granular systems and rocks. The samples include a consolidated sandstone, an idealized multiscale material and an outcrop limestone. In Fig. 3 we show the images of network subsets. The networks in Fig. 3 contain a small subset ($< 10\%$) of the full image volume obtained from the tomogram. However the information obtained is very rich in detail. The difference in the network structure for these four samples is visually dramatic. The more complex samples (bidisperse and limestone) can exhibit pores of extremely high coordination ($Z > 100$) and large volume weighted coordination numbers $Z > 30$ reflecting the high connectivity of the larger pores. Clearly the use of a simple cubic lattice gives a poor topological description of these systems.

Implication of Pore Topology to IP Properties As discussed in Sect. “Trapping Thresholds for TIP”, percolation and trapping thresholds have been related to the mean coordination number of the lattice [23]. Simulations of ordinary percolation on a range of simple lattices has led to the approximate relationships that $p_c^{\text{bond}} = 1.5/Z$ and $p_c^{\text{site}} = 2/Z$ [52] and $p_c = p_0(Z - 1)^{-a}$. Results on regular lattices [50] showed that these predictions however did not give good correlations for $Z < 6$. As extensive topological analysis of rock core material from a range of ge-

Invasion Percolation, Table 2

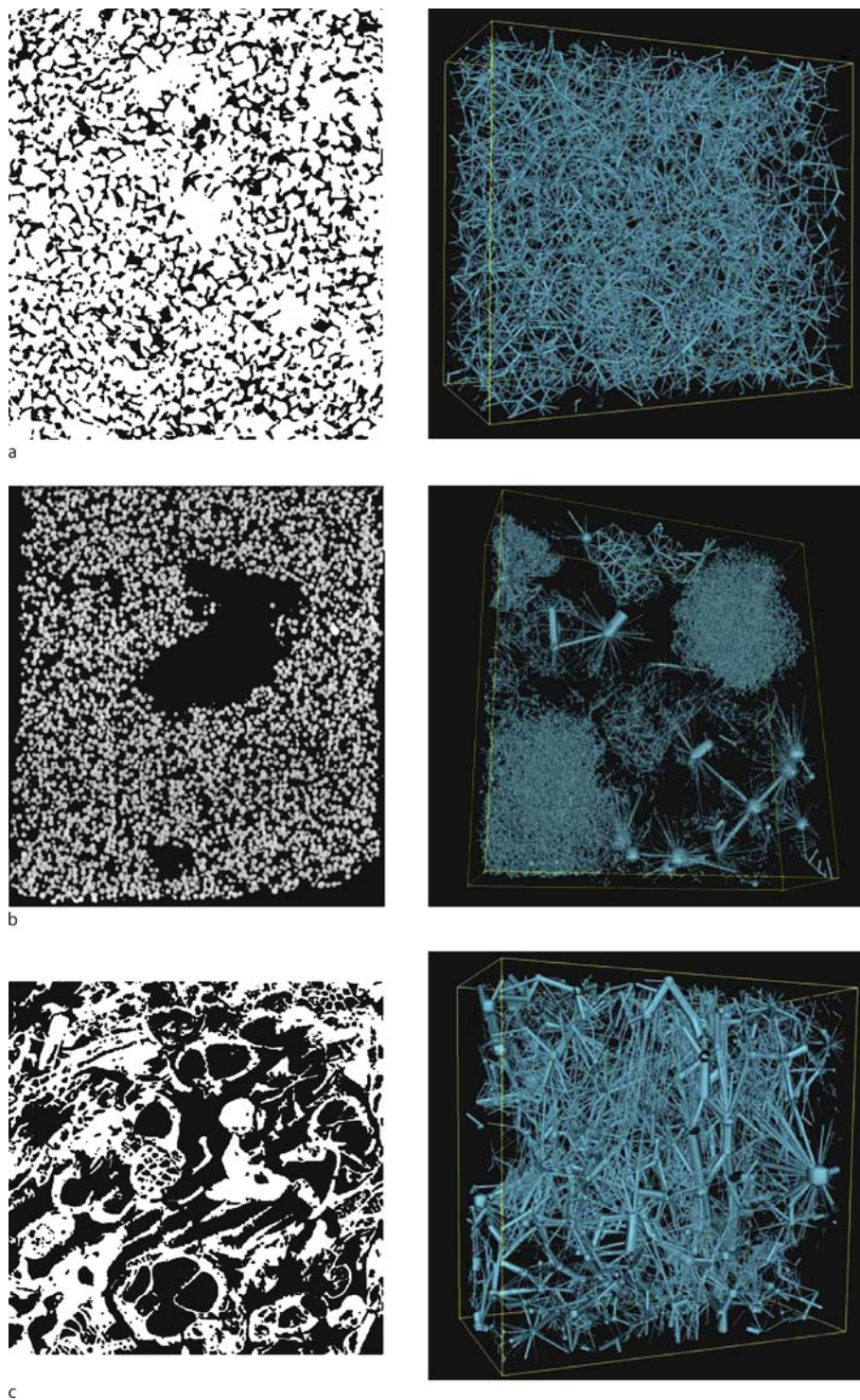
Details of the network structure for the three samples shown in Fig. 3. Z_m gives the mean coordination number, Z_w the volume weighted mean and Z_{max} the maximal pore coordination number. The mean $(\frac{R_p}{R_t})_m$ and volume weighted $(\frac{R_p}{R_t})_w$ pore to throat aspect ratios are also given

Sample	Z_m	Z_w	Z_{max}	$(\frac{R_p}{R_t})_m$	$(\frac{R_p}{R_t})_w$
Bead Pack	5.1	7.2	19	2.3	3.5
Sandstone	5.4	9.0	49	2.9	4.0
BiDisperse	3.6	31.4	227	3.0	30.3
Limestone	5.6	30.4	372	6.5	20.3



Invasion Percolation, Figure 2

a 2D slice of a 3D image of a sand pack along with the b results of a pore partitioning of the sample. c shows a small subset of the resultant network of pores and throats in 3D and d the coordination number distribution. e and f give geometric information including the pore size distribution and the pore tortuosity



Invasion Percolation, Figure 3

Images (*left*) and networks (*right*) of four samples. a Castlegate sandstone, b bidisperse sample and c Mt. Gambier limestone. The size of the pores and throats reflects their actual size in the partitioning of the 3D image. The variation in structure across the 3 samples is dramatic

ological settings has shown that the mean coordination number Z is often $Z < 6$ and this is of importance in the prediction of residual phase saturations in porous rocks. Moreover, given the broad distributions in Z observed in real rock materials, is the mean coordination number sufficient to predict the trapping threshold on real pore network topologies? These questions have been explored recently by a number of researchers.

Mean Coordination Number Suding and Ziff [61] first showed the importance of lattice topology other than mean coordination number on percolation thresholds of OP in two dimensions. They considered 11 Archimedian lattices with identical mean coordination numbers. They showed that the mean coordination number alone was not sufficient to predict p_c . Predictions for p_c on these lattices varied from 0.55–0.80. Stochastic networks of rock material had previously [25] been generated to solely match the coordination number of rocks by diluting sites on a regular cubic lattice until the remaining connected component had the desired mean coordination number. Analysis of the trapping threshold on a diluted cubic network and Fontainebleau sandstone however gave very different thresholds. There was clearly a need to compare networks in 3D with precisely controlled topology. Shepard developed a 4 part algorithm to generate a network with a specified coordination number distribution [56]. Use of this algorithm allowed one to generate stochastic networks with matching of mean coordination and the coordination number distribution. Sok et al. [56] generated an ordered (diamond) network ($Z = 4$) and a range of stochastic networks with an identical mean coordination number $Z = 4$ and different standard deviations in Z ; $\sigma(Z) = 0.0001, 1.0, 2.0$. The trapping threshold for TIP on the lattices differed strongly from the diamond network (see Table 3). The bond TIP threshold for the diamond network and the stochastic network with $\sigma(Z) = 0.001$ varied strongly, while the site TIP threshold varied. Variation of the coordination number distribution led to strong differences in the resultant TIP thresholds. These results showed that thresholds cannot be correlated solely to dimension and coordination number of the network.

Invasion Percolation, Table 3
Threshold values for the four coordinated lattices

Regular Lattice	0.290	0.417
Stochastic: $\sigma = 0.001$	0.298	0.426
Stochastic: $\sigma = 1.0$	0.328	0.429
Stochastic: $\sigma = 2.0$	0.375	0.437

Coordination Number Distribution and Higher Order Measures Rock networks display a broad distribution of coordination number and the presence of long range topological bonds. Recent results have shown [32,56] that honoring the full coordination number distribution does lead to better prediction of the trapping IP thresholds than matching the coordination number alone. However differences have still been noted. This result shows that one might require a complete description of the network topology to accurately predict trapping thresholds. One study [56] introduced higher order topological quantities; ring size, coordination sequence and topological bond length as measures of relevance to the prediction of trapping thresholds. Comparison of these topological properties between rock networks and stochastic networks showed clear differences. A second study [32] showed that rock networks exhibited a large proportion of topological long bonds; bonds connecting two pores which are not nearest neighbors. A procedure to add topological long bonds was then developed. The introduction of topological long bonds led to a better match to the higher order measures and a marked effect on the prediction of trapping thresholds. Overall the results illustrate the importance of network topology on the accurate prediction of trapping phase thresholds.

Influence of Pore Geometry on IP Properties The resultant percolation saturations in IP are strongly dependent on the geometric properties of rock, in particular the distributions of sizes and shapes of pores [30,31,71]. For an uncorrelated system in which random numbers are assigned to each site/bond, we would expect for IP that most small/large random numbers are accepted as part of the invading cluster, while the large/small numbers remain part of the defending phase at the final saturation. Unlike OP, where one observes a sharp acceptance profile, in TIP there is a transition region [71]. From the acceptance profile and the distribution of pore volumes one can estimate the trapped saturation for various pore size distributions.

Influence of Pore Scale Geometric Correlations on IP Properties One can further illustrate the effect of local pore geometry/correlation in real rock structures by randomly rearranging pore volumes on the rock network structure. By doing this one is preserving the topology of the network; only considering the effect of randomizing the pore bodies. The effect of this rearrangement was illustrated for a set of Fontainebleau sandstone samples in [56]. The site-based trapping thresholds were $STIP < 0.25$ for the actual rock samples, while the rearranged sample exhibited trapping thresholds $STIP > 0.45$. These differences are con-

sistent with the presence of local correlated heterogeneity in the pore network at the smallest scales. A method to measure and introduce the spatial heterogeneity observed in the rock structure at a pore scale is lacking; at present the best representations of rock microstructure would be based on true 3D realizations; either from reconstruction method [5] or from direct measurement of 3D structure via X-ray computed tomography [15,16,57].

Effect of Correlated Heterogeneity on IP

In this section we consider the application of IP to the description of multiphase fluid flow in porous media at scales larger than individual pores. Early work with network models and invasion percolation concentrated on macroscopically homogeneous porous media built using independently generated random numbers to assign pore and throat sizes. Correspondingly most of the testing of the models was conducted on laboratory porous media made from relatively uniform glass spheres or from uniformly etched glass plates. However it became apparent when strategies for improved oil recovery were attempted in the field that recoveries did not match expectations, largely because of heterogeneity at all scales. This led to studies of heterogeneity and evidence that long-range correlations in properties down to the centimeter [46] and the pore scale [26,73] exist in many, if not most, porous sedimentary rocks.

Measurement of Heterogeneity

The earliest work that considered correlated substrates used two-dimensional multifractal lattices as a means of incorporating heterogeneity [39,40].

Subsequently there have been many studies that have improved understanding of the long-range correlations in rock properties that exist from the pore scale to the kilometer scale. To describe these correlations, fractional Brownian motion (fBm) is one of the models that has been used as a model for the underlying reservoir heterogeneity [44]. Fractional Brownian motion is straightforward to generate, hence it has been an obvious candidate for studies of percolation in correlated property maps [3,13,14,48].

Invasion percolation with trapping on substrates with a spatially correlated threshold distribution resulting from the mapping of a self-affine surface has also been studied by Wagner et al. [67].

Analysis of sandstones [27] on rock samples of a few centimeters in extent suggests that correlated heterogeneity exists down to the pore scale and that at this scale correlations persist at scales up to several pore lengths. For rock samples at these scales a more appropriate model of corre-

lated heterogeneity is one which introduces a cutoff length scale below which correlations persist and above which the system behaves like a random material.

Effect of Heterogeneity on IP

Fractal Dimension The effect of the correlated heterogeneity measured in naturally occurring sedimentary rocks on IP was first studied in 2D [40,48,67]; the fluid displacement patterns for lattices with long range correlations based on multifractal, fBm and fLm models were compared to uncorrelated networks. The cluster fractal dimension for NTIP was found to be compact for almost all systems exhibiting correlated heterogeneity while TIP exhibited fractal saturation patterns for multifractal systems and fBm models with a Hurst exponent $H < 0.5$. In all studies with fBm model correlations the cluster fractal dimension was observed to increase with $H < 0.5$; for $H > 0.5$ all clusters were compact. This result was verified later [28] on large lattices for fBm correlated lattices. Similar results were obtained on 3D lattices [28]. Other fractal dimensions were also studied; for $H < 0.5$ the fractal dimensions of the hull, minimal path, backbone and invasion front are all fractal with dimensions that depend on H [3,28,40].

The effect of correlated heterogeneity with a finite cutoff length scale for the extent of correlations (correlated heterogeneity at a smaller scale and uncorrelated above some cutoff length scale) has also been considered [28]. In these cases clusters were fractal at length scale above the cutoff length with fractal dimensions that are the same as those in normal IP models. For smaller length scales the clusters' structures are similar to those observed for fully correlated lattices.

Thresholds Studies in two dimensions have considered the effect of correlated heterogeneity on breakthrough saturations [3,14,40] – results on finite lattices showed that correlation has a significant effect on the percolation threshold and that the threshold is no longer unique but depends on the spanning rule employed [38]. Residual saturations S_r in 3D systems found that the introduction of correlation leads to a large reduction in residual saturation with increasing H [29]. For uncorrelated cubic lattices $S_r = 0.34$, while $S_r = 0.25, 0.22$ and 0.18 for fBm correlated lattices with $H = 0.2, 0.5$ and 0.8 respectively. Correlations with a finite cutoff also led to changes in S_r ; in particular the residual saturation was found to exhibit a minimal value for finite cutoff lengths; the increase in saturations at large cutoff length scales was due to the possibility of trapping very large regions of the defend-

ing fluid at larger cutoffs. Small scale correlations therefore have a profound effect on resultant residuals, even at large scales. Analysis of the scaling behavior of the variance of residual saturations shows that the measurements of S_r must be made on samples that are at least ten times larger than the extent of correlated heterogeneity. Measurements on porous rock samples with extensive correlations would lead to a wide variety of S_r being measured. This has been observed experimentally [49].

Trapped Cluster Distribution Introduction of correlated heterogeneity has a strong effect on the resultant distribution of clusters of the trapped defending fluid [29]. For small-scale correlations one observes the presence of larger clusters of trapped phase. For large scale correlations one or two trapped clusters can account for much of the trapped defending phase saturation. This has important implications to recovery of fluids from rocks under tertiary displacements where a third phase is injected to further reduce the saturation of the defending fluid. The presence of larger residual clusters may make it easier to reconnect and recover the trapped phase.

Stratification and Relative Permeability Extending the work on substrates generated from fractional Brownian motion, Paterson et al. [49] simulated invasion percolation with anisotropic correlations that were introduced to simulate the stratification often apparent in sedimen-

tary rocks. The anisotropy was created by compressing the property maps in one direction and then deleting layers so the the grid spacing in each direction is equal. In their study invasion percolation was being used as a step to calculate relative permeability in porous media. The anisotropic heterogeneity led to different sets of relative permeability curves parallel and perpendicular to the bedding direction. This was consistent with experimental observations that relative permeability at a given saturation is greater for flow parallel to the bedding when compared with flow perpendicular to the bedding.

IP in Fractured Systems

Fractured systems differ from granular porous media in that representation by a network of pores (sites) connected via throats (bonds) disappears. Instead a fracture can be represented by a single aperture of varying thickness. Nevertheless, surface roughness within the fracture creates variable capillary pressure in two-phase flow, so it is still possible to apply the invasion percolation concept.

To model flow in fractal fractures, Wagner et al. [65, 68] used invasion percolation with trapping on two-dimensional substrates with a correlated distribution of invasion thresholds. Hence their simulations are essentially two-dimensional porous-media flow with heterogeneity. They simulated displacement in the void space between one fBm surface and one plane surface, and studied how

Invasion Percolation, Algorithm 1

Mathematica algorithm for generating invasion percolation clusters from a single starting cell. Code is from [20]. In Mathematica version 6 `Table[Random[], {4}]` can be replaced with `RandomReal[{0, 1}, 4]`

Invasion Percolation, Algorithm 2

Mathematica code that can be used to plot clusters generated from the `invasion[]` function

fractal scaling behavior depends on the surface roughness. Simulations on fractures consisting of a fBm surface and its displaced replica displayed a cross-over phenomenon. For displacements longer than the correlation length the flow patterns were found to have the properties of ordinary IP clusters grown on uncorrelated substrates.

In a separate line of investigation, Glass et al. [21] used invasion percolation to model experiments in a synthetic horizontal fracture made from two plates of roughened glass in contact. The aperture field for this fracture was measured using a light absorption technique. To apply invasion percolation they analyzed the correct curvature to use for the fluid invasion steps. A subsequent study by Glass et al. [22] developed simulations for a fracture patterned on measured data from a block of welded tuff.

Invasion percolation simulations were also matched to experiments by Amundsen et al. [1]. Their experiments involved two different rough bottom plates with a smooth planar top plate. One of the bottom plates was plastic milled to a self-affine surface with a Hurst exponent of 0.8. The other bottom plate was textured glass measured using a light absorption technique like Glass et al. [21].

Future Directions

The simple invasion percolation model provides a very realistic simulation of the slow fluid-fluid displacement processes within porous materials. The utility of the model and its variants to important applications are numerous. They include the understanding of contaminant migration in soils crucial to the successful implementation of groundwater remediation strategies. The ability to predict the transport of contaminants in soils will impact on the understanding of water quality issues. After primary oil recovery, more than 50% of the original oil in place remains unrecovered; a significant volume fraction of the pore space occupied by oil and gas is unrecovered because it is bypassed in the rock by the combined effects of the natural water drive mechanism, capillary forces and rock

heterogeneity. Realistic estimation of recoveries is a central problem in the development of new fields and in the development of improved oil recovery methods in existing fields. The further development of the invasion percolation model and its variants coupled with a more realistic structural characterization of the pore structure of porous materials will play a crucial role in the development of an understanding of these important problems.

Bibliography

1. Amundsen H, Wagner G, Oxaal U, Meakin P, Feder J, Jøssang T (1999) Slow two-phase flow in artificial fractures: Experiments and simulations. *Water Resour Res* 35:2619–2626
2. Arns CH, Sakellariou A, Senden TJ, Sheppard AP, Sok RM, Pinczewski WV, Knackstedt MA (2005) Digital core laboratory: Petrophysical analysis from 3D images. *Petrophysics* 46(4):260–277
3. Babadagli T (2000) Invasion percolation in correlated porous media. *Physica A* 285:248–258
4. Babalievski F (1998) Cluster counting: The Hoshen–Kopelman algorithm vs. spanning tree approaches. *Int J Mod Phys C* 9:43–60
5. Bakke S, Øren P (1997) 3-D pore-scale modelling of sandstones and flow simulations in the pore networks. *SPE J* 2:136–149
6. Blunt MJ (1997) Pore level modeling of the effects of wettability. *SPE J* 2:494–510
7. Blunt MJ, Scher H (1995) Pore-level model of wetting. *Phys Rev E* 52:6387–6403
8. Blunt MJ, Jackson MD, Piri M, Valvatne PH (2002) Detailed physics, predictive capabilities and macroscopic consequences for pore-network models of multiphase flow. *Adv Water Resour* 25:1069–1089
9. Chandler R, Koplik J, Lerman K, Willemsen J (1982) Capillary displacement and percolation in porous media. *J Fluid Mech* 119:249–267
10. Chen JD, Wada N (1986) Visualisation of immiscible displacement in a three dimensional transparent porous medium. *Exp Fluids* 4:336–338
11. Cieplak M, Maritan A, Banavar JR (1994) Optimal paths and domain walls in the strong disorder limit. *Phys Rev Lett* 72:2320–2323
12. Dobrin R, Duxbury P (2001) Minimum spanning trees on random networks. *Phys Rev Lett* 86:5076–5079

13. Du C, Xu B, Yortsos YC, Chaouche M, Rakotomalala N, Salin D (1995) Correlation of occupation profiles in invasion percolation. *Phys Rev Lett* 74:694–697
14. Du C, Satik C, Yortsos Y (1996) Percolation in a fractional Brownian motion lattice. *AIChE J* 42:2392–2394
15. Dunsmuir JH, Ferguson SR, D'Amico KL (1991) Design and operation of an imaging X-ray detector for microtomography. *IOP Conf Ser* 121:257–261
16. Flannery BP, Deckman HW, Roberge WG, D'Amico KL (1987) Three-dimensional X-ray microtomography. *Science* 237:1439–1444
17. Franzese G, Cataudella V, Coniglio A (1998) Invaded cluster dynamics for frustrated models. *Phys Rev E* 57:88–93
18. Fredrich J, Menendez B, Wong TF (1995) Imaging the pore structure of geomaterials. *Science* 268:276–279
19. Galam S, Mauger A (1996) Universal formulas for percolation thresholds. *Phys Rev E* 53:2177–2180
20. Gaylord RJ, Wellin PR (1994) Computer simulations with mathematics: Explorations in complex physical and biological systems. *TELOS/Springer*, New York
21. Glass RJ, Nicholl MJ, Yarrington L (1998) A modified invasion percolation model for low-capillary number immiscible displacements in horizontal rough-walled fractures: Influence of local in-plane curvature. *Water Resour Res* 34(12):3215–3234
22. Glass RJ, Nicholl MJ, Rajaram H, Andre B (2004) Development of slender transport pathways in unsaturated fractured rock: Simulation with modified invasion percolation. *Geophys Res Lett* 31:L06502
23. Heiba A, Sahimi M, Scriven L, Davis H (1992) Percolation theory of two-phase relative permeability. *SPE Reserv Engin* 7:123–132
24. Hoshen J, Kopelman R (1976) Percolation and cluster sizes. *Phys Rev B* 14:3438
25. Ioannidis MA, Kwiczen MJ, Chatzis I, MacDonald IF, Dullien FAL (1997) Comprehensive pore structure characterization using 3D computer reconstruction and stochastic modeling. In: *SPE Annual Technical Conference and Exhibition held in San Antonio, Texas, USA, 1997*
26. Knackstedt MA, Sheppard AP, Pinczewski WV (1998) Simulation of mercury porosimetry on correlated grids: Evidence for extended correlated heterogeneity at the pore scale in rocks. *Phys Rev E* 58:6923–6926
27. Knackstedt MA, Sheppard AP, Pinczewski WV (1998) Simulation of mercury porosimetry on correlated grids: Evidence for extended correlated heterogeneity at the pore scale in rocks. *Phys Rev E Rapid Communications* 58:R6923–R6926
28. Knackstedt MA, Sahimi M, Sheppard AP (2000) Invasion percolation with long-range correlations: First-order phase transition and nonuniversal scaling properties. *Phys Rev E* 61:4920–4934
29. Knackstedt MA, Marrink S, Sheppard AP, Pinczewski W, Sahimi M (2001) Invasion percolation on correlated and elongated lattices: Implications for the interpretation of residual saturations in rock cores. *Transp Porous Media* 44:465–485
30. Larson R, Scriven LE, Davis HT (1977) Percolation theory of residual phases in porous media. *Nature* 268:409–413
31. Larson R, Scriven LE, Davis HT (1991) Percolation theory of two-phase flow in porous media. *Chem Eng Sci* 36:57–73
32. Lee J-Y, Robins V, Sok RM, Sheppard AP, Pinczewski W, Knackstedt MA (2004) Effect of topology on relative permeability. *Transp Porous Media* 55:21–46
33. Lenormand R, Bories S (1980) Description d'un mecanisme de connexion de liaison destin l'tude du drainage avec pigeage en milieu poreux. *CR Acad Sci Paris B* 291:279
34. Lenormand R, Bories S (1985) Fractal patterns from chemical dissolution. *Physicochem Hydro* 6:497
35. Lenormand R, Zarcane C (1985) Invasion percolation in an etched network; measurement of a fractal dimension. *Phys Rev Lett* 54:2226–2229
36. Lindquist B, Lee SM, Coker D (1996) Medial axis analysis of void structure in three-dimensional tomographic images of porous media. *J Geophys Res* 101B:8297–8310
37. Lindquist WB, Venkatarangan A, Dunsmuir J, Wong TF (2000) Pore and throat size distributions measured from synchrotron X-ray tomographic images of fontainebleau sandstones. *J Geophys Res* 105B:21508
38. Marrink SJ, Paterson L, Knackstedt M (2000) Definition of percolation thresholds on self-affine surfaces. *Physica A* 280:207–214
39. Meakin P (1988) Invasion percolation and invading Eden growth on multifractal lattices. *J Phys A: Math Gen* 21:3501–3522
40. Meakin P (1991) Invasion percolation on substrates with correlated disorder. *Physica A* 173:305–324
41. Meakin P, Feder J, Frette V, Jøssang T (1992) Invasion percolation in a destabilizing gradient. *Phys Rev A* 46:3357–3368
42. Meakin P, Wagner G, Frette V, Feder J, Jøssang T (1995) Fractals and secondary migration. *Fractals* 3:799–806
43. Meakin P, Wagner G, Vedvik A, Amundsen H, Feder J, Jøssang T (2000) Invasion percolation and secondary migration: experiments and simulations. *Mar Pet Geol* 17:777–795
44. Molz FJ, Liu HH, Szulga J (1997) Fractional Brownian motion and fractional Gaussian noise in subsurface hydrology: A review, presentation of fundamental properties, and extensions. *Water Resour Res* 33:2273–2286
45. Øren P, Bakke S, Arntzen OJ (1998) Extending predictive capabilities to network models. *SPE J* 3:324–336
46. Painter S (2001) Flexible scaling model for use in random field simulation of hydraulic conductivity. *Water Resour Res* 37:1155–1163
47. Paterson L (1998) Trapping thresholds in ordinary percolation. *Phys Rev E* 58:7137–7140
48. Paterson L, Painter S, Knackstedt MA, Pinczewski WV (1996) Patterns of fluid flow in naturally heterogeneous rocks. *Physica A* 233:619–628
49. Paterson L, Painter S, Zhang X, Pinczewski WV (1998) Simulating residual saturation and relative permeability in heterogeneous formations. *SPE J* 3:211–218
50. Paterson L, Sheppard AP, Knackstedt MA (2002) Trapping thresholds in invasion percolation. *Phys Rev E* 66:056122
51. Porto M, Havlin S, Schwarzer S, Bunde A (1997) Optimal path in strong disorder and shortest path in invasion percolation with trapping. *Phys Rev Lett* 79:4060–4063
52. Sahimi M (1994) *Applications of percolation theory*, 1st edn. Taylor Francis, London
53. Sheppard AP, Knackstedt MA, Pinczewski WV, Sahimi M (1999) Invasion percolation: New algorithms and universality classes. *J Phys A Lett* 32:L521–L529
54. Sheppard AP, Knackstedt MA, Pinczewski WV, Sahimi M (1999) Invasion percolation: New algorithms and universality classes. *J Phys A: Math Gen* 32:L521–L529
55. Sheppard AP, Sok RM, Averdunk H (2005) Improved pore net-

- work extraction methods. 19th International Symposium of the SCA, SCA, Toronto, 2005
56. Sok RM, Knackstedt MA, Sheppard AP, Pinczewski W, Lindquist WB, Venkatarangan A, Paterson L (2002) Direct and stochastic generation of network models from tomographic images; effect of topology on two phase flow properties. *Transp Porous Media* 46:345–372
 57. Spanne P, Thovert J, Jacquin J, Lindquist WB, Jones K, Adler PM (1994) Synchrotron computed microtomography of porous media: Topology and transports. *Phys Rev Lett* 73:2001–2004
 58. Stark C (1991) An invasion percolation model of drainage network evolution. *Nature* 352:423
 59. Stauffer D, Aharony A (1994) Introduction to percolation theory, 2nd edn. Taylor Francis, London
 60. Stokes JP, Weitz D, Gollub J, Dougherty A, Robbins M, Chaikin P, Lindsay H (1986) Interfacial stability of immiscible displacement in a porous medium. *Phys Rev Lett* 57:2226–2229
 61. Suding PN, Ziff RM (1999) Site percolation thresholds for archimedean lattices. *Phys Rev E* 60:275–283
 62. Thovert J-F, Salles J, Adler P (1993) Computerised characterization of the geometry of real porous media: Their description, analysis and interpretation. *J Microsc* 170:65–79
 63. Tsimpanogiannis IN, Yortsos YC (2004) The critical gas saturation in a porous medium in the presence of gravity. *J Colloid Interface Sci* 270:388–395
 64. Vedvik A, Wagner G, Oxaal U, Feder J, Meakin P, Jøssang T (1998) Fragmentation transition for invasion percolation in hydraulic gradients. *Phys Rev Lett* 80:3065–3068
 65. Wagner G, Amundsen H, Oxaal U, Meakin P, Feder J, Jøssang T (2000) Slow two-phase flow in single fractures: Fragmentation, migration, and fractal patterns simulated using invasion percolation models. *Pure Appl Geophys* 157:621–635
 66. Wagner G, Meakin P, Feder J, Jøssang T (1997) Buoyancy-driven invasion percolation with migration and fragmentation. *Physica A* 245:217–230
 67. Wagner G, Meakin P, Feder J, Jøssang T (1997) Invasion percolation on self-affine topographies. *Phys Rev E* 55:1698–1703
 68. Wagner G, Meakin P, Feder J, Jøssang T (1999) Invasion percolation in fractal fractures. *Physica A* 264:321–337
 69. Wilkinson D (1984) Percolation model of immiscible displacement in the presence of buoyancy forces. *Phys Rev A* 30:520–531
 70. Wilkinson D (1986) Percolation effects in immiscible displacement. *Phys Rev A* 34:1380–1391
 71. Wilkinson D, Willemsen JF (1983) Invasion percolation: A new form of percolation theory. *J Phys A: Math Gen* 16:3365–3376
 72. Xu B, Yortsos YC, Salin D (1998) Invasion percolation with viscous forces. *Phys Rev E* 57:739–751
 73. Yuan H (1991) Pore-scale heterogeneity from mercury porosimetry data. *SPE Form Eval* 6:233–242

Inverse Scattering Transform and the Theory of Solitons

TUNCAY AKTOSUN

Department of Mathematics, University of Texas
at Arlington, Arlington, USA

Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Inverse Scattering Transform](#)

[The Lax Method](#)

[The AKNS Method](#)

[Direct Scattering Problem](#)

[Time Evolution of the Scattering Data](#)

[Inverse Scattering Problem](#)

[Solitons](#)

[Future Directions](#)

[Bibliography](#)

Glossary

AKNS method A method introduced by Ablowitz, Kaup, Newell, and Segur in 1973 that identifies the nonlinear partial differential equation (NPDE) associated with a given first-order system of linear ordinary differential equations (LODEs) so that the initial value problem (IVP) for that NPDE can be solved by the inverse scattering transform (IST) method.

Direct scattering problem The problem of determining the scattering data corresponding to a given potential in a differential equation.

Integrability A NPDE is said to be integrable if its IVP can be solved via an IST.

Inverse scattering problem The problem of determining the potential that corresponds to a given set of scattering data in a differential equation.

Inverse scattering transform A method introduced in 1967 by Gardner, Greene, Kruskal, and Miura that yields a solution to the IVP for a NPDE with the help of the solutions to the direct and inverse scattering problems for an associated LODE.

Lax method A method introduced by Lax in 1968 that determines the integrable NPDE associated with a given LODE so that the IVP for that NPDE can be solved with the help of an IST.

Scattering data The scattering data associated with a LODE usually consists of a reflection coefficient which is a function of the spectral parameter λ , a finite number of constants λ_j that correspond to the poles of the transmission coefficient in the upper half complex plane, and the bound-state norming constants whose number for each bound-state pole λ_j is the same as the order of that pole. It is desirable that the potential in the LODE is uniquely determined by the corresponding scattering data and vice versa.

Soliton The part of a solution to an integrable NPDE due to a pole of the transmission coefficient in the upper half complex plane. The term soliton was introduced by Zabusky and Kruskal in 1965 to denote a solitary wave pulse with a particle-like behavior in the solution to the Korteweg-de Vries (KdV) equation.

Time evolution of the scattering data The evolution of the scattering data from its initial value $S(\lambda, 0)$ at $t = 0$ to its value $S(\lambda, t)$ at a later time t .

Definition of the Subject

A general theory to solve NPDEs does not seem to exist. However, there are certain NPDEs, usually first order in time, for which the corresponding IVPs can be solved by the IST method. Such NPDEs are sometimes referred to as integrable evolution equations. Some exact solutions to such equations may be available in terms of elementary functions, and such solutions are important to understand nonlinearity better and they may also be useful in testing accuracy of numerical methods to solve such NPDEs.

Certain special solutions to some of such NPDEs exhibit particle-like behaviors. A single-soliton solution is usually a localized disturbance that retains its shape but only changes its location in time. A multi-soliton solution consists of several solitons that interact nonlinearly when they are close to each other but come out of such interactions unchanged in shape except for a phase shift.

Integrable NPDEs have important physical applications. For example, the KdV equation is used to describe [14,23] surface water waves in long, narrow, shallow canals; it also arises [23] in the description of hydro-magnetic waves in a cold plasma, and ion-acoustic waves in anharmonic crystals. The nonlinear Schrödinger (NLS) equation arises in modeling [24] electromagnetic waves in optical fibers as well as surface waves in deep waters. The sine-Gordon equation is helpful [1] in analyzing the magnetic field in a Josephson junction (gap between two superconductors).

Introduction

The first observation of a soliton was made in 1834 by the Scottish engineer John Scott Russell at the Union Canal between Edinburgh and Glasgow. Russell reported [21] his observation to the British Association of the Advancement of Science in September 1844, but he did not seem to be successful in convincing the scientific community. For example, his contemporary George Airy, the influential mathematician of the time, did not believe in the existence of solitary water waves [1].

The Dutch mathematician Korteweg and his doctoral student de Vries published [14] a paper in 1895 based on de Vries' Ph.D. dissertation, in which surface waves in shallow, narrow canals were modeled by what is now known as the KdV equation. The importance of this paper was not understood until 1965 even though it contained as a special solution what is now known as the one-soliton solution.

Enrico Fermi in his summer visits to the Los Alamos National Laboratory, together with J. Pasta and S. Ulam, used the computer named Maniac I to computationally analyze a one-dimensional dynamical system of 64 particles in which adjacent particles were joined by springs where the forces also included some nonlinear terms. Their main goal was to determine the rate of approach to the equipartition of energy among different modes of the system. Contrary to their expectations there was little tendency towards the equipartition of energy but instead the almost ongoing recurrence to the initial state, which was puzzling. After Fermi died in November 1954, Pasta and Ulam completed their last few computational examples and finished writing a preprint [11], which was never published as a journal article. This preprint appears in Fermi's Collected Papers [10] and is also available on the internet [25].

In 1965 Zabusky and Kruskal explained [23] the Fermi-Pasta-Ulam puzzle in terms of solitary wave solutions to the KdV equation. In their numerical analysis they observed "solitary-wave pulses", named such pulses "solitons" because of their particle-like behavior, and noted that such pulses interact with each other nonlinearly but come out of interactions unaffected in size or shape except for some phase shifts. Such unusual interactions among solitons generated a lot of excitement, but at that time no one knew how to solve the IVP for the KdV equation, except numerically. In 1967 Gardner, Greene, Kruskal, and Miura presented [12] a method, now known as the IST, to solve that IVP, assuming that the initial profile $u(x, 0)$ decays to zero sufficiently rapidly as $x \rightarrow \pm\infty$. They showed that the integrable NPDE, i. e. the KdV equation,

$$u_t - 6uu_x + u_{xxx} = 0, \quad (1)$$

is associated with a LODE, i. e. the 1-D Schrödinger equation

$$-\frac{d^2\psi}{dx^2} + u(x, t)\psi = k^2\psi, \quad (2)$$

and that the solution $u(x, t)$ to (1) can be recovered from the initial profile $u(x, 0)$ as explained in the diagram given in Sect. "Inverse Scattering Transform". They also ex-

plained that soliton solutions to the KdV equation correspond to a zero reflection coefficient in the associated scattering data. Note that the subscripts x and t in (1) and throughout denote the partial derivatives with respect to those variables.

In 1972 Zakharov and Shabat showed [24] that the IST method is applicable also to the IVP for the NLS equation

$$iu_t + u_{xx} + 2u|u|^2 = 0, \quad (3)$$

where i denotes the imaginary number $\sqrt{-1}$. They proved that the associated LODE is the first-order linear system

$$\begin{cases} \frac{d\xi}{dx} = -i\lambda\xi + u(x, t)\eta, \\ \frac{d\eta}{dx} = i\lambda\eta - \overline{u(x, t)}\xi, \end{cases} \quad (4)$$

where λ is the spectral parameter and an overline denotes complex conjugation. The system in (4) is now known as the Zakharov–Shabat system.

Soon afterwards, again in 1972 Wadati showed in a one-page publication [22] that the IVP for the modified Korteweg–de Vries (mKdV) equation

$$u_t + 6u^2u_x + u_{xxx} = 0, \quad (5)$$

can be solved with the help of the inverse scattering problem for the linear system

$$\begin{cases} \frac{d\xi}{dx} = -i\lambda\xi + u(x, t)\eta, \\ \frac{d\eta}{dx} = i\lambda\eta - u(x, t)\xi. \end{cases} \quad (6)$$

Next, in 1973 Ablowitz, Kaup, Newell, and Segur showed [2,3] that the IVP for the sine-Gordon equation

$$u_{xt} = \sin u,$$

can be solved in the same way by exploiting the inverse scattering problem associated with the linear system

$$\begin{cases} \frac{d\xi}{dx} = -i\lambda\xi - \frac{1}{2}u_x(x, t)\eta, \\ \frac{d\eta}{dx} = i\lambda\eta + \frac{1}{2}u_x(x, t)\xi. \end{cases}$$

Since then, many other NPDEs have been discovered to be solvable by the IST method.

Our review is organized as follows. In the next section we explain the idea behind the IST. Given a LODE known to be associated with an integrable NPDE, there are two primary methods enabling us to determine the

corresponding NPDE. We review those two methods, the Lax method and the AKNS method, in Sect. “The Lax Method” and in Sect. “The AKNS Method”, respectively. In Sect. “Direct Scattering” we introduce the scattering data associated with a LODE containing a spectral parameter and a potential, and we illustrate it for the Schrödinger equation and for the Zakharov–Shabat system. In Sect. “Time Evolution of the Scattering Data” we explain the time evolution of the scattering data and indicate how the scattering data sets evolve for those two LODEs. In Sect. “Inverse Scattering Problem” we summarize the Marchenko method to solve the inverse scattering problem for the Schrödinger equation and that for the Zakharov–Shabat system, and we outline how the solutions to the IVPs for the KdV equation and the NLS equation are obtained with the help of the IST. In Sect. “Solitons” we present soliton solutions to the KdV and NLS equations. A brief conclusion is provided in Sect. “Future Directions”.

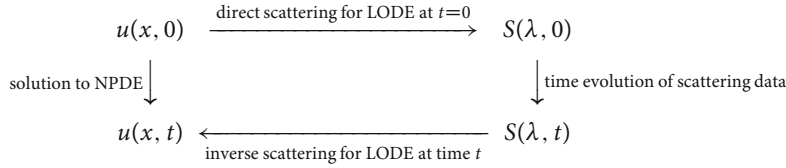
Inverse Scattering Transform

Certain NPDEs are classified as integrable in the sense that their corresponding IVPs can be solved with the help of an IST. The idea behind the IST method is as follows: Each integrable NPDE is associated with a LODE (or a system of LODEs) containing a parameter λ (usually known as the spectral parameter), and the solution $u(x, t)$ to the NPDE appears as a coefficient (usually known as the potential) in the corresponding LODE. In the NPDE the quantities x and t appear as independent variables (usually known as the spatial and temporal coordinates, respectively), and in the LODE x is an independent variable and λ and t appear as parameters. It is usually the case that $u(x, t)$ vanishes at each fixed t as x becomes infinite so that a scattering scenario can be created for the related LODE, in which the potential $u(x, t)$ can uniquely be associated with some scattering data $S(\lambda, t)$. The problem of determining $S(\lambda, t)$ for all λ values from $u(x, t)$ given for all x values is known as the direct scattering problem for the LODE. On the other hand, the problem of determining $u(x, t)$ from $S(\lambda, t)$ is known as the inverse scattering problem for that LODE.

The IST method for an integrable NPDE can be explained with the help of Diagram 1.

In order to solve the IVP for the NPDE, i. e. in order to determine $u(x, t)$ from $u(x, 0)$, one needs to perform the following three steps:

- (i) Solve the corresponding direct scattering problem for the associated LODE at $t = 0$, i. e. determine the initial scattering data $S(\lambda, 0)$ from the initial potential $u(x, 0)$.



Inverse Scattering Transform and the Theory of Solitons, Diagram 1
The method of inverse scattering transform

- (ii) Time evolve the scattering data from its initial value $S(\lambda, 0)$ to its value $S(\lambda, t)$ at time t . Such an evolution is usually a simple one and is particular to each integrable NPDE.
- (iii) Solve the corresponding inverse scattering problem for the associated LODE at fixed t , i. e. determine the potential $u(x, t)$ from the scattering data $S(\lambda, t)$.

It is amazing that the resulting $u(x, t)$ satisfies the integrable NPDE and that the limiting value of $u(x, t)$ as $t \rightarrow 0$ agrees with the initial profile $u(x, 0)$.

The Lax Method

In 1968 Peter Lax introduced [15] a method yielding an integrable NPDE corresponding to a given LODE. The basic idea behind the Lax method is the following. Given a linear differential operator \mathcal{L} appearing in the spectral problem $\mathcal{L}\psi = \lambda\psi$, find an operator \mathcal{A} (the operators \mathcal{A} and \mathcal{L} are said to form a Lax pair) such that:

- (i) The spectral parameter λ does not change in time, i. e. $\lambda_t = 0$.
- (ii) The quantity $\psi_t - \mathcal{A}\psi$ remains a solution to the same linear problem $\mathcal{L}\psi = \lambda\psi$.
- (iii) The quantity $\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L}$ is a multiplication operator, i. e. it is not a differential operator.

From condition (ii) we get

$$\mathcal{L}(\psi_t - \mathcal{A}\psi) = \lambda(\psi_t - \mathcal{A}\psi), \quad (7)$$

and with the help of $\mathcal{L}\psi = \lambda\psi$ and $\lambda_t = 0$, from (7) we obtain

$$\begin{aligned}
 \mathcal{L}\psi_t - \mathcal{L}\mathcal{A}\psi &= \lambda\psi_t - \mathcal{A}(\lambda\psi) = \partial_t(\lambda\psi) - \mathcal{A}\mathcal{L}\psi \\
 &= \partial_t(\mathcal{L}\psi) - \mathcal{A}\mathcal{L}\psi = \mathcal{L}_t\psi + \mathcal{L}\psi_t - \mathcal{A}\mathcal{L}\psi, \quad (8)
 \end{aligned}$$

where ∂_t denotes the partial differential operator with respect to t . After canceling the term $\mathcal{L}\psi_t$ on the left and right hand sides of (8), we get

$$(\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L})\psi = 0,$$

which, because of (iii), yields

$$\mathcal{L}_t + \mathcal{L}\mathcal{A} - \mathcal{A}\mathcal{L} = 0. \quad (9)$$

Note that (9) is an evolution equation containing a first-order time derivative, and it is the desired integrable NPDE. The equation in (9) is often called a compatibility condition.

Having outlined the Lax method, let us now illustrate it to derive the KdV equation in (1) from the Schrödinger equation in (2). For this purpose, we write the Schrödinger equation as $\mathcal{L}\psi = \lambda\psi$ with $\lambda := k^2$ and

$$\mathcal{L} := -\partial_x^2 + u(x, t), \quad (10)$$

where the notation $:=$ is used to indicate a definition so that the quantity on the left should be understood as the quantity on the right hand side. Given the linear differential operator \mathcal{L} defined as in (10), let us try to determine the associated operator \mathcal{A} by assuming that it has the form

$$\mathcal{A} = \alpha_3 \partial_x^3 + \alpha_2 \partial_x^2 + \alpha_1 \partial_x + \alpha_0, \quad (11)$$

where the coefficients α_j with $j = 0, 1, 2, 3$ may depend on x and t , but not on the spectral parameter λ . Note that $\mathcal{L}_t = u_t$. Using (10) and (11) in (9), we obtain

$$(\) \partial_x^5 + (\) \partial_x^4 + (\) \partial_x^3 + (\) \partial_x^2 + (\) \partial_x + (\) = 0, \quad (12)$$

where, because of (iii), each coefficient denoted by $(\)$ must vanish. The coefficient of ∂_x^5 vanishes automatically. Setting the coefficients of ∂_x^j to zero for $j = 4, 3, 2, 1$, we obtain

$$\begin{aligned}
 \alpha_3 &= c_1, \quad \alpha_2 = c_2, \quad \alpha_1 = c_3 - \frac{3}{2}c_1 u, \\
 \alpha_0 &= c_4 - \frac{3}{4}c_1 u_x - c_2 u,
 \end{aligned}$$

with c_1, c_2, c_3 , and c_4 denoting arbitrary constants. Choosing $c_1 = -4$ and $c_3 = 0$ in the last coefficient in (12) and setting that coefficient to zero, we get the KdV equation in (1). Moreover, by letting $c_2 = c_4 = 0$, we obtain the operator \mathcal{A} as

$$\mathcal{A} = -4\partial_x^3 + 6u\partial_x + 3u_x. \quad (13)$$

For the Zakharov–Shabat system in (4), we proceed in a similar way. Let us write it as $\mathcal{L}\psi = \lambda\psi$, where the linear differential operator \mathcal{L} is defined via

$$\mathcal{L} := i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \partial_x - i \begin{bmatrix} 0 & u(x, t) \\ u(x, t) & 0 \end{bmatrix}.$$

Then, the operator \mathcal{A} is obtained as

$$\mathcal{A} = 2i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \partial_x^2 - 2i \begin{bmatrix} 0 & u \\ \bar{u} & 0 \end{bmatrix} \partial_x - i \begin{bmatrix} -|u|^2 & u_x \\ \bar{u}_x & |u|^2 \end{bmatrix}, \quad (14)$$

and the compatibility condition (9) gives us the NLS equation in (3).

For the first-order system (6), by writing it as $\mathcal{L}\psi = \lambda\psi$, where the linear operator \mathcal{L} is defined by

$$\mathcal{L} := i \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \partial_x - i \begin{bmatrix} 0 & u(x, t) \\ u(x, t) & 0 \end{bmatrix},$$

we obtain the corresponding operator \mathcal{A} as

$$\mathcal{A} = -4 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \partial_x^3 - 6 \begin{bmatrix} u^2 & -u_x \\ u_x & u^2 \end{bmatrix} \partial_x - \begin{bmatrix} 6uu_x & -3u_{xx} \\ 3u_{xx} & 6uu_x \end{bmatrix},$$

and the compatibility condition (9) yields the mKdV equation in (5).

The AKNS Method

In 1973 Ablowitz, Kaup, Newell, and Segur introduced [2,3] another method to determine an integrable NPDE corresponding to a LODE. This method is now known as the AKNS method, and the basic idea behind it is the following. Given a linear operator X associated with the first-order system $v_x = Xv$, we are interested in finding an operator \mathcal{T} (the operators \mathcal{T} and X are said to form an AKNS pair) such that:

- (i) The spectral parameter λ does not change in time, i. e. $\lambda_t = 0$.
- (ii) The quantity $v_t - \mathcal{T}v$ is also a solution to $v_x = Xv$, i. e. we have $(v_t - \mathcal{T}v)_x = X(v_t - \mathcal{T}v)$.
- (iii) The quantity $X_t - \mathcal{T}_x + X\mathcal{T} - \mathcal{T}X$ is a (matrix) multiplication operator, i. e. it is not a differential operator.

From condition (ii) we get

$$\begin{aligned} v_{tx} - \mathcal{T}_x v - \mathcal{T}v_x &= Xv_t - X\mathcal{T}v \\ &= (Xv)_t - X_t v - X\mathcal{T}v \\ &= (v_x)_t - X_t v - X\mathcal{T}v \\ &= v_{xt} - X_t v - X\mathcal{T}v. \end{aligned} \quad (15)$$

Using $v_{tx} = v_{xt}$ and replacing $\mathcal{T}v_x$ by $\mathcal{T}Xv$ on the left side and equating the left and right hand sides in (15), we obtain

$$(X_t - \mathcal{T}_x + X\mathcal{T} - \mathcal{T}X)v = 0,$$

which in turn, because of (iii), implies

$$X_t - \mathcal{T}_x + X\mathcal{T} - \mathcal{T}X = 0. \quad (16)$$

We can view (16) as an integrable NPDE solvable with the help of the solutions to the direct and inverse scattering problems for the linear system $v_x = Xv$. Like (9), the compatibility condition (16) yields a nonlinear evolution equation containing a first-order time derivative. Note that X contains the spectral parameter λ , and hence \mathcal{T} also depends on λ as well. This is in contrast with the Lax method in the sense that the operator \mathcal{A} does not contain λ .

Let us illustrate the AKNS method by deriving the KdV equation in (1) from the Schrödinger equation in (2). For this purpose we write the Schrödinger equation, by replacing the spectral parameter k^2 with λ , as a first-order linear system $v_x = Xv$, where we have defined

$$v := \begin{bmatrix} \psi_x \\ \psi \end{bmatrix}, \quad X := \begin{bmatrix} 0 & u(x, t) - \lambda \\ 1 & 0 \end{bmatrix}.$$

Let us look for \mathcal{T} in the form

$$\mathcal{T} = \begin{bmatrix} \alpha & \beta \\ \rho & \sigma \end{bmatrix},$$

where the entries α, β, ρ , and σ may depend on x, t , and λ . The compatibility condition (16) yields

$$\begin{bmatrix} -\alpha_x - \beta + \rho(u - \lambda) & u_t - \beta_x + \sigma(u - \lambda) - \alpha(u - \lambda) \\ -\rho_x + \alpha - \sigma & -\sigma_x + \beta - \rho(u - \lambda) \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}. \quad (17)$$

The (1, 1), (2, 1), and (2, 2)-entries in the matrix equation in (17) imply

$$\beta = -\alpha_x + (u - \lambda)\rho, \quad \sigma = \alpha - \rho_x, \quad \sigma_x = -\alpha_x. \quad (18)$$

Then from the (1, 2)-entry in (17) we obtain

$$u_t + \frac{1}{2}\rho_{xxx} - u_x\rho - 2\rho_x(u - \lambda) = 0. \quad (19)$$

Assuming a linear dependence of ρ on the spectral parameter and hence letting $\rho = \lambda\zeta + \mu$ in (19), we get

$$2\zeta_x\lambda^2 + \left(\frac{1}{2}\zeta_{xxx} - 2\zeta_x u + 2\mu_x - u_x\zeta\right)\lambda + \left(u_t + \frac{1}{2}\mu_{xxx} - 2\mu_x u - u_x\mu\right) = 0.$$

Equating the coefficients of each power of λ to zero, we have

$$\begin{aligned}\zeta &= c_1, \quad \mu = \frac{1}{2}c_1 u + c_2, \\ u_t - \frac{3}{2}c_1 u u_x - c_2 u_x + \frac{1}{4}c_1 u_{xxx} &= 0, \quad (20)\end{aligned}$$

with c_1 and c_2 denoting arbitrary constants. Choosing $c_1 = 4$ and $c_2 = 0$, from (20) we obtain the KdV equation given in (1). Moreover, with the help of (18) we get

$$\begin{aligned}\alpha &= u_x + c_3, \\ \beta &= -4\lambda^2 + 2\lambda u + 2u^2 - u_{xx}, \\ \rho &= 4\lambda + 2u, \quad \sigma = c_3 - u_x,\end{aligned}$$

where c_3 is an arbitrary constant. Choosing $c_3 = 0$, we find

$$\mathcal{T} = \begin{bmatrix} u_x & -4\lambda^2 + 2\lambda u + 2u^2 - u_{xx} \\ 4\lambda + 2u & -u_x \end{bmatrix}.$$

As for the Zakharov–Shabat system in (4), writing it as $v_x = \mathcal{X}v$, where we have defined

$$\mathcal{X} := \begin{bmatrix} -i\lambda & u(x, t) \\ -\overline{u(x, t)} & i\lambda \end{bmatrix},$$

we obtain the matrix operator \mathcal{T} as

$$\mathcal{T} = \begin{bmatrix} -2i\lambda^2 + i|u|^2 & 2\lambda u + iu_x \\ -2\lambda\bar{u} + i\bar{u}_x & 2i\lambda^2 - i|u|^2 \end{bmatrix},$$

and the compatibility condition (16) yields the NLS equation in (3).

As for the first-order linear system (6), by writing it as $v_x = \mathcal{X}v$, where

$$\mathcal{X} := \begin{bmatrix} -i\lambda & u(x, t) \\ -u(x, t) & i\lambda \end{bmatrix},$$

we obtain the matrix operator \mathcal{T} as

$$\mathcal{T} = \begin{bmatrix} -4i\lambda^3 + 2i\lambda u^2 & 4\lambda^2 u + 2i\lambda u_x - u_{xx} - 2u^3 \\ -4\lambda^2 u + 2i\lambda u_x + u_{xx} + 2u^3 & 4i\lambda^3 - 2i\lambda u^2 \end{bmatrix}$$

and the compatibility condition (16) yields the mKdV equation in (5).

As for the first-order system $v_x = \mathcal{X}v$, where

$$\mathcal{X} := \begin{bmatrix} -i\lambda & -\frac{1}{2}u_x(x, t) \\ \frac{1}{2}u_x(x, t) & i\lambda \end{bmatrix},$$

we obtain the matrix operator \mathcal{T} as

$$\mathcal{T} = \frac{i}{4\lambda} \begin{bmatrix} \cos u & \sin u \\ \sin u & -\cos u \end{bmatrix}.$$

Then, the compatibility condition (16) gives us the sine-Gordon equation

$$u_{xt} = \sin u.$$

Direct Scattering Problem

The direct scattering problem consists of determining the scattering data when the potential is known. This problem is usually solved by obtaining certain specific solutions, known as the Jost solutions, to the relevant LODE. The appropriate scattering data can be constructed with the help of spatial asymptotics of the Jost solutions at infinity or from certain Wronskian relations among the Jost solutions. In this section we review the scattering data corresponding to the Schrödinger equation in (2) and to the Zakharov–Shabat system in (4). The scattering data sets for other LODEs can similarly be obtained.

Consider (2) at fixed t by assuming that the potential $u(x, t)$ belongs to the Faddeev class, i.e. $u(x, t)$ is real valued and $\int_{-\infty}^{\infty} dx (1 + |x|) |u(x, t)|$ is finite. The Schrödinger equation has two types of solutions; namely, scattering solutions and bound-state solutions. The scattering solutions are those that consist of linear combinations of e^{ikx} and e^{-ikx} as $x \rightarrow \pm\infty$, and they occur for $k \in \mathbf{R} \setminus \{0\}$, i.e. for real nonzero values of k . Two linearly independent scattering solutions f_l and f_r , known as the Jost solution from the left and from the right, respectively, are those solutions to (2) satisfying the respective asymptotic conditions

$$\begin{aligned}f_l(k, x, t) &= e^{ikx} + o(1), \\ f_l'(k, x, t) &= ike^{ikx} + o(1), \quad x \rightarrow +\infty, \quad (21)\end{aligned}$$

$$\begin{aligned}f_r(k, x, t) &= e^{-ikx} + o(1), \\ f_r'(k, x, t) &= -ike^{-ikx} + o(1), \quad x \rightarrow -\infty,\end{aligned}$$

where the notation $o(1)$ indicates the quantities that vanish. Writing their remaining spatial asymptotics in the form

$$f_l(k, x, t) = \frac{e^{ikx}}{T(k, t)} + \frac{L(k, t) e^{-ikx}}{T(k, t)} + o(1), \quad x \rightarrow -\infty, \quad (22)$$

$$f_r(k, x, t) = \frac{e^{-ikx}}{T(k, t)} + \frac{R(k, t) e^{ikx}}{T(k, t)} + o(1), \quad x \rightarrow +\infty,$$

we obtain the scattering coefficients; namely, the transmission coefficient T and the reflection coefficients L and R , from the left and right, respectively.

Let \mathbf{C}^+ denote the upper half complex plane. A bound-state solution to (2) is a solution that belongs to $L^2(\mathbf{R})$ in the x variable. Note that $L^2(\mathbf{R})$ denotes the set of complex-valued functions whose absolute squares are integrable on the real line \mathbf{R} . When $u(x, t)$ is in the Faddeev class, it is known [5,7,8,9,16,17,18,19] that the number of bound states is finite, the multiplicity of each bound state is one, and the bound-state solutions can occur only at certain k -values on the imaginary axis in \mathbf{C}^+ . Let us use N to denote the number of bound states, and suppose that the bound states occur at $k = i\kappa_j$ with the ordering $0 < \kappa_1 < \dots < \kappa_N$. Each bound state corresponds to a pole of T in \mathbf{C}^+ . Any bound-state solution at $k = i\kappa_j$ is a constant multiple of $f_l(i\kappa_j, x, t)$. The left and right bound-state norming constants $c_{lj}(t)$ and $c_{rj}(t)$, respectively, can be defined as

$$c_{lj}(t) := \left[\int_{-\infty}^{\infty} dx f_l(i\kappa_j, x, t)^2 \right]^{-1/2},$$

$$c_{rj}(t) := \left[\int_{-\infty}^{\infty} dx f_r(i\kappa_j, x, t)^2 \right]^{-1/2},$$

and they are related to each other through the residues of T via

$$\text{Res}(T, i\kappa_j) = i c_{lj}(t)^2 \gamma_j(t) = i \frac{c_{rj}(t)^2}{\gamma_j(t)}, \quad (23)$$

where the $\gamma_j(t)$ are the dependency constants defined as

$$\gamma_j(t) := \frac{f_l(i\kappa_j, x, t)}{f_r(i\kappa_j, x, t)}. \quad (24)$$

The sign of $\gamma_j(t)$ is the same as that of $(-1)^{N-j}$, and hence $c_{rj}(t) = (-1)^{N-j} \gamma_j(t) c_{lj}(t)$.

The scattering matrix associated with (2) consists of the transmission coefficient T and the two reflection coefficients R and L , and it can be constructed from $\{\kappa_j\}_{j=1}^N$

and one of the reflection coefficients. For example, if we start with the right reflection coefficient $R(k, t)$ for $k \in \mathbf{R}$, we get

$$T(k, t) = \left(\prod_{j=1}^N \frac{k + i\kappa_j}{k - i\kappa_j} \right) \times \exp \left(\frac{1}{2\pi i} \int_{-\infty}^{\infty} ds \frac{\log(1 - |R(s, t)|^2)}{s - k - i0^+} \right), \quad k \in \mathbf{C}^+ \cup \mathbf{R},$$

where the quantity $i0^+$ indicates that the value for $k \in \mathbf{R}$ must be obtained as a limit from \mathbf{C}^+ . Then, the left reflection coefficient $L(k, t)$ can be constructed via

$$L(k, t) = - \frac{\overline{R(k, t)} T(k, t)}{T(k, t)}, \quad k \in \mathbf{R}.$$

We will see in the next section that $T(k, t) = T(k, 0)$, $|R(k, t)| = |R(k, 0)|$, and $|L(k, t)| = |L(k, 0)|$.

For a detailed study of the direct scattering problem for the 1-D Schrödinger equation, we refer the reader to [5,7,8,9,16,17,18,19]. It is important to remember that $u(x, t)$ for $x \in \mathbf{R}$ at each fixed t is uniquely determined [5,7,8,9,16,17,18] by the scattering data $\{R, \{\kappa_j\}, \{c_{lj}(t)\}\}$ or one of its equivalents. Letting $c_j(t) := c_{lj}(t)^2$, we will work with one such data set, namely $\{R, \{\kappa_j\}, \{c_j(t)\}\}$, in Sect. “Time Evolution of the Scattering Data” and Sect. “Inverse Scattering Problem”.

Having described the scattering data associated with the Schrödinger equation, let us briefly describe the scattering data associated with the Zakharov–Shabat system in (4). Assuming that $u(x, t)$ for each t is integrable in x on \mathbf{R} , the two Jost solutions $\psi(\lambda, x, t)$ and $\phi(\lambda, x, t)$, from the left and from the right, respectively, are those unique solutions to (4) satisfying the respective asymptotic conditions

$$\psi(\lambda, x, t) = \begin{bmatrix} 0 \\ e^{i\lambda x} \end{bmatrix} + o(1), \quad x \rightarrow +\infty,$$

$$\phi(\lambda, x, t) = \begin{bmatrix} e^{-i\lambda x} \\ 0 \end{bmatrix} + o(1), \quad x \rightarrow -\infty. \quad (25)$$

The transmission coefficient T , the left reflection coefficient L , and the right reflection coefficient R are obtained

via the asymptotics

$$\begin{aligned}\psi(\lambda, x, t) &= \left[\frac{L(\lambda, t) e^{-i\lambda x}}{T(\lambda, t)} \right] + o(1), \quad x \rightarrow -\infty, \\ \phi(\lambda, x, t) &= \left[\frac{e^{-i\lambda x}}{T(\lambda, t)} \right] + o(1), \quad x \rightarrow +\infty.\end{aligned}\quad (26)$$

The bound-state solutions to (4) occur at those λ values corresponding to the poles of T in \mathbb{C}^+ . Let us use $\{\lambda_j\}_{j=1}^N$ to denote the set of such poles. It should be noted that such poles are not necessarily located on the positive imaginary axis. Furthermore, unlike the Schrödinger equation, the multiplicities of such poles may be greater than one. Let us assume that the pole λ_j has multiplicity n_j . Corresponding to the pole λ_j , one associates [4,20] n_j bound-state norming constants $c_{js}(t)$ for $s = 0, 1, \dots, n_j - 1$. We assume that, for each fixed t , the potential $u(x, t)$ in the Zakharov–Shabat system is uniquely determined by the scattering data $\{R, \{\lambda_j\}, \{c_{js}(t)\}\}$ and vice versa.

Time Evolution of the Scattering Data

As the initial profile $u(x, 0)$ evolves to $u(x, t)$ while satisfying the NPDE, the corresponding initial scattering data $S(\lambda, 0)$ evolves to $S(\lambda, t)$. Since the scattering data can be obtained from the Jost solutions to the associated LODE, in order to determine the time evolution of the scattering data, we can analyze the time evolution of the Jost solutions with the help of the Lax method or the AKNS method.

Let us illustrate how to determine the time evolution of the scattering data in the Schrödinger equation with the help of the Lax method. As indicated in Sect. “The Lax Method”, the spectral parameter k and hence also the values κ_j related to the bound states remain unchanged in time. Let us obtain the time evolution of $f_l(k, x, t)$, the Jost solution from the left. From condition (ii) in Sect. “The Lax Method”, we see that the quantity $\partial_t f_l - \mathcal{A} f_l$ remains a solution to (2) and hence we can write it as a linear combination of the two linearly independent Jost solutions f_l and f_r as

$$\begin{aligned}\partial_t f_l(k, x, t) - (-4\partial_x^3 + 6u\partial_x + 3u_x) f_l(k, x, t) \\ = p(k, t) f_l(k, x, t) + q(k, t) f_r(k, x, t),\end{aligned}\quad (27)$$

where the coefficients $p(k, t)$ and $q(k, t)$ are yet to be determined and \mathcal{A} is the operator in (13). For each fixed t , assuming $u(x, t) = o(1)$ and $u_x(x, t) = o(1)$ as $x \rightarrow +\infty$ and using (21) and (22) in (27) as $x \rightarrow +\infty$, we get

$$\begin{aligned}\partial_t e^{ikx} + 4\partial_x^3 e^{ikx} &= p(k, t) e^{ikx} \\ &+ q(k, t) \left[\frac{1}{T(k, t)} e^{-ikx} + \frac{R(k, t)}{T(k, t)} e^{ikx} \right] + o(1).\end{aligned}\quad (28)$$

Comparing the coefficients of e^{ikx} and e^{-ikx} on the two sides of (28), we obtain

$$q(k, t) = 0, \quad p(k, t) = -4ik^3.$$

Thus, $f_l(k, x, t)$ evolves in time by obeying the linear third-order PDE

$$\partial_t f_l - \mathcal{A} f_l = -4ik^3 f_l. \quad (29)$$

Proceeding in a similar manner, we find that $f_r(k, x, t)$ evolves in time according to

$$\partial_t f_r - \mathcal{A} f_r = 4ik^3 f_r. \quad (30)$$

Notice that the time evolution of each Jost solution is fairly complicated. We will see, however, that the time evolution of the scattering data is very simple. Letting $x \rightarrow -\infty$ in (29), using (22) and $u(x, t) = o(1)$ and $u_x(x, t) = o(1)$ as $x \rightarrow -\infty$, and comparing the coefficients of e^{ikx} and e^{-ikx} on both sides, we obtain

$$\partial_t T(k, t) = 0, \quad \partial_t L(k, t) = -8ik^3 L(k, t),$$

yielding

$$T(k, t) = T(k, 0), \quad L(k, t) = L(k, 0) e^{-8ik^3 t}.$$

In a similar way, from (30) as $x \rightarrow +\infty$, we get

$$R(k, t) = R(k, 0) e^{8ik^3 t}. \quad (31)$$

Thus, the transmission coefficient remains unchanged and only the phases of the reflection coefficients change as time progresses.

Let us also evaluate the time evolution of the dependency constants $\gamma_j(t)$ defined in (24). Evaluating (29) at $k = i\kappa_j$ and replacing $f_l(i\kappa_j, x, t)$ by $\gamma_j(t) f_r(i\kappa_j, x, t)$, we get

$$\begin{aligned}f_r(i\kappa_j, x, t) \partial_t \gamma_j(t) + \gamma_j(t) \partial_t f_r(i\kappa_j, x, t) \\ - \gamma_j(t) \mathcal{A} f_r(i\kappa_j, x, t) = -4\kappa_j^3 \gamma_j(t) f_r(i\kappa_j, x, t).\end{aligned}\quad (32)$$

On the other hand, evaluating (30) at $k = ik_j$, we obtain

$$\begin{aligned} \gamma_j(t) \partial_t f_r(ik_j, x, t) - \gamma_j(t) \mathcal{A} f_r(ik_j, x, t) \\ = 4\kappa_j^3 \gamma_j(t) f_r(ik_j, x, t). \end{aligned} \quad (33)$$

Comparing (32) and (33) we see that $\partial_t \gamma_j(t) = -8\kappa_j^3 \gamma_j(t)$, or equivalently

$$\gamma_j(t) = \gamma_j(0) e^{-8\kappa_j^3 t}. \quad (34)$$

Then, with the help of (23) and (34), we determine the time evolutions of the norming constants as

$$c_{lj}(t) = c_{lj}(0) e^{4\kappa_j^3 t}, \quad c_{rj}(t) = c_{rj}(0) e^{-4\kappa_j^3 t}.$$

The norming constants $c_j(t)$ appearing in the Marchenko kernel (38) are related to $c_{lj}(t)$ as $c_j(t) := c_{lj}(t)^2$, and hence their time evolution is described as

$$c_j(t) = c_j(0) e^{8\kappa_j^3 t}. \quad (35)$$

As for the NLS equation and other integrable NPDEs, the time evolution of the related scattering data sets can be obtained in a similar way. For the former, in terms of the operator \mathcal{A} in (14), the Jost solutions $\psi(\lambda, x, t)$ and $\phi(\lambda, x, t)$ appearing in (25) evolve according to the respective linear PDEs

$$\psi_t - \mathcal{A}\psi = -2i\lambda^2 \psi, \quad \phi_t - \mathcal{A}\phi = 2i\lambda^2 \phi.$$

The scattering coefficients appearing in (26) evolve according to

$$\begin{aligned} T(\lambda, t) &= T(\lambda, 0), \\ R(\lambda, t) &= R(\lambda, 0) e^{4i\lambda^2 t}, \\ L(\lambda, t) &= L(\lambda, 0) e^{-4i\lambda^2 t}. \end{aligned} \quad (36)$$

Associated with the bound-state pole λ_j of T , we have the bound-state norming constants $c_{js}(t)$ appearing in the Marchenko kernel $\Omega(y, t)$ given in (41). Their time evolution is governed [4] by

$$\begin{aligned} [c_{j(n_j-1)}(t) \quad c_{j(n_j-2)}(t) \quad \dots \quad c_{j0}(t)] \\ = [c_{j(n_j-1)}(0) \quad c_{j(n_j-2)}(0) \quad \dots \quad c_{j0}(0)] e^{-4iA_j^2 t}, \end{aligned} \quad (37)$$

where the $n_j \times n_j$ matrix A_j appearing in the exponent is defined as

$$A_j := \begin{bmatrix} -i\lambda_j & -1 & 0 & \dots & 0 & 0 \\ 0 & -i\lambda_j & -1 & \dots & 0 & 0 \\ 0 & 0 & -i\lambda_j & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -i\lambda_j & -1 \\ 0 & 0 & 0 & \dots & 0 & -i\lambda_j \end{bmatrix}.$$

Inverse Scattering Problem

In Sect. “Direct Scattering Problem” we have seen how the initial scattering data $S(\lambda, 0)$ can be constructed from the initial profile $u(x, 0)$ of the potential by solving the direct scattering problem for the relevant LODE. Then, in Sect. “Time Evolution of the Scattering Data” we have seen how to obtain the time-evolved scattering data $S(\lambda, t)$ from the initial scattering data $S(\lambda, 0)$. As the final step in the IST, in this section we outline how to obtain $u(x, t)$ from $S(\lambda, t)$ by solving the relevant inverse scattering problem. Such an inverse scattering problem may be solved by the Marchenko method [5,7,8,9,16,17,18,19]. Unfortunately, in the literature many researchers refer to this method as the Gel’fand–Levitan method or the Gel’fand–Levitan–Marchenko method, both of which are misnomers. The Gel’fand–Levitan method [5,7,16,17,19] is a different method to solve the inverse scattering problem, and the corresponding Gel’fand–Levitan integral equation involves an integration on the finite interval $(0, x)$ and its kernel is related to the Fourier transform of the spectral measure associated with the LODE. On the other hand, the Marchenko integral equation involves an integration on the semi-infinite interval $(x, +\infty)$, and its kernel is related to the Fourier transform of the scattering data.

In this section we first outline the recovery of the solution $u(x, t)$ to the KdV equation from the corresponding time-evolved scattering data $\{R, \{\kappa_j\}, \{c_j(t)\}\}$ appearing in (31) and (35). Later, we will also outline the recovery of the solution $u(x, t)$ to the NLS equation from the corresponding time-evolved scattering data $\{R, \{\lambda_j\}, \{c_{js}(t)\}\}$ appearing in (36) and (37).

The solution $u(x, t)$ to the KdV equation in (1) can be obtained from the time-evolved scattering data by using the Marchenko method as follows:

- (a) From the scattering data $\{R(k, t), \{\kappa_j\}, \{c_j(t)\}\}$ appearing in (31) and (35), form the Marchenko kernel Ω defined via

$$\Omega(y, t) := \frac{1}{2\pi} \int_{-\infty}^{\infty} dk R(k, t) e^{iky} + \sum_{j=1}^N c_j(t) e^{-\kappa_j y}. \quad (38)$$

- (b) Solve the corresponding Marchenko integral equation

$$\begin{aligned} K(x, y, t) + \Omega(x + y, t) \\ + \int_x^{\infty} dz K(x, z, t) \Omega(z + y, t) = 0, \\ x < y < +\infty, \end{aligned} \quad (39)$$

and obtain its solution $K(x, y, t)$.

(c) Recover $u(x, t)$ by using

$$u(x, t) = -2 \frac{\partial K(x, x, t)}{\partial x}. \quad (40)$$

The solution $u(x, t)$ to the NLS equation in (3) can be obtained from the time-evolved scattering data by using the Marchenko method as follows:

(i) From the scattering data $\{R(\lambda, t), \{\lambda_j\}, \{c_{js}(t)\}\}$ appearing in (36) and (37), form the Marchenko kernel Ω as

$$\begin{aligned} \Omega(y, t) := & \frac{1}{2\pi} \int_{-\infty}^{\infty} d\lambda R(\lambda, t) e^{i\lambda y} \\ & + \sum_{j=1}^N \sum_{s=0}^{n_j-1} c_{js}(t) \frac{y^s}{s!} e^{i\lambda_j y}. \end{aligned} \quad (41)$$

(ii) Solve the Marchenko integral equation

$$\begin{aligned} K(x, y, t) - \overline{\Omega(x+y, t)} + \int_x^{\infty} dz \\ \times \int_x^{\infty} ds K(x, s, t) \Omega(s+z, t) \overline{\Omega(z+y, t)} = 0, \\ x < y < +\infty, \end{aligned}$$

and obtain its solution $K(x, y, t)$.

(iii) Recover $u(x, t)$ from the solution $K(x, y, t)$ to the Marchenko equation via

$$u(x, t) = -2K(x, x, t).$$

(iv) Having determined $K(x, y, t)$, one can alternatively get $|u(x, t)|^2$ from

$$|u(x, t)|^2 = 2 \frac{\partial G(x, x, t)}{\partial x},$$

where we have defined

$$G(x, y, t) := - \int_x^{\infty} dz \overline{K(x, z, t)} \Omega(z+y, t).$$

Solitons

A soliton solution to an integrable NPDE is a solution $u(x, t)$ for which the reflection coefficient in the corresponding scattering data is zero. In other words, a soliton solution $u(x, t)$ to an integrable NPDE is nothing but a reflectionless potential in the associated LODE. When the reflection coefficient is zero, the kernel of the relevant Marchenko integral equation becomes separable. An integral equation with a separable kernel can be solved explicitly by transforming that linear equation into a system of linear algebraic equations. In that case, we get exact solutions to the integrable NPDE, which are known as soliton solutions.

For the KdV equation the N -soliton solution is obtained by using $R(k, t) = 0$ in (38). In that case, letting

$$\begin{aligned} X(x) &:= [e^{-\kappa_1 x} \quad e^{-\kappa_2 x} \quad \dots \quad e^{-\kappa_N x}], \\ Y(y, t) &:= \begin{bmatrix} c_1(t) e^{-\kappa_1 y} \\ c_2(t) e^{-\kappa_2 y} \\ \vdots \\ c_N(t) e^{-\kappa_N y} \end{bmatrix}, \end{aligned}$$

we get $\Omega(x+y, t) = X(x) Y(y, t)$. As a result of this separability the Marchenko integral equation can be solved algebraically and the solution has the form $K(x, y, t) = H(x, t) Y(y, t)$, where $H(x, t)$ is a row vector with N entries that are functions of x and t . A substitution in (39) yields

$$K(x, y, t) = -X(x) \Gamma(x, t)^{-1} Y(y, t), \quad (42)$$

where the $N \times N$ matrix $\Gamma(x, t)$ is given by

$$\Gamma(x, t) := I + \int_x^{\infty} dz Y(z, t) X(z), \quad (43)$$

with I denoting the or $N \times N$ identity matrix. Equivalently, the (j, l) -entry of Γ is given by

$$\Gamma_{jl} = \delta_{jl} + \frac{c_j(0) e^{-2\kappa_j x + 8\kappa_j^3 t}}{\kappa_j + \kappa_l},$$

with δ_{jl} denoting the Kronecker delta. Using (42) in (40) we obtain

$$\begin{aligned} u(x, t) &= 2 \frac{\partial}{\partial x} [X(x) \Gamma(x, t)^{-1} Y(x, t)] \\ &= 2 \operatorname{tr} \frac{\partial}{\partial x} [Y(x, t) X(x) \Gamma(x, t)^{-1}], \end{aligned}$$

where tr denotes the matrix trace (the sum of diagonal entries in a square matrix). From (43) we see that $-Y(x, t) X(x)$ is equal to the x -derivative of $\Gamma(x, t)$ and hence the N -soliton solution can also be written as

$$\begin{aligned} u(x, t) &= -2 \operatorname{tr} \frac{\partial}{\partial x} \left[\frac{\partial \Gamma(x, t)}{\partial x} \Gamma(x, t)^{-1} \right] \\ &= -2 \frac{\partial}{\partial x} \left[\frac{\frac{\partial}{\partial x} \det \Gamma(x, t)}{\det \Gamma(x, t)} \right], \end{aligned} \quad (44)$$

where \det denotes the matrix determinant. When $N = 1$, we can express the one-soliton solution $u(x, t)$ to the KdV equation in the equivalent form

$$u(x, t) = -2\kappa_1^2 \operatorname{sech}^2(\kappa_1 x - 4\kappa_1^3 t + \theta),$$

with $\theta := \log \sqrt{2\kappa_1/c_1(0)}$.

Let us mention that, using matrix exponentials, we can express [6] the N -soliton solution appearing in (44) in various other equivalent forms such as

$$u(x, t) = -4Ce^{-Ax+8A^3t} \Gamma(x, t)^{-1} A \Gamma(x, t)^{-1} e^{-Ax} B,$$

where

$$\begin{aligned} A &:= \text{diag}\{\kappa_1, \kappa_2, \dots, \kappa_N\}, \\ B^\dagger &:= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix}, \\ C &:= \begin{bmatrix} c_1(0) & c_2(0) & \dots & c_N(0) \end{bmatrix}. \end{aligned} \quad (45)$$

Note that a dagger is used for the matrix adjoint (transpose and complex conjugate), and B has N entries. In this notation we can express (43) as

$$\Gamma(x, t) = I + \int_x^\infty dz e^{-zA} B C e^{-zA} e^{8tA^3}.$$

As for the NLS equation, the well-known N -soliton solution (with simple bound-state poles) is obtained by choosing $R(\lambda, t) = 0$ and $n_j = 1$ in (41). Proceeding as in the KdV case, we obtain the N -soliton solution in terms of the triplet A, B, C with

$$A := \text{diag}\{-i\lambda_1, -i\lambda_2, \dots, -i\lambda_N\}, \quad (46)$$

where the complex constants λ_j are the distinct poles of the transmission coefficient in \mathbb{C}^+ , B and C are as in (45) except for the fact that the constants $c_j(0)$ are now allowed to be nonzero complex numbers. In terms of the matrices $P(x, t)$, M , and Q defined as

$$\begin{aligned} P(x, t) &:= \text{diag}\{e^{2i\lambda_1 x + 4i\lambda_1^2 t}, e^{2i\lambda_2 x + 4i\lambda_2^2 t}, \dots, e^{2i\lambda_N x + 4i\lambda_N^2 t}\}, \\ M_{jl} &:= \frac{i}{\lambda_j - \bar{\lambda}_l}, \quad Q_{jl} := \frac{-i\bar{c}_j c_l}{\bar{\lambda}_j - \lambda_l}. \end{aligned}$$

we construct the N -soliton solution $u(x, t)$ to the NLS equation as

$$u(x, t) = -2B^\dagger [I + P(x, t)^\dagger Q P(x, t) M]^{-1} P(x, t)^\dagger C^\dagger, \quad (47)$$

or equivalently as

$$u(x, t) = -2B^\dagger e^{-A^\dagger x} \Gamma(x, t)^{-1} e^{-A^\dagger x + 4i(A^\dagger)^2 t} C^\dagger, \quad (48)$$

where we have defined

$$\begin{aligned} \Gamma(x, t) &:= I + \left[\int_x^\infty ds (C e^{-As - 4iA^2 t})^\dagger (C e^{-As - 4iA^2 t}) \right] \\ &\quad \times \left[\int_x^\infty dz (e^{-Az} B) (e^{-Az} B)^\dagger \right]. \end{aligned} \quad (49)$$

Using (45) and (46) in (49), we get the (j, l) -entry of $\Gamma(x, t)$ as

$$\Gamma_{jl} = \delta_{jl} - \sum_{m=1}^N \frac{\bar{c}_j c_l e^{i(2\lambda_m - \bar{\lambda}_j - \bar{\lambda}_l)x + 4i(\lambda_m^2 - \bar{\lambda}_j^2)t}}{(\lambda_m - \bar{\lambda}_j)(\lambda_m - \bar{\lambda}_l)}.$$

Note that the absolute square of $u(x, t)$ is given by

$$\begin{aligned} |u(x, t)|^2 &= \text{tr} \left[\frac{\partial}{\partial x} \left(\Gamma(x, t)^{-1} \frac{\partial \Gamma(x, t)}{\partial x} \right) \right] \\ &= \frac{\partial}{\partial x} \left[\frac{\frac{\partial}{\partial x} \det \Gamma(x, t)}{\det \Gamma(x, t)} \right]. \end{aligned}$$

For the NLS equation, when $N = 1$, from (47) or (48) we obtain the single-soliton solution

$$u(x, t) = \frac{-8\bar{c}_1 (\text{Im}[\lambda_1])^2 e^{-2i\bar{\lambda}_1 x - 4i(\bar{\lambda}_1)^2 t}}{4(\text{Im}[\lambda_1])^2 + |c_1|^2 e^{-4x(\text{Im}[\lambda_1]) - 8t(\text{Im}[\lambda_1]^2)}},$$

where Im denotes the imaginary part.

Future Directions

There are many issues related to the IST and solitons that cannot be discussed in such a short review. We will briefly mention only a few.

Can we characterize integrable NPDEs? In other words, can we find a set of necessary and sufficient conditions that guarantee that an IVP for a NPDE is solvable via an IST? Integrable NPDEs seem to have some common characteristic features [1] such as possessing Lax pairs, AKNS pairs, soliton solutions, infinite number of conserved quantities, a Hamiltonian formalism, the Painlevé property, and the Bäcklund transformation. Yet, there does not seem to be a satisfactory solution to their characterization problem.

Another interesting question is the determination of the LODE associated with an IST. In other words, given an integrable NPDE, can we determine the corresponding LODE? There does not yet seem to be a completely satisfactory answer to this question.

When the initial scattering coefficients are rational functions of the spectral parameter, representing the time-evolved scattering data in terms of matrix exponentials results in the separability of the kernel of the Marchenko integral equation. In that case, one obtains explicit formulas [4,6] for exact solutions to some integrable NPDEs and such solutions are constructed in terms of a triplet of constant matrices A, B, C whose sizes are $p \times p$, $p \times 1$, and $1 \times p$, respectively, for any positive integer p . Some special

cases of such solutions have been mentioned in Sect. “Solitons”, and it would be interesting to determine if such exact solutions can be constructed also when p becomes infinite.

Bibliography

Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolution equations and inverse scattering. Cambridge University Press, Cambridge
2. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1973) Method for solving the sine–Gordon equation. *Phys Rev Lett* 30:1262–1264
3. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1974) The inverse scattering transform–Fourier analysis for nonlinear problems. *Stud Appl Math* 53:249–315
4. Aktosun T, Demontis F, van der Mee C (2007) Exact solutions to the focusing nonlinear Schrödinger equation. *Inverse Problems* 23:2171–2195
5. Aktosun T, Klaus M (2001) Chapter 2.2.4, Inverse theory: problem on the line, In: Pike ER, Sabatier PC (eds) *Scattering*. Academic Press, London, pp 770–785
6. Aktosun T, van der Mee C (2006) Explicit solutions to the Korteweg–de Vries equation on the half-line. *Inverse Problems* 22:2165–2174
7. Chadan K, Sabatier PC (1989) Inverse problems in quantum scattering theory. 2nd edn. Springer, New York
8. Deift P, Trubowitz E (1979) Inverse scattering on the line. *Commun Pure Appl Math* 32:121–251
9. Faddeev LD (1967) Properties of the S -matrix of the one-dimensional Schrödinger equation. *Amer Math Soc Transl (Ser. 2)* 65:139–166
10. Fermi E (1965) Collected papers, vol II: United States, 1939–1954. University of Chicago Press, Chicago
11. Fermi E, Pasta J, Ulam S (1955) Studies of non linear problems, I. Document LA-1940, Los Alamos National Laboratory
12. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. *Phys Rev Lett* 19:1095–1097
13. Gel’fand IM, Levitan BM (1955) On the determination of a differential equation from its spectral function. *Amer Math Soc Transl (Ser. 2)* 1:253–304
14. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular channel and on a new type of long stationary waves. *Phil Mag* 39:422–443
15. Lax PD (1968) Integrals of nonlinear equations of evolution and solitary waves. *Commun Pure Appl Math* 21:467–490
16. Levitan BM (1987) Inverse Sturm–Liouville problems. Science VNU Press, Utrecht
17. Marchenko VA (1986) Sturm–Liouville operators and applications. Birkhäuser, Basel
18. Melin A (1985) Operator methods for inverse scattering on the real line. *Commun Pure Appl Math* 10:677–766
19. Newton RG (1983) The Marchenko and Gel’fand–Levitan methods in the inverse scattering problem in one and three dimensions. In: Bednar JB, Redner R, Robinson E, Weglein A (eds) *Conference on inverse scattering: theory and application*. SIAM, Philadelphia, pp 1–74
20. Olmedilla E (1987) Multiple pole solutions of the nonlinear Schrödinger equation. *Phys D* 25:330–346
21. Russell JS (1845) Report on waves, Report of the 14th meeting of the British Association for the Advancement of Science. John Murray, London, pp 311–390
22. Wadati M (1972) The exact solution of the modified Korteweg–de Vries equation. *J Phys Soc Jpn* 32:1681
23. Zabusky NJ, Kruskal MD (1965) Interaction of “solitons” in a collisionless plasma and the recurrence of initial states. *Phys Rev Lett* 15:240–243
24. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Soviet Phys JETP* 34:62–69
25. <http://www.osti.gov/accomplishments/pdf/A80037041/A80037041.pdf>

Books and Reviews

- Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia
- Aktosun T (2004) Inverse scattering transform, KdV, and solitons, In: Ball JA, Helton JW, Klaus M, Rodman L (ed), *Current trends in operator theory and its applications*. Birkhäuser, Basel, pp 1–22
- Aktosun T (2005) Solitons and inverse scattering transform, In: Clemence DP, Tang G (eds) *Mathematical studies in nonlinear wave propagation*, Contemporary Mathematics, vol 379. Amer Math Soc, Providence, pp 47–62
- Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) Solitons and nonlinear wave equations. Academic Press, London
- Drazin PG, Johnson RS (1988) Solitons: an introduction. Cambridge University Press, Cambridge
- Lamb GL Jr (1980) Elements of soliton theory. Wiley, New York
- Novikov S, Manakov SV, Pitaevskii LP, Zakharov VE (1984) Theory of solitons. Consultants Bureau, New York
- Scott AC, Chu FYF, McLaughlin D (1973) The soliton: a new concept in applied science. *Proc IEEE* 61:1443–1483

Investment Decision Making in Finance, Models of

JØRGEN VITTING ANDERSEN

Institut Non Linéaire de Nice (UMR CNRS 6618),

Université de Nice-Sophia Antipolis, Valbonne, France

Article Outline

Glossary

Definition of the Subject

Introduction

Rational Decision Making

Irrational Decision Making

Investment Decision Making in a Complex Environment

Future Directions

Bibliography

Glossary

Decision making A human cognitive process which leads to a course of action among a set of choices.

Rational expectations Expectations that give the best guess of the future (the optimal forecast) using all available information.

Behavioral finance A research field that applies human and social cognitive and emotional biases to understanding how decision making affects market prices, returns and allocation of resources.

Definition of the Subject

A general definition of “decision making” is that it is a human cognitive process which leads to a course of action among a set of choices. A decision process gives rise to a choice which can be an action or an opinion. In the context of modeling investment decision making in finance, by far the majority of models assume that decision making is a reasoning process which is *rational*. Classic examples of rational decision models in finance describe how risk-willing investors should react to ensure a certain return, and have been formalized in models such as the Markovitz mean-variance formulation [26], the Capital Asset Price Model [24,34] as well as the Arbitrage Pricing Theory (all defined in this article). Rational decision making is at the cornerstone of the foundation for modern financial thinking. Nonetheless, over the last two decades or so, models of irrational reasoning have led to the appearance of a new field in finance called “Behavioral Finance” (also defined in this article). More recently, decision making in a complex environment has been a very active research field, notably in interdisciplinary approaches drawing on both physics and finance.

Introduction

Historically, *rational* decision making has been at the very core of financial models describing how one should deal with a variety of problems in finance such as, for example, portfolio allocation [26], pricing of financial assets [24,34], valuation of firms and of capital costs [27], pricing of options [8] and even (somewhat less intuitively) the creation of financial bubbles [39]. The aforementioned examples all led to Nobel prizes in economy: in 1985 (valuation of firms), in 1990 (portfolio theory and asset pricing), in 1997 (new method to price derivatives) and in 2002 (for laboratory experiments as a tool in empirical economic analysis).

The biggest advantage about the assumption of rational expectations is clearly that it allows one to progress and actually do some calculations on how to do investment de-

cision making in finance, as illustrated by the mentioned Nobel prizes in economy. The disadvantage is, however, that it is not clear at all whether one thereby has obtained a framework that to any degree describes the reality of financial markets. As will be seen Sect. “[Critiques of Markovitz Portfolio Theory and the CAPM](#)”, the strongest objection against rational expectations is the assumption that *all* market participants, if given access to a complete information set, would come to same conclusion with regard to price.

Questions about the validity of the assumption of rational expectations in decision making and the theories based upon it, had already surfaced in the 1980s. Criticism was notably raised by a succession of discoveries of anomalies, particularly the evidence of excess volatility of returns [36]. Since the reported excess volatility raised some of the very first empirically founded questions related to the efficient market theory, it became the subject of harsh academic disputes. Following the problems made clear by the puzzle of “excess volatility”, competing theories of rational decision making and the efficient market surfaced in the 1990s, most notably in the field now known as behavioral finance. The emerging field of behavioral finance in turn lead to more recent theories which view financial markets as a complex system composed of many actors with different goals and often lead by irrational decision making.

Rational Decision Making

In the following, the concept of rational expectations in finance will be defined in terms of how one should determine the price of one single asset, depending on the money one is going to receive in the future for holding that asset. As will be seen, in order to know today’s price of an asset, the money that one is going to receive in one year’s time, two year’s time etc. (called the future cash flow) has to be expressed in the present time value of money, that is, it has to be discounted. After the introduction of the concept of rational expectations has been applied to one single asset, the concept will be generalized to also take into account decision making based on risk, encountered in a situation in which one has to deal with a portfolio of assets.

Rational Expectations

According to the efficient market theory, the price of an asset at time t , P_t , should equal the expectation value conditioned on all available information, I_t , at time t of all future cash flows (coming from the asset) discounted to the present time value of money. If we call P_t^* the sum over all

future discounted cash flows, one has:

$$P_t = E_t(P_t^* | I_t) \quad (1)$$

with E_t denoting the expectation value. Rational expectations theory, therefore, defines the price as being identical to the best guess of the future (the optimal forecast) that uses all available information. This amounts to assuming that people do not make systematic errors when predicting the future, and that deviations from perfect foresight are only random. One should notice that P^* is *not* known at time t , but has to be forecast in order to assign a value to the price of the asset at time t , P_t . Note that (1) dictates that all movements in the stock market must have their origin in some new information about the fundamental value P_t^* , one of the facts of the efficient market theory that seems hard to believe in practice.

Equivalently one can write (1) as

$$P_t = \sum_{\tau=t+1}^{\infty} \frac{D(\tau)}{(1+r(\tau))^{(\tau-t)}} \quad (2)$$

where $D(\tau)$ is the cash flow (dividends) at time τ , the factor $(1+r(\tau))^{(\tau-t)}$ is the discount factor transforming money at time τ into present value money at time t , and $r(\tau)$ is the interest rate. That is, (2) says that the price of an asset today should be calculated from the sum over all future cash flows discounted to the present day value of money. One should notice that despite an apparently elegant and appealing explanation offered by (1),(2), the equations have profound implications that one has to counterbalance against a more realistic description of facts.

In fact, under strict assumptions of efficiency, the financial markets would be “dead” without any trades occurring. As seen from (2) the price of an asset should change only with the arrival of new information of either future dividends or interest rates. One way out of this apparent paradox is to notice that risk is *not* taken into account in the derivation above. Informationally efficient markets are an impossibility, because if markets are perfectly efficient, the return of gathering information would be nil. Therefore the degree of market inefficiency could be seen as due to the cost of gathering information as well as due to investors being risk-loving or risk-averse.

The strongest objection against rational expectations theory is, however, the assumption that *all* market participants, if given access to a complete information set, would come to same conclusion with regard to price as, e.g., illustrated by (2). This seems very far from reality, to say the least. For more detail on this discussion see Sect. “[Critiques of Markovitz Portfolio Theory and the CAPM](#)”.

The remarkable fact is that rational expectations theory is the key building block, the very foundation that underpins modern portfolio and pricing theories like the CAPM, APT and versions thereof. Thinking that most risk management theories use rational expectations at their core seems somehow a paradox in that it amounts to making a very risky ansatz at the very foundation!

As noted in [30,33], since rational expectations have to be unchanged in all futures (it is without cost to change an expectation therefore if an expectation is expected to change in the future, rationality will require a revision now) rationality requires that for any $0 < j < n$:

$$E_t[E_{t+j}(P_{t+n})] = E_t(P_{t+n}). \quad (3)$$

Ex post (3) takes the following form for the special case $j = 1$:

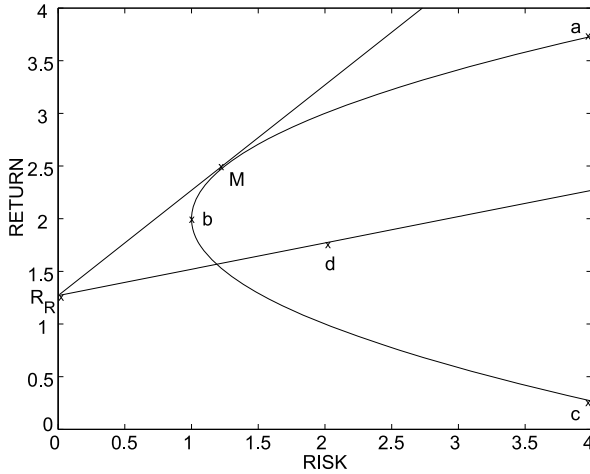
$$E_{t+1} = E_t + \epsilon_{t+1} \quad (4)$$

where ϵ_{t+1} has mean zero and covariance $\text{cov}(\epsilon_{t+j}, \epsilon_t) = 0$ for $j \neq 0$. When the process ϵ is independent and identically distributed, (4) describes a random walk which is non-stationary. Therefore non-stationarity can be seen as a direct consequence of rationality [30].

Markovitz Portfolio Theory and the Capital Asset Price Model (CAPM)

One of the most fundamental tenets of economic theory and practice is that returns above the so-called riskless rate come with increased risks. This is the basis of Markovitz's portfolio theory [26] and of the Capital Asset Pricing Model (CAPM) [24,34]. Reciprocally, investors want to be compensated for taking risk, that is, they want to earn a return high enough to make them comfortable with the level of risk they are assuming.

Markovitz's portfolio theory considers investments over *one* time period, i. e., an investor selects a portfolio at time t that produces a return at time $t + 1$. It is implicitly assumed that the probability distribution function of the returns of the assets in the portfolio do not change between time t and $t + 1$. In Markovitz's model, investors are risk-averse, and when choosing among portfolios they take into account only the return of the portfolio (the mean) and the risk (estimated by the variance) of their one-period investment. To illustrate in practice how Markovitz's method works, imagine for a moment the decision an investor faces by investing a fixed amount of money in, say, the 40 assets of the French CAC40 index. One investment decision could be to put 1% of the fixed investment amount into asset 1, 3% into asset 2, ..., 4% into asset 40. Calculating the return and risk over one time period (say a year)



Investment Decision Making in Finance, Models of, Figure 1
Risk versus return of a portfolio

for this specific choice of portfolio would give *one* point, say the point *d*, within the parabola bounded by “*a-b-c*” in Fig. 1. Another decision choice would, of course, give rise to a different point within this risk-return diagram, and so by probing all possible decision choices one probes the area of the parabola bounded by “*a-b-c*”. Now assuming all investors are rational, they all see the same optimal choice of investment, namely the so called “efficient frontier”, which consists of the part *a-b* on the parabola. The part *a-b* on the parabola corresponds to all those portfolios which, for a given fixed value of risk, give the largest return or, equivalently, for a given fixed value of return (mean) have the smallest amount of risk (variance). Therefore, the Markovitz approach is often called the standard “mean-variance” approach.

Sharpe and Lintner added a key assumption to the work of Markovitz by assuming borrowing and lending at a risk free rate which is the same for all investors and does not depend on the amount borrowed or lent. The introduction of borrowing/lending at a risk free rate by Lintner and Sharpe turns the efficient set into a straight line [16]. If all funds are invested in the risk free asset, i. e., all funds are lent at the risk free rate R_f , one gets a portfolio at the point R_f in Fig. 1, i. e., a portfolio with zero variance and a risk free rate of return R_f . Instead, progressively investing a percentage $1 - x$ of all funds into the risky assets, one would follow the line $R_f - d$ with the point *d* corresponding to being invested 100% in a portfolio of risky assets. Points to the right of *d* correspond to investing more than 100% by borrowing at the risk free rate. Formally this can be written:

$$R_p = xR_f + (1 - x)R_d \quad (5)$$

$$E(R_p) = xR_f + (1 - x)E(R_d) \quad (6)$$

$$\sigma(R_p) = (1 - x)\sigma(R_d) \quad (7)$$

with R_p denoting the return of the portfolio, E the expectation value and σ the standard deviation. But the portfolio *d* in Fig. 1 is not mean-variance-efficient. In order to get a mean-variance-efficient portfolio with the largest return per unit of risk free borrowed money (the largest slope of the line), one swings the line from R_f up and to the left as far as possible, which gives the tangency portfolio *M*. Since all investors agree completely about the distribution of returns, they all see the same optimal solution and combine the same risky tangency portfolio *M* with risk free lending/borrowing. Because all investors have the same portfolio *M*, this has to be the value weighted market portfolio, that is, each asset’s weight in the market portfolio *M* must be given by the total market value of all outstanding units of the asset divided by the total market value of all risky assets. This, then, in turn gives a benchmark for the returns of each individual asset, since an asset which has the same time series of returns as the market portfolio *M* should give exactly the same return as the market. Stated in another way:

$$E(R_i) = E(R_f) + [E(R_M) - E(R_f)]\beta_i, \quad i = 1, \dots, N \quad (8)$$

$$\beta_i = \frac{\text{cov}(R_i, R_M)}{\sigma(R_M)} \quad (9)$$

where $\text{cov}(R_i, R_M)$ is the covariance of the return of asset *i* with the market return, and $\sigma(R_M)$ is the variance of the market return. Eq. (9) is the famous CAPM relation, which allows us to price an asset, $E(R_i)$, from the knowledge of a certain “market portfolio” and the risk free return R_f . The “market portfolio” should, in principle, include not only all traded financial assets, but also consumer durables, real estate and human capital, a task impossible to estimate in practice. Even a more limited view of the “market portfolio” as consisting only of financial assets would still mean including all quoted assets worldwide, an exercise hardly manageable in practice. Instead, a typical choice in the financial literature is to take U.S. common stocks, but this is, strictly speaking, not a “market portfolio” in the CAPM sense.

Finally, we should mention another influential asset pricing theory that also uses the assumption of rational expectations. The Arbitrage Pricing Theory (APT) was introduced by the economist Stephen Ross in 1976 [31], and can be seen as an extension of the CAPM, since the price of an asset is given by a number of economical factors (in contrast to *one* factor in the CAPM given by the market).

Specifically, the APT assumes that the returns of a given asset are determined by a number of macroeconomic factors F_j (assumed to be random variables with zero mean)

$$E(R_i) = R_r + \sum_{j=1}^n \beta_{i,j} P_j \quad (10)$$

$$R_i = E(R_i) + \sum_{j=1}^n \beta_{i,j} F_j + \epsilon_i \quad (11)$$

with P_j the risk premium of factor j , $\beta_{i,j}$ the sensitivity of asset i to factor j and ϵ_i an idiosyncratic noise term with mean zero. (11) says that the uncertainty of the return of asset i is given by a linear combination of n economic factors. In the APT, the portfolio choice is no longer unique, since different investors will hold portfolios with specific choices of $\beta_{i,j}$. In this sense, the assumptions of the APT are closer to real investment strategies, but like the CAPM it also suffers from the crucial claim of rational expectations, see the discussion in Sect. “Critiques of Markovitz Portfolio Theory and the CAPM”.

Critiques of Markovitz Portfolio Theory and the CAPM

One problem within the framework of CAPM and of Markovitz’s portfolio theory is the notion of risk, which is described by the standard deviation of the return. This is a one dimensional measure, whereas the fundamental decision problem in principle has an infinite number of dimensions, given only by full knowledge of the probability distribution of the portfolio returns.

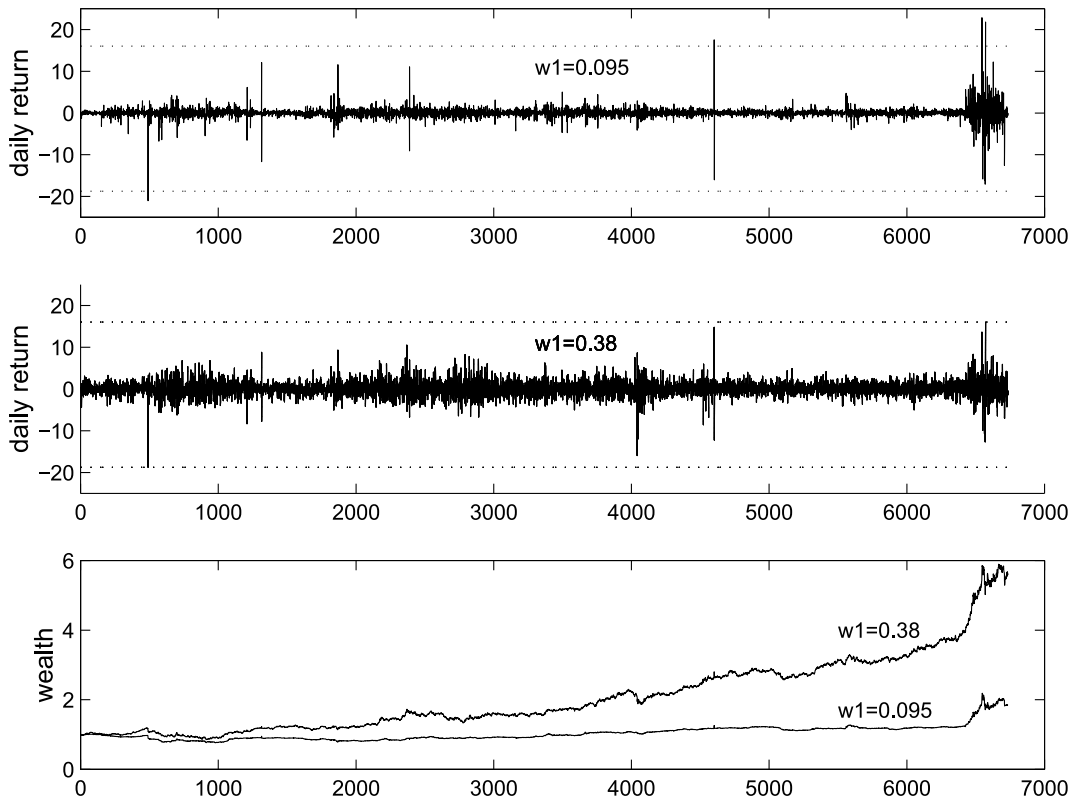
In order to illustrate this point, consider Fig. 2 which illustrates a portfolio of just two assets. The topmost plot correspond to a portfolio chosen according to Markovitz’s theory, whereas the plot in the middle row shows a portfolio which has been chosen so that it does not correspond to minimum variance, but rather to the minimum of higher order moments (for a detailed description see [2]). Comparing the two plots, it is clear by simple visual inspection that the Markovitz portfolio does a good job *most* of the time in ensuring that only small risks (i. e., fluctuations of the portfolio) are taken. Similarly, choosing the weights in the portfolio by minimizing higher order moments, it is clear that one opens oneself up for taking more small and intermediate size risks (seen by, in average, larger fluctuations compared to the Markovitz solution). However, and this is the important point, by minimizing the higher order moments of the portfolio distribution, one minimizes large events that takes place in the tail of the distribution. So by accepting more small and intermediate risks compared to the Markovitz solution, one avoids taking the big

risks. Specifically, this can be seen by the band contained by the dotted lines, which are the same in the topmost and middle plots. Even though the portfolio showing the Markovitz solution is constructed so that small and intermediate risks are contained *most* of the time, it actually gives a false impression of control, since one thereby becomes susceptible to the big risks seen by the events that exceed the bounds of the dotted lines in the topmost plot. On the contrary, accepting more small and intermediate risks, as for the portfolio seen in the middle plot, one steers free of the same big risks that were encountered in Markovitz’s solution. It is remarkable that, by decreasing large risks, one can, at the same time, profit from such a choice of portfolio. This is seen in the lower plot of Fig. 2, where a fourfold gain in comparison with the Markovitz solution is obtained for the portfolio that limits large risks.

The strongest objection against the CAPM and the Markovitz solution, however, lies in its foundation, since it is assumed that *all* market participants if given access to a complete information set would come to the same conclusion with regard to price. That is, there is complete agreement among investors about the joint distribution of asset returns from t to $t + 1$. In physics, such an assumption goes under the name of a “mean-field” theory, a somewhat surprising application for a system (the financial market) whose complexity by far exceeds the most complex system encountered in the realm of physics. The fact that the hypothesis of rational expectations, with its implication of “complete agreement among investors”, lays the groundwork for much of the modern work done in the theory of finance, seems problematic. A recent and very clear criticism of this point is illustrated by the following excerpts from an article by Phillip Ball (editor of Nature) published in Financial Times, 29th of October 2006:

It is easy to mock economic theory. Any fool can see that the world of neoclassical economics, which dominates the academic field today, is a gross caricature in which every trader or company acts in the same self-interested way rational, cool, omniscient. The theory has not foreseen a single stock market crash and has evidently failed to make the world any fairer or more pleasant.

The usual defense is that you have to start somewhere. But mainstream economists no longer consider their core theory to be a start. The tenets are so firmly embedded that economists who think it is time to move beyond them are cold-shouldered. It is a rigid dogma. To challenge these ideas is to invite blank stares of incomprehension you might as well be telling a physicist that gravity does not exist.



Investment Decision Making in Finance, Models of, Figure 2

Daily returns and cumulative wealth (in percentage) for a portfolio composed of two assets, Chevron stock (with weight w_1) and Malaysian ringgit (with weight $1 - w_1$). The topmost plot shows daily return for a Markovitz mean-variance portfolio ($w_1 = 0.095$), whereas the second plot shows returns for a portfolio that minimizes higher order moments (for a detailed explanation see [2]). Cumulative wealth for both choices of portfolio is shown in the bottom plot

That is disturbing because these things matter. Neo-classical idiocies persuaded many economists that market forces would create a robust post-Soviet economy in Russia (corrupt gangster economies do not exist in neoclassical theory). Neoclassical ideas favouring unfettered market forces may determine whether Britain adopts the euro, how we run our schools, hospitals and welfare system. If mainstream economic theory is fundamentally flawed, we are no better than doctors diagnosing with astrology...

Paul Ormerod, author of *The Death of Economics*, argues that one of the most limiting assumptions of neoclassical theory is that agent behavior is fixed: people in markets pursue a single goal regardless of what others do. The only way one person can influence another's choices is via the indirect effect of trading on prices. Yet it is abundantly clear that herding – irrational, copycat buying and selling – provokes market fluctuations.

There are ways of dealing with the variety and irrationality of real agents in economic theory. But not in mainstream economics journals, because the models defy neoclassical assumptions.

There is no other "science" in such a peculiar state. A demonstrably false conceptual core is sustained by inertia alone. This core, "the Citadel", remains impregnable while its adherents fashion an increasingly baroque fantasy. As Alan Kirman, a progressive economist, said: "No amount of attention to the walls will prevent the Citadel from being empty."

Irrational Decision Making

Whereas rational decision making, especially with the idea of the efficient market theory, was well established in the academic world of finance in the 1970s, questions about the validity of the theory surfaced in the 1980s, prompted by a succession of discoveries of anomalies, notably the

evidence of excess volatility of returns [36]. Since the reported excess volatility raised some of the very first empirically founded questions related to the efficient market theory, it became the subject of harsh academic disputes. The idea behind it will be explained in detail in this section. Following the problems made clear by the puzzle of “excess volatility”, competing theories to rational decision making and the efficient market theory surfaced in the 1990s, most notably by the field now known as behavioral finance. The emerging field of behavioral finance, in turn, led to more recent theories such as feedback and herding, as well as models of interaction between smart money and ordinary investors.

“Excess Volatility”

Created by Irrational Decision Making

To illustrate the problem of so-called “excess volatility”, notice that it follows from (1) that $P_t^* = P_t + U_t$ where U_t is a forecast error. In this equation P_t is known at time t , whereas U_t, P_t^* are not known. When making a forecast at time t , one is assumed to have access to all available information. Since the forecast is assumed to be optimal, U_t therefore has to be uncorrelated with any information variable at time t . P_t , however, is itself known information at time t , so P_t and U_t must be uncorrelated. But this means that

$$\sigma^2(P_t^*) = \sigma^2(P_t) + \sigma^2(U_t) \quad (12)$$

where σ^2 denotes the variance (the standard deviation squared) of a variable. Since the variance of a variable is always positive, this means that:

$$\sigma^2(P_t^*) > \sigma^2(P_t). \quad (13)$$

(13) gives a “handle” to test the efficient market theory. We can use historical price data to construct P_t^* from the actual dividend paid from time t till the present time, discounted by the actual interest rates, and then compare the variance of P_t^* to the variance of the price P_t over the same time period. In ([37]), historical data was analyzed in this manner from year $t = 1871$ till year 2002 of the Standard & Poor’s Composite Stock Price Index. Clear evidence showed that (13) was violated, thus giving rise to a remarkable amount of controversy within the academic community. This so called “excess volatility” is still one of the often debated “puzzles” in finance, and gave one of the first clear empirical demonstrations that one must go beyond the rational decision making framework to gain a proper understanding of financial data.

Example of a Model

That Can Not Be Solved Using Rational Expectations

It should be noted that even within a theoretical framework, some problems are simply not solvable using rational expectations. This was pointed out by B. Arthur [5] in what has been coined “the El Farol bar problem”. The El Farol bar problem is a game in which N people must decide independently each week whether to show up (or not) at their favorite bar, with the constraint that the bar has only $L < N/2$ chairs. Since each person will only be happy when seated, people use last week’s attendance to predict future attendance. If a person predicts that the number of people that will attend the bar is larger than L , that person will stay home. It is important to notice that the game describes adaptive behavior, as the people adapt their behavior to prior attendance. Also, it gives a clear illustration of a case where rational expectations *cannot* be used to find a solution. This can be seen as follows: suppose that a rational expectation prediction machine existed and that all agents possessed a copy of it. If, in the example given, the machine predicts that a number larger than L will attend the bar, nobody will show up, thereby negating the prediction of the prediction machine. The El Farol bar problem, in turn, led to a new parsimonious formulation of financial markets seen as a complex system and now known under the name “the Minority Game”, see Sect. “Investment Decision Making in a Complex Environment”.

Behavioral Finance

Decision making became a research topic in the field of psychology in the 1950s in the work of Edwards [15] as well as that of H. A. Simon, who introduced the concept of decision making based on *bounded* rationality [38]. It was, however, not until the work of Daniel Kahneman and Amos Tversky (deceased in 1996) that results from cognitive psychology found their way into economy and finance.

Prospect Theory Prospect theory, introduced by Kahneman and Tversky [22] (and for which Kahneman received the Nobel prize in 2002), takes into account the fact that people in general are unable to fully analyze situations that involve economic and probabilistic judgments. Specifically, prospect theory makes three assumptions which deviate from the standard neoclassical framework of economic rational decision making:

- When investing, people are often sensitive to some reference level. Say an investment decision is likely to be influenced by the absolute wealth possessed by an in-

vestor. The satisfaction/dissatisfaction by a gain/loss of an investor is related to the wealth of an investor and not just the absolute amount gained/lost.

- A rational person would act symmetrically with respect to gains versus losses of same size, that is, a rational person would be equally pleased/annoyed. People, however, appear to be more averse to losses than attracted by gains of the same size. This appears to be especially true for small losses/gains: even the slightest loss is conceived as very bad, whereas a small gain is not considered equally satisfying.
- People have the tendency to overweigh events that occur very rarely and to underweigh events which occur with large probabilities. E.g., people keep on buying lottery tickets despite the fact that their chance of winning is almost nil!

Anchoring and Other Irrational Beliefs Used in Decision Making

Anchoring is a term used in psychology to describe the case in which human decision making relies (anchors) on just one piece of (often irrelevant) information. One of the first observations of anchoring was reported in an experiment by Tversky and Kahneman. In [41], two groups of test persons were shown to give different mean estimates of the percentage of African nations in the United Nations, depending on the specific anchor of percentage suggested by the experimenters to the two groups. Evidence for human anchoring has since been reported in many completely different domains such as e.g., customer inertia in brand switching [44] (old brand price acting as an anchor), whereas other evidence comes from studies on on-line auctions [13] (people bid more for an item the higher the “buy-now” price) and anchoring in real estate prices [28] (subjects’ appraisal values depend on an arbitrary posted listing price of the house). In the context of financial markets, anchoring has been observed via the so called “disposition effect” [35], [21], which is the tendency for people to sell assets that have gained value and keep assets that have lost value. As noted in [43], conclusive tests using real market data are usually difficult because the investors’ expectations, as well as individual decisions, cannot be controlled or easily observed. In experimental security trading, however, subjects were observed to sell winners and keep losers [43].

In [1], another method was analyzed which rigorously tests for anchoring in the decision making of humans when they trade in the stock market. The test, heavily inspired by interdisciplinary approaches, was introduced in a trading algorithm by L. Gil [20], which was inspired by the way biological motors work by exploiting favor-

able Brownian fluctuations to generate directed forces and move¹. In [1], weekly data of the CAC40 and the Dow Jones index were analyzed to test for the claim that market participants, by actively following the price, thereby create a subjective reference (anchor) and memory of when an asset is “cheap” or “expensive”. Several studies on the persistence of human memory have reported that sleep, as well as post-training wakefulness before sleep, play an important role in the offline processing and consolidation of memory [29]. It therefore makes sense to think that conscious as well as unconscious mental processes influence the judgments of people who specialize in active trading on a day-to-day basis. The out-of-sample profit from the market-neutral trading algorithm (with transaction costs taking into account) on the CAC40 index as well on the Dow Jones index, gives evidence that anchoring does, indeed, play a dominant role on the weekly pricing of the Dow Jones and CAC40 stock markets.

Examples of other human beliefs that can lead to irrational decision making are (see e.g., ([7]) for a longer discussion)

- **Overconfidence.** People are overconfident when it comes to decision making. A typical bias is that people have the tendency to ascribe any good outcomes to their own talents, while blaming bad outcomes on external circumstances.
- **Hindsight bias.** The tendency to think one predicted what had happened, whereas in reality it was only realized after the fact.
- **Framing.** Refers to the case where the *exact same problem* has different outcomes in decision making, depending on how the problem is described.
- **Confirmation bias.** The tendency to search for or interpret information in a way that confirms one’s preconceptions.
- **Wishful Thinking.** Most people display unrealistically high self-esteem. Typically 90% of all car drivers consider themselves above average, whereas only 50% should think so.

Investment Decision Making in a Complex Environment

Decision making in a complex environment refers to the case in which the decision of an investor is directly influenced by the outcome of actions taken by other decision makers. A day trader in a stock market is one such

¹Similar ideas were also introduced in [40] where it was shown how increments of uncorrelated time series can be predicted with a universal 75% probability of success.

example, since the decision of when to enter/exit a position depends on price trajectories created by other day traders. The El Farol bar game mentioned in Sect. “[Example of a Model that can not be Solved Using Rational Expectations](#)” is another example, since a person will use the prior attendance of other people in the decision making on whether or not to go to the bar.

Some of the first approaches by economists to model decision making in a complex environment were put forth by A. S. Kyle [23], J. A. Frankel and K. A. Froot [17] and J. Bradford de Long, A. Schleifer, L. H. Summers and R. J. Waldmann [10]. Kyle studied how quickly private information of a given commodity is incorporated into market prices via a dynamic model with three types of traders: insider trader, noise traders and market makers. Frankel and Froot introduced a model of the currency markets with three types of actors: fundamentalists, chartists and portfolio managers. As noted in their abstract, their motivation to go beyond the standard rational expectation theory was that “... the proportion of exchange rate movements that can be explained even *after* the fact, using contemporaneous macroeconomic variables, is disturbingly low”. Bradford de long et al. came to the conclusion, in a model of rational and noise traders, that noise traders in fact can survive and dominate the market in the long run.

More recently, models of decision making in complex environments has become a major research field for people working, notably, in statistical physics. Some of the first work that sparked a lot of interest in this field was e. g., [11,25], which viewed the marketplace as a self-organizing complex system. Later work [12] extended the El Farol game to study adaptive behavior in financial markets. The Minority Game (MG) was introduced by two physicists, Y.C. Zhang and D. Challet (1997) as a “minimal” model for financial markets (it now has its own website: www.unifr.ch/econophysics [12]).

The MG describes a model in which some agents (market participants) trade in a given market by placing buy/sell orders. A decision is made to enter a position in the market based on the last m price directions of the market. As in the El Farol game, adaptation is a part of the game, since the agents always choose the best performing strategy (defined below) among a pool of strategies. As the market changes, the strategies used by the agents will change. This, in turn, will lead to new changes of the market, illustrating feedback which is thought to be an important part in real market price dynamics. The main idea behind the MG is to introduce as parsimonious a model of financial market dynamics as possible, while sorting out the most important aspects of price formation in a market. Three parameters were suggested in such a description:

- N Number of agents (market participants). This parameter takes into account the assumption that in order to get a proper description of a market, one need only know the price impact of the most important players, such as, e. g., major banks, hedge funds, pension funds, brokerage houses, etc.
- m “Memory” of the strategies (assumed in the simplest version of the MG to be the same for all strategies) used by the agents. The idea is that decision making is based on strategies that use the last m price movements. This corresponds to agents using technical analysis. In its simplest version the MG, therefore, considers only short time scales, where the impact of changes in the fundamental price can be ignored.
- s_i Number of strategies held by agent i . Market participants try out different strategies in order to perform optimally. As the market changes, the strategy which is optimal will eventually also change.

The strategies in the MG describe how an investment decision is made depending on the way the market has performed over the last m time steps. E.g., if $m = 3$, so that only the last three time steps (say days) are used in decision making, a strategy will tell what to do in all possible cases of how the market performed. Formally, one can represent the performance of the market by assigning, say 0 to a down movement of the market, and 1 if the market moved up. So if the market went down, down, down over the last three time days, then this can be represented by a binary string (000). If, on the other hand, the market first went up, then down and then up, one can formally represent this by the binary string (101). Using such a representation, one has 2^m different price histories (i. e., 2^m different ways to describe how the market performed over the last m time steps). A strategy, then, is defined in terms of how to make a decision for each of the 2^m possible ways the market can have gone over the last m time steps. Specifically a strategy might look like:

Investment Decision Making in Finance, Models of, Table 1

Price history	Decision
000	1
001	0
010	0
011	1
100	0
101	0
110	1
111	0

This specific strategy tells you to buy if the market over the last m time steps went down, down, down, to sell if the market over the last three time steps went down, down, up, etc. However, since each price history must give a decision whether to buy or sell (i. e., 0/1) the total number of such strategies is therefore given by $S = 2^{2^m}$. This is a *very* big number even for a moderate number of time steps that one might use in the decision process. E.g., using daily market data, and considering an investment decision of whether to enter/exit (i. e., buy/sell) the market over two weeks (i. e. $m = 10$ days), the total number of possible strategies (i. e., decisions) goes like $2^{2^{10}} \approx 10^{200}$ which by far exceeds the total number ($\approx 10^{80}$) of elementary particles (i.e, quarks (the constituents of neutrons/protons), leptons, bosons) thought to exist in the whole universe! Despite the highly simplified framework of technical analysis, this huge number still gives you a hint as to the tremendous complexity a financial investment decision exhibit, and foretells some of the difficulties to be encountered when analyzing the model.

The market mechanism introduced in the MG is a minority game: strategies which took the minority action were rewarded, whereas the majority of optimal strategies that were actually used by the agents lose. At every time step, each agent uses his most successful (among the s) strategies. If one calls the decision at time t of the j th strategy of agent i for a_i^j and the optimal strategy at time t of agent i for a_i^* , then the payoff function of any given strategy is taken as:

$$u_i^j(t) = -a_i^j(t)A(t), \quad A(t) = \sum_i a_i^*(t). \quad (14)$$

The price dynamics, $p(t)$, is then expressed in terms of the excess demand:

$$\log p(t+1) = \log p(t) + A(t)/\lambda \quad (15)$$

with λ the “depth” (or liquidity) of the market. The MG captures some behavioral phenomena that are thought to be of importance in financial markets, such as:

- competition
- frustration
- adaptability

and

- evolution.

Since the introduction of the Minority Game many extensions have been suggested (the history of the work done on this model can be found at the website in [12]). E.g., it seems natural to model the case with agents trying to

profit from market movements instead of striving always be in the minority. To do so, they will take a position at time t at, say, price $p(t)$, but will know only at time $t+1$ whether that decision meant a profit or a loss [3,19]. From (15), the return is proportional to $A(t)$, so the payoff function (14) in that case becomes $u_i^j(t) = a_i^j(t-1)A(t)$, i. e., the payoff function changes sign and is evaluated over two time steps. It turns out that the solution in that case *can* be solved using rational expectations [32], in contrast to the MG where no such solution is available. Without any constraint in buying power, it becomes rational for the agents to try to figure out what the other agents will do, i. e. the solution reflects Keynes’s beauty queen contest. Interestingly bubbles in this case are created spontaneously, something missing in the MG.

Future Directions

Beyond doubt, any proper description of how decision making in finance really takes place will have to involve aspects from all three of the above-mentioned decision making contexts, that is, rational decision making, irrational decision making, and decision making in a complex environment. Also, a multi disciplinary effort seems to be an absolute necessity. Financial markets are made up of humans, not robots with perfect insight, and with the complexity and psychology often involved in human decision making, we face an enormous challenge in the development of new tools to do proper risk management.

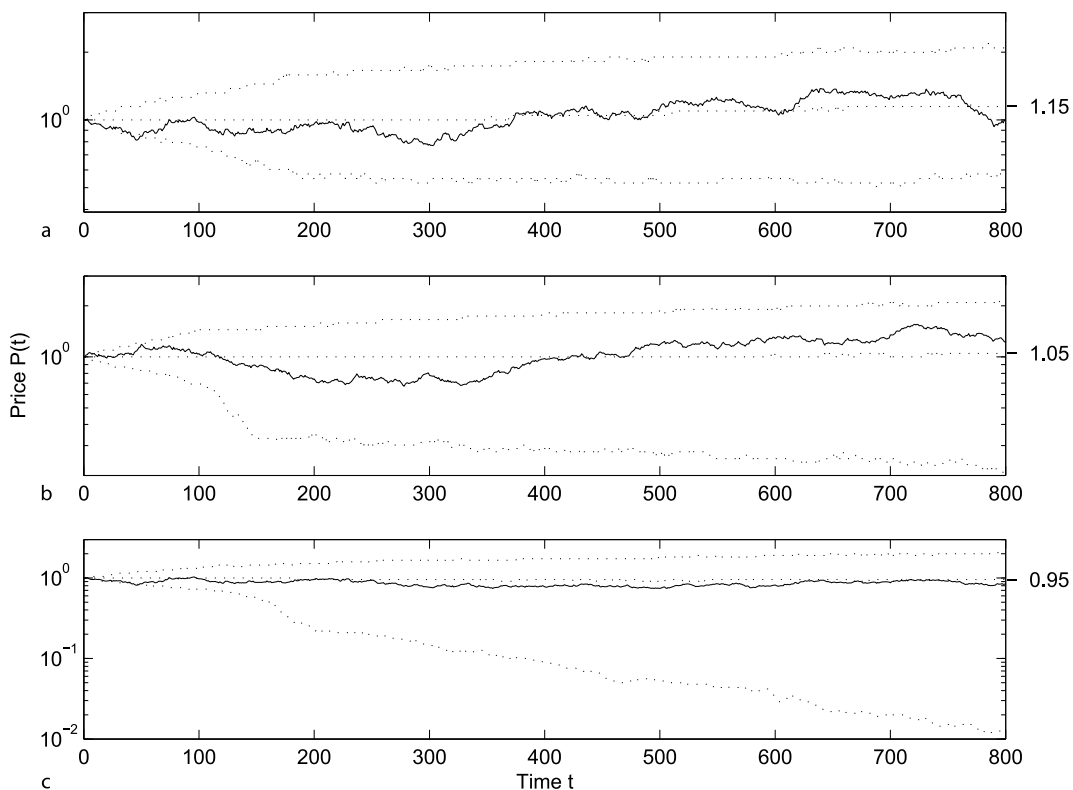
Behavioral finance laid the first bricks for interdisciplinary approaches between psychology and finance. Econophysics has made attempts to further bridge finance with new research methods found in another field. Such approaches, where new ideas come across to an already established field, however, do not always happen without resistance from researchers using conventional methods. For the time being, the relevance of econophysics is under debate (see e. g. [6,14,18]) just as behavioral finance was when it first appeared. Whereas criticism is good and constructive, exclusion is bad. As mentioned in a discussion about econophysics by the progressive economist B. Arthur in [6]: “They tried to publish in mainstream economics journals, but were rebutted. It’s a shame and a scandal that the journals haven’t opened their pages to them”. Talking about future directions, this does not seem like the path to take.

In general, it would be very interesting to gain firmer understanding of how aggregation of individual behavior can give rise to measurable effects in a population in general, and in financial markets in particular. It would also be interesting to model specific human traits on a micro

scale and study the emergence of a dynamics with observable or even predictable effects on a macro scale. The hope would be to reproduce in models many of the mechanisms reported at work in behavioral finance. One step in this direction was made in [4], where it was shown how consensus (called “decoupling” in [4]), and thereby predictability, could emerge due to mutual influence of the price in a commonly traded asset, among a group of agents who had initially different opinions. In [42], a very original study was done on self-referential behavior and the impact it had on overreaction and conventions in financial markets.

From the point of view of a financial market as a complex system, it would be interesting to sort out some of the important factors where new ways of investment decisions could have determining impact on the price dynamics of the market. In [45], it was suggested that the exponential growth of hedge funds which has happened since the start of the twenty-first century could have a catastrophic impact on the *long term* trends of financial markets. The ex-

planation for this is the role of traditional market players of long-only mutual funds versus hedge funds which take both short and long positions. In [45], it was argued that financial markets since their very origin, and only till very recently, have been in a state of “broken symmetry” which favored long term growth instead of contraction. The reason for this “broken symmetry” into a long term “bull phase” is the historical, almost complete, dominance by long-only players in financial markets. Dangers connected to short trading can be seen in Fig. 3, which shows simulations of a financial market with varying percentages of market participants taking short positions. As illustrated in the figure, two predictions come out concerning the impact of the new hedge fund players in the markets: over long time horizons, one should see an increase in volatility and, more importantly, an increase in the probability for periods with “bear” markets. If, as is often speculated, one has a spillover from the financial markets into the economy (through the so called “wealth effect”), the introduction of hedge funds that accelerated from the beginning



Investment Decision Making in Finance, Models of, Figure 3

Fat solid line: price $P(t)$ generated from an agent based model (see [45] for a definition of the model) **Thin dotted lines** represent the 5%, 50% and 95% quantiles (from bottom to top) respectively, i. e., at every time t , out of the 1000 different initial configurations, only 50 got below the 5% quantile line, and similarly for the other quantile lines. **a** The fraction of agents allowed to take short positions, $\rho = 0$, **b** $\rho = 0.2$ and **c** $\rho = 0.4$

of this century could have dire consequences. This message, however, is something which is difficult to communicate to researchers solidly implanted in rational expectation theory.

Bibliography

Primary Literature

- Andersen JV (2007) Detecting anchoring in financial markets. Submitted to: Quant Financ. The article can be retrieved from the site: <http://arXiv.org/abs/0705.3319>. Accessed 2007
- Andersen JV, Sornette D (2001) Have your cake and eat it too: increasing returns while lowering large risks! *J Risk Finance*, p 70; Sornette D, Andersen JV, Simonetti P (2000) Minimizing volatility increases large risks. *Int J Theor Appl Financ* 3(3):523–535; Sornette D, Simonetti P, Andersen JV (2000) φ^q -field theory for portfolio optimization: fat tails and nonlinear correlations. *Phys Rep* 335:19–92
- Andersen JV, Sornette D (2003) The $\$$ -game. *Eur Phys J B* 31:141
- Andersen JV, Sornette D (2006) A Mechanism for Pockets of Predictability in Complex Adaptive Systems. *Europhys Lett* 70:697; The findings of the article is described in: Buchanan M (2005) Too much information. *New Sci* 85:32–35
- Arthur B (1994) Bounded rationality and inductive behavior. *Am Econ Rev, Papers and Proceedings* 84:406
- Ball P (2006) Culture Crash. *Nature* 441:686–688
- Barberis N, Thaler R (2003) A Survey of Behavioral Finance. *Handbook of the Economics of Finance*. North Holland, Amsterdam, pp 1053–1128
- Black F, Scholes M (1973) The pricing of options and corporate liabilities. *J Political Econ* 81:637–654; Black F (1989) How we came up with the option formula. *J Portf Manag* 15:4–8; Merton RC (1973) Theory of rational option pricing. *Bell J Econ Manag Sci* 4:141–183
- Bofinger P, Schmidt R (2004) Should One Rely on Professional Exchange Rate Forecasts? An Empirical Analysis of Professional Forecasts for the ϵ /US $\$$ -Rate. Discussion Paper Series - Centre For Economic Policy Research London, ISSU 4235
- Bradford de Long J, Schleifer A, Summers LH, Waldmann RJ (1990) The survival of noise traders in financial markets. *J Bus* 64(1):1–19
- Caldarelli G, Marsili M, Zhang YC (1997) A prototype model of stock exchange. *Europhys Lett* 40:479–484
- Challet D, Zhang Y-C (1997) Emergence of cooperation and organization in an evolutionary game. *Physica A* 246:407–418; To follow the literature of the MG see e.g. the special website www.unifr.ch/econophysics/minority/minority.html, and the recent book: Challet D, Marsili M, Zhang Y-C (2004) *Minority Games: Interacting Agents In Financial Markets*. Oxford University Press, Oxford
- Dodonova A, Khoroshilov Y (2004) Anchoring and transaction utility: evidence from on-line auctions. *Appl Econ Lett* 11:307
- Durlauf SN (2005) Complexity and empirical economics. *Econ J* 115:F225–F243
- Edwards W (1954) Behavioral decision theory. *Annu Rev Psychol* 12:473–98
- Fama EF, French KR (2004) The capital asset pricing model: theory and evidence. *J Econ Perspect* 18(3):25–46
- Frankel JA, Froot KA (1990) Chartists, fundamentalists and the demand for dollars. In: Courakis AS, Taylor MP (eds) *Private Behavior and Government Policy in Interdependent Economies*. Oxford University press, Oxford, pp 73–126
- Gallegatti M, Ken S, Lux T, Ormerod P (2006) Worrying trends in econophysics. *Physica A* 370:1–6
- Giardina I, Bouchaud J-P (2003) Bubbles, crashes and intermittency in agent based models. *Eur Phys J B* 31:421–437
- Gil L (2007) A simple algorithm based on fluctuations to play the market. [arXiv:0705.2097](http://arXiv.org/abs/0705.2097)
- Heilmann K, Laeger V, Oehler A (2000) The Disposition Effect: Evidence about the Investors Aversion to Realize Losses. In: *Proceedings of the 25th Annual Colloquium, Baden/Vienna, 12–16 July 2000*. IAREP, Wien
- Kahneman D, Tversky A (1979) Prospect theory: An analysis of decision under risk. *Econometrica* 47:263–291; Kahneman D, Tversky A (1972) Subjective probability: A judgement of representativeness. *Cogn Psychol* 3:430–454; Kahneman D, Tversky A (1973) On the psychology of prediction. *Psychol Rev* 80:237–251
- Kyle AS (1985) Continuous auctions and insider trading. *Econometrica* 53(6):1315–1336
- Lintner J (1965) The valuation of risk assets and selection of risky investments in stock portfolios and capital budgets. *Rev Econ Stat* 47(1):13–37
- Lux T, Marchesi M (1999) Scaling and criticality in a stochastic multi-agent model of a financial market. *Nature* 397:498–500
- Markovitz H (1959) *Portfolio selection: Efficient diversification of investments*. John Wiley and Sons, New York
- Modigliani F, Miller MH (1958) The cost of capital, corporation finance and the theory of investment. *Am Econ Rev* 48:261–297
- Northcraft GB, Neale MA (1987) Experts, amateurs, and real estate: an anchoring and adjustment perspective on property pricing decisions. *Organ Behav Hum Decis Process* 39:84
- Peigneux P, Orban P, Balteau E, Degueldre C, Luxen A (2006) Offline Persistence of Memory-Related Cerebral Activity during Active Wakefulness. *PLoS Biology* 4(4):e100 [doi:10.1371/journal.pbio.0040100](https://doi.org/10.1371/journal.pbio.0040100)
- Roll R (2002) Rational infinitely lived asset prices must be non-stationary. *J Bank Financ* 26(6):1093–1097
- Ross S (1976) The arbitrage theory of capital asset pricing. *J Econ Theory* 13(3):341–360
- Roszczynska M, Nowak A, Kamieniarz D, Solomon S, Andersen JV (2008) Detecting speculative bubbles created in experiments via decoupling in agent based models. Available at <http://arxiv.org/abs/0806.2124>
- Samuelson PA (1965) Proff that properly anticipated prices fluctuate randomly. *Ind Manag Rev* 6:41–49
- Sharpe WF (1964) Capital asset prices: a theory of market equilibrium under conditions of risk. *J Financ* 19(3):425–442
- Shefrin HM, Statman M (1985) The disposition to sell winners too early and rise losers too long. *J Financ* 40:777
- Shiller R (1981) Do Stock Prices Move Too much to be Justified by Subsequent Changes in Dividends? *Am Econ Rev* 71:421–426
- Shiller RJ (2003) From Efficient Market Theory to Behavioral Finance. *J Econ Perspect, Am Econ Assoc* 17(1):83–104. http://papers.ssrn.com/abstract_id=349660
- Simon HA (1955) A behavioral model of rational choice.

- Q J Econ 69(1):99–118; Simon HA (1956) Rational choice and the structure of the environment. Psychol Rev 63:129–138
39. Smith VL (1962) An experimental study of competitive market behavior. J Political Econ 70:111–137; Smith VL (1965) Experimental auction markets and the Walrasian hypothesis. J Political Econ 73:387–393; Smith VL (1976) Experimental economics: Induced value theory. Am Econ Rev, Papers and Proceedings, pp 274–279; Plott C, Smith VL (1978) An experimental examination of two exchange institutions. Rev Econ Stud 45:133–153
 40. Sornette D, Andersen JV (2000) Increments of uncorrelated time series can be predicted with a universal 75% probability of success. Int J Mod Phys C 11:713
 41. Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. Science 185:1124
 42. Wyart M, Bouchaud J-P (2007) Self-referential behavior, overreaction and conventions in financial markets. J Econ Behav Organ 63:1–24
 43. Weber M, Camerer CF (1998) The disposition effect in securities trading: An experimental analysis. J Econ Behav Organ 33:167
 44. Ye G (2004) Inertia Equity: The impact of anchoring price in brand switching. SSRN-id548862
 45. Andersen JV (2005) Could Short Selling Make Financial Markets Tumble? Int J Theor App Fin 8(4):509–521

Books and Reviews

- Bouchaud JP, Potters M (2000) Theory of Financial Risks. Cambridge University Press, Cambridge
- Johnson NF, Jefferies P, Hui PM (2003) Financial Market Complexity. Oxford University Press, Oxford
- Mantegna RN, Stanley HE (2000) An Introduction to Econophysics: Correlations and Complexity in Finance. Cambridge University Press, Cambridge
- McCauley JL (2004) Dynamics of Markets: Econophysics and Finance. Cambridge University Press, Cambridge
- Roehner BM (2002) Patterns of Speculation: A Study in Observational Econophysics. Cambridge University Press, Cambridge
- Sornette D (2003) Why Stock Markets Crash (Critical Events in Complex Financial Systems). Princeton University Press, Princeton

Isomorphism Theory in Ergodic Theory

CHRISTOPHER HOFFMAN

Department of Mathematics, University of Washington,
Seattle, USA

Article Outline

Glossary
 Definition of the Subject
 Introduction
 Basic Transformations
 Basic Isomorphism Invariants
 Basic Tools

Isomorphism of Bernoulli Shifts
 Transformations Isomorphic to Bernoulli Shifts
 Transformations not Isomorphic to Bernoulli Shifts
 Classifying the Invariant Measures of Algebraic Actions
 Finitary Isomorphisms
 Flows
 Other Equivalence Relations
 Non-invertible Transformations
 Factors of a Transformation
 Actions of Amenable Groups
 Future Directions
 Bibliography

Glossary

Almost everywhere A property is said to hold **almost everywhere (a.e.)** if the set on which the property does not hold has measure 0.

Bernoulli shift A Bernoulli shift is a stochastic process such that all outputs of the process are independent.

Conditional measure For any measure space (X, \mathcal{B}, μ) and σ -algebra $\mathcal{C} \subset \mathcal{B}$ the conditional measure is a \mathcal{C} -measurable function g such that $\mu(C) = \int_C g \, d\mu$ for all $C \in \mathcal{C}$.

Coupling of two measure spaces A coupling of two measure spaces (X, μ, \mathcal{B}) and (Y, ν, \mathcal{C}) is a measure γ on $X \times Y$ such that $\gamma(B \times Y) = \mu(B)$ for all $B \in \mathcal{B}$ and $\gamma(X \times C) = \nu(C)$ for all $C \in \mathcal{C}$.

Ergodic measure preserving transformation A measure preserving transformation is ergodic if the only invariant sets $(\mu(A \Delta T^{-1}(A)) = 0)$ have measure 0 or 1.

Ergodic theorem The pointwise ergodic theorem says that for any measure preserving transformation (X, \mathcal{B}, μ) and T and any L^1 function f the time average

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n f(T^i(x))$$

converges a.e. If the transformation is ergodic then the limit is the space average, $\int f \, d\mu$ a.e.

Geodesic A geodesic on a Riemannian manifold is a distance minimizing path between points.

Horocycle A horocycle is a circle in the hyperbolic disk which intersects the boundary of the disk in exactly one point.

Invariant measure Likewise a measure μ is said to be invariant with respect to (\mathbf{X}, T) provided that $\mu(T^{-1}(A)) = \mu(A)$ for all measurable $A \in \mathcal{B}$.

Joining of two measure preserving transformations

A joining of two measure preserving transformations (\mathbf{X}, T) and (\mathbf{Y}, S) is a coupling of \mathbf{X} and \mathbf{Y} which is invariant under $T \times S$.

Markov shift A Markov shift is a stochastic process such that the conditional distribution of the future outputs $(\{x_n\}_{n>0})$ of the process conditioned on the last output (x_0) is the same as the distribution conditioned on all of the past outputs of the process $(\{x_n\}_{n\leq 0})$.

Measure preserving transformation A measure preserving transformation consists of a probability space (X, T) and a measurable function $T: X \rightarrow X$ such that $\mu(T^{-1}(A)) = \mu(A)$ for all $A \in \mathcal{B}$.

Measure theoretic entropy A numerical invariant of measure preserving transformations that measures the growth in complexity of measurable partitions refined under the iteration of the transformation.

Probability space A probability space $\mathbf{X} = (X, \mu, \mathcal{B})$ is a measure space such that $\mu(B) = 1$.

Rational function, rational map A rational function $f(z) = g(z)/h(z)$ is the quotient of two polynomials. The degree of $f(z)$ is the maximum of the degrees of $g(z)$ and $h(z)$. The corresponding rational maps $T_f: z \mapsto f(z)$ on the Riemann sphere \mathbb{C} are a main object of study in complex dynamics.

Stochastic process A stochastic process is a sequence of measurable functions $\{x_n\}_{n \in \mathbb{Z}}$ (or outputs) defined on the same measure space, \mathbf{X} . We refer to the value of the functions as outputs.

Definition of the Subject

Our main goal in this article is to consider when two measure preserving transformations are in some sense different presentations of the same underlying object. To make this precise we say two measure preserving maps (\mathbf{X}, T) and (\mathbf{Y}, S) are **isomorphic** if there exists a measurable map $\phi: X \rightarrow Y$ such that

- (1) ϕ is measure preserving,
- (2) ϕ is invertible almost everywhere and
- (3) $\phi(T(x)) = S(\phi(x))$ for almost every x .

The main goal of the subject is to construct a collection of invariants of a transformation such that a necessary condition for two transformations to be isomorphic is that the invariant be the same for both transformations. Another goal of the subject is to solve the much more difficult problem of constructing invariants such that the invariants being the same for two transformations is a sufficient condition for the transformation to be isomorphic. Finally we apply these invariants to many natural classes of transformation to see which of them are (or are not) isomorphic.

Introduction

In this article we look at the problem of determining of which measure preserving transformations are isomorphic. We look at a number of isomorphism invariants, the most important of which is the Kolmogorov–Sinai entropy. The central theorem in this field is Ornstein’s proof that any two Bernoulli shifts of the same entropy are isomorphic. We also discuss some of the consequences of this theorem, which transformations are isomorphic to Bernoulli shifts as well as generalizations of Ornstein’s theory.

Basic Transformations

In this section we list some of the basic classes of measure preserving transformations that we study in this article.

Bernoulli shifts Some of the most fundamental transformations are the Bernoulli shifts. A **probability vector** is a vector $\{p_i\}_{i=1}^n$ such that $\sum_{i=1}^n p_i = 1$ and $p_i \geq 0$ for all i . Let $\mathbf{p} = \{p_i\}_{i=1}^n$ be a probability vector. The **Bernoulli shift** corresponding to \mathbf{p} has state space $\{1, 2, \dots, n\}^{\mathbb{Z}}$, the shift operator $T(x)_i = x_{i+1}$. To specify the measure we only need to specify it on **cylinder sets**

$$A = \{x \in X: x_i = a_i \ \forall i \in \{m, \dots, k\}\}$$

for some $m \leq k \in \mathbb{Z}$ and a sequence $a_m, \dots, a_k \in \{1, \dots, n\}$. The measure on cylinder sets is defined by

$$\mu\{x \in X: x_i = a_i \text{ for all } i \text{ such that } m \leq i \leq k\} = \prod_{i=m}^k p_{a_i}.$$

For any $d \in \mathbb{N}$ if $\mathbf{p} = (1/d, \dots, 1/d)$ we refer to $\text{Bernoulli}_{\mathbf{p}}$ as the **Bernoulli d shift**.

Markov shifts A **Markov shift** on state n symbols is defined by an $n \times n$ matrix, M , such that $\{M(i, j)\}_{j=1}^n$ is a probability vector for each i . The Markov shift is a measure preserving transformation with state space $\{1, 2, \dots, n\}^{\mathbb{Z}}$, transformation $T(x)_i = x_{i+1}$

$$\mu\{x_1 = a_1 \mid x_0 = a_0\} = \mu\{x_1 = a_1 \mid x_0 = a_0, x_{-1} = a_{-1}, x_{-2} = a_{-2}, \dots\}$$

for all choices of $a_i, i \leq 1$. Let $\mathbf{m} = \{m(i)\}_{i=1}^n$ be a vector such that $M\mathbf{m} = \mathbf{m}$. Then an invariant measure is defined by setting the measure on cylinder sets to be $A = \{x: x_0 = a_0, x_1 = a_1, \dots, x_n = a_n\}$ is given by $\mu(A) = m(a_0) \prod_{i=1}^n M(a_{i-1}, a_i)$.

Shift maps More generally the **shift map** σ is the map $\sigma: \mathbb{N}^{\mathbb{Z}} \rightarrow \mathbb{N}^{\mathbb{Z}}$ where $\sigma(x)_i = x_{i+1}$ for all $x \in \mathbb{N}^{\mathbb{Z}}$ and $i \in \mathbb{Z}$. We also let σ designate the shift map on $\mathbb{N}^{\mathbb{N}}$. For each measure that is invariant under the shift map there is a corresponding measure defined on $\mathbb{N}^{\mathbb{N}}$ that is invariant under the shift map. Let μ be an invariant measure under the shift map. For any measurable set of $A \subset \mathbb{N}^{\mathbb{N}}$ we define \tilde{A} on $\mathbb{N}^{\mathbb{Z}}$ by

$$\tilde{A} = \{\dots, x_{-1}, x_0, x_1, \dots : x_0, x_1, \dots \in A\}.$$

Then it is easy to check that $\tilde{\mu}$ defined by $\tilde{\mu}(\tilde{A}) = \mu(A)$ is invariant. If the original transformation was a Markov or Bernoulli shift then refer to the resulting transformations as a **one sided Markov shifts** or **one sided Bernoulli shift** respectively.

Rational maps of the Riemann sphere We say that $f(z) = g(z)/h(z)$ is a **rational function of degree** $d \geq 2$ if both $g(z)$ and $h(z)$ are polynomials with $\max(\deg(g(z)), \deg(h(z))) = d$. Then f induces a natural action on the Riemann sphere $T_f: z \rightarrow f(z)$ which is a d to one map (counting with multiplicity). In Subsect. “[Rational Maps](#)” we shall see that for every rational function f there is a canonical measure μ_f such that T_f is a measure preserving transformation.

Horocycle flows The horocycle flow acts on $SL(2, \mathbb{R})/\Gamma$ where Γ is a discrete subgroup of $SL(2, \mathbb{R})$ such that $SL(2, \mathbb{R})/\Gamma$ has finite Haar measure. For any $g \in SL(2, \mathbb{R})$ and $t \in \mathbb{R}$ we define the horocycle flow by

$$h_t(\Gamma g) = \Gamma g \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix}.$$

Matrix actions Another natural class of actions is given by the action of matrices on tori. Let M be an invertible $n \times n$ integer valued matrix. We define $T_M: [0, 1)^n \rightarrow [0, 1)^n$ by $T_M(x)_i = M(x)_i \bmod 1$ for all i , $1 \leq i \leq n$. It is easy to check that if M is surjective then Lebesgue measure is invariant under T_M . T_M is a $|\det(M)|$ to one map. If $n = 1$ then M is an integer and we refer to the map as times M .

The $[T, T^{-1}]$ transformations Let (X, T) be any invertible measure preserving transformation. Let σ be the shift operator on $Y = \{-1, 1\}^{\mathbb{Z}}$. The space Y comes equipped with the Bernoulli $(1/2, 1/2)$ product measure ν .

The $[T, T^{-1}]$ transformation is a map on $Y \times X$ which preserves $\nu \times \mu$. It is defined by

$$T, T^{-1}(y, x) = (S(y), T^{y_0}(x)).$$

Induced transformations Let (X, T) be a measure preserving transformation and let $A \subset X$ with $0 < \mu(A) < 1$. The **transformation induced by** A , (A, T_A, μ_A) , is defined as follows. For any $x \in A$

$$T_A(x) = T^{n(x)}(x)$$

where $n(x) = \inf\{m > 0: T^m(x) \in A\}$. For any $B \subset A$ we have that $\mu_A(B) = \mu(B)/\mu(A)$.

Basic Isomorphism Invariants

The main purpose of isomorphism theory is to classify which pairs of measure preserving transformation are isomorphic and which are not isomorphic. One of the main ways that we can show that two measure preserving transformation are not isomorphic is using isomorphism invariants. An **isomorphism invariant** is a function f defined on measure preserving transformations such that if (X, T) is isomorphic to (Y, S) then $f((X, T)) = f((Y, S))$.

A measure preserving action is said to be **ergodic** if $\mu(A) = 0$ or 1 for every A with

$$\mu(A \Delta T^{-1}(A)) = 0.$$

A measure preserving action is said to be **weak mixing** if for every measurable A and B

$$\frac{1}{n} \sum_{n=1}^{\infty} |\mu(A \cap T^{-n}(B)) - \mu(A)\mu(B)| = 0.$$

A measure preserving action is said to be **mixing** if for every measurable A and B

$$\lim_{n \rightarrow \infty} \mu(A \cap T^{-n}(B)) = \mu(A)\mu(B).$$

It is easy to show that all three of these properties are isomorphism invariants and that

$$\begin{aligned} (X, T) \text{ is mixing} &\Rightarrow (X, T) \text{ is weak mixing} \\ &\Rightarrow (X, T) \text{ is ergodic.} \end{aligned}$$

We include one more definition before introducing an even stronger isomorphism invariant. The **action of a group** (or a semigroup) G on a probability space (X, μ, \mathcal{B}) is a family of measure preserving transformations $\{f_g\}_{g \in G}$ such that for any $g, h \in G$ we have that $f_g(f_h(x)) = f_{g+h}(x)$ for almost every $x \in X$. Thus any invertible measure preserving transformation (X, T) induces a \mathbb{Z} action by $f_n(x) = T^n(x)$.

We say that a group action (X, T_g) is **mixing of all orders** if for every $n \in \mathbb{N}$ and every collection of set

$A_1, \dots, A_n \in \mathcal{B}$ then

$$\mu(A_1 \cap T_{g_2}(A_2) \cap T_{g_2+g_3}(A_3) \dots \cap T_{g_2+g_3+\dots g_n}(A_n)) \rightarrow \prod_{i=1}^n \mu(A_i)$$

as the g_i go to infinity.

Basic Tools

Birkhoff's ergodic theorem states that for every measure preserving action the limits

$$\hat{f}(x) = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=0}^{n-1} f(T^k x)$$

exists for almost every x and if (X, T) is ergodic then the limit is $\hat{f} = \int f d\mu$ [2].

A **partition** P of X is a measurable function defined on X . (For simplicity we often assume that a partition is a function to \mathbb{Z} or some subset of \mathbb{Z} .) We write P_i for $P^{-1}(i)$. For any partition P of (X, T) define the partition $T^i P$ by $P \circ T^i$. Thus for invertible T this is given by $T^i P(x) = P(T^{-i}(x))$. Then define $(P)_T = \bigvee_{i \in \mathbb{Z}} T^i P$. Thus $(P)_T$ is the smallest σ -algebra which contains $T^i(P_j)$ for all $i \in \mathbb{Z}$ and $j \in \mathbb{N}$. We say that $(P)_T$ is the **σ -algebra generated by P** . A partition P is a **generator** of (X, T) if $(P)_T = \mathcal{B}$. Many measure preserving transformations come equipped with a natural partition.

Rokhlin's theorem For any measure preserving transformation (X, T) , any $\epsilon > 0$ and any $n \in \mathbb{N}$ there exists $A \subset X$ such that $T^{-i}(A) \cap T^{-j}(A)$ for all $0 \leq i < j \leq n$ and $\mu(\bigcup_{i=0}^n T^i(A)) > 1 - \epsilon$. Moreover for any finite partition P of X we can choose A such that $\mu(P_i \cap A) = \mu(A)\mu(P_i)$ for all $i \in \mathbb{N}$ [63].

Shannon-McMillan-Breiman theorem [3,57] For any measure preserving system (X, T) and any $\epsilon > 0$ there exists $n \in \mathbb{N}$ and a set G with $\mu(G) > 1 - \epsilon$ with the following property. For any sequence $g_1, \dots, g_n \in \mathbb{N}$ let $g = \bigcap_{i=1}^n P(T^{-i}(x)) = g_i$. Then if $\mu(g \cap G) > 0$ then

$$\mu(g) \in \left(2^{h((X, T)) - \epsilon}, 2^{h((X, T)) + \epsilon} \right).$$

Krieger generator theorem If $H((X, T)) < \infty$ then there exists a finite partition P such that $(P)_T = \mathcal{B}$. Thus every measure preserving transformation with finite entropy is isomorphic to a shift map on finitely many symbols [33].

Measure-theoretic entropy Entropy was introduced in physics by Rudolph Clausius in 1854. In 1948 Claude

Shannon introduced the concept to information theory. Consider a process that generates a string of data of length n . The entropy of the process is the smallest number h such that you can condense the data to a string of zeroes and one of length hn and with high probability you can reconstruct the original data from the string of zeroes and ones. Thus the entropy of a process is the average amount of information transmitted per symbol of the process.

Kolmogorov and Sinai introduced the concept of entropy to ergodic theory in the following way [31,32]. They defined the entropy of a partition Q is defined to be

$$H(Q) = - \sum_{i=1}^k \mu(Q_i) \log \mu(Q_i).$$

The measure-theoretic entropy of a dynamical system (X, T) with respect to a partition $Q: X \rightarrow \{1, \dots, k\}$ is then defined as

$$h(X, T, Q) = \lim_{N \rightarrow \infty} \frac{1}{N} H\left(\bigvee_{n=1}^N T^{-n} Q\right).$$

Finally, the measure-theoretic entropy of a dynamical system (X, T) is defined as

$$h((X, T)) = \sup_Q h(X, T, Q)$$

where the supremum is taken over all finite measurable partitions. A theorem of Sinai showed that if Q is a generator of (X, T) then $h(T) = h(T, Q)$ [73].

This shows that for every measure preserving function (X, T) there is an associated entropy $h(T) \in [0, \infty]$. It is easy to show from the definition that entropy is an isomorphism invariant.

We say that (Y, S) is a **factor** of (X, T) if there exists a map $\phi: X \rightarrow Y$ such that

- (1) ϕ is measure preserving and
- (2) $\phi(T(x)) = S(\phi(x))$ for almost every x .

Each factor (Y, S) can be associated with $\phi^{-1}(C)$, which is an invariant sub σ -algebra of \mathcal{B} . We say that (Y, S) is trivial if Y consists of only one point. We say that a transformation (X, T) has **completely positive entropy** if every non-trivial factor of (X, T) has positive entropy.

Isomorphism of Bernoulli Shifts

Kolmogorov-Sinai

A long standing open question was for which \mathbf{p} and \mathbf{q} are Bernoulli $_p$ and Bernoulli $_q$ isomorphic. In particular are the

Bernoulli 2 shift and the Bernoulli 3 shift isomorphic. Both of these transformations have completely positive entropy and all other isomorphism invariants which were known at the time are the same for the two transformations. The first application of the Kolmogorov–Sinai entropy was to show that the answer to this question is no.

Fix a probability vector \mathbf{p} . The transformation $\text{Bernoulli}_{\mathbf{p}}$ has $Q_{\mathbf{p}}: x \rightarrow x_0$ as a generating partition. By Sinai's theorem

$$H(\text{Bernoulli}_{\mathbf{p}}) = H(Q_{\mathbf{p}}) = \sum_{i=1}^n -p_i \log_2(p_i).$$

Thus the Bernoulli 2 shift (with entropy 1) is not isomorphic to the Bernoulli 3 shift (with entropy $\log_2(3)$).

Sinai also made significant progress toward showing that Bernoulli shifts with the same entropy are isomorphic by proving the following theorem.

Theorem 1 [72] *If (\mathbf{X}, T) is a measure preserving system of entropy h and (\mathbf{Y}, S) is a Bernoulli shift of entropy $h' \leq h$ then (\mathbf{Y}, S) is a factor of (\mathbf{X}, T) .*

This theorem implies that if \mathbf{p} and \mathbf{q} are probability vectors and $H(\mathbf{p}) = H(\mathbf{q})$ then $\text{Bernoulli}_{\mathbf{p}}$ is a factor in $\text{Bernoulli}_{\mathbf{q}}$ and $\text{Bernoulli}_{\mathbf{q}}$ is a factor in $\text{Bernoulli}_{\mathbf{p}}$. Thus we say that $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are **weakly isomorphic**.

Explicit Isomorphisms

The other early progress on proving that Bernoulli shifts with the same entropy are isomorphic came from Moshalkin. He considered pairs of probability vectors \mathbf{p} and \mathbf{q} with $H(\mathbf{p}) = H(\mathbf{q})$ and all of the p_i and q_i are related by some algebraic relations. For many such pairs he was able to prove that the two Bernoulli shifts are isomorphic. In particular he proved the following theorem.

Theorem 2 [40] *Let $\mathbf{p} = (1/4, 1/4, 1/4, 1/4)$ and $\mathbf{q} = (1/2, 1/8, 1/8, 1/8, 1/8, 1/8)$. Then $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are isomorphic.*

Ornstein

The central theorem in the study of isomorphisms of measure preserving transformations is Ornstein's isomorphism theorem.

Theorem 3 [46] *If \mathbf{p} and \mathbf{q} are probability vectors and $H(\mathbf{p}) = H(\mathbf{q})$ then the Bernoulli shifts $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are isomorphic.*

To see how central this is to the field most of the rest of this article is a summary of:

- (1) The proof of Ornstein's theorem,
- (2) The consequences of Ornstein's theorem,
- (3) The generalizations of Ornstein's theorem, and
- (4) How the properties that Ornstein's theorem implies that Bernoulli shifts must have differ from the properties of every other class of transformations.

The key to Ornstein's proof was the introduction of the finitely determined property. To explain the finitely determined property we first define the Hamming distance of length n between sequences $x, y \in \mathbb{N}^{\mathbb{Z}}$ by

$$\bar{d}_n(x, y) = 1 - \frac{|\{k \in \{1, \dots, n\}: x_k = y_k\}|}{n}.$$

Let (\mathbf{X}, T) and (\mathbf{Y}, S) be a measure preserving transformation and let P and Q be finite partitions of X and Y respectively.

We say that (\mathbf{X}, T) and P and (\mathbf{Y}, S) and Q are within δ in n distributions if

$$\sum_{(a_1, \dots, a_n) \in \mathbb{Z}^n} \left| \mu \left(\left\{ x: P \left(T^i(x) \right) = a_i \ \forall i = 1, \dots, n \right\} \right) - \nu \left(\left\{ y: Q \left(T^i(y) \right) = a_i \ \forall i = 1, \dots, n \right\} \right) \right| < \delta.$$

A process (\mathbf{X}, T) and P are **finitely determined** if for every ϵ there exist n and δ such that if (\mathbf{Y}, S) and Q are such that

- (1) (\mathbf{X}, T) and P and Y and Q are within δ in n distributions and
- (2) $|H(X, P) - H(Y, Q)| < \delta$

then there exists a joining γ of X and Y such that for all m

$$\int_{x, y} \bar{d}_m(x, y) d\gamma(x, y) < \epsilon.$$

A transformation (\mathbf{X}, T) is finitely determined if it is finitely determined for every finite partition P .

It is fairly straightforward to show that Bernoulli shifts are finitely determined. Ornstein used this fact along with the Rokhlin lemma and the Shannon–McMillan–Breiman theorem to prove a more robust version of Theorem 1.

To describe Ornstein's proof we use the a description due to Rothstein. We say that for a joining γ of (X, μ) and (Y, ν) that $P \subseteq_{\epsilon, \gamma} C$ if there exists a partition \tilde{P} of C such that

$$\sum_i \gamma(P_i \Delta \tilde{P}_i) < \epsilon.$$

If $P \subset_{\epsilon, \gamma} C$ for all $\epsilon > 0$ then it is possible to show that there exists a partition \tilde{P} of C such that

$$\sum_i \gamma(P_i \triangle \tilde{P}_i) = 0$$

and we write $P \subset_{\gamma} C$. If $P \subset_{\gamma} C$ then (X, T) is a factor of (Y, S) by the map ϕ that sends $y \rightarrow x$ where $P(T^i x) = \tilde{P}(S^i y)$ for all i .

In this language Ornstein proved that if (X, T) is finitely determined, P is a generating partition of X and $h((Y, S)) \geq h((X, T))$ then for every $\epsilon > 0$ the set of joinings γ such that $P \subset_{\gamma, \epsilon} C$ is an open and dense set. Thus by the Baire category theorem there exists γ such that $P \subset_{\gamma} C$. This reproves Theorem 1.

Moreover if (X, T) and (Y, S) are finitely determined, $h((X, T)) = h((Y, S))$ and P and Q are generating partitions of X and Y then by the Baire category theorem there exists γ such that $P \subset_{\gamma} C$ and $Q \subset_{\gamma} B$. Then the map ϕ that sends $y \rightarrow x$ where $P(T^i x) = \tilde{P}(S^i y)$ for all i is an isomorphism.

Properties of Bernoullis

Now we define the very weak Bernoulli property which is the most effective property for showing that a measure preserving transformation is isomorphic to a Bernoulli shift.

Given X and a partition P define the past of x by

$$P_{\text{past}}(x) = \{x' : T^i P(x') = T^i P(x) \forall i \in \mathbb{N}\}$$

and denote the measure μ conditioned on $P_{\text{past}}(x)$ by $\mu|_{P_{\text{past}}(x)}$. Define

$$\bar{d}_{n, x, \mu} = \inf_{\gamma} \int \bar{d}_n(x, x') d\gamma(x, x'),$$

where the inf is taken over all γ which are couplings of $\mu|_{P_{\text{past}}(x)}$ and μ . Also define

$$\bar{d}_{n, P_{\text{past}}, (X, T)} = \int \bar{d}_{n, x, \mu} d\mu.$$

We say that (X, T) and P are very weak Bernoulli if for every $\epsilon > 0$ there exists n such that $\bar{d}_{n, P_{\text{past}}, (X, T)} < \epsilon$. We say that (X, T) is very weakly Bernoulli if there exists a generating partition P such that (X, T) and P are very weak Bernoulli.

Ornstein and Weiss were able to show that the very weak Bernoulli property is both necessary and sufficient to be isomorphic to a Bernoulli shift.

Theorem 4 [45, 53] *For transformations (X, T) the following conditions are equivalent:*

- (1) (X, T) is finitely determined,
- (2) (X, T) is very weak Bernoulli and
- (3) (X, T) is isomorphic to a Bernoulli shift.

Using the fact that a transformation is finitely determined or very weak Bernoulli is equivalent to it being isomorphic to a Bernoulli shift we can prove the following theorem.

Theorem 5 [44]

- (1) If (X, T^n) is isomorphic to a Bernoulli shift then (X, T) is isomorphic to a Bernoulli shift.
- (2) If (X, T) is a factor of a Bernoulli shift then (X, T) is isomorphic to a Bernoulli shift.
- (3) If (X, T) is isomorphic to a Bernoulli shift then there exists a measure preserving transformation (Y, S) such that (Y, S^n) is isomorphic to (X, T) .

Rudolph Structure Theorem

An important application of the very weak Bernoulli condition is the following theorem of Rudolph.

Theorem 6 [67] *Let (X, T) be isomorphic to a Bernoulli shift, G be a compact Abelian group with Haar measure μ_G and $\sigma : X \rightarrow G$ be a measurable map. Then let $S : X \times G \rightarrow X \times G$ be defined by*

$$S(x, g) = (T(x), g + \sigma(x)).$$

Then $(X \times G, S, \mu \times \mu_G)$ is isomorphic to a Bernoulli shift.

Transformations Isomorphic to Bernoulli Shifts

One of the most important features of Ornstein's isomorphism theory is that it can be used to check whether specific transformations (or families of transformations) are isomorphic to Bernoulli shifts. The finitely determined property is the key to the proof of Ornstein's theorem and the proof of many of the consequences listed in Subsect. "Properties of Bernoullis". However if one wants to show a particular transformation is isomorphic to a Bernoulli shift then the very weak Bernoulli property is more useful. There have been many classes of transformations that have been proven to be isomorphic to a Bernoulli shift. Here we mention two.

The first class are the Markov chains. Friedman and Ornstein proved that if a Markov chain is mixing then it is isomorphic to a Bernoulli shift [9]. The second are automorphisms of $[0, 1]^n$. Let M be any $n \times n$ matrix with integer coefficients and $|\det(M)| = 1$. If none of the eigenvalues $\{\lambda_i\}_{i=1}^n$ of M are roots of unity then Katznelson proved that T_M is isomorphic to the Bernoulli shift with entropy $\sum_{i=1}^n \max(0, \log(\lambda_i))$ [29].

Transformations not Isomorphic to Bernoulli Shifts

Recall that a measure preserving transformation (X, T) has **completely positive entropy** if for every nontrivial (Y, S) which is a factor of (X, T) we have that $H((Y, S)) > 0$. It is easy to check that Bernoulli shifts have completely positive entropy. It is natural to ask if the converse true? We shall see that the answer is an emphatic no. While the isomorphism class of Bernoulli shifts is given by just one number, the situation for transformations of completely positive entropy is infinitely more complicated.

Ornstein constructed the first example of a transformation with completely positive entropy which is not isomorphic to a Bernoulli shift [47]. Ornstein and Shields built upon this construction to prove the following theorem.

Theorem 7 [50] *For every $h > 0$ there is an uncountable family of completely positive entropy transformations which all have entropy h but no two distinct members of the family are isomorphic.*

Now that we see there are many isomorphic transformations that have completely positive entropy it is natural to ask if (X, T) is not isomorphic to a Bernoulli shift then is there any reasonable condition we can put on (Y, S) that implies the two transformations are isomorphic. For example if (X, T^2) and (Y, S^2) are completely positive entropy transformations which are isomorphic does that necessarily imply that (X, T) and (Y, S) are isomorphic? The answer turns out to be no [66]. We could also ask if (X, T) and (Y, S) are completely positive entropy transformations which are weakly isomorphic does that imply that (X, T) and (Y, S) are isomorphic? Again the answer is no [17]. The key insight to answering questions like this is due to Rudolph who showed that such questions about the isomorphism of transformations can be reduced to questions about conjugacy of permutations.

Rudolph's Counterexample Machine

Given any transformation (X, T) and any permutation π in S_n (or $S_{\mathbb{N}}$) we can define the transformation (X^n, T_π, μ^n, B) by

$$\begin{aligned} T_\pi(x_1, x_2, \dots, x_n) \\ = (T(x_{\pi(1)}), T(x_{\pi(2)}), \dots, T(x_{\pi(n)})) \end{aligned}$$

where μ^n is the direct product of n copies of μ . Rudolph introduced the concept of a transformation having **minimal self joinings**. If a transformation has minimal self

joinings then for every π it is possible to list all of the measures on X^n which are invariant under T_π .

If there exists an isomorphism ϕ between T and (Y, S) then there is a corresponding measure on $X \times Y$ which is supported on points of the form $(x, \phi(x))$ and has marginals μ and ν . Thus if we know all of the measures on $X \times Y$ which are invariant under $T \times S$ then we know all of the isomorphisms between (X, T) and (Y, S) . Using this we get the following theorem.

Theorem 8 [65] *There exists a nontrivial transformation with minimal self joinings. For any transformation (X, T) with minimal self joinings the corresponding transformation T_{π_1} is isomorphic to T_{π_2} if and only if the permutations π_1 and π_2 are conjugate.*

There are two permutations on two elements, the flip $\pi_1 = (12)$ and the identity $\pi_2 = (1)(2)$. For both permutations, the square of the permutation is the identity. Thus there are two distinct permutations whose square is the same. Rudolph showed that this fact can be used to generate two transformations which are mixing that are not isomorphic but their squares are isomorphic. The following theorem gives more examples of the power of this technique.

Theorem 9 [65]

- (1) *There exists measure preserving transformations (X, T) and (Y, S) which are weakly isomorphic but not isomorphic.*
- (2) *There exists measure preserving transformations (X, T) and (Y, S) which are not isomorphic but (X, T^k, μ) is isomorphic to (Y, S^k, ν) for every $k > 1$, and*
- (3) *There exists a mixing transformation with no non trivial factors.*

If (X, T) has minimal self joinings then it has zero entropy. However Hoffman constructed a transformation with completely positive entropy that shares many of the properties of transformations with minimal self joinings listed above.

Theorem 10 [17]

- (1) *There exist measure preserving transformations (X, T) and (Y, S) which both have completely positive entropy and are weakly isomorphic but not isomorphic.*
- (2) *There exist measure preserving transformations (X, T) and (Y, S) which both have completely positive entropy and are not isomorphic but (X, T^k, μ) is isomorphic to (Y, S^k, ν) for every $k > 1$.*

T, T Inverse

All of the transformations that have completely positive entropy but are not isomorphic to a Bernoulli shift described above are constructed by a process called cutting and stacking. These transformations have little inherent interest outside of their ergodic theory properties. This led many people to search for a “natural” example of such a transformation. The most natural examples are the $[T, T^{-1}]$ transformation and many other transformations derived from it. It is easy to show that the $[T, T^{-1}]$ transformation has completely positive entropy [39]. Kalikow proved that for many T the corresponding $[T, T^{-1}]$ transformation is not isomorphic to a Bernoulli shift.

Theorem 11 [24] *If $h(T) > 0$ then the $[T, T^{-1}]$ transformation is not isomorphic to a Bernoulli shift.*

The basic idea of Kalikow’s proof has been used by many others. Katok and Rudolph used the proof to construct smooth measure preserving transformations on infinite differentiable manifolds which have completely positive entropy but are not isomorphic to Bernoulli shifts [27,68]. Den Hollander and Steif did a thorough study of the ergodic theory properties of $[T, T^{-1}]$ transformations where T is simple random walk on a wide family of graphs [4].

Classifying the Invariant Measures of Algebraic Actions

This problem of classifying all of the invariant measures like Rudolph did with his transformation with minimal self joinings comes up in a number of other settings. Ratner characterized the invariant measures for the horocycle flow and thus characterized the possible isomorphisms between a large family of transformations generated by the horocycle flow [59,60,61]. This work has powerful applications to number theory.

There has also been much interest in classifying the measures on $[0, 1)$ that are invariant under both times 2 and times 3. Furstenberg proved that the only closed infinite set on the circle which is invariant under times 2 and times 3 is $[0, 1)$ itself and made the following measure theoretic conjecture.

Conjecture 1 [10] *The only nonatomic measure on $[0, 1)$ which is invariant under times 2 and times 3 is Lebesgue measure.*

Rudolph improved on the work of Lyons [36] to provide the following partial answer to this conjecture.

Theorem 12 [69] *The only measure on $[0, 1)$ which is invariant under multiplication by 2 and by 3 and has positive entropy under multiplication by 2 is Lebesgue measure.*

Johnson then proved that for all relatively prime p and q that p and q can be substituted for 2 and 3 in the theorem above.

This problem can be generalized to higher dimensions by studying the actions of commuting integer matrices of determinant greater than one on tori. Katok and Spatzier [28] and Einsiedler and Lindenstrauss [5] obtained results similar to Rudolph’s for actions of commuting matrices.

Finitary Isomorphisms

By Ornstein’s theorem we know that there exists an isomorphism between any two Bernoulli shifts (or mixing Markov shifts) of the same entropy. There has been much interest in studying how “nice” the isomorphism can be. By this we mean can ϕ be chosen so that the map $x \rightarrow (\phi(x))_0$ is continuous and if so what is its best possible modulus of continuity?

We say that a map ϕ from $\mathbb{N}^{\mathbb{Z}}$ to $\mathbb{N}^{\mathbb{Z}}$ is **finitary** if in order to determine $(\phi(x))_0$ we only need to know finitely many coordinates of x . More precisely if for almost every x there exists $m(x)$ such that

$$\mu\left(\left\{x': x'_i = x_i \text{ for all } |i| \leq m(x)\right.\right. \\ \left.\left. \text{and } (\phi(x'))_0 \neq \phi(x)_0\right\}\right) = 0.$$

We say that $m(x)$ has **finite t th moment** if $\int m(x)^t d\mu < \infty$ and that ϕ has **finite expected coding length** if the first moment of $m(x)$ is finite.

Keane and Smorodinsky proved the following strengthening of Ornstein’s isomorphism theorem.

Theorem 13 [30] *If \mathbf{p} and \mathbf{q} are probability vectors with $H(\mathbf{p}) = H(\mathbf{q})$ then the Bernoulli shifts $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are isomorphic and there exists an isomorphism ϕ such that ϕ and ϕ^{-1} are both finitary.*

The nicest that we could hope ϕ to be is if both ϕ and ϕ^{-1} are finitary and have finite expected coding length. Schmidt proved that this happens only in the trivial case that \mathbf{p} and \mathbf{q} are rearrangements of each other.

Theorem 14 [70] *If \mathbf{p} and \mathbf{q} are probability vectors and the Bernoulli shifts $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are isomorphic and there exists an isomorphism ϕ such that ϕ and ϕ^{-1} are both finitary and have finite expected coding time then \mathbf{p} is a rearrangement of \mathbf{q} .*

The best known result about this problem is the following theorem of Harvey and Peres.

Theorem 15 [14] *If \mathbf{p} and \mathbf{q} are probability vectors with $H(\mathbf{p}) = H(\mathbf{q})$ then $\sum_i (p_i)^2 \log(p_i) = \sum_i (q_i)^2 \log(q_i)$ if and only if the Bernoulli shifts $\text{Bernoulli}_{\mathbf{p}}$ and $\text{Bernoulli}_{\mathbf{q}}$ are isomorphic and there exists an isomorphism ϕ such that ϕ and ϕ^{-1} are both finitary and have finite one half moment.*

Flows

A **flow** is a measure preserving action of the group \mathbb{R} on a measure space (X, T) . A **cross section** is any measurable set in $C \subset X$ such that for almost every x

$$0 < \inf \{t: T_t(x) \in C\} < \infty.$$

For any flow (X, T) and $\{T_t\}_{t \in \mathbb{R}}$ and any cross section C we define the **return time map for C** $R: C \rightarrow C$ as follows. For any $x \in C$ define

$$t(x) = \inf \{t: T_t(x) \in C\}$$

then set $R(x) = T_{t(x)}(x)$. There is a standard method to project the probability measure μ on X to an invariant probability measure μ_C on C as well as the σ -algebra \mathcal{B} on X to a σ -algebra \mathcal{B}_C on C such that $(C, \mu_C, \mathcal{B}_C)$ and R is a measure preserving transformation.

First we show that there is a natural analog of Bernoulli shifts for flows.

Theorem 16 [45] *There exists a flow (X, T) and $\{T_t\}_{t \in \mathbb{R}}$ such that for every $t > 0$ the map (X, T) and T_t is isomorphic to a Bernoulli shift. Moreover for any $h \in (0, \infty]$ there exists (X, T) and $\{T_t\}_{t \in \mathbb{R}}$ such that $h(T_1) = h$.*

We say that such a flow $(X, \{f_t\}_{t \in \mathbb{R}}, \mu, \mathcal{B})$ is a **Bernoulli flow**.

This next version of Ornstein's isomorphism theorem shows that up to isomorphism and a change in time (considering the flow \mathbf{X} and $\{T_{ct}\}$ instead of \mathbf{X} and $\{T_t\}$) there are only two Bernoulli flows, one with positive but finite entropy and one with infinite entropy.

Theorem 17 [45] *If (X, T) and $\{T_t\}_{t \in \mathbb{R}}$ and (Y, S) $\{S_t\}_{t \in \mathbb{R}}$ are Bernoulli flows and $h(T_1) = h(S_1)$ then they are isomorphic.*

As in the case of actions of \mathbb{Z} there are many natural examples of flows that are isomorphic to the Bernoulli flow. The first is for geodesic flows. In the 1930s Hopf proved that geodesic flows on compact surfaces of constant negative curvature are ergodic [22]. Ornstein and Weiss extended Hopf's proof to show that the geodesic flow is also Bernoulli [52].

The second class of flows comes from billiards on a square table with one circular bumper. The state space X consists of all positions and velocities for a fixed speed. The flow T_t is frictionless movement for time t with elastic collisions. This flow is also isomorphic to the Bernoulli flow [11].

Other Equivalence Relations

In this section we will discuss a number of equivalence relations between transformations that are weaker than isomorphism. All of these equivalence relations have a theory that is parallel to Ornstein's theory.

Kakutani Equivalence

We say that two transformations (X, T) and (Y, S) are **Kakutani equivalent** if there exist subsets $A \subset X$ and $B \subset Y$ such that (T_A, A, μ_A) and (S_B, B, ν_B) are isomorphic. This is equivalent to the existence of a flow (X, T) and $\{T_t\}_{t \in \mathbb{R}}$ with cross sections C and C' such that the return time maps of C and C' are isomorphic to (X, T) and (Y, S) respectively.

Using the properties of entropy of the induced map we have that if (X, T) and (Y, S) are Kakutani equivalent then either $h((X, T)) = h((Y, S)) = 0$, $0 < h((X, T))$, $h((Y, S)) < \infty$ or $h((X, T)) = h((Y, S)) = \infty$.

In general the answer to the question of which pairs of measure preserving transformations are isomorphic is quite poorly understood. But if one of the transformations is a Bernoulli shift then Ornstein's theory gives a fairly complete answer to the question. A similar situation exists for Kakutani equivalence. In general the answer to the question of which pairs of measure preserving transformations are Kakutani equivalent is also quite poorly understood. But the more specialized question of which transformations are isomorphic to a Bernoulli shift has a more satisfactory answer.

Feldman constructed a transformation (X, T) which has completely positive entropy but (X, T) is not Kakutani equivalent to a Bernoulli shift. Ornstein, Rudolph and Weiss extended Feldman's work to construct a complete theory of the transformations that are Kakutani equivalent to a Bernoulli shift [49] for positive entropy transformations and a theory of the transformations that are Kakutani equivalent to an irrational rotation [49] for zero entropy transformations. (The zero entropy version of this theorem had been developed independently (and earlier) by Katok [26].)

They defined two class of transformations called **loosely Bernoulli** and **finitely fixed**. The definitions of these properties are the same as the definitions of very

weak Bernoulli and finitely determined except that the \bar{d} metric is replaced by the \bar{f} metric. For $x, y \in \mathbb{N}^{\mathbb{Z}}$ we define

$$\bar{f}_n(x, y) = 1 - \frac{k}{n}$$

where k is the largest number such that sequences $1 \leq i_1 < i_2 < \dots < i_k \leq n$ and $1 \leq j_1 < j_2 < \dots < j_k \leq n$ such that $x_{i_l} = y_{j_l}$ for all $j, 1 \leq j \leq k$. (In computer science this metric is commonly referred to as the edit distance.) Note that $\bar{d}_n(x, y) \geq \bar{f}_n(x, y)$.

They proved the following analog of Theorem 5.

Theorem 18 *For transformations (X, T) with $h((X, T)) > 0$ the following conditions are equivalent:*

- (1) (X, T) is finitely fixed,
- (2) (X, T) is loosely Bernoulli,
- (3) (X, T) is Kakutani equivalent to a Bernoulli shift and
- (4) There exists a Bernoulli flow Y and $\{F_t\}_{t \in \mathbb{R}}$ and a cross section C such that the return time map for C is isomorphic to (X, T) .

Restricted Orbit Equivalence

Using the \bar{d} metric we got a theory of which transformations are isomorphic to a Bernoulli shift. Using the \bar{f} metric we got a strikingly similar theory of which transformations are Kakutani equivalent to a Bernoulli shift. Rudolph showed that it is possible to replace the \bar{d} metric (or the \bar{f} metric) with a wide number of other metrics and produce parallel theories for other equivalence relations. For instance, for each of these theories we get a version of Theorem 5. This collection of theories is called restricted orbit equivalence [64].

Non-invertible Transformations

The question of which noninvertible measure preserving transformations are isomorphic turns out to be quite different from the same question for invertible transformations. In one sense it is easier because of an additional isomorphism invariant.

For any measure preserving transformation (X, T) the probability measure $\mu|_{T^{-1}(x)}$ on $T^{-1}(x)$ is defined for almost every $x \in X$. (If (X, T) is invertible then this measure is trivial as $|\{T^{-1}(x)\}| = 1$ and $\mu|_{T^{-1}(x)}(T^{-1}(x)) = 1$ for almost every x .) It is easy to check that if ϕ is an isomorphism from (X, T) to (Y, S) then for almost every x and $x' \in T^{-1}(x)$ we have

$$\mu|_{T^{-1}(x)}(x') = \nu|_{S^{-1}(\phi(x))}(\phi(x')) .$$

From this we can easily see that if $\mathbf{p} = \{p_i\}_{i=1}^n$ and $\mathbf{q} = \{q_i\}_{i=1}^m$ are probability vectors then the corresponding one sided Bernoulli shifts are isomorphic only if $m = n$ and there is a permutation $\pi \in S_n$ such that $p_{\pi(i)} = q_i$ for all i . (In this case we say \mathbf{p} is a **rearrangement** of \mathbf{q} .) If \mathbf{p} is a rearrangement of \mathbf{q} then it is easy to construct an isomorphism between the corresponding Bernoulli shifts. Thus the analog of Ornstein's theorem for Bernoulli endomorphisms is trivial. However we will see that there still is an analogous theory classifying the class of endomorphism that are isomorphic to Bernoulli endomorphisms.

We say that an endomorphism is **uniformly d to 1** if for almost every x we have that $|\{T^{-1}(x)\}| = d$ and $\mu|_{T^{-1}(x)}(y) = 1/d$ for all $y \in T^{-1}(x)$. Hoffman and Rudolph defined two classes of noninvertible transformations called **tree very weak Bernoulli** and **tree finitely determined** and proved the following theorem.

Theorem 19 *The following three conditions are equivalent for uniformly d to 1 endomorphisms.*

- (1) (X, T) is tree very weak Bernoulli,
- (2) (X, T) is tree finitely determined, and
- (3) (X, T) is isomorphic to the one sided Bernoulli d shift.

Jong extended this theorem to say that if there exists a probability vector \mathbf{p} such that for almost every x the distribution of $\mu|_{T^{-1}(x)}$ is the same as the distribution of \mathbf{p} then (X, T) is isomorphic to the one sided Bernoulli d shift if and only if it is tree finitely determined and if and only if it is tree very weak Bernoulli [23].

Markov Shifts

We saw that mixing Markov chains are isomorphic if they have the same entropy. As we have seen there are additional isomorphism invariants for noninvertible transformations. Ashley, Marcus and Tuncel managed to classify all one sided mixing Markov chains up to isomorphism [1].

Rational Maps

Rational maps are the main object of study in complex dynamics. For every rational function $f(z) = p(z)/q(z)$ there is a nonempty compact set J_f which is called the **Julia set**. Roughly speaking this is the set of points for which every neighborhood acts "chaotically" under repeated iterations of f .

In order to consider rational maps as measure preserving transformations we need to specify an invariant measure. The following theorem of Gromov shows that for ev-

ery rational map there is one canonical measure to consider.

Theorem 20 [12] *For every f rational function of degree d there exists a unique invariant measure μ_f of maximal entropy. We have that $h(\mu_f) = \log_2 d$ and $\mu_f(J_f) = 1$.*

The properties of this measure were studied by Freire, Lopes and Mañé [8]. Mañé that analysis to prove the following theorem.

Theorem 21 [38] *For every rational function f of degree d there exists n such that (C, f^n, μ_f) (where $f^n(z) = f(f(\dots f(z)))$ is composition) is isomorphic to the one sided Bernoulli d^n shift.*

Heicklen and Hoffman used the tree very weak Bernoulli condition to show that we can always take n to be one.

Theorem 22 [16] *For every rational function f of degree $d \geq 2$ the corresponding map $((C, \mu_f, B), T_f)$ is isomorphic to the one sided Bernoulli d shift.*

Differences with Ornstein's Theory

Unlike Kakutani equivalence and restricted orbit equivalence which are very close parallels to Ornstein's theory, the theory of which endomorphisms are isomorphic to a Bernoulli endomorphism contains some significant differences. One of the principal results of Ornstein's isomorphism theory is that if (X, T) is an invertible transformation and (X, T^2) is isomorphic to a Bernoulli shift then (X, T) is also isomorphic to a Bernoulli shift. There is no corresponding result for noninvertible transformations.

Theorem 23 [19] *There is a uniformly two to one endomorphism (X, T) which is not isomorphic to the one sided Bernoulli 2 shift but (X, T^2) is isomorphic to the one sided Bernoulli 4 shift.*

Factors of a Transformation

In this section we study the relationship between a transformation and its factors. There is a natural way to associate a factor of (X, T) with a sub σ -algebra of \mathcal{B} . Let (Y, S) be a factor of (X, T) with factor map $\phi: X \rightarrow Y$. Then the σ -algebra associated with (Y, S) is $\mathcal{B}_Y = \phi^{-1}(C)$. Thus the study of factors of a transformation is the study of its sub σ -algebras.

Almost every property that we have discussed above has an analog in the study of factors of a transformation. We give three such examples. We say that two factors C and \mathcal{D} of (X, T) are **relatively isomorphic** if there exists an isomorphism $\psi: X \rightarrow X$ of (X, T) with itself such that $\psi(C) = \mathcal{D}$. We say that (X, T) has **relatively completely**

positive entropy with respect to C if every factor \mathcal{D} which contains C has $h(\mathcal{D}) > h(C)$. We say that C is **relatively Bernoulli** if there exists a second factor \mathcal{D} which is independent of C and $\mathcal{B} = C \vee \mathcal{D}$.

Thouvenot defined properties of factors called **relatively very weak Bernoulli** and **relatively finitely determined**. Then he proved an analog of Theorem 3. This says that a factor being relatively Bernoulli is equivalent to it being relatively finitely determined (and also equivalent to it being relatively very weak Bernoulli).

The **Pinsker algebra** is the maximal σ -algebra \mathcal{P} such that $h(\mathcal{P}) = 0$. The Pinsker conjecture was that for every measure preserving transformation (X, T) there exists a factor C such that

- (1) C is independent of the Pinsker algebra \mathcal{P}
- (2) $\mathcal{B} = C \vee \mathcal{P}$ and
- (3) (X, C) has completely positive entropy.

Ornstein found a counterexample to the Pinsker conjecture [48]. After Thouvenot developed the relative isomorphism theory he came up with the following question which is referred to as the weak Pinsker conjecture.

Conjecture 2 *For every measure preserving transformation (X, T) and every $\epsilon > 0$ there exist invariant σ -algebras $C, \mathcal{D} \subset \mathcal{B}$ such that*

- (1) C is independent of \mathcal{D}
- (2) $\mathcal{B} = C \vee \mathcal{D}$
- (3) (X, T, μ, \mathcal{D}) is isomorphic to a Bernoulli shift
- (4) $h((X, T, \mu, C)) < \epsilon$.

There is a wide class of transformations which have been proven to satisfy the weak Pinsker conjecture. This class includes almost all measure preserving transformations which have been extensively studied.

Actions of Amenable Groups

All of the discussion above has been about the action of a single invertible measure preserving transformation (actions of \mathbb{N} and \mathbb{Z}) or flows (actions of \mathbb{R}). We now consider more general group actions. If we have two actions S and T on a measure space (X, μ) which commute ($S(T(x)) = T(S(x))$ for almost every x) then there is an action of \mathbb{Z}^2 on (X, μ) given by $f_{(n,m)}(x) = S^n(T^m(x))$. A natural question to ask is do there exist a version of entropy theory and Ornstein's isomorphism theory for actions of two commuting automorphisms. More generally for each of the results discussed above we can ask what is the largest class of groups such that an analogous result is

true. It turns out that for most of the properties described above the right class of groups is discrete amenable groups.

A **Følner sequence** F_n in a group G is a sequence of subsets F_n of G such that for all $g \in G$ we have that $\lim_{n \rightarrow \infty} |g(F_n)|/|F_n| = 1$. A countable group is **amenable** if and only if it has a Følner sequence.

For nonamenable groups it is much more difficult to generalize Birkhoff's ergodic theorem [41,42]. Lindenstrauss proved that for every discrete amenable group there is an analog of the ergodic theorem [35]. For every amenable group G and every probability vector \mathbf{p} we can define a Bernoulli action of G . There are also analogs of Rokhlin's lemma and the Shannon–McMillan–Breiman theorem for actions of all discrete amenable groups [33,51,54]. Thus we have all of the ingredients to prove a version of Ornstein's isomorphism theorem.

Theorem 24 *If \mathbf{p} and \mathbf{q} are probability vectors and $H(\mathbf{p}) = H(\mathbf{q})$ then the Bernoulli actions of G corresponding to \mathbf{p} and \mathbf{q} are isomorphic.*

Also all of the aspects of Rudolph's theory of restricted orbit equivalence can be carried out for actions of amenable groups [25].

Differences Between Actions of \mathbb{Z} and Actions of Other Groups

Although generalizations of Ornstein theory and restricted orbit equivalence carry over well to the actions of discrete amenable groups there do turn out to be some significant differences between the possible actions of \mathbb{Z} and those of other discrete amenable groups.

Many of these have to do with the generalization of Markov shifts. For actions of \mathbb{Z}^2 these are called Markov random fields. By the result of Friedman and Ornstein if a Markov chain is mixing then it has completely positive entropy and it is isomorphic to a Bernoulli shift. Mixing Markov random fields can have very different properties. Ledrappier constructed a \mathbb{Z}^2 action which is a Markov random field and is mixing but has zero entropy [34]. Even more surprising even though it is mixing it is not mixing of all orders. The existence of a \mathbb{Z} action which is mixing but not mixing of all orders is one of the longest standing open questions in ergodic theory [13].

Even if we try to strengthen the hypothesis of Friedman and Ornstein's theorem to assume that the Markov random field has completely positive entropy we will not succeed as there exists a Markov random field which has completely positive entropy but is not isomorphic to a Bernoulli shift [18].

Future Directions

In the future we can expect to see progress of isomorphism theory in a variety of different directions. One possible direction for future research is better understand the properties of finitary isomorphisms between various transformations and Bernoulli shifts described in Sect. "Finitary Isomorphisms". Another possible direction would be to find a theory of equivalence relations for Bernoulli endomorphisms analogous to the one for invertible Bernoulli transformations described in Sect. "Other Equivalence Relations".

As the subject matures the focus of research in isomorphism theory will likely shift to connections to other fields. Already there are deep connections between isomorphism theory and both number theory and statistical physics. Finally one hopes to see progress made on the two dominant outstanding conjectures in the field: Thouvenot weak Pinsker conjecture (Conjecture 2) and Furstenberg's conjecture (Conjecture 1) about measures on the circle invariant under both the times 2 and times 3 maps. Progress on either of these conjectures would invariably lead the field in exciting new directions.

Bibliography

1. Ashley J, Marcus B, Tuncel S (1997) The classification of one-sided Markov chains. *Ergod Theory Dynam Syst* 17(2):269–295
2. Birkhoff GD (1931) Proof of the ergodic theorem. *Proc Natl Acad Sci USA* 17:656–660
3. Breiman L (1957) The individual ergodic theorem of information theory. *Ann Math Statist* 28:809–811
4. Den Hollander F, Steif J (1997) Mixing E properties of the generalized T, T^{-1} -process. *J Anal Math* 72:165–202
5. Einsiedler M, Lindenstrauss E (2003) Rigidity properties of \mathbb{Z}^d -actions on tori and solenoids. *Electron Res Announc Amer Math Soc* 9:99–110
6. Einsiedler M, Katok A, Lindenstrauss E (2006) Invariant measures and the set of exceptions to Littlewood's conjecture. *Ann Math* (2) 164(2):513–560
7. Feldman J (1976) New K -automorphisms and a problem of Kakutani. *Isr J Math* 24(1):16–38
8. Freire A, Lopes A, Mañé R (1983) An invariant measure for rational maps. *Bol Soc Brasil Mat* 14(1):45–62
9. Friedman NA, Ornstein DS (1970) On isomorphism of weak Bernoulli transformations. *Adv Math* 5:365–394
10. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem in Diophantine approximation. *Math Syst Theory* 1:1–49
11. Gallavotti G, Ornstein DS (1974) Billiards and Bernoulli schemes. *Comm Math Phys* 38:83–101
12. Gromov M (2003) On the entropy of holomorphic maps. *Enseign Math* (2) 49(3–4):217–235
13. Halmos PR (1950) *Measure theory*. D Van Nostrand, New York
14. Harvey N, Peres Y. An invariant of finitary codes with finite expected square root coding length. *Ergod Theory Dynam Syst*, to appear

15. Heicklen D (1998) Bernoullis are standard when entropy is not an obstruction. *Isr J Math* 107:141–155
16. Heicklen D, Hoffman C (2002) Rational maps are d -adic Bernoulli. *Ann Math* (2) 156(1):103–114
17. Hoffman C (1999) A K counterexample machine. *Trans Amer Math Soc* 351(10):4263–4280
18. Hoffman C (1999) A Markov random field which is K but not Bernoulli. *Isr J Math* 112:249–269
19. Hoffman C (2004) An endomorphism whose square is Bernoulli. *Ergod Theory Dynam Syst* 24(2):477–494
20. Hoffman C (2003) The scenery factor of the $[T, T^{-1}]$ transformation is not loosely Bernoulli. *Proc Amer Math Soc* 131(12):3731–3735
21. Hoffman C, Rudolph D (2002) Uniform endomorphisms which are isomorphic to a Bernoulli shift. *Ann Math* (2) 156(1):79–101
22. Hopf E (1971) Ergodic theory and the geodesic flow on surfaces of constant negative curvature. *Bull Amer Math Soc* 77: 863–877
23. Jong P (2003) On the isomorphism problem of p -endomorphisms. Ph D thesis, University of Toronto
24. Kalikow SA (1982) T, T^{-1} transformation is not loosely Bernoulli. *Ann Math* (2) 115(2):393–409
25. Kammeyer JW, Rudolph DJ (2002) Restricted orbit equivalence for actions of discrete amenable groups. *Cambridge tracts in mathematics*, vol 146. Cambridge University Press, Cambridge
26. Katok AB (1975) Time change, monotone equivalence, and standard dynamical systems. *Dokl Akad Nauk SSSR* 223(4):789–792; in Russian
27. Katok A (1980) Smooth non-Bernoulli K -automorphisms. *Invent Math* 61(3):291–299
28. Katok A, Spatzier RJ (1996) Invariant measures for higher-rank hyperbolic abelian actions. *Ergod Theory Dynam Syst* 16(4):751–778
29. Katznelson Y (1971) Ergodic automorphisms of T^n are Bernoulli shifts. *Isr J Math* 10:186–195
30. Keane M, Smorodinsky M (1979) Bernoulli schemes of the same entropy are finitarily isomorphic. *Ann Math* (2) 109(2):397–406
31. Kolmogorov AN (1958) A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl Akad Nauk SSSR* (NS) 119:861–864; in Russian
32. Kolmogorov AN (1959) Entropy per unit time as a metric invariant of automorphisms. *Dokl Akad Nauk SSSR* 124:754–755; in Russian
33. Krieger W (1970) On entropy and generators of measure-preserving transformations. *Trans Amer Math Soc* 149:453–464
34. Ledrappier F (1978) Un champ markovien peut être d'entropie nulle et mélangeant. *CR Acad Sci Paris Sér A–B* 287(7):A561–A563; in French
35. Lindenstrauss E (2001) Pointwise theorems for amenable groups. *Invent Math* 146(2):259–295
36. Lyons R (1988) On measures simultaneously 2- and 3-invariant. *Isr J Math* 61(2):219–224
37. Mañé R (1983) On the uniqueness of the maximizing measure for rational maps. *Bol Soc Brasil Mat* 14(1):27–43
38. Mañé R (1985) On the Bernoulli property for rational maps. *Ergod Theory Dynam Syst* 5(1):71–88
39. Meilijson I (1974) Mixing properties of a class of skew-products. *Isr J Math* 19:266–270
40. Meshalkin LD (1959) A case of isomorphism of Bernoulli schemes. *Dokl Akad Nauk SSSR* 128:41–44; in Russian
41. Nevo A (1994) Pointwise ergodic theorems for radial averages on simple Lie groups. I. *Duke Math J* 76(1):113–140
42. Nevo A, Stein EM (1994) A generalization of Birkhoff's pointwise ergodic theorem. *Acta Math* 173(1):135–154
43. Ornstein DS (1973) A K automorphism with no square root and Pinsker's conjecture. *Adv Math* 10:89–102
44. Ornstein D (1970) Factors of Bernoulli shifts are Bernoulli shifts. *Adv Math* 5:349–364
45. Ornstein D (1970) Two Bernoulli shifts with infinite entropy are isomorphic. *Adv Math* 5:339–348
46. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. *Adv Math* 4:337–352
47. Ornstein DS (1973) An example of a Kolmogorov automorphism that is not a Bernoulli shift. *Adv Math* 10:49–62
48. Ornstein DS (1973) A mixing transformation for which Pinsker's conjecture fails. *Adv Math* 10:103–123
49. Ornstein DS, Rudolph DJ, Weiss B (1982) Equivalence of measure preserving transformations. *Mem Amer Math Soc* 37(262). American Mathematical Society
50. Ornstein DS, Shields PC (1973) An uncountable family of K -automorphisms. *Adv Math* 10:63–88
51. Ornstein D, Weiss B (1983) The Shannon–McMillan–Breiman theorem for a class of amenable groups. *Isr J Math* 44(1): 53–60
52. Ornstein DS, Weiss B (1973) Geodesic flows are Bernoullian. *Isr J Math* 14:184–198
53. Ornstein DS, Weiss B (1974) Finitely determined implies very weak Bernoulli. *Isr J Math* 17:94–104
54. Ornstein DS, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. *J Anal Math* 48:1–141
55. Parry W (1981) Topics in ergodic theory. *Cambridge tracts in mathematics*, vol 75. Cambridge University Press, Cambridge
56. Parry W (1969) Entropy and generators in ergodic theory. *WA Benjamin*, New York
57. Petersen K (1989) Ergodic theory. *Cambridge studies in advanced mathematics*, vol 2. Cambridge University Press, Cambridge
58. Pinsker MS (1960) Dynamical systems with completely positive or zero entropy. *Dokl Akad Nauk SSSR* 133:1025–1026; in Russian
59. Ratner M (1978) Horocycle flows are loosely Bernoulli. *Isr J Math* 31(2):122–132
60. Ratner M (1982) Rigidity of horocycle flows. *Ann Math* (2) 115(3):597–614
61. Ratner M (1983) Horocycle flows, joinings and rigidity of products. *Ann Math* (2) 118(2):277–313
62. Ratner M (1991) On Raghunathan's measure conjecture. *Ann Math* (2) 134(3):545–607
63. Halmos PR (1960) Lectures on ergodic theory. *Chelsea Publishing*, New York
64. Rudolph DJ (1985) Restricted orbit equivalence. *Mem Amer Math Soc* 54(323). American Mathematical Society
65. Rudolph DJ (1979) An example of a measure preserving map with minimal self-joinings, and applications. *J Anal Math* 35:97–122
66. Rudolph DJ (1976) Two nonisomorphic K -automorphisms with isomorphic squares. *Isr J Math* 23(3–4):274–287
67. Rudolph DJ (1983) An isomorphism theory for Bernoulli free Z -skew-compact group actions. *Adv Math* 47(3):241–257
68. Rudolph DJ (1988) Asymptotically Brownian skew products

- give non-loosely Bernoulli K -automorphisms. *Invent Math* 91(1):105–128
69. Rudolph DJ (1990) $\times 2$ and $\times 3$ invariant measures and entropy. *Ergod Theory Dynam Syst* 10(2):395–406
 70. Schmidt K (1984) Invariants for finitary isomorphisms with finite expected code lengths. *Invent Math* 76(1):33–40
 71. Shields P (1973) The theory of Bernoulli shifts. Chicago lectures in mathematics. University of Chicago Press, Chicago
 72. Sinaĭ JG (1962) A weak isomorphism of transformations with invariant measure. *Dokl Akad Nauk SSSR* 147:797–800; in Russian
 73. Sinaĭ J (1959) On the concept of entropy for a dynamic system. *Dokl Akad Nauk SSSR* 124:768–771; in Russian
 74. 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. Conference on ergodic theory and topological dynamics, Kibbutz Lavi, 1974. *Isr J Math* 21(2–3):177–207; in French
 75. Thouvenot J-P (1975) Une classe de systèmes pour lesquels la conjecture de Pinsker est vraie. Conference on ergodic theory and topological dynamics, Kibbutz Lavi, 1974. *Isr J Math* 21(2–3):208–214; in French