

## T

## Thermodynamics of Computation

H. JOHN CAULFIELD, LEI QIAN  
Fisk University, Nashville, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Thermodynamics  
Computer Equivalents of the First and Second Laws  
The Thermodynamics of Digital Computers  
Analog and Digital Computers  
Natural Computing  
Quantum Computing  
Optical Computing  
Thermodynamically Inspired Computing  
Cellular Array Processors  
Conclusions  
Future Directions  
Bibliography

### Glossary

**Thermodynamics** Literally accounting for the impact of temperature and energy flow. It is a branch of physics that describes how temperature, energy, and related properties affect behavior of an object or event.

**First law of thermodynamics** Also known as the law of *energy conservation*. It states that the energy remains constant in an isolated system.

**Second law of thermodynamics** The second law of thermodynamics asserts that the *entropy* of an isolated system never decreases with time.

**Entropy** A measure of the disorder or unavailability of energy within a closed system.

**Computing** Any activity with input/output patterns mapped onto real problems to be solved.

**Turing machine** An abstract computing model that was invented by Alan Turing in 1936, long before the first electronic computer was invented, to serve as an idealized model for computing. A Turing machine has a tape that is unbounded in both directions, a read-write head and a finite set of instructions. At each step, the head may modify the symbol on the tape right under the head, change the state of the head, and then move on the tape one unit to the left or right. Although extremely simple, Turing machines can solve any problem that can be solved by any computers that could possibly be constructed (see **Church–Turing thesis**).

**The Church–Turing thesis** A combined hypothesis about the nature of computable functions. It states that *any function that is naturally regarded as computable is computable by a Turing Machine*. In early 20th century, various computing models such as Turing Machine,  $\lambda$ -calculus and recursive functions are invented. It was proved that Turing Machine,  $\lambda$ -calculus and recursive functions are equally powerful. Alonzo Church and Alan Turing independently raised the thesis that any *naturally computable function* must be a recursive function or, equivalently, be computed by a Turing Machine, or be a  $\lambda$ -definable function. In other words, it is not possible to build a computing device that is more powerful than those machines with simplest computing mechanisms. Note that Church–Turing thesis is not a provable or refutable conjecture because “*naturally computable function*” is not a rigorous definition. There is no way to prove or refute it. But it has been accepted by nearly all mathematicians today.

**Analog** An analog computer is often negatively defined as a computer that is not digital. More properly, it is a computer that uses quantities that can be made proportional to the amount of signal detected. That analogy between a real number and a physical property gives the name “analog.” Many problems arise in analog computing, because it is so difficult to obtain

a large number of distinguishable levels and because noise builds up in cascaded computations. But, being unencoded, it can always be run faster than its digital counterpart.

**Digital** The name derives from digits (the fingers). Digital computing works with discrete items like fingers. Most digital computing is binary – using 0 and 1. Because signals are restored to a 1 or a 0 after each operation, noise accumulation is not much of a problem. And, of course, digital computers are much more flexible than analog computers that tend to be rather specialized.

**Quantum computer** A quantum computer is a computing device that makes direct use of quantum mechanical phenomena such as superposition and entanglement. Theoretically, quantum computers can achieve much faster speed than traditional computers. It can solve some NP hard or even exponentially hard problem within linear time. Due to many technical difficulties, no practical quantum computer using entanglement has been built up to now.

### Definition of the Subject

Thermodynamics is a very general field that helps to define the physics limits that impose on all stuff. When people study an abstract computing model such as a Turing Machine, thermodynamics does not impact their behavior because nothing physical is been done. However, real computers must be made of some physical material, material that has a temperature. The physical stuff must have at least two readily distinguishable states that can be interpreted to two different values. In the process of computing, the stuff should be able to change the state under control. The change of the state must obey the laws of thermodynamics. For example, there is a minimum amount of energy associated with a bit of information. Information is limited by the entropy change  $k \log 2$  and the energy is  $kT \log 2$ , where  $k$  is the Boltzmann's constant and  $T$  is the absolute temperature of the device and its surroundings. Mysteriously, information and entropy seem linked even though one is physical and the other is conceptual. Those limits are vital to know, as it allows designers to stop seeking improvements as they reach or near those limits. At the same time, thermodynamics indicates where computers do not come near what is possible. Such aspects of computers represent potentially important fields to explore.

### Introduction

Thermodynamics – the science of the changes in energy over time – is one of the most fundamental sciences. Arguably, the first and second laws of thermodynamics are

the scientific principles most scientists would label as the most fundamental of all the laws of science. No one ever expects them to be violated. We will explore the following aspects of the relationship between thermodynamics and computing

- The energetics and corresponding temporal properties of classical digital computers.
- Possible modifications such as conservative computing, analog computing, and quantum computing.
- Thermodynamically-inspired computing.

Unsurprisingly, we will find that those topics are all related.

Before we can do any of those things, however, we must review thermodynamics itself and do so in terms suited for this discussion

### Thermodynamics

Thermodynamics originated as a set of general rules characterizing the effects of system parameters such as energy and temperature over time. Somewhat later, thermodynamics came to be partitioned into two broad classes – the classical thermodynamics (now called equilibrium thermodynamics as it predicts the final system parameters that result from some perturbing effect) and nonequilibrium thermodynamics (what happens if energy input to a system continues over time so equilibrium is never obtained).

The fact that thermodynamics applies to quantities averaged over the system and deals only with probabilistic results, means that no physical law is broken if entropy within the system decreases due to either of two reasons: the occurrence of an improbable event that is allowable but fleeting or the occurrence of a state that is not reachable by equilibrium thermodynamics but is reachable with continuing input of energy. The latter is an example of nonequilibrium thermodynamics. This also connects nonequilibrium thermodynamics with topics such as nonlinear dynamics, chaos, emergence, and so forth. Two concepts from nonlinear dynamics are especially interesting: attractor states and emergence. Attractor states are system states such that if the state of the system is near a given attractor (within its “attractor basin”), successive states tend to move toward that attractor. Some of the attractor states in systems far from equilibrium can be highly organized. The properties of highly organized nonequilibrium thermodynamic states (states can be fixed conditions, periodic events, or even more complex situations called strange attractors that are somewhat predictable and necessarily chaotic in the technical sense of that word. There is generally a control parameter such as power input which de-

termines the system performance. For some values systems often develop excessive sensitivity to the values of the control parameter leading to long term unpredictability of the deterministic result – a condition called chaos. Systems that are too unorganized like an ideal gas, are not very interesting from a behavior viewpoint. Systems that are fully organized like a crystal also have limited interest. But between those extremes lie regions of high complexity and great interest. One of the favorite and well justified sayings of complexity theorists is that life arises on the edge of chaos. You are a system that shows spectacular organization if kept away from equilibrium with continuing input of energy. The properties that arise far from equilibrium are said to be emergent and are generally unpredictable until discovered experimentally. A general feature of such organized complexity in systems maintained well away from equilibrium is self organization. That is the order observed does not have to be created by a scientist. Instead it arises spontaneously under the correct circumstances.

Turing machines can be accounted for by conventional thermodynamics, but some computer such as neural networks and cellular array processors can become complex enough to exhibit interesting and useful emergent properties that must be discussed in terms of nonlinear thermodynamics. Such self organized complex processor behavior has also been called “synergistic.”

The most complex computer we know is the human brain. It not only arises on the edge of chaos, but also uses and consumes chaos. Over-regular temporal behavior of the brain is called epilepsy and is not a desirable state.

Two of the first and most influential people to explore nonequilibrium thermodynamics were Warren Weaver and Ilya Prigogine.

Weaver [1] discussed three ways people can understand the world: simple systems, statistical methods (conventional thermodynamics), and nonequilibrium organized complexity [thermodynamics]. Organized complexity is compatible with classical thermodynamics from an overall perspective but can be achieved locally in space and time.

Prigogine’s Nobel-Prize-winning work on nonequilibrium thermodynamics is explained beautifully in his popular books [2,3,4].

Let us now review very briefly what thermodynamics deals with and why it is important to computers.

Temperature,  $T$ , is a measure of the driving force for energy. There can be no systematic energy flow from a lower temperature region to a higher temperature region.

Entropy,  $S$ , is a measure of the disorder. The second law of thermodynamics, perhaps science’s most sacred

law, says that entropy never decreases. Indeed it nearly always increases. Note that entropy is essentially an abstract quantity. Indeed, it is the negative of information as defined by Shannon. Therefore information is sometimes called negentropy – negative entropy. There is a metaphysical puzzle here that has been more ignored than solved. How does a mathematical concept like information come to govern physics as does entropy?

Boltzmann’s constant  $k$  symbolizes the relationship between information and physics. It appears in purely physical equations such as the ideal gas law. It also appears in equations about information. Leo Szilard suggested that writing a memory does not produce entropy; it is possible to store information into a memory without increasing entropy. However, entropy is produced in every case that the memory is *erased*. It turns out that the (minimum) entropy created by erasing one bit is given by

$$S \text{ per erased bit} = k \ln 2,$$

and the number  $\ln 2 \approx 0.69$  is the natural logarithm of 2. What’s going on here? How are information and the physical world related?

Information relates to the answers to yes or no questions. Entropy is physical entity not necessarily related to questions and answers. Yet the mathematics of both is the same. Information is the negative of entropy. What are we to make of this metaphysically? One view (originating in the mind and writing of John Archibald Wheeler) is that information is primary. The slogan is “It from Bit.” Wheeler’s metaphysics is extremely appealing, because it requires no external scaffolding. But it strikes many as not quite convincing. There is an opposite approach that begins with stuff and energy. The physical world is made of relationships between parts. The universe is self organized and self defined. Stuff is differentiation (yes and no) within the whole. Another view is that physics (stuff, energy, motion) is primary. Information in a brain or a computer is present only if some stuff is used and some energy is expended in generating, transmitting, or storing it. In that case, information relates to the probability that a given record would appear by chance if all records were equally probable. The less likely the chance appearance, the more information is present. In this way, information ties back to Bayes’s Theorem. Neither caricature of a position is complete enough to be fair, but perhaps this footnote will stimulate some readers to dig deeper. This is not a metaphysics chapter, so we take no side in this dispute. But, some of us are physicists and physicists are drawn to physics by the hope of finding answers not just equations. We cannot help asking such questions, so we think noting this unresolved metaphysical problem is appropriate.

Note also, that no metaphysical problem can ever be fully resolved, as the criteria for truth of metaphysics lie in the domain of metametaphysics (Ockham's razor, etc.) and so on ad infinitum.

There is a fundamental relationship between the change  $\Delta E$  in the physical quantity energy and the change  $\Delta S$  in the mathematical quantity entropy, namely

$$\Delta E = T \Delta S.$$

What is remarkable about thermodynamics is that it tells us nothing about how things actually work. It tells us instead about what cannot be done. It is the science of fundamental limits. It is, in that sense, a negative field. Physicists sometimes have to work very hard to find the detailed ways nature enforces these limits in any particular problem.

When energy ceases to have any net flow, the system is said to be at equilibrium. Until the latter part of the 20th century, thermodynamics meant near-equilibrium behavior. Far from equilibrium, very interesting things can occur. Order can arise from disorder at least locally. The second law is satisfied by creating even more disorder somewhere in the universe. Life exists far from equilibrium and seems to increase order. Indeed it does – locally. But it must take energy from its environment and cause a global increase in energy.

The laws of thermodynamics define and thus forbid magic – something from nothing (the first law) or effects without causes (the second law). We will interpret all magic forbidding laws as “thermodynamic” even if the apply to nonphysical things such as information

### Computer Equivalents of the First and Second Laws

We propose two laws of information inspired by the two laws of thermodynamics just described.

**First law: Information is not increased in a computer.** A computer cannot create information. The output result is strictly determined by the data and the instructions. If we were smart enough and had unlimited resources (time and memory), we would know that  $1 + 2 + 3 + 4 + 5 = 15$ . We only need a computer, because computer can compute faster and more reliably. We may not have known the answer now, but we could if we had enough time, memory and patience. The result is not new but derived information.

**Second law: The energy required to destroy one bit of information is not smaller than the temperature  $T$  multiplied by some minimum amount of information  $\Delta S$ .** This second law applies for conservative as well as dissipative logic – terms to be described later. This

second law, like the second law of thermodynamics, deals with irreversible acts.

### The Thermodynamics of Digital Computers

Just as all digital computers are equivalent to a Turing machine, all computations in digital computers can be thought of as being performed by Boolean logic gates. If we understand the thermodynamics of a Boolean logic gate, we understand the thermodynamics of any digital computer.

All readers of this chapter will recognize that Boolean logic gates such as *AND*, *OR*, *NOR*, and so forth convert two one-bit inputs into a single one-bit output. Immediately, we notice

- A Boolean logic gate has less information out than we put in.
- Such a logic gate is irreversible.

From the two laws of computing, we conclude.

- Since information must be conserved, one bit of information must somehow have been converted into some noninformative state.
- The amount of energy required is related to the information content of a single bit multiplied by the operating temperature  $T$ .

The person who first realized and determined the amount of energy needed to destroy one bit of information is

$$\Delta E = kT \log 2$$

was an IBM scientist named Landauer. Here  $k$  is the Boltzmann constant. The average energy of a molecule at temperature  $T$  is about  $kT$ , so this makes some intuitive sense as well.

This is really a very small amount of energy – molecules of that energy are striking you right now and you can't even feel them. Two things make this energy remark important. First, most current gates require millions of  $kT$ s per operation. This is no accident. It is required as we will see later when we discuss analog computing. Second, the number of operations per second is now in the billions. So trillions of  $kT$ s are dissipated in current computers each second. They get very hot and consume a great amount of power.

Landauer and fellow IBM physicist Bennett then took a totally unexpected and brilliant direction of research. What would happen, they asked, if we could use logic gates that do not destroy information? Then there would be no thermodynamic minimum energy requirement for such



a gate's operation. Perhaps the energy cost of computing could be reduced or even eliminated.

A trivial example from arithmetic will illustrate these concepts for readers not familiar with them.

Here is a conventional (dissipative) arithmetic expression:

$$2 + 3 = 5.$$

It would be more informative to write

$$2 + 3 \Rightarrow 5.$$

If we know the data (2,3) and the instructions (+), the result (5) is determined. But this is irreversible. If we know the result (5) and the instruction to be inverted (+), we cannot recover the data (2, 3). The data could have been (3, 2), (1, 4), etc. Here is an information conserving arithmetic expression:

$$2, 3 \textcircled{R} 3, 2.$$

The operator  $\textcircled{R}$  means reverse the order of the data. Such an operation is clearly conservative of information. And, clearly, it is reversible:

$$3, 2 \textcircled{R} 2, 3.$$

An  $\textcircled{R}$  gate would have no inevitable energy price. It destroys no information (our first law is satisfied).

Here is a not such trivial example. Consider the operation:

$$a, b \textcircled{C} a + b, a - b$$

This operation is also conservative of information because with  $a + b$  and  $a - b$ , one can easily recover the original  $a$  and  $b$ .

Since that time, a number of conservative, reversible logic gates have been proposed and implemented. We can say some things in general about such gates.

- They will have equal numbers of inputs and outputs (usually three).
- The changes effected by the gates must either rearrange the data (like the  $\textcircled{R}$  operator in our arithmetic example) or mix them (like the  $\textcircled{C}$  in the second example). All electronic gates so far proposed involve rearrangements. In optical logic, reversible mixing is common.

Stringing conservative logic gates together can allow us to produce sought after results. Inevitably, however, we will also have computed things we do not care about – called garbage.

Garbage is inevitable in reversible computing, because the components must retain the information normally destroyed by a conventional Boolean logic gate. Reversibility eliminates one problem (the inevitable dissipation of information) but creates another (the inevitable appearance of garbage bits).

Nevertheless, we must detect the desired answer, and that must cost at least  $kT \log 2$ . The zero energy possibility applies only to the logic operations, not the entering of data into the system or the extraction of data from the system. Failure to understand this has caused more than a few to proclaim that zero-energy logic operations are impossible.

We enter data at the input ports and wait for an answer to reach the output ports. As the system is reversible, not all occurrences lead in the “right direction.” Temperature-dependent random walk will eventually produce the answer. But with a little energy bias in the forward direction, we will get the answer faster at the cost of expending energy – the more energy, the faster the results are likely to arrive. One might expect that something akin to the energy-time uncertainty principle might apply here, so very low energy gates might require a very long time to operate. Indeed that has usually been the case. An exception will be noted in the case of certain optical elements.

## Analog and Digital Computers

Because digital computers and computation have been so successful, they have influenced how we think about both computers as machines and computation as a process – so much so, it is difficult today to reconstruct what analog computing was all about... It is a history in which digital machines can do things ‘better’ and ‘faster’ than other machines... However, what is at stake here are not matters of speed or precision. Rather, it is an argument about what can be rendered and understood through a machine that does computation [5].

All computers involve the operation of some physical system and the interpretation of physically measurable quantities in numerical terms. The most obvious interpretation is

$$N = cQ.$$

The number  $N$  is the measured quantity  $Q$  multiplied by some conversion coefficient. In classical physics  $Q$  is a continuous variable. Computers using such numbers are called analog computers.

Analog computers have several profound disadvantages relative to digital computers.

- **Sensitivity.** By this we mean the ability to distinguish small difference in large numbers. Analog computers tend to have trouble being able to distinguish, say, 47 from 48. Digital computers only have to distinguish between 0 and 1. It encodes bigger numbers in terms of 0s and 1s.
- **Inability to compute complex or even real numbers in most cases,** because most useful quantities  $Q$  are usually nonnegative. This is usually “solved” by encryption of real or complex numbers as multiple nonnegative numbers.
- **Cumulative errors.** When the output from one analog computation is used as an input to another, the errors are carried forward and accumulated. This effect is made even worse by the fact that some problems amplify the input errors significantly. In linear algebra problems, do instance, that error amplification factor is called the condition number of the matrix. That number is never less than 1 and often exceedingly large (ill conditioned matrices) or even infinite (singular matrices).
- **Inflexibility.** This chapter was written using a word processor in a digital computer. Making an analog word processor would require almost inconceivably complex.

Some distinct advantages come from being continuous in space or time [6].

In addition, it is widely believed that anything an analog computer can compute can also be computed by a digital computer. In fact, that belief is a corollary of what is called the strong Church–Turing thesis. It asserts that anything that can be computed in any way (including analog computations) can also be computed by a Turing machine. The more widely accepted form of the Church–Turing thesis simply asserts that anything that is sensibly computable by any computer can be computed by a Turing machine. There are some who believe they have discovered exceptions to the Church–Turing thesis and many more who doubt that. This is not the place to examine that controversy. It is mentioned here for two reasons.

First, it suggests that analog computers cannot do anything a digital computer cannot do. There is no advantage in principle. There may be practical advantages to analog computing when it suffices as we have already suggested, though. But there is much more to computing than simply getting the right answer. Speed, cost, and power consumption are examples. Sometimes, analog computing can win on those criteria.

Second, the Church–Turing thesis is thermodynamic in the generalized sense that it sets presumably immutable

constraints but does not provide information on how those constraints are enforced.

So why would anyone bother with analog computers? There are many reasons. Here are some obvious ones.

- A binary gate is an analog gate followed by a nonlinear operator (a quantizer) that sets low signals to 0 and high signals to 1 or vice versa. So the difference between analog and binary computing lies in two things – the number of quantization levels (2 for binary and many more for analog) and how often measurements and corrections are made (after each operation for binary computers and after the whole computation for analog). So, we can argue that all computers are analog at their hearts.
- But encryption in terms of many binary operations, they are inevitably slower and more power hungry than digital computers. We could digitize to more than two values. It takes far fewer ternary values than binary values to represent a large number, but the more quantization levels, the more probable it is that we will make an error. It is generally assumed that two is the optimum number overall.

But there are more subtle effects that sometimes favor analog over digital.

Consider the representation of physically measured time signal in a digital computer. The sampling theorem tells us how frequently we must sample the signal IF we know its bandwidth. But suppose we do not know. We can do either of two dangerous things: apply a bandpass filter and risk losing information or guess the bandwidth and risk producing aliasing. So discretization of an analog signal has a significant probability of giving misleading results.

Iterative digital solution of linear algebra problems converge only for condition numbers limited by the digitization itself, while analog solutions have no such limit because the condition number itself is undefined.

A more general treatment can be found in [7].

For chaotic systems, discretization can lead to wildly different results with different levels of discretization [8,9,10].

So what can we say about the thermodynamics of analog computers? The operation occurs in three steps:

1. **Preparation.** The apparatus is set up so that the input parameters are the ones we choose and the physics will carry out the desired transformations.
2. **Time Evolution.** The physical system evolves through time.

3. **Readout.** The desired physical parameters are measured.

Preparation is necessary in digital computing as well and has not been treated in any thermodynamic discussion. Nor will we treat it here. The last two steps must somehow be governed by the two laws of computing propounded earlier. That is, the system cannot produce what we interpret as information beyond that we inserted in the preparation. The underlying physics may be exceedingly complex, but the informatics must be as described by the first law. The readout requires some minimum energy. In a digital computer with  $L$  levels the Landauer analysis says we must dissipate  $kT \log L$  in energy. Here is at least a qualitative explanation of the fact that we need far more than  $kT$  to operate a binary logic gate. We need to measure to far more than just two levels in order to make the binary decisions with sufficiently low error rate.

Analog computing is an emerging field, in part because of interest in what is called “natural computing.”

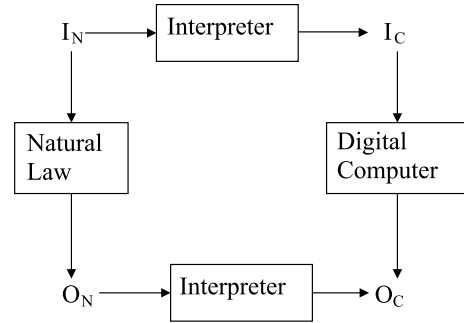
### Natural Computing

In serial non Von Neumann machines the interconnections between processors have negligible impact on system space and energy consumption. But concurrency (e. g. parallelism) requires more interconnections and more attention to their space and energy consumption. And those problems become worse as the number of concurrent processors increases. For large enough numbers of processors, interconnections are no longer negligible and can become limiting.

Two approaches have been proposed to ameliorate such problems: optical interconnection and quantum computing. Optical interconnection may be useful, because optical paths can cross in space without effect. That is many optical paths can share the same space. Quantum computing is of interest because  $N$  entangled states can probe  $2^N$  paths. Both avoid massive interconnection with wires.

Richard Feynman anticipated both approaches. One of Feynman’s goals in suggesting quantum computing was to solve the wiring problem. Another motivation also guided him. He knew that the accuracy with which the results of a quantum experiment could be predicted depended on the number of path integrals evaluated, so (in a sense) each natural event in quantum mechanics is equivalent to infinitely many calculations. The only way to calculate quantum events with perfect accuracy would be to use a quantum computer.

Let us describe a natural computation. We begin with a physical system  $P$  that operates on an input  $I_N$  to pro-



**Thermodynamics of Computation, Figure 1**

Nature converts input  $I_N$  into output  $O_N$  by natural law. The digital computer performing some fixed algorithm converts input  $I_C$  into output  $O_C$ . If, for all allowable  $I_N$ , we have a rule that converts it to an equivalent  $I_C$  and if the output of the natural computer  $I_N$  can be converted to the output  $I_C$ , then we have in the physical operator a “natural computer”

duce an output  $O_N$ . It does this naturally without human assistance.

Now suppose we wish to use that physical process  $P$  to calculate some result. If there is a computer operation  $C$  that operates on an input  $I_C$  to produce an output  $O_C$ , and if there are mappings

$$M1 : I_N \leftrightarrow I_C$$

$$M2 : O_N \leftrightarrow O_C$$

Then we can declare that the  $P$  and  $C$  are congruent and thus compute the same results if the diagram in Fig. 1 commutes.

That is if nature in doing what it will do automatically allows us to establish the starting state  $I_N$  (state preparation) in a way that can be viewed using  $M1$  as representing the computer input  $I_C$  and if the output  $O_N$  of the natural process and if  $M2$  carries that to the very output that a conventional computer produces with input, then the two processes are congruent and we have a natural computer.

Certainly, the most widely studied form of natural computing at this moment is the quantum computer.

### Quantum Computing

The behaviors of matter in very small scale obey laws of quantum mechanics, which are so different from what we usually experience as to be essentially beyond our understanding. We can write the equations and perform the experiments but the results inevitably seem weird. Despite this, it is possible to harness some of these effects for computing. Quantum computers usually utilize superposition

(a principal of quantum mechanism which claims that while the state of any object is unknown, it is actually in all possible states simultaneously, as long as we don't measure it) and quantum entanglement (a quantum mechanical phenomenon in which the quantum states of two or more objects have to be described with reference to each other, even though the individual objects may be spatially separated) to perform computation [11,12].

In traditional digital computers, information is carried by numerous binary digits, which are called bits. In quantum computing, information is carried by particles which carry quantum states. Such quantum logic bit is called qubit. A qubit can be expressed by

$$|i\rangle = a|0\rangle + b|1\rangle$$

where  $|0\rangle$  and  $|1\rangle$  are two orthonormal states. The coefficients  $a$  and  $b$  are the normalized probability amplitudes (in general, complex-valued numbers) of states  $|0\rangle$  and  $|1\rangle$  that satisfy the condition  $|a|^2 + |b|^2 = 1$ . These two orthonormal states can be polarization in orthogonal directions, two oppositely directed spins of atoms, etc. It seems that a qubit carries partially information of  $|0\rangle$  and partially information of  $|1\rangle$ . However, when a qubit is measured (which is necessary if we want to read the output), the qubit "collapses" to either state 0 or state 1 with the probabilities  $|a|^2$  or  $|b|^2$ . This process is irreversible. Whence a qubit collapses to a state, there is no way to recover the original state. On the other hand, it is also not possible to clone a quantum state without destroying the original one.

A very weird but extremely useful phenomenon in quantum world is **entanglement**. It is the coherence among two or more objects even though these objects are spatially separated. Consider a 2-qubit state:  $(|0\rangle|0\rangle + |0\rangle|1\rangle + |1\rangle|0\rangle + |1\rangle|1\rangle)/2$ . It can be decomposed to  $((|0\rangle + |1\rangle)/\sqrt{2})(|0\rangle + |1\rangle)/\sqrt{2}$ . That means we have 50% probability to read  $|0\rangle$  and 50% probability to read  $|1\rangle$  when we measure the first qubit. We will get same probability for the second qubit. The measured value of the first qubit is independent of that of the second qubit. However, consider the state:  $1/\sqrt{2}(|0\rangle|1\rangle - |1\rangle|0\rangle)$ . It is easy to prove that one cannot decompose the state into a form  $(a|0\rangle + b|1\rangle)(c|0\rangle + d|1\rangle)$ . Actually, if we measure the first qubit, we have 50% probability to read  $|1\rangle$  and 50% chance to read  $|0\rangle$ . However, if we read  $|0\rangle$  when the first qubit is measured, we **always** read  $|1\rangle$  for the second qubit; if we read  $|1\rangle$  for the first qubit, we **always** read  $|0\rangle$  for the second qubit. The measurement to one of them will affect the state of the other qubit immediately no matter how far they are from each other.

Consider a quantum system with  $n$  qubits. If there is no entanglement, it can be decomposed into the product of  $n$  1-qubit systems, or  $2n$  independent numbers. However, if entanglement is allowed, then there are  $2^n$  independent coefficients and therefore can hold  $2^n$  independent values. So it can operate tremendous number of values in parallel. It is possible to solve exponentially hard or NP complete problems in polynomial or even linear time. In 1994, Peter Shor of AT&T devised an algorithm for factoring an integer  $N$  in  $O((\log N)^3)$  time and  $O(\log N)$  space with a quantum computer. The security of RSA, the most widely used public-key cryptography method is based on the assumption that factoring a large integer is not computationally feasible. A quantum computer can break it easily.

Even though quantum computers are much more powerful than traditional digital computers in terms of speed, it still obeys Church-Turing thesis. In other words, no quantum computer can solve a problem that a traditional computer cannot. Quantum computers can still be simulated by traditional computers. The difference is that a traditional computer may have to spend millions of years to solve a problem that a quantum computer can solve in seconds. Therefore, the first law of computing – information cannot increase in a computer – is still true because a quantum computer has the same power as a traditional computer in terms of computability.

Just as Boolean logic gates play a central role in traditional computers, quantum logic gates are the heart of quantum computers. Unlike Boolean logic gates, all quantum logic gates are reversible. More precisely, a quantum logic gate can be represented by a unitary matrix  $M$  (an  $n \times n$  matrix  $M$  is unitary if  $M^+M = I_n$  where  $M^+$  is the conjugate transpose of  $M$ ). So if  $M$  is the matrix that represents the quantum logic gate,  $M^+$  represents the reverse operation of the gate.

While it is possible to define quantum logic gates for any number of qubits, the most common quantum logic gates operate on spaces of one or two qubits. For example, Hadamard gate and phase shifter gates are single-qubit gate and CNOT (controlled-NOT) gate is a two-qubit gate. With Hadamard gates, phase shifter gates and CNOT gates, one can do any operation that can be done by a quantum computer. We call sets of quantum gates like this by **universal quantum gates**.

Because quantum logic gates are all reversible, quantum computers do not destroy any information until outputs are measured. Therefore, quantum computers may consume less energy than traditional computers in theory.

While quantum computers have tremendous power and may consume much less energy, they are extremely difficult to be built. A quantum computer must maintain

coherence among its qubits long enough to perform an algorithm. Preventing qubits from interacting with the environment (which will cause decoherence) is extremely hard. We can get around the difficulty by giving up the entanglements. However, the quantum computer is no longer as powerful if we do so. Without entanglements, the state space of  $n$ -qubits only has  $2n$  dimensions rather than  $2^n$  dimensions. Exponentially hard or NP complete problems cannot be computed by these computers within polynomial time (assume  $P \neq NP$ ) [13]. However, these computers use reversible logic gates and therefore have advantage in energy consuming.

In conclusion, even though quantum computers use a quite different model, two laws of information still apply to them.

## Optical Computing

Optical computing has been studied for many decades [14]. This work has led to conclusions that seem obvious in retrospect; namely that

- Optical computing can never compete with electronic computing for general purpose computing.
- Optical computing has legitimate roles in niches defined by the characteristic that the information being processed is already in the optical domain, e. g. in optical communication or optical storage.
- Two types of complexity arise. One is the complexity of the optical effects themselves, e. g. optics often has three or more activities going on at each location, so if the processes are nonlinear (e. g. photorefractives or cross talk among active elements) then complexity factors are likely to arise [15].

This is nonequilibrium thermodynamics at work. Not to be confused with this is the savings in temporal coherence accomplished by use of space, time, and fanin complexity to consume computational complexity [16,17].

Optical computing is limited by thermodynamics in a unique sense. Let us suppose that we wish to measure an output signal that is either 1 or 0 in amplitude. Either for speed or power consumption, we seek to measure as few photons as possible to decide which of those two states (0 or 1) is present. Because the photons do not (without encouragement) come evenly spaced in time of arrival, we must detect many photons to be sure that no detected photon means a logical zero rather than a statistical extreme when a logical 1 is intended. With normal statistics, it takes about 88 detected photons to distinguish between what was intended as a logical 1 and what was intended

as a zero. It is possible to make a light source with substantially uniform emission rates. This is called squeezed light. Naively, squeezing can overcome that problem to an amazing extent, but the squeezed state is very fragile and many of the central activities in an optical computer destroy the order squeezing put in before it can be detected [18].

## Thermodynamically Inspired Computing

The approach of systems to equilibrium may become a model for computing. The most prominent of these are called simulated annealing, the Hopfield net, and the Boltzmann machine. Basically, they are ways to encode a quantity to be minimized as a sort of pseudo energy and then simulating some energy minimization activity. And, of course hybridization of these methods make the clean division among them impossible.

Annealing of materials is an ancient art. Domains in a metal sword, for instance, may be randomly disposed to one another, making them brittle. The goal of annealing is to get them aligned, so they behave more like a single crystal and thus hardened. This is done by a method sometimes called “heat and beat.” The material is heated to a high temperature where domains changes are fairly easy and then beat with a hammer to encourage such movement. Then the material is cooled to a little lower temperature where one hopes the gains so far are not rerandomized and beat again. The art is in the cooling schedule and in the beating. In thermodynamic terms, we are trying to get the sword in its minimum energy state where no beating against the opponent will cause it to relax into a lower energy state. Unlike more familiar optimization methods such as steepest descent or Newton–Raphson that find only a local extremum (one near the starting state), simulated annealing aims at finding the global extremum independently of starting state. It does this by pure imitation of annealing. Some monotonic figure of merit for the problem is chosen. It might be the RMS difference between the output vector of a matrix-vector multiplication and a target vector. Minimizing it is analogous to minimizing the sword’s energy. Driving it to or near zero by adjusting the input vector is a way of solving the ubiquitous  $A\mathbf{x} = \mathbf{b}$  problem for  $\mathbf{x}$ . In early stages, big random steps (chosen from a thermodynamic expression governing probabilities as a function of energy and “temperature”) are used. The perturbed system is then “cooled” in the sense of allowing the starting fictitious  $T$  value to decrease slight, and the process is repeated. This is very computer intensive but very effective. More information can be found in many places including [19,20,21].



A Hopfield net [22] is very similar to simulated annealing in that it uses guided stochastic methods to minimize a user-defined energy (Figure of merit). It is based on adjusting connections of a rather simple binary connected neural network by picking a starting state and adjusting one interconnection at a time.

A Boltzmann machine is a more thermodynamically inspired version of a Hopfield network [23].

One more thermodynamically-inspired means for computation needs to be mentioned, namely stochastic resonance [24]. It is a creative use of noise. Noise is usually thought of as the enemy of signal. But if in some nonlinear operation the signal is so small that it is not useful, adding a certain amount of noise to the input may help. Indeed, there is an optimal amount of noise for each case. Too little or too much noise are both bad, but there is a noise value in resonance with the problem which give optimal results.

This is not the place for detailed discussions of these thermodynamically-inspired methods. Rather, our intent is to show that thermodynamics has impacted not only physical computers but also the way those computers are used to achieve certain goals.

### Cellular Array Processors

Cellular array processors (some prefer “cellular automata”) are discrete binary valued elements in fixed locations or arrays [25]. Usually started with random inputs, they can be observed to produce quite interesting space-time histories as each cell is updated according to a fixed rule using its current value and those of its immediate neighbors. Under some circumstances, those space-time patterns simply terminate or generate chaos. Often, however, the system self organizes to produce highly organized patterns. One “application” is so-called “Artificial Life” which began with the immensely popular set of evolution rules for a two-dimensional array called “the Game of Life.” For the few who have not played with it, we urge you to do so. An online version is available free at <http://www.bitstorm.org/gameoflife/>. Many brilliant computer scientists have become enthralled with it. For instance, Stephen Wolfram analyzed it in terms of thermodynamics [26].

### Conclusions

Thermodynamics

- Governs all computing.
- Has inspired conservative, reversible computing.
- Is not easy to apply to non Turing computers such as analog and quantum.
- Has directly inspired some optimization methods.

In the end, computers are at least partially physical in the classical sense, so they are ultimately limited by thermodynamics. The presence of many virtual operations in some forms of quantum computing can reduce the price of physicality dramatically but cannot eliminate it [27].

### Future Directions

Classical equilibrium thermodynamics places limits on what can be accomplished under different conditions. More importantly, it tells us when we are approaching a limit state. Things often get exceedingly difficult in such cases, so effort might be better spent where there is greater room for improvement.

Optical computing may allow the reaching of the energy required for reversible computing soon. As there is no fundamental limit but negative energy is not defined, we can say that zero-energy computing is the limit. Doing the computation by interferometry of beams passing through passive optics requires no energy. At last, we will have an example of how the choice of medium can be critical. That kind of zero-energy computation is not possible for electronics. Early versions of zero-energy gates have already been described and made [28,29]. That work might not have been undertaken except for the “permission” for zero-energy systems granted by thermodynamics.

### Bibliography

#### Primary Literature

1. Weaver W (1948) Science and complexity. *Am Sci* 36:536–541
2. Nicolis G, Prigogine I (1977) Self-organization in nonequilibrium systems. Wiley, New York
3. Prigogine I, Stengers I (1984) Order out of chaos. Bantam Books, New York
4. Prigogine I, Stengers I, Toffler A (1986) Order out of chaos: Man's new dialogue with nature. Flamingo, London
5. Nyce JM (1996) Guest editor's introduction. *IEEE Ann Hist Comput* 18:3–4
6. Orponen P (1997) A survey of continuous-time computation theory. In: Du DZ, Ko KI (eds) *Advances in Algorithms, Languages, and Complexity*. Kluwer, Dordrecht, pp 209–224
7. Pour-El MB, Richards JI (1989) Computability in Analysis and Physics. Springer, Berlin
8. Grantham W, Amit M A (1990) Discretization chaos – Feedback control and transition to chaos. In: *Control and dynamic systems*, vol 34. *Advances in control mechanics*. Pt. 1 (A91–50601 21–63). Academic Press, San Diego, pp 205–277
9. Herbst BM, Ablowitz MJ (1989) Numerically induced chaos in the nonlinear Schrödinger equation. *Phys Rev Lett* 62: 2065–2068
10. Ogorzalek MJ (1997) Chaos and complexity in nonlinear electronic circuits. *World Sci Ser Nonlinear Sci Ser A* 22
11. Nielsen MA, Chuang IL (2000) Quantum Computation and Quantum Information. Cambridge University Press, New York

12. Bouwmeester D, Ekert A, Zeilinger A (2001) *The Physics of Quantum Information*. Springer, Berlin
13. Caulfield HJ, Qian L (2006) The other kind of quantum computing. *Int J Unconv Comput* 2(3):281–290
14. HJ Caulfield, Vikram CS, Zavalin A (2006) Optical logic redux. *Opt Int J Light Electron Opt* 117:199–209
15. Caulfield HJ, Kukhtarev N, Kukhtareva T, Schamschula MP, Banarjee P (1999) One, two, and three-beam optical chaos and self organization effects in photorefractive materials. *Mat Res Innov* 3:194–199
16. Caulfield HJ, Brasher JD, Hester CF (1991) Complexity issues in optical computing. *Opt Comput Process* 1:109–113
17. Caulfield HJ (1992) Space – time complexity in optical computing. *Multidimens Syst Signal Process* 2:373–378
18. Shamir J, Caulfield HJ, Crowe DG (1991) Role of photon statistics in energy-efficient optical computers. *Appl Opt* 30:3697–3701
19. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220(4598):671–680
20. Cerny V (1985) A thermodynamical approach to the travelling salesman problem: an efficient simulation algorithm. *J Optim Theory Appl* 45:41–51
21. Das A, Chakrabarti BK (eds) (2005) *Quantum Annealing and Related Optimization Methods*. Lecture Notes in Physics, vol 679. Springer, Heidelberg
22. Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci USA* 79:2554–2558
23. Hinton GE, Sejnowski TJ (1986) Learning and Relearning in Boltzmann Machines. In: Rumelhart DE, McClelland JL (eds) and the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. vol 1. Foundations. Cambridge MIT Press, Cambridge, pp 282–317
24. Benzi R, Parisi G, Sutera A, Vulpiani A (1983) A theory of stochastic resonance in climatic change. *SIAM J Appl Math* 43:565–578
25. Ilachinski A (2001) *Cellular Automata: A Discrete Universe*. World Scientific, Singapore
26. Wolfram S (1983) Statistical mechanics of cellular automata. *Rev Mod Phys* 55:601–644
27. Kupiec SA, Caulfield HJ (1991) Massively parallel optical PLA. *Int J Opt Comput* 2:49–62
28. Caulfield HJ, Soref RA, Qian L, Zavalin A, Hardy J (2007) Generalized optical logic elements – GOLEs. *Opt Commun* 271:365–376
29. Caulfield HJ, Soref RA, Vikram CS (2007) Universal reconfigurable optical logic with silicon-on-insulator resonant structures. *Photonics Nanostruct* 5:14–20

## Books and Reviews

- Bennett CH (1982) The thermodynamics of computation a review. *Int J Theor Phys* 21:905–940
- Bernard W, Callen HB (1959) Irreversible thermodynamics of non-linear processes and noise in driven systems. *Rev Mod Phys* 31:1017–1044
- Bub J (2002) *Maxwell's Demon and the thermodynamics of computation*. arXiv:quant-ph/0203017
- Casti JL (1992) *Reality Rules*. Wiley, New York
- Deutsch D (1985) *Quantum Theory, the Church-Turing Principle*

and the Universal Quantum Computer. *Proc Roy Soc Lond Ser A Math Phys Sci* 400:97–117

Karplus WJ, Soroka WW (1958) *Analog Methods: Computation and Simulation*. McGraw Hill, New York

Leff HS, Rex AF (eds) (1990) *Maxwell's Demon: Entropy, Information, Computing*. Princeton University Press, Princeton

Zurek WH (1989) Algorithmic randomness and physical entropy. *Phys Rev Abstr* 40:4731–4751

## Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation

CAI-ZHUANG WANG<sup>1</sup>, GUN-DO LEE<sup>2</sup>, KAI-MING HO<sup>1</sup>

<sup>1</sup> Ames Laboratory and Department of Physics and Astronomy, Iowa State University, Ames, USA

<sup>2</sup> School of Materials Science and Engineering and Inter-university Semiconductor Research Center (ISRC), Seoul National University, Seoul, Korea

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Tight-Binding Potentials for Carbon](#)

[TBMD Simulations of Cage and Tube Formation](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

## Glossary

**Tight-binding method** A method to calculate electronic structure of a condensed matter system (solid or liquid) through modeling the interactions (overlap and hopping matrix elements) between the atomic orbitals on each atom in the system.

**Molecular dynamics** An atomistic simulation method for studying the structure and dynamics of a condensed system (solid or liquid) based on given interatomic potentials and on classical equations of motion (Newton or Langevin).

**Tight-binding molecular dynamics** Molecular dynamics simulations using interatomic potentials described by the tight-binding method.

### Environment-dependent tight-binding potential

A tight-binding potential model in which the interatomic hopping matrix elements between a pair of atoms in the solid or liquid phases depend not only on

the distance between the two atoms, but also on the geometry of neighboring atoms.

**Stone–Wales transformation** A common topological transformation in two-dimensional carbon networks (e.g., carbon nanotube, carbon fullerene). The Stone–Wales transformation proceeds by rotating a bond between two carbon atoms by 90 degree which changes the topology of the network.

### Definition of the Subject

Tight-binding molecular dynamics has emerged as a useful method for studying the structural, dynamical, and electronic properties of materials. In this article, we will review the development of tight-binding potentials for carbon systems and the accuracy of the tight-binding potentials for studying nanostructures such as carbon fullerenes and carbon nanotubes. We will also review atomistic simulations using tight-binding molecular dynamics to study the nucleation and formation of carbon fullerenes and single-walled carbon nanotubes. Several formation pathways and nucleation mechanisms have been discussed including nucleation from the gas phase, formation by zipping graphitic patches, growth via coalescence, junction formation by self-healing of vacancies, and transformation from nanodiamonds.

### Introduction

Atomistic modeling and simulation of materials has attracted continuous interests in physics, chemistry, and materials science. In the past four decades, first-principles simulation methods based on density functional theory and formalism [23,37] (e.g., Car–Parrinello method [5], VASP [39,40,41,42,43]) have been well developed and widely used in computer simulation studies of materials [7,8]. However, due to heavy computational workload, first-principles methods are prohibitive for studying systems with increasing complexity. Most of molecular dynamics (MD) simulations using the first-principles density functional formalism for force calculations are still limited to a small number of atoms (a few hundred atoms) and a short time period (a few hundred ps) on commonly available computers. On the other hand, quite a number of empirical interatomic potentials suitable for large scale simulations have been developed for covalent materials such as silicon and carbon [3,4,9,13,32,77,81,82,83]. These potentials include three-body interactions and are usually fitted to energy-volume phase diagrams and other properties of the crystalline structures obtained by first-principles calculations or from experiments. Although simulations with classical potentials are fast, these empirical

potentials do not always give correct descriptions for structures and properties that are not explicitly included in the fitting database [1,11]. Electronic structure information cannot be obtained, nor can we expect these classical potentials to give accurate descriptions of phenomena where quantum mechanical interference effects are essential (e.g., Jahn–Teller distortions around vacancies, conjugated  $\pi$ -state effects in carbon systems, electronic entropy at high temperatures). There is a large class of problems in the emerging area of nano science and technology that requires more atoms than first-principles techniques can handle and demands more accuracy than classical potentials can provide. Thus it is imperative to have a scheme that is powerful enough to treat several thousand atoms (tens of thousand to million atoms with massively parallel computers) while being accurate enough so that we can trust the results.

In the past two decades, a scheme for molecular dynamics simulation based on a simplified quantum mechanical description of interatomic forces, i.e., tight-binding molecular dynamics (TBMD), has been developed [31,46,62,71,74,86,87]. This method bridges the gap between classical-potential simulations and the first-principles Car–Parrinello method. In the same spirit as the Car–Parrinello scheme [5], TBMD incorporates electronic structure into molecular dynamics simulation through a tight-binding Hamiltonian  $H_{\text{TB}}$ . Using the Hellmann–Feynman theorem, the quantum mechanical many-body nature of the interatomic forces are taken into account naturally by calculating the quantum state of the electrons at each MD time step. Since the TBMD scheme usually uses a minimal basis set for the electronic structure calculations and the Hamiltonian matrix elements are parametrized or tabulated, larger numbers of atoms can be tackled within the present computer capabilities. One of the distinctive features of this scheme in comparison with other empirical schemes is that all the parameters in the model can be obtained theoretically. It is therefore very useful for studying novel nano-materials where experimental data may not be readily available. The scheme has been demonstrated to be a powerful method for studying various structural, dynamical and electronic properties of covalent systems as well as metallic systems.

In this article, we will review the development of tight-binding models for carbon and applications of tight-binding molecular dynamics to the study of the formation of carbon nanostructures such as fullerenes and carbon nanotubes. The discovery of carbon fullerenes [38,44] and carbon nanotubes [24] have opened a new field in the science and technology of nano-scale materials. Carbon fullerenes and nanotubes have been shown to have var-

ious novel properties due to their size and dimensionality. Carbon fullerenes are zero-dimensional and carbon nanotubes are one-dimensional nano scale materials. Among many techniques used in this very dynamic research field, tight-binding molecular dynamics simulations have made notable contributions to the understanding of various aspects of the structures, formation mechanism, as well as the mechanical and electronic properties of these novel nanomaterials. In this article, we will focus ourselves on the simulation studies of fullerene and nanotube formation.

This article is organized as follows: In the next section, we will review the development of two popular tight-binding potentials for carbon, i.e., the transferable two-center tight-binding potential and the environment-dependent tight-binding potential developed by the authors and co-workers. The accuracy of the potentials for simulation studies of carbon-based materials will also be discussed. In Sect. “TBMD Simulations of Cage and Tube Formation”, we will review tight-binding molecular dynamics simulations of various atomic processes of the formations of fullerenes and single-walled carbon nanotubes. Finally, concluding remarks are given in Sect. “Future Directions”.

### Tight-Binding Potentials for Carbon

The expression for binding energy of a system with  $M$  atoms and  $N$  valence electrons in tight-binding molecular dynamics is given by

$$E_{\text{binding}} = E_{\text{bs}} + E_{\text{rep}} \quad (1)$$

The first term on the right hand side of Eq. (1) is the band structure energy which is equal to the sum of the one-electron eigenvalues  $\varepsilon_i$  of the occupied states given by a tight-binding Hamiltonian  $H_{\text{TB}}$ ,

$$E_{\text{bs}} = \sum_i f_i \varepsilon_i \quad (2)$$

where  $f_i$  is the electron occupation (Fermi–Dirac) function and  $\sum_i f_i = N$ . The tight-binding Hamiltonian  $H_{\text{TB}}$  is constructed following the scheme proposed by Slater and Koster [75]. For systems containing  $s$  and  $p$  orbitals, such as carbon, the tight-binding Hamiltonian  $H_{\text{TB}}$  consists of four types of hopping integrals  $h_{ss\sigma}$ ,  $h_{sp\sigma}$ ,  $h_{pp\sigma}$ , and  $h_{pp\pi}$  as well as two on-site atomic energies  $e_s$  and  $e_p$ . The hopping integrals are dependent on the interatomic distances and in general also on the binding environment of the structures. The on-site atomic energies are also dependent on the bonding environment of the atoms. These hopping integrals and on-site energies are treated

as parameters in the tight-binding scheme. In the following subsections, we will discuss how these parameters are determined.

The second term on the right hand site of Eq. (1) is a repulsive energy usually expressed as the sum of short-ranged pairwise interactions

$$E_{\text{rep}} = \sum_{i,j} \phi(r_{i,j}) \quad (3)$$

or a functional of pairwise interactions

$$E_{\text{rep}} = \sum_i F_i \left( \sum_j \phi(r_{i,j}) \right) \quad (4)$$

where  $F$  is a functional which, for example, can be a 4th order polynomial as will be discussed in the following subsections.

### Two-Center Tight-Binding Potential

The most commonly used two-center tight-binding potential for carbon was developed by Xu, Wang, Chan, and Ho (XWCH) in the early 1990s [96] following the transferable scaling function proposed by Goodwin, Skinner, and Pettifor (GSP) for Si [16]. The dependence of the TB hopping parameters and the pairwise potential on the interatomic separation is given by:

$$h_{\alpha}(r) = h_{\alpha}(r_0) \left( \frac{r_0}{r} \right)^n \exp \left[ n \left\{ - \left( \frac{r}{r_c} \right)^{n_c} + \left( \frac{r_0}{r_c} \right)^{n_c} \right\} \right] \quad (5)$$

and

$$\phi(r) = \phi_0 \left( \frac{d_0}{r} \right)^m \exp \left[ m \left\{ - \left( \frac{r}{d_c} \right)^{m_c} + \left( \frac{d_0}{d_c} \right)^{m_c} \right\} \right] \quad (6)$$

where  $r_0$  denotes the nearest-neighbor atomic separations in diamond and  $h_{\alpha}(r_0)$ ,  $n$ ,  $n_c$ ,  $r_c$ ,  $\phi_0$ ,  $m$ ,  $d_0$ ,  $d_c$ , and  $m_c$  are parameters. Unlike the GSP model, the scaling parameters can be different for different hopping integrals and the pairwise repulsion term  $\phi(r)$ . Moreover, for the convenience of molecular-dynamics simulation, the scaling function  $h(r)$  and the pair potential  $\phi(r)$  are also required to go smoothly to zero at some designated cutoff distance. This was achieved by replacing the tail of  $h(r)$  with a third order polynomial  $t_h(r - r_1)$  whose coefficients are determined by requiring that the connection of  $h(r)$  and  $t_h(r)$  at  $r_1$ , (the matching point,  $r_1 \leq r_m$ ), be smooth up to the first derivative, and that  $t_h(r)$  and its first derivative be zero

**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 1**

Parameters for the functions  $h(r)$  and  $\phi(r)$  of the XWCH tight-binding model

$n$	$n_c$	$r_c(\text{\AA})$	$r_0(\text{\AA})$	$r_1(\text{\AA})$	
2.0	6.5	2.18	1.536329	2.45	
$\phi_0(\text{eV})$	$m$	$m_c$	$d_c(\text{\AA})$	$d_0(\text{\AA})$	$d_1(\text{\AA})$
8.18555	3.30304	8.6655	2.1052	1.64	2.57

at  $r_m$ . The same procedure is used to determine  $t_\phi(r - d_1)$ , which replaces the tail of  $\phi(r)$ . The expression for the repulsive energy  $E_{\text{rep}}$  in this model takes the functional form of Eq. (4) with  $F(x)$  being a 4th order polynomial in terms of  $x = \sum_j \phi(r_{ij})$ .

The potential is given a relatively short cutoff distance ( $r_m$  and  $d_m$ ) of 2.6 Å. The parameters in the model are chosen primarily by fitting first-principles density functional calculation results of energy-volume curves of different carbon polytypes: diamond, graphite, linear chain, simple cubic, and face-centered cubic structures, with special emphasis on the diamond, graphite, and linear chain structures. Additional checks were made to ensure that the model gives reasonable results for the electronic band structure, elastic moduli, and phonon frequencies in the diamond and graphite structures, although these properties do not enter explicitly into the fitting procedure.

The resulting  $sp^3$  tight-binding parameters were:  $E_s = -2.99$  eV,  $E_p = 3.71$  eV,  $h_{ss\sigma}(r_0) = -5.0$  eV,  $h_{sp\sigma}(r_0) = 4.7$  eV,  $h_{pp\sigma}(r_0) = 5.5$  eV, and  $h_{pp\pi}(r_0) = -1.55$  eV. The scaling parameters for  $h(r)$  and  $\phi(r)$ , the coefficients for the tail functions  $t_h(r - r_1)$  and  $t_\phi(r - d_1)$ , and the coefficients for the repulsive energy polynomial function  $F(x) = \sum_{n=0}^4 c_n x^n$ , with  $x = \sum_j \phi(r_{ij})$  are given in Table 1 and Table 2 respectively.

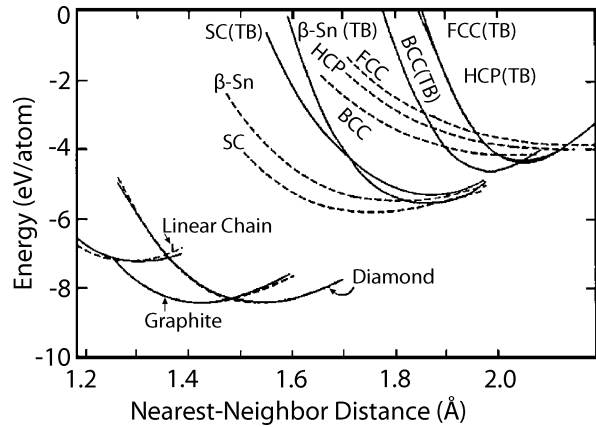
To correctly handle the effects of charge transfer in disordered systems, particularly in the presence of dangling bonds, a Hubbard-like term,

$$H_u = \frac{1}{2} u (q_i - q_i^0)^2 \quad (7)$$

**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 2**

Coefficients of the polynomial functions  $t_h(r - r_1)$ ,  $t_\phi(r - d_1)$ , and  $F(x)$  in the XWCH TB model. See also the text for details

	$t_s(r - r_1)$	$t_\phi(r - d_1)$	$F(x)$
$c_0$	$6.7392620074314 \times 10^{-3}$	$2.2504290109 \times 10^{-8}$	$-2.5909765118191$
$c_1$	$-8.1885359517898 \times 10^{-2}$	$-1.4408640561 \times 10^{-6}$	$0.5721151498619$
$c_2$	$0.1932365259144$	$2.10433033744 \times 10^{-5}$	$-1.7896349903996 \times 10^{-3}$
$c_3$	$0.3542874332380$	$6.60243902262 \times 10^{-5}$	$2.3539221516757 \times 10^{-5}$
$c_4$			$-1.24251169551587 \times 10^{-7}$



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 1**

Binding energy versus nearest-neighbor atomic distance for carbon in different crystalline structures calculated using the XWCH tight-binding potential (solid lines) are compared with the results from first-principles density functional (LDA) calculations (dashed lines). (From Ref. [96])

was added to the tight-binding Hamiltonian  $H_{\text{TB}}$ , where  $q_i$  is the Mulliken population at atomic site  $i$  and  $q_i^0$  is the number of valence electrons of atom  $i$ . The parameter  $u$  is taken to be 4 eV for carbon.

The XWCH carbon potential has been shown to have good transferability when applied to a variety of crystal structures with low coordinations (i. e., carbon chain, graphene, and diamond). This can be seen from Fig. 1 and Tables 3 and 4 where the energies, vibrational and elastic properties for different coordinated crystalline structures obtained from this model are compared with first-principles calculations and experimental data. Description of higher-coordinated metallic structures is only qualitative with this potential. Applications in the molecular-dynamics study of liquid and amorphous phases of carbon [64,89,90,91,92] indicated that the potential is fairly good for describing lower-coordinated carbon systems over a wide range of bonding environments.



### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 3

Elastic constants, phonon frequencies and Grüneisen parameters of diamond calculated from the XWCH-TB model [96] and the environment-dependent TB (EDTB) model [78] are compared with experimental results [59,60]. Elastic constants are in units of  $10^{12}$  dyn/cm<sup>2</sup> and the phonon frequencies are in terahertz

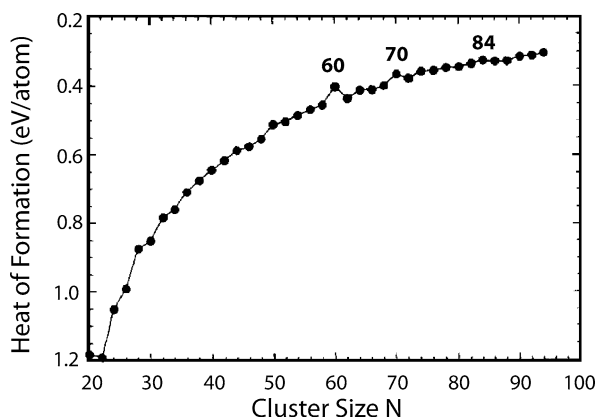
	XWCH	EDTB	Experiment
$a(\text{\AA})$	3.555	3.585	3.567
$B$	4.56	4.19	4.42
$c_{11} - c_{12}$	6.22	9.25	9.51
$c_{44}$	4.75	5.55	5.76
$\nu_{LTO}(\Gamma^-)$	37.80	41.61	39.90
$\nu_{TA(X)}$	22.42	25.73	24.20
$\nu_{TO(X)}$	33.75	32.60	32.0
$\nu_{LA(X)}$	34.75	36.16	35.5
$\gamma_{LTO}(\Gamma^-)$	1.03	0.93	0.96
$\gamma_{TA(X)}$	-0.16	0.30	
$\gamma_{TO(X)}$	1.10	1.50	
$\gamma_{LA(X)}$	0.62	0.98	

### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 4

Elastic constants, phonon frequencies and Grüneisen parameters of graphite calculated from the XWCH-TB model [96] and the environment-dependent TB (EDTB) model [78] are compared with experimental results [14]. Elastic constants are in units of  $10^{12}$  dyn/cm<sup>2</sup> and the phonon frequencies are in terahertz

	XWCH	EDTB	Experiment
$c_{11} - c_{12}$	8.40	8.94	8.80
$E_{2g_2}$	49.92	48.99	47.46
$A_{2u}$	29.19	26.07	26.04
$\gamma(E_{2g_2})$	2.00	1.73	1.63
$\gamma(A_{2u})$	0.10	0.05	

The reliability of this potential for fullerene calculations has been further tested by studying the ground-state geometries and energies of carbon fullerenes. The fully relaxed  $C_{60}$  molecule obtained by the XWCH tight-binding potential [88] has icosahedral symmetry and bond lengths of 1.40 Å and 1.46 Å respectively for the double and single bonds, which agree very well with the first-principles density functional calculation results of 1.40 Å and 1.45 Å. The tight-binding calculation for  $C_{60}$  yields a cohesive energy of 0.41 eV per atom relative to graphite and HOMO-LUMO energy separation of 1.61 eV, agree very well with the LDA results of 0.40 eV per atom and 1.71 eV respectively. The potential also describes well the vibrational modes of  $C_{60}$  and  $C_{70}$  fullerenes [88]. Furthermore, the potential has been employed successfully to determine the ground state geometries of every even-numbered carbon



### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 2

Heat of formation of carbon fullerenes relative to that of graphite as a function of fullerene size obtained from the TBMD calculations. (From Ref. [102])

fullerenes ranging from  $C_{20}$  to  $C_{102}$  [99,100,101,102]. The heats of formation as a function of the cluster size obtained from the calculations are plotted in Fig. 2. While the energies of small fullerenes grow rapidly as the cluster size increases, the energies of large fullerenes which have no adjacent pentagons are found to increase at a slower rate. The most interesting feature as shown in Fig. 2 is that  $C_{60}$ ,  $C_{70}$ , and  $C_{84}$ , which correspond to “magic” numbers in the abundance peaks in the carbon cluster beam mass spectrum, are energetically more favorable than their neighbors. Typical accuracy of the XWCH tight-binding potential in describing the energetics of large fullerenes can also be seen from Table 5 where the relative energies of some  $C_{78}$ ,  $C_{82}$ , and  $C_{84}$  isomers obtained from the tight-binding calculations are compared with the results from ab-initio calculations using the density functional method within the local density approximation (LDA). Although the energy ordering obtained by tight-binding calculations are not in perfect agreement with the results of LDA calculations (see the case of  $C_{82}$ ), tight-binding calculations predict the lower-energy isomers correctly.

The XWCH tight-binding potential has also been demonstrated to have good accuracy for single-walled carbon nanotube calculations. Ozaki et al. have shown that the Young’s moduli of the zigzag (17,0) and armchair (10,10) nanotubes are 988 and 973 GPa, respectively, as calculated by this potential [69]. These values are very close to the first-principles results of 1 TPa [67]. The formation energy and activation energy of the topological pentagon-heptagon-heptagon-pentagon (5-7-7-5) defect in a single-walled carbon nanotube under tensile strain

### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 5

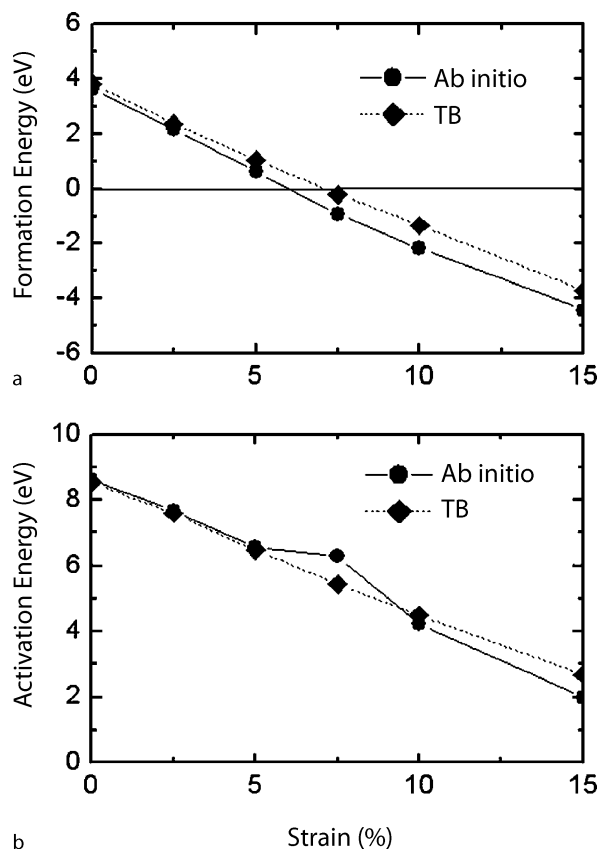
The relative energies of some  $C_{78}$ ,  $C_{82}$ , and  $C_{84}$  fullerene isomers obtained by the XWCH tight-binding potential are compared with the results from the LDA calculations. The energies are in the unit of eV/molecule

Cluster Size	Isomer Symmetry	$\Delta E$ (TB)	$\Delta E$ (LDA)
$C_{78}$			
	$C_{2v}'$	0.000	0.000
	$D_{3h}'$	0.087	0.160
	$C_{2v}$	0.284	0.210
	$D_3$	0.324	0.361
	$D_{3h}$	0.913	1.200
$C_{82}$			
	$C_2$	0.000	0.000
	$C_s$	0.076	0.123
	$C_{2v}'$	0.166	0.312
	$C_{2v}''$	0.191	0.418
	$C_{s'}$	0.226	0.385
	$C_{s''}$	0.229	0.467
	$C_{2v}$	0.285	0.656
	$C_{3v}$	0.566	1.287
	$C_{3v}'$	0.731	1.025
$C_{84}$			
	$D_2$ -22	0.000	0.000
	$D_{2d}$ -23	0.033	0.025
	$C_2$ -11	0.277	0.370
	$D_2$ -1	1.915	2.150

have been studied by Zhao et al. [104] using the tight-binding potential and first-principles calculations. Fig. 3 shows that the results from the tight-binding calculations are in very good agreement with the first-principles results. It should be noted that the potential has very short cutoff distance of 2.6 Å, which makes it inaccurate for describing the interaction between graphite layers. Therefore, the potential is not suitable for studying multi-walled carbon nanotubes.

### Environment-Dependent Tight-Binding Potential

Although the XWCH potential is accurate for the structures and properties of lower-coordinated covalent carbon systems, it fails to give good descriptions for the higher-coordinated carbon structures. We noted that the accuracy and transferability of the XWCH model are limited by the approximations inherent in the Slater and Koster formulation of the tight-binding theory. One severe approximation is the assumption of a fixed minimal basis set independent of the bonding environment of the atom. Experience from first principles calculations showed that a fixed



### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 3

Comparison between ab initio and tight-binding (XWCH) results on the strain-dependent of a formation energies and b activation energies of a (5-7-7-5) topological defect in a (5,5) carbon nanotube. Ab initio results are in solid circles and tight-binding results are in solid diamonds. (From Ref. [104])

minimal basis set optimized for a given atomic geometry usually will not give accurate results for total energies when the atomic geometry of the system is changed. Minimal basis sets need to have the flexibility to deform according to the bonding environment of the atom on which they are based so that they can describe accurately the electronic bonding behavior for different atomic structures. Another major limitation in the Slater and Koster theory is the use of the two-center approximation. Such an approximation is more justified when the system are strongly covalent bonded (e.g., lower-coordinated structures of carbon). The approximation becomes poorer for metallic systems. Contributions beyond pairwise interactions need to be included. Furthermore, the Lowdin procedure used to construct the orthogonal basis set may also result in additional structure-dependent contributions to the two-

center hopping integrals because the overlap matrices are different for different structures. Molecular dynamics simulation of complex carbon structures (e.g., liquid, amorphous, clusters, surfaces) requires that tight-binding potential models must be able to describe the system accurately in various coordination and bonding environments.

A tight-binding potential for carbon that include the environment dependence for off-diagonal as well as diagonal matrix elements have been developed by Tang, Wang, Chan, and Ho [78]. An orthogonal  $sp^3$  basis set is used and the hopping parameters and the pairwise repulsive potential are expressed as

$$h(r_{ij}) = \alpha_1 R_{ij}^{-\alpha_2} \exp[-\alpha_3 R_{ij}^{\alpha_4}] (1 - S_{ij}) \quad (8)$$

In this expression,  $h(r_{ij})$  denotes the possible types of interatomic hopping parameters  $h_{ss\sigma}$ ,  $h_{sp\sigma}$ ,  $h_{pp\sigma}$ ,  $h_{pp\pi}$ , and the pairwise repulsive potential  $\phi(r_{ij})$  between atoms  $i$  and  $j$ .  $r_{ij}$  is the real distance and  $R_{ij}$  is a scaled distance between atoms  $i$  and  $j$ .  $S_{ij}$  is a screening function. The parameters  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$ , and  $\alpha_4$  and parameters for the bond-length scaling function  $R_{ij}$  and the screening function  $S_{ij}$  can be different for different hopping parameters and the pairwise repulsive potential. Note that expression Eq. (8) reduces to the two-center form of the previous subsection if we set  $R_{ij} = r_{ij}$  and  $S_{ij} = 0$ .

The screening function  $S_{ij}$  is expressed as a hyperbolic tangent (tanh) function with argument  $\xi_{ij}$  given by

$$\xi_{ij} = \beta_1 \sum_l \exp[-\beta_2 (\frac{r_{il} + r_{jl}}{r_{ij}}) \beta_3] \quad (9)$$

where  $\beta_1$ ,  $\beta_2$ , and  $\beta_3$  are adjustable parameters. Note that  $\xi_{ij}$  depends not only on the distance between atoms  $i$  and  $j$ , but also on the positions of the neighbors of atoms  $i$  and  $j$ . Maximum screening effect occurs when the atom  $l$  is sitting on the line connecting the atoms  $i$  and  $j$  (i.e.,  $r_{il} + r_{jl}$  is minimum). This approach allows us to distinguish between the first and further neighbor interactions without explicit specification. This is well-suited for molecular dynamics simulations when one encounters configurations where it may be difficult to define exactly which atoms are first neighbors and which atoms are second neighbors.

The bond-length scaling function scales the distance between two atoms according to their effective coordination numbers. Longer effective bond lengths are assumed for higher coordinated atom pairs. The strength of the hopping parameters between atoms  $i$  and  $j$  are therefore dependent on the coordination number of the atoms: weaker interaction strength for higher-coordinated structures. The scaling between the real and effective inter-

atomic distance is given by

$$R_{ij} = r_{ij}(1 + \delta\Delta) \quad (10)$$

where  $\Delta = \frac{1}{2}[(\frac{n_i - n_0}{n_0}) + (\frac{n_j - n_0}{n_0})]$  is the fractional coordination number relative to the coordination number of the diamond structure  $n_0$ , averaging between the coordination numbers  $n_i$  and  $n_j$  of atoms  $i$  and  $j$ .

The coordination number is modeled by a smooth function,  $n_i = \sum_j (1 - S_{ij})$  with  $S_{ij}$  has the form of the screening function described above. By choosing the parameters  $\beta_1$ ,  $\beta_2$  and  $\beta_3$  to be 2.0, 0.0478, and 7.16 respectively, the coordination number calculated from this smooth function is 2.08639, 3.17678, 4.41022, 6.23620, 10.38529, and 11.89829 for the linear-chain, graphite, diamond, simple cubic, bcc, and fcc structures, respectively. These values give a reasonable representation of the effective coordinations in these structures.

Besides the hopping parameters, the diagonal matrix elements are also dependent on the bonding environments. The expression for the diagonal matrix elements is

$$e_{\lambda,i} = e_{\lambda,0} + \sum_j \Delta e_{\lambda}(r_{ij}) \quad (11)$$

where  $\Delta e_{\lambda}(r_{ij})$  takes the same expression as Eq. (8),  $\lambda$  denotes the two types of orbitals ( $s$  or  $p$ ).  $e_{s,0}$  and  $e_{p,0}$  are the on-site energies of a free atom which were chosen to be  $-6.041$  and  $1.024$  eV, respectively.

Finally, the repulsive energy term  $E_{rep}$  is expressed in a form similar to the XWCH model [96] discussed above.

The parameters in the model are determined by first fitting to the self-consistent first-principles density functional calculations results of electronic band structures and then the cohesive energy versus volume curves of diamond,  $\beta$ -tin, simple cubic, bcc, and fcc structures respectively. These parameters are listed in Tables 6 and 7. The cutoff distance for the interatomic interaction in this potential is  $3.3 \text{ \AA}$ .

As shown in Fig. 4, the environment-dependent tight-binding (EDTB) potential model describes very well not only the covalent structures, but also the higher-coordinated metallic structures. The model also describes the electronic band structures of various carbon crystals with a reasonable accuracy as one can see from Fig. 5. The elastic constants and phonon frequencies in the diamond and graphite structures from the environment-dependent tight-binding model are also improved over the XWCH two-center model as one can see from Tables 3 and 4.

The EDTB carbon potential has been applied successfully to the studies of structures and properties of amorphous carbon [18,19,72], surfaces [18,93], nanodia-

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 6

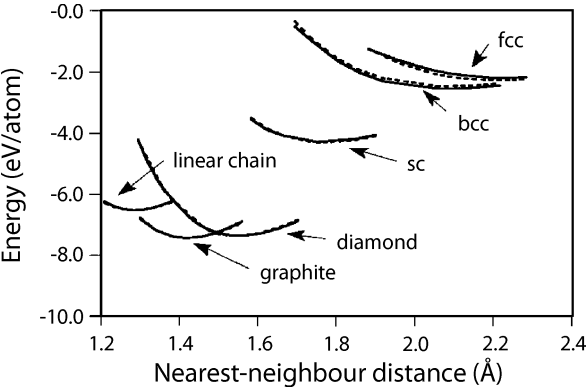
The parameters of the EDTB model for carbon. The TB hopping integrals are in the unit of eV and the interatomic distances are in the unit of Å.  $\phi$  is dimensionless

	$\alpha_1$	$\alpha_2$	$\alpha_3$	$\alpha_4$	$\beta_1$	$\beta_2$	$\beta_3$	$\delta$
$V_{ss\sigma}$	-8.9491	0.8910	0.1580	2.7008	2.0200	0.2274	4.7940	0.0310
$V_{sp\sigma}$	8.3183	0.6170	0.1654	2.4692	1.3000	0.2274	4.7940	0.0310
$V_{pp\sigma}$	11.7955	0.7620	0.1624	2.3509	1.0400	0.2274	4.7940	0.0310
$V_{pp\pi}$	-5.4860	1.2785	0.1383	3.4490	0.2000	8.5000	4.3800	0.0310
$\phi$	30.0000	3.4905	0.00423	6.1270	1.5035	0.205325	4.1625	0.002168
$\Delta e_s, \Delta e_p$	0.1995275	0.029681	0.19667	2.2423	0.055034	0.10143	3.09355	0.272375

Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Table 7

The coefficients (in unit of eV) of the polynomial function  $F(x)$

$c_0$	$c_1$	$c_2$	$c_3$	$c_4$
12.201499972	0.583770664	$0.336418901 \times 10^{-3}$	$-0.5334093735 \times 10^{-4}$	$0.7650717197 \times 10^{-6}$



Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 4

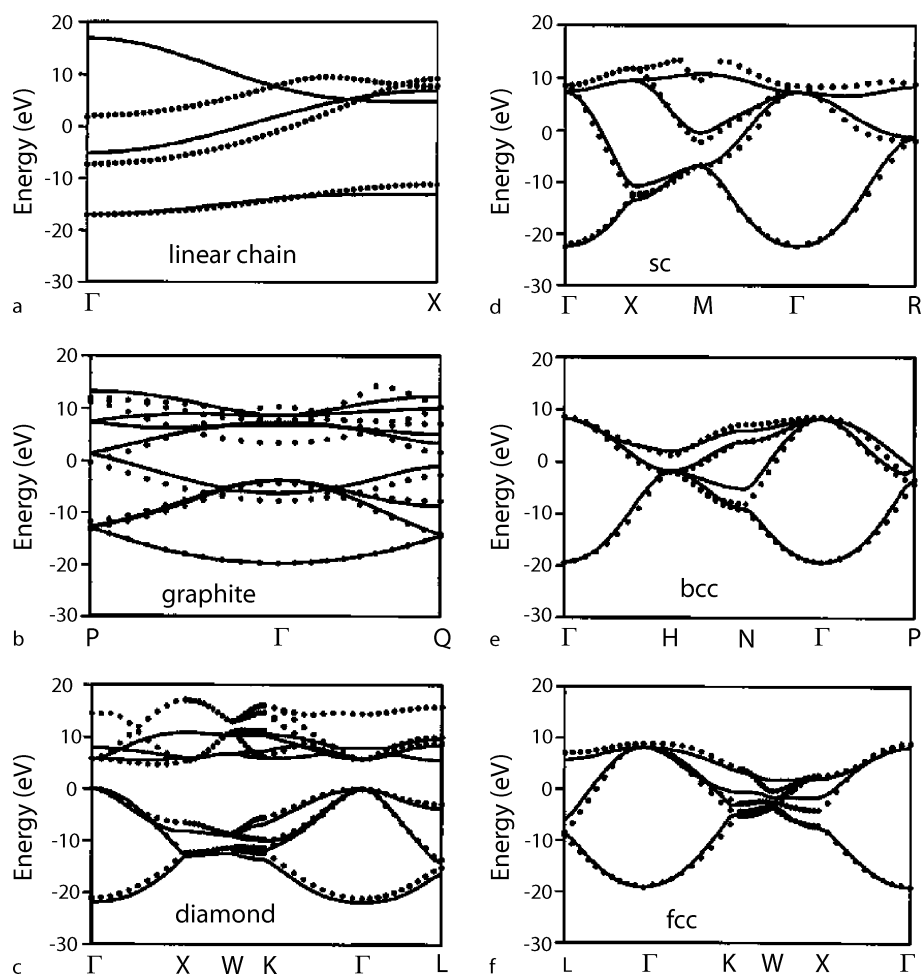
Binding energy versus nearest-neighbor atomic distance for carbon in different crystalline structures calculated using the environment-dependent tight-binding potential (solid lines) are compared with the results from first-principles density functional (GGA) calculations (dashed lines). (From Ref. [78])

mond [52,72], as well as nanotubes [34,35,67]. In particular, Wang and Ho have applied the potential to simulate the atomic process of the structural transformation at diamond surfaces under intense ultrafast laser irradiation [93]. Interesting insights have been gained from the simulation study which shows that the transformation pathways from the diamond surface into graphite sheets are different depending on whether the driving force of the transformation is “thermal” or “non-thermal”. Such different behaviors are illustrated in Figs. 6 and 7. Very recently, Lee et al. [52] also used the EDTB to study the atomic process of the transformation of nanodiamond to tube-like fullerene cages and formation of carbon nano-

tube junctions. This simulation will be discussed in more detail later in this review.

### TBMD Simulations of Cage and Tube Formation

Although carbon fullerenes and carbon nanotubes of various sizes, lengths, topologies and chiralities have been observed in experiments, the nucleation mechanism and formation process of such remarkable nano-scale materials are still not well understood. On the other hand, it has been well demonstrated that the mechanical and electronic properties of these nanomaterials are sensitive to the size and chirality of the structures. For example, it is well known that single-walled carbon nanotube can either be metallic or semiconducting depending on the size and helicity of the tubes [20,63,66,73,95]. It has also been shown that the band-gap of the nanotubes can be modified by elastic deformation of the tubes [33,54,67,70,97] and by the introduction of topological defects that make metal-semiconductor or semiconductor-semiconductor junctions [10,12]. Recently, “X”, “Y”, and “T”-shaped molecular junctions between the single-walled carbon nanotubes (SWCNT) have been created experimentally by electron beam irradiation of crossed SWCNTs [79]. These junctions of SWCNT could act as nano scale multiterminal electronic devices. Understanding the formation process of the carbon nanostructures at the atomic-scale is therefore highly desirable and is important for the structural control and mass production of these nanomaterials. In this section, we will review tight-binding molecular dynamics simulation studies that investigate several pathways to control catalyst-free growth of these nanoscale materials.



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 5**

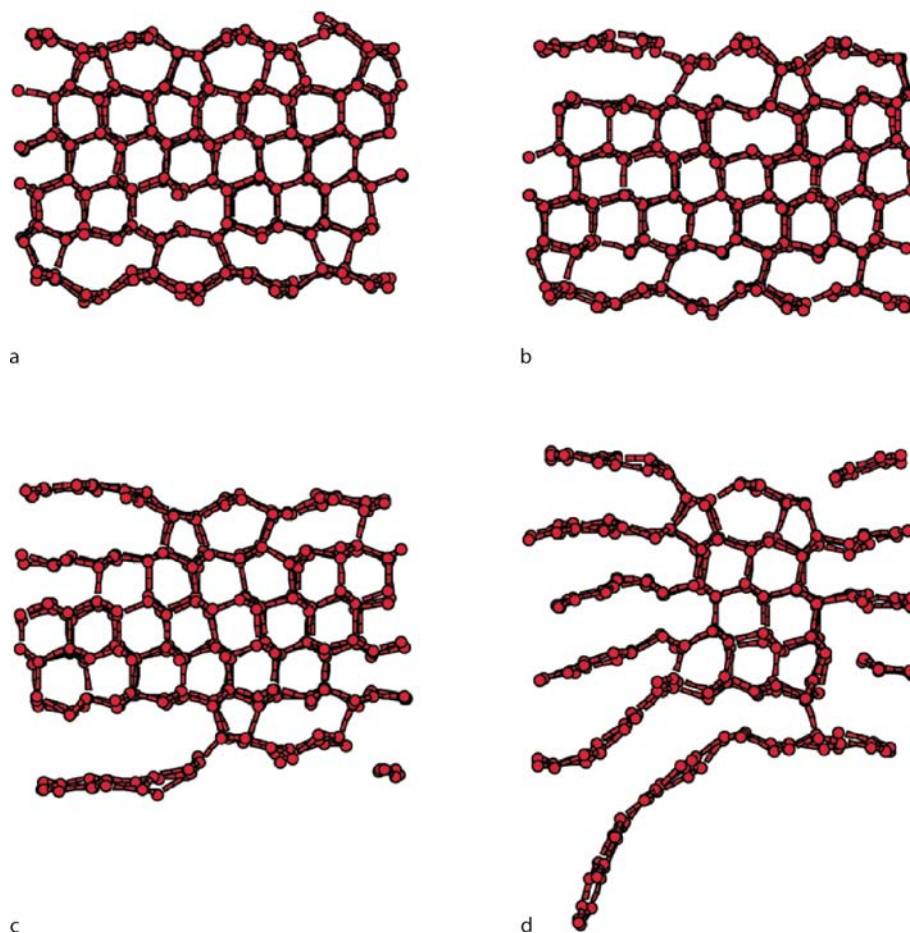
The electronic energy bands of various crystalline structures of carbon calculated using the environment-dependent tight-binding potential (solid curves) are compared with the first-principles LDA calculation results (dots). The Fermi levels are located at  $E = 0$  eV. (From Ref. [78])

### Formation by Gas Phase Condensation

In order to understand the nucleation mechanism and growth condition of  $C_{60}$  buckyball, Wang et al. have performed molecular dynamics simulations using the XWCH carbon tight-binding potential to study the formation of  $C_{60}$  from gaseous carbon atoms [94]. They enclosed 60 carbon atoms in a hollow sphere of radius  $R$  with specular reflection imposed when the atoms hit the inner surface of the sphere. They start the simulation by heating the carbon atoms to very high temperatures (10000 K) in a larger sphere ( $R = 9.22\text{\AA}$ ). The carbon atoms under this condition are found to be gas like. Then they gradually reduce the temperature and the radius of the sphere. When the temperature is reduced to 6000 K within

a sphere of radius  $5.3\text{\AA}$ , they found that polygonal rings nucleate rather rapidly from the originally loose linear-chained cluster (Fig. 8a–c). By reducing the sphere radius gradually from  $5.3$  to  $3.832\text{\AA}$  while keeping the temperature at 6000 K, they found that a close cage structure is forming in the process of simulation as one can see from Fig. 8d–f. The structure of the  $C_{60}$  cages obtained from this simulation are similar to that of buckyball but with many defects due to the rapidity of the compression and cooling in the simulation. These defects include seven- and four-membered rings and adjacent pentagons. In a later study, Laszlo has performed TBMD simulation to investigate the formation of  $C_{60}$  in helium atmosphere using the same XWCH carbon tight-binding potential and obtained similar results [48]. Action-derived molecular dy-





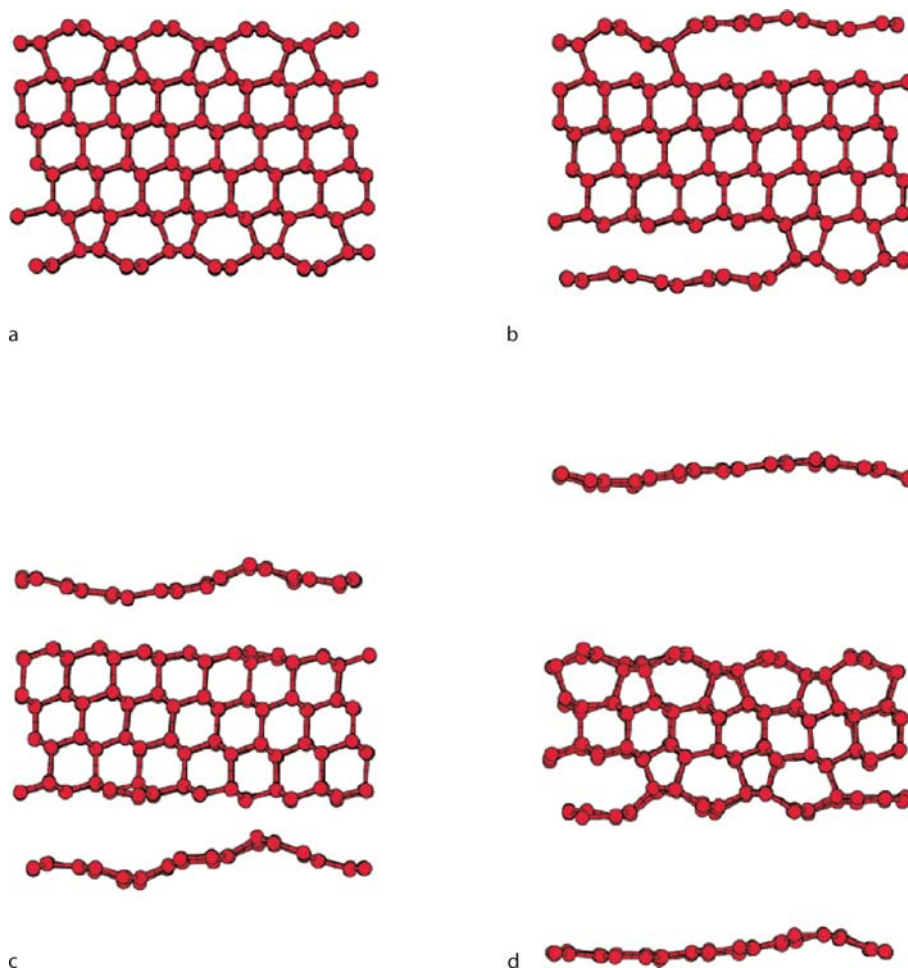
**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 6**

Graphitization of the diamond (111) surface via the thermal process. The snapshot pictures are taken from the tight-binding molecular dynamics simulation in which the electrons and the ions are thermal equilibrated at 2700 K. The plots show the side view of the simulation unit cell which is a 12-layer slab with two (111) surfaces (the top and bottom layers). Periodic boundary conditions are imposed in the plans perpendicular to the surface. Graphitization is found to occur through the formation of graphite-diamond interfaces (see d). The whole process takes about 3 ps. (From Ref. [93])

namics simulation using the XWCH carbon tight-binding potential suggested that existence of chains in the precursor models of tangled polycyclics and open cages is beneficially for the formation of  $C_{60}$  molecule [28]. The formation mechanism of  $C_{60}$  by  $C_2$  assembly has also been investigated extensively by Irle et al. using density functional tight binding (DFTB) and QM/MD simulations [76,78,105].

Tight-binding molecular dynamics simulations were also performed by Oh and Lee [68] to understand how the carbon atoms from the gas phase are adsorbed on the carbon nanotube edge to either grow further or close up the tube by forming a dome-like cap. They start the simulations with a carbon nanotube that is capped at the bot-

tom but open at the top. A number of carbon atoms in a gas phase are initially placed above the open end of the tube. The gaseous carbon atoms are confined to stay within a sphere of diameter 14 Å. The center of the confining sphere is located at the axis of the tube and is about 7 Å above the open tube edge. The carbon atoms are randomly placed in the sphere and the simulations were performed at a temperature of 2000 K. The simulations showed that once a pentagon is formed at the open edge of the tube by the deposited carbon atoms, the top of the tube closes rapidly. Some of the simulation results can be seen from Fig. 9. Irle et al. have also investigated the formation of carbon cages from the tube precursor using DFTB molecular dynamics simulations [25,106].



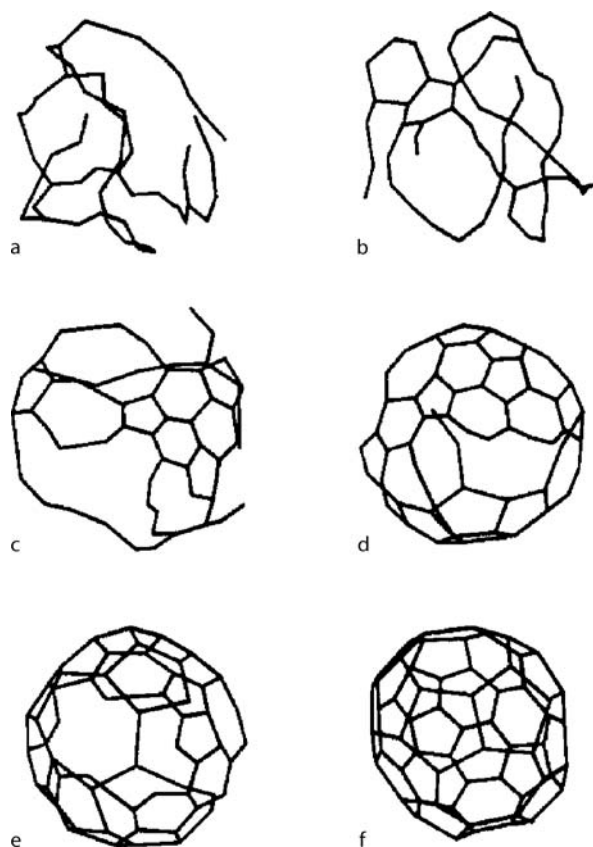
**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 7**

Graphitization of the diamond (111) surface due to the effects of hot electron plasma (no-thermal process). The snapshot pictures are taken from the present tight-binding molecular dynamics simulation in which the electronic temperature is raised to 15,000 K and the ions are evolved freely. The orientation of the simulation unit cell is the same as specified in Fig. 6. Note that the graphitization takes place in a layer-by-layer fashion. The slab is graphitized completely within 500 fs of simulation time. (From Ref. [93])

### Nucleation from Graphitic Patches

Carbon nanotubes can be viewed as graphitic sheets with hexagonal lattice that have been wrapped up into a seamless cylinder. Although the energy barrier for rolling a big piece of graphitic sheet into a tube is very high, two small graphitic patches that zip along their edges and open up into a tubular segment could be energetically favorable and kinetically possible. Such tube-like rings can serve as a nucleus for further growth of nanotubes. Using the XWCH potential, Zhang and Crespi [103] have performed tight-binding molecular dynamics simulations on double-layered graphitic patches as illustrated in Fig. 10 and showed that a nucleus of nanotube can be formed from the sim-

ple planar patches once their edges are connected to each other by additional bridging carbon atoms. Their simulations suggested that there is no energy barrier for nucleation of a nanotube of diameter less than 23 Å (see Fig. 11). For a tube of diameter of 34 Å, the energy barrier for the tube formation would be large. Therefore, nucleation of a large tube is energetically disfavored via the graphitic patches connection mechanism. Kawai et al. [29] have studied the graphitic patch mechanism in great detail by performing extensive tight-binding molecular dynamics simulations with different initial stacking of graphitic patches and with different temperatures and velocities for collision between the graphitic patches. From hundreds of independent simulations, they found that the nuclei of



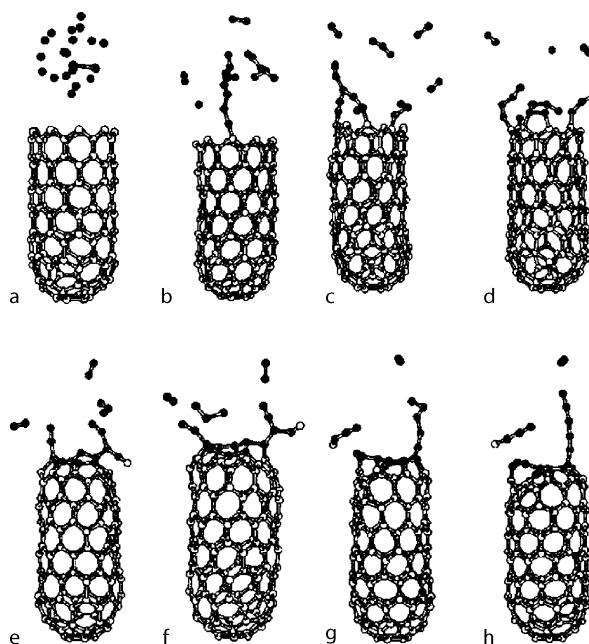
**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 8**

Perspective view of  $C_{60}$  cage formation process at  $T = 6000$  K. The carbon atoms are connected by *straight lines* when the interatomic distances are less than  $1.8 \text{ \AA}$ . a, b, and c are snapshots at 2.8 ps, 4.2 ps, and 5.6 ps respectively and with the sphere radius  $R = 5.32 \text{ \AA}$ . d, e, and f are typical snapshot when the sphere radius is reduced to  $4.61 \text{ \AA}$ ,  $3.90 \text{ \AA}$ ,  $3.832 \text{ \AA}$  respectively. Note a closed cage forms when  $R = 3.832 \text{ \AA}$ . The total simulation time is around 30 ps. (From Ref. [94])

nanotubes, nanohorns, and nanocages can be generated depending on the initial stacking manners and interlayer distances, temperatures, as well as the initial velocities for the collision. Some of their simulation results are shown in Fig. 12.

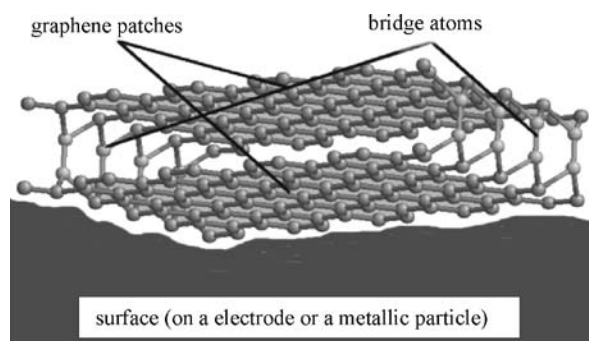
### Growth via Coalescence

Another possible pathway for forming large fullerene cages and nanotubes is due to coalescence of fullerene cages. In 1992, Zhang et al. [98] performed tight-binding molecular dynamics to study the collision of two  $C_{60}$  buckyballs. They showed that with a proper choice of initial velocities, two  $C_{60}$  buckyball can be fused to-



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 9**

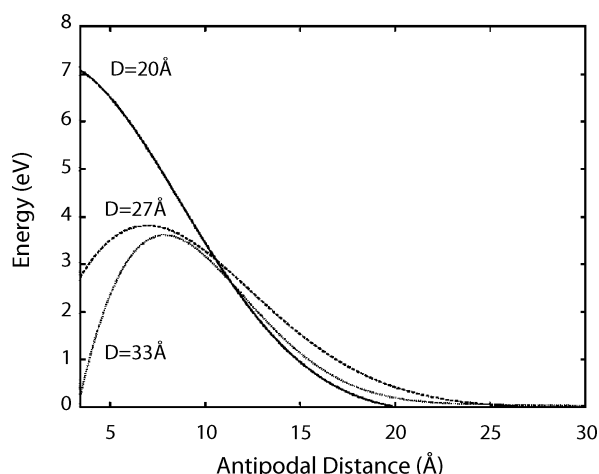
Typical snapshot (side view) of gaseous carbon adsorption on the open edge of a zigzag nanotube during the TBMD simulations at  $T = 2000$  K. a initial configuration  $t = 0$ , b  $t = 0.84$  ps, c  $t = 1.96$  ps, d  $t = 2.52$  ps, e  $t = 3.64$  ps, f  $t = 4.2$  ps, g optimized structure at  $T = 0$  K from the simulation geometry at 5.6 ps. The bonds are drawn when the distance is within  $1.9 \text{ \AA}$ . (From Ref. [68])



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 10**

A schematic diagram of the nucleation model in which a nanotube forms via edge-mediated opening of a double-layered graphitic patch. (From Ref. [103])

gether to form a elongated  $C_{120}$  fullerene cage. Coalescence of  $C_{60}$  buckyballs to form nanotubes have been recently observed in experiments. Luzzi and Smith [58] showed that electron-beam irradiation on an array of



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 11**

The energy as a function of a reaction coordinate (i.e., the distance between opposite inner surface as the tube nucleus opens) for nucleation graphitic patches of varying diameter when opened and of length 1 nm. The smallest patches ( $D = 20 \text{ Å}$ ) have no barrier to opening. (From Ref. [103])

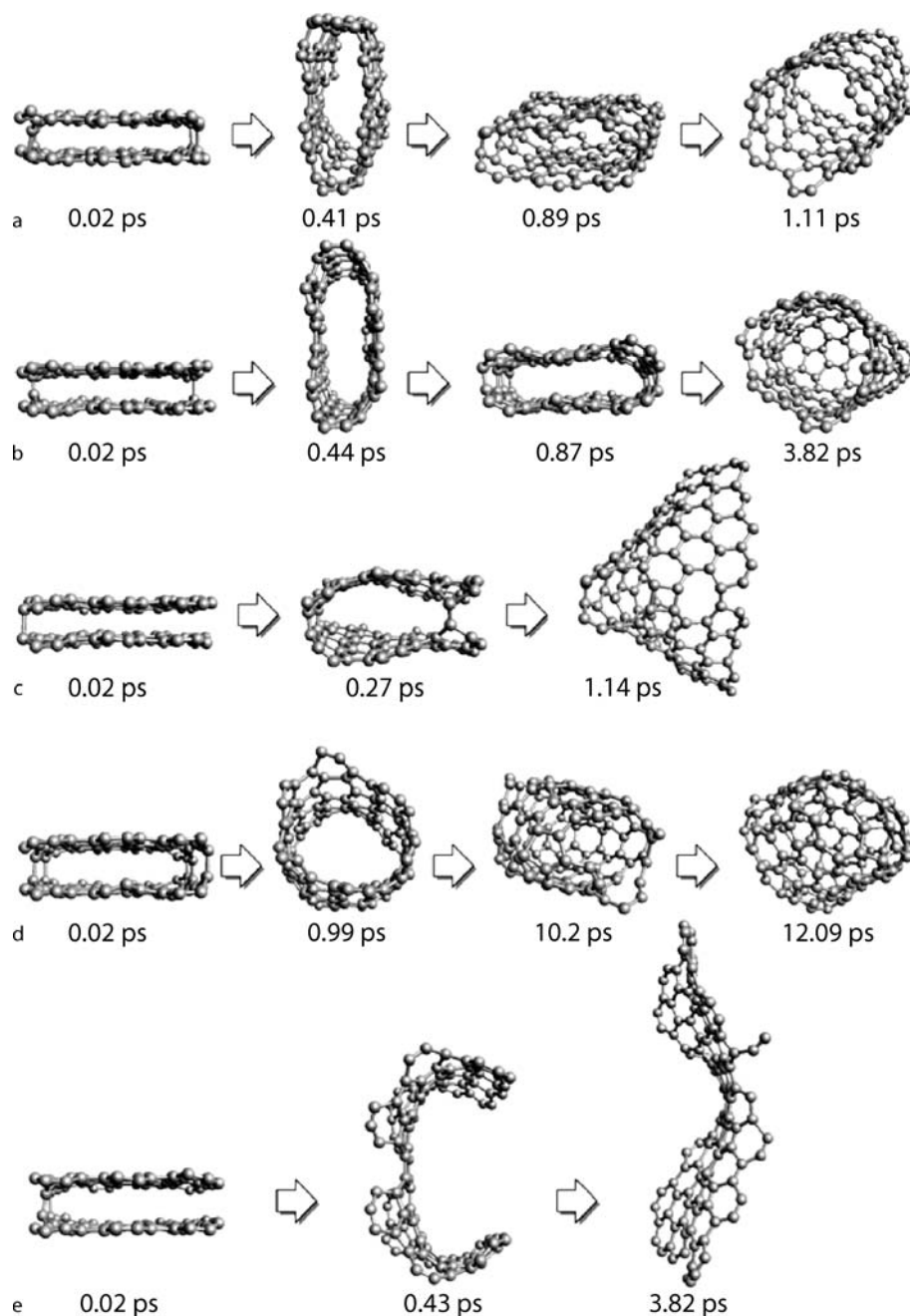
fullerene molecules inside a single-walled carbon nanotube [76] called nanopeapods causes the coalescence of the fullerene to make a carbon nanotube inside the bigger tube. Iijima and co-workers baked nanopeapods up to 1473 K for 14 hours. They also found that the inside fullerenes start to coalesce at 1073 K and complete transformation to a single-wall nanotube at 1473 K [2]. Kim et al. have recently performed tight-binding molecular dynamics to study the dynamics of fullerene coalescence. They showed that the dynamical process of forming a  $C_{120}$  nanocapsule with various chiral indexes of (5,5), (10,0), (6,6), and (12,0) through coalescence of two  $C_{60}$  buckyball as shown in Fig. 13 consists of successive Stone–Wales-type bond rotations with a transition energy barrier of about 8–12 eV [36]. Hernandez et al. have also performed TBMD simulation using the XWCH potential to study fullerene coalescence inside a (10, 10) nanotube (diameter ca. 1.38 nm) at 1500 K [22]. Their study have demonstrated that fullerene molecules can easily coalesce inside SWNTs via a process driven by either thermal annealing or electron irradiation. The resulting structure consists of a corrugated tubule nested inside the original SWNT.

The coalescence of two smaller nanotubes to make a bigger nanotube has also been observed to occur by Nikolaev et al. and Terrones et al. when irradiated by electrons or ions at elevated temperatures [65,80]. The coalescence was considered to be induced by vacancies or

other defects. Terrones et al. have performed tight-binding molecular dynamics simulations to study the coalescence and showed that two (10,10) nanotubes coalesce to make a (20,20) nanotube when defects are introduced [80]. The simulation was performed at 1273 K in order to accelerate the creation of interlinks and surface reconstruction between the two smaller tubes. The total simulation time is 150 ps and the coalescence of the two smaller tubes was observed to occur after 100 ps via the “zipping” mechanism. They also suggested that coalescence would not occur if the two nanotubes had the same chirality, because of the required global rearrangement of the bonding network. Recently, Kawai et al. [30] performed tight-binding molecular dynamics to investigate the coalescence of small carbon nanotubes, that is, combinations of the armchair (3,3), zigzag (5,0), and chiral (4,2) nanotubes. They found that two small nanotubes having the same or different chirality can coalesce without initially introducing atomic defects to enhance the reaction. They also found that the chiral index of the coalesced nanotube can be expressed as a vector sum of the indexes of the original nanotubes. These simulation results also suggested that the chirality of nanotube can be changed through chemical reaction. The simulations were performed with initial temperature ranging from 500–2500 K but a thermostat was not used during the simulation. Corresponding kinetic energy of translational velocity (center-of-mass velocity) ranged from 0.09 to 0.52 eV/atom is introduced to facilitate the collision and coalescence of the small tubes. Some of their simulations on the coalescence process are shown in Fig. 14.

Local coalescence between two nanotubes is believed to play an important role in making nanotube junctions. It has been shown by experiment that “X”, “Y”, and “T” nanotube junctions can be created by controlled electron beam exposure of crossed tubes at elevated temperatures [79]. Some of these junctions are shown in Fig. 15. Terrones et al. have performed tight-binding molecular dynamics simulation to study the formation mechanism of the junctions [79]. They showed that vacancies and interstitials created under electron beam exposure can trigger the local coalescence of the tubes by forming the various junctions between the two nanotubes. The simulation was performed at 1273 K with two (8,8) nanotubes crossing each other perpendicularly. In order to include the irradiation effects, 20 atoms were removed randomly from the nanotubes to create vacancies in the crossing region. The coalescence process during the simulation is shown in Fig. 16. Similar simulations were also performed by Menon et al. using a different tight binding carbon potential [61].





**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 12**

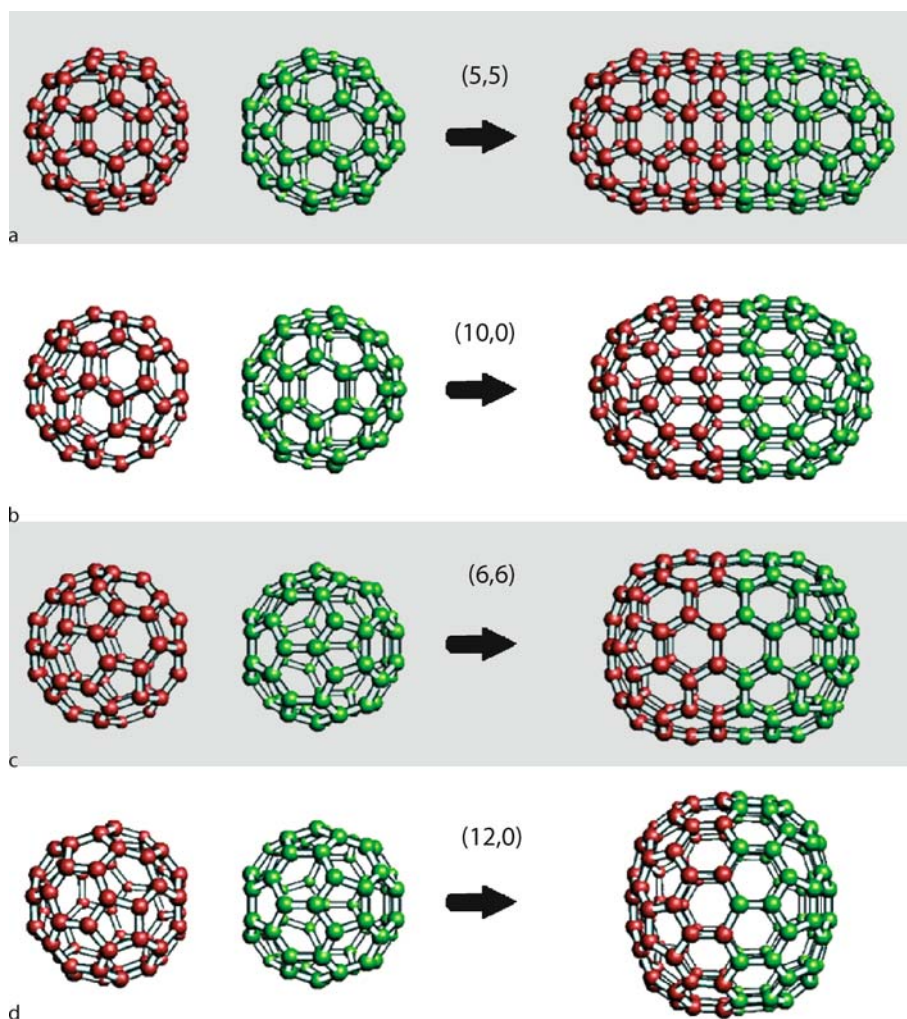
The time evolution of two reacting graphitic patches starting with A-A staking but with different interlayer distances  $D_0$  and different temperatures: a  $D_0=2.1$  Å,  $T_0=2000$ K, b  $D_0=2.1$  Å,  $T_0=1000$ K, c  $D_0=2.3$  Å,  $T_0=2000$ K, d  $D_0=1.9$  Å,  $T_0=1000$ K, e  $D_0=1.9$  Å,  $T_0=1000$ K. Note that d and e have the same starting condition but different end product. (From Ref. [29])

### Junction Formation by Self-Healing of Vacancies

In the previous subsection, we noted that the formation of carbon nanotube junctions usually involves vacancy de-

fects induced by irradiations. Very recently, the detailed atomistic processes of vacancy reconstruction which lead to the formation of nanotube junctions have been investigated by tight-binding molecular dynamics simulations





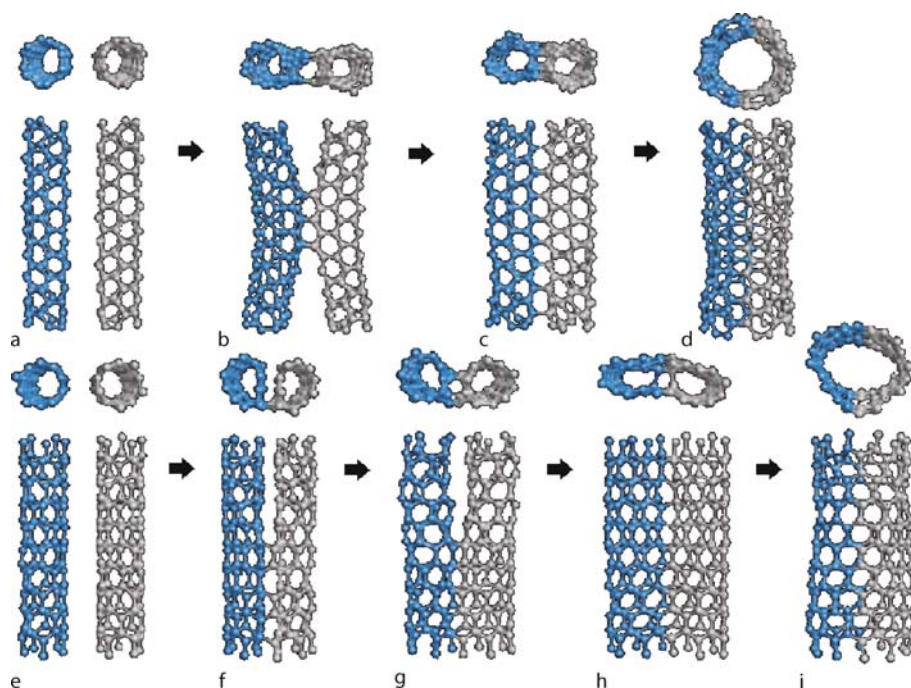
**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 13**

Initial and final states for forming various  $C_{120}$  nanocapsules as precursors for fullerene-based nanotube growth. (From Ref. [36])

based on the environment-dependent tight-binding carbon potential developed by Tang et al. [49,50,51,78]. The original EDTB carbon potential [78] is modified by incorporating an angle dependence factor into the repulsive energy to describe correctly the diffusion of an adatom and a vacancy in carbon nanotubes and graphene [49,50,51]. It has been shown that a carbon nanotube semiconductor-metal intramolecular junction can be formed by self-assembly of vacancy defects after the tube has been subjected to electron or ion irradiation [51].

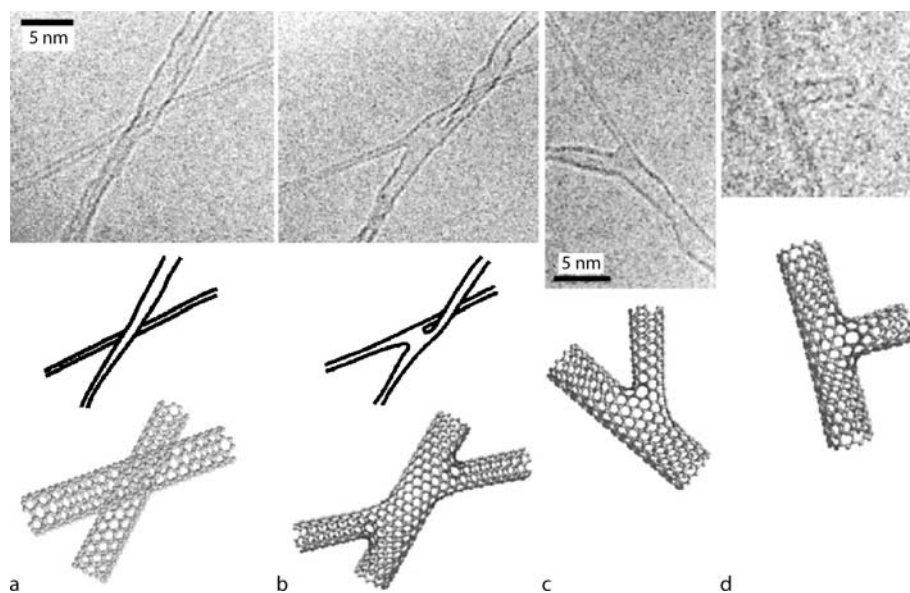
Figure 17 shows the atomic details of vacancy reconstruction in a (16,0) SWCNT with a six-vacancy hole by the TBMD simulation [51]. The TBMD simulation is performed starting from a relaxed six-vacancy hole geometry as shown in Fig. 17a. In the early stage of the simulation,

the SWCNT is heated up to high temperature through a constant-temperature molecular dynamics simulation. It was found that rearrangement of carbon atoms around the vacancy hole starts to occur near 4500K at the simulation time of 18 ps through the rotation of carbon dimers, i. e., Stone–Wales transformation. After 19 ps of the simulation time, three hexagons at the lower left corner of the vacancy hole (those containing the atoms 1–4 in Fig. 17b) are recombined into a pentagon-octagon-pentagon defect by successive Stone–Wales transformations of the dimers 1–2 and 3–4 as shown in Fig. 17b. After the simulation time of 20 ps, another two hexagons (containing the atoms 5–7) on the other side of the vacancy hole are also reconstructed into one pentagon and one heptagon by the Stone–Wales transformation of the dimer 5–6. In order to prevent the



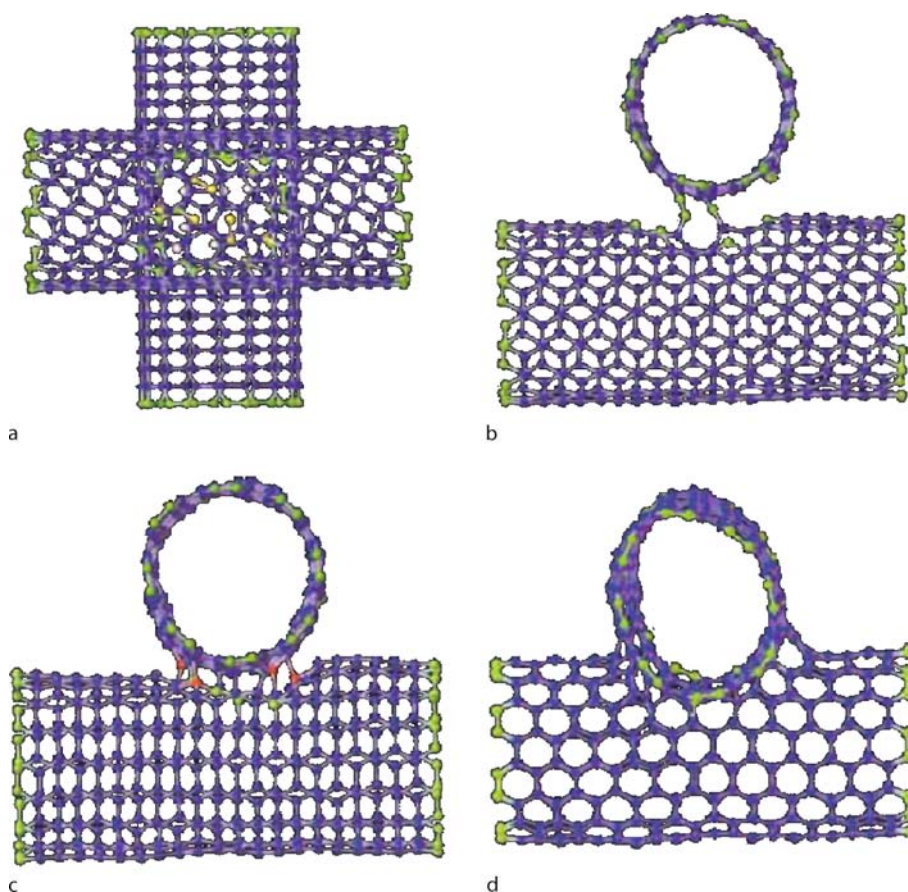
Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 14

Coalescence process of the same two achiral nanotubes by colliding with each other (top and side views). The figures a–d are the snapshots of the reaction for (3,3) + (3,3), and the figures e–i are those of reaction process for (5,0) + (5,0). The resulting structures were identified as armchair (6,6) and zigzag (10,0) nanotube respectively. No defects are introduced before the collision. (From Ref. [30])



Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 15

Various carbon nanotube junctions created by experiment. (From Ref. [79])



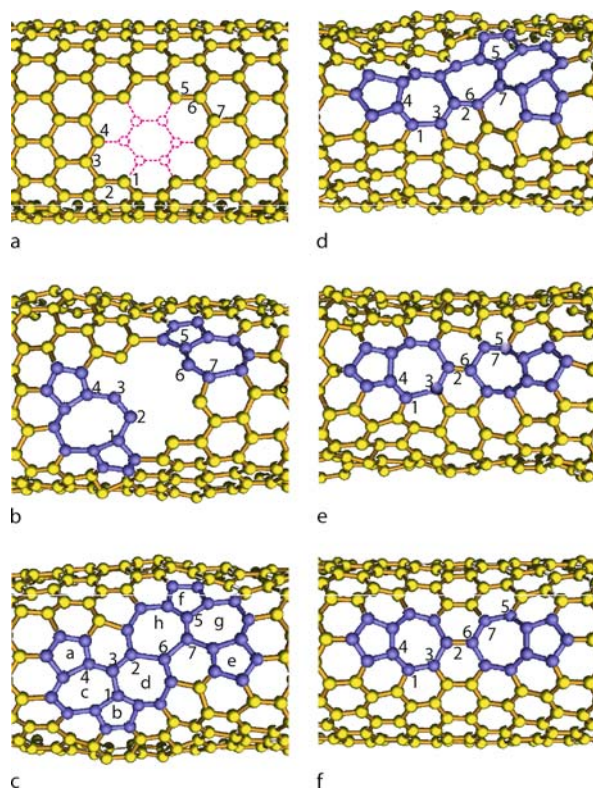
**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 16**

Sequences of merging between two crossing (8,8) carbon nanotubes into a unique X-like junction. **a** TBMD simulation starts with the random creation of 20 vacancies in the lattice of the two tubes in the localized neighboring region (top view). **b** After 10 ps, two links between the two defective tubes are formed via carbon chain (side view). **c** After 100 ps, the connection between the two tubes is established, although some  $sp^3$ -carbon atoms (red) and dangling bonds (green) are still remaining. **d** After 220 ps, surface reconstruction occurs and the carbon system approach an X junction. The reconstructed surface contains six heptagons, one octagon, one pentagon, and two dangling bonds. (From Ref. [79])

evaporation of carbon atoms, the system is then cooled down to 3,000 K for 4 ps and the vacancy hole is healed during this simulation period as shown in Fig. 17c. The structure immediately after the healing process consists of four pentagons and four heptagons with a two-fold rotation symmetry. The pentagon a and b and the heptagon c and d are related to the pentagon e and f and the heptagon g and h, respectively through the 2-fold axis which goes through the center of the carbon bond between atom 2 and 6. After the simulation time of 24 ps, the system is heated up again to 4,500 K for 7 ps and another structural reconstruction among the defects is observed. As shown in Fig. 17c and Fig. 17d, as the result of a Stone–Wales transformation of the dimer 1–3, the two heptagons (c and d in Fig. 17c) and one pentagon (b in Fig. 17c) on the left

side of the 2-fold symmetry axis are transformed into three hexagons while one hexagonal ring containing the carbon atom 3 is transformed into a heptagonal ring. Finally a pentagon-heptagon pair defect, which has been observed in the experiment after the irradiation [21], is emerged through the reconstruction process. Since the dimer 5–7 is equivalent to the dimer 1–3 due to the 2-fold symmetry at the stage of Fig. 17c, the dimer 5–7 is expected to undergo a similar Stone–Wales transformation. Indeed, after 41 ps of simulation time, the Stone–Wales transformation happens to the carbon dimer 5–7. Consequently another pentagon-heptagon pair defect is formed at the right side of the 2-fold axis in the same way as the formation of the previous pentagon-heptagon pair on the left side of the 2-fold axis. The structure with two pentagon-heptagon pairs



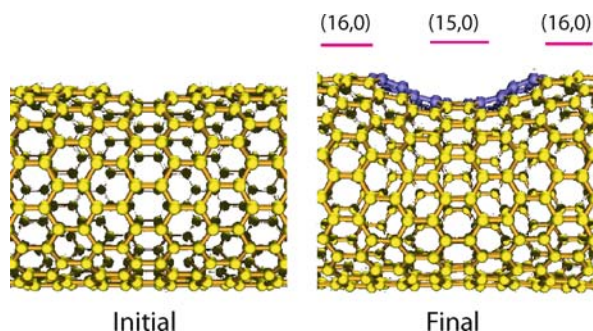


**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 17**

Atomic processes in the TBMD simulation for the (16, 0) SWCNT with six vacancies. **a** 0 K (at time  $t = 0$  ps), **b**  $\sim 4,500$  K ( $t \approx 20.2$  ps), **c**  $\sim 3,100$  K ( $t \approx 23.2$  ps), **d**  $\sim 4,400$  K ( $t \approx 32.3$  ps), **e**  $\sim 4,700$  K ( $t \approx 41.5$  ps), **f**  $\sim 90$  K ( $t \approx 53.9$  ps). The carbon atoms on the rear side of the tube are concealed in figures in order to see the reconstruction of vacancies more clearly. *Dotted circles* in **a** indicate the positions of the six carbon vacancies in the perfect (16, 0) SWCNT. *Yellow colors* indicate carbon atoms and bonds in hexagonal rings. *Blue colors* indicate carbon atoms and bonds in non-hexagonal rings. See the text for small letters in **c** and numbers. (From Ref. [51])

in Fig. 17e is very stable energetically and can sustain its shape without any changes for more than 20 ps in the simulation even at a temperature 4,500 K. At the final stage of the simulation, the system is gradually cooled down to 0 K in 12.5 ps and the structure with two pentagon-heptagon pair defects is found to maintain without any additional reconstruction as shown in Fig. 17f.

Figure 18 shows the front view of the initial and final structure from the TBMD simulation. The vacancy hole in the initial structure is healed up in the final structure and the radius of the tube in the middle section is reduced. The diameter and chirality in the center part of the final structure is found to be (15, 0), which is one of the metallic



**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 18**

Front views of initial and final structure from TBMD simulation for (16, 0) SWCNT with six vacancies. The initial structure corresponds to Fig. 17a. The final structure corresponds to Fig. 17f. (From Ref. [51])

SWCNTs. In order to understand the effects of the vacancy cluster size on the formation of junctions, they have also performed the TBMD simulation to study the junction formation dynamics of a (16, 0) SWCNT containing a hole of ten vacancies. The formation of two pentagon-heptagon pair defects is also observed, with the mechanism similar to that in the simulation of the (16, 0) SWCNT with six vacancies discussed earlier in this subsection. The most interesting difference between the simulation results of the ten and six vacancies is that the length of the (15, 0) tube section is longer with ten vacancies. These simulation results demonstrate that intramolecular semiconductor-metal junctions of SWCNTs can be produced by irradiation followed by a proper annealing which allow various vacancy defects generated by the irradiation to reconstruct into the pentagon-heptagon pairs at the junction. These simulations also suggest a mechanism for synthesis of carbon nanotube semiconductor-metal intramolecular junctions with specific locations and controlled sizes and show the possibility of application to nanoelectronic devices.

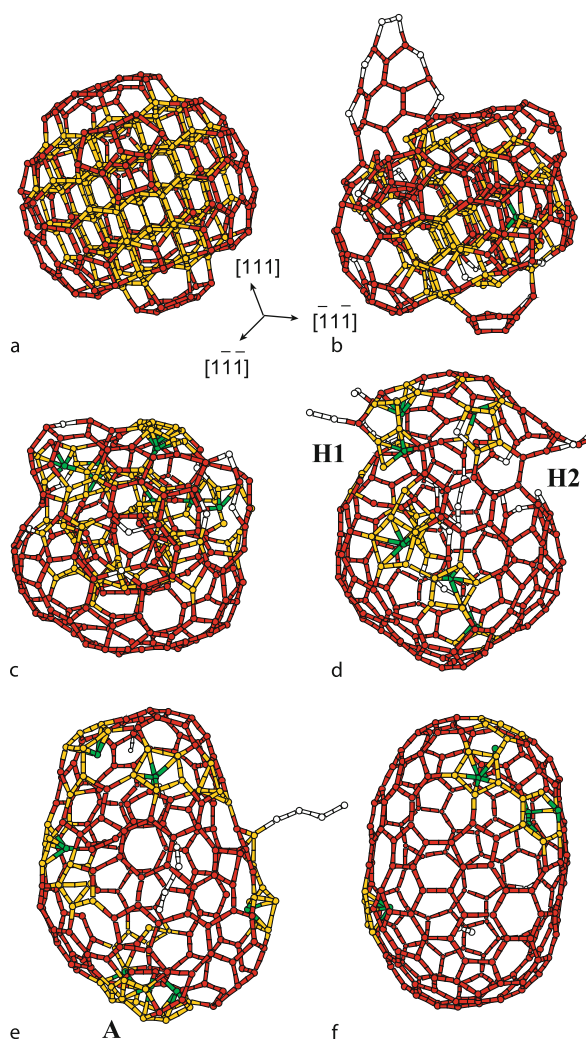
### Transformation from Nanodiamond

Nanometer-sized diamonds have been found in interstellar dust [53], solid detonation products [17], and diamond-like films [6]. Recently, Raty et al. [72] performed ab initio calculations and tight-binding molecular dynamics simulations to study the structure of a nanodiamond and found that the carbon nanoparticle consist of a diamond core and a reconstructed fullerene-like surface. Experiments have shown that diamond nanoparticles of diameter  $\sim 5$  nm can be transformed into spherical and polyhedron carbon onions at high temperatures [45,84,85].

Using the environment-dependent carbon tight-binding potential developed by Tang et al. [78], Lee et al. have performed tight-binding molecular dynamics simulations to study the structural transformation of nanodiamond at high temperature. The simulations show that upon annealing up to 2500 K, a 1.4 nm-diameter nanodiamond is transformed into a cage structure that looks like a single-walled capped nanotube [52].

The simulation was started with a bulk-terminated carbon cluster of 275 atoms within a sphere of diameter of 1.4 nm cut from bulk diamond. This cluster is relaxed using the steepest descent method with the environment-dependent tight-binding carbon potential. The cluster structure after the relaxation is similar to that of the previous ab initio calculation [72]. Due to the surface reconstruction of nanodiamonds, there are graphite-like fragments present at the first atomic layer of the (111) facets, together with the formation of pentagons which link the graphene fragments to the underneath atoms (Fig. 19a). The graphite-like fragments at the (111), ( $\bar{1}\bar{1}1$ ), ( $\bar{1}1\bar{1}$ ), and ( $1\bar{1}\bar{1}$ ) surfaces consist of 3 pentagons and 3 hexagons and those at the ( $\bar{1}11$ ), (11 $\bar{1}$ ), (111), and ( $\bar{1}\bar{1}\bar{1}$ ) surfaces consist of only 3 pentagons. Starting from this relaxed cluster geometry, tight-binding molecular dynamics simulation was performed to investigate the structural transformation of the nanodiamond at high temperatures.

Figure 19 displays snapshots of the system during the structural transformation into a capped nanotube. First the nanodiamond cluster was heated up to about 2,500 K by constant-temperature molecular dynamics simulations. Near 2,500 K, as shown in Fig. 19b, the (111) surface layer of the nanodiamond begins to graphitize after a simulation time of 3 picoseconds (ps), the exfoliation of the graphitized (111) layer occurs by breaking the bonds between graphene fragments and the underneath “core” atoms. This resembles the graphitization process of the (111) surface of bulk diamond induced by nanosecond laser pulses [93]. As the simulation continues, the graphitized layer evaporates, breaking down into carbon dimers one-by-one from the end of layer. Similarly, the ( $\bar{1}\bar{1}\bar{1}$ ) surface layer consisting of three pentagons undergoes the same exfoliation and evaporation process as that of the (111) surface. At about 18 ps, as shown in Fig. 19c, the graphitization process extends to the entire cluster surface. The “core” atoms and the surface “shell” atoms start to separate at the bottom side of the cluster. As the bonds between the “core” and “shell” atoms start to break up, the cluster begins to inflate like a bubble. At this stage, if the thermostat is maintained to keep the system at the constant temperature of 2,500 K, the whole cluster would completely evaporate within a simulation time of 45 ps. To prevent

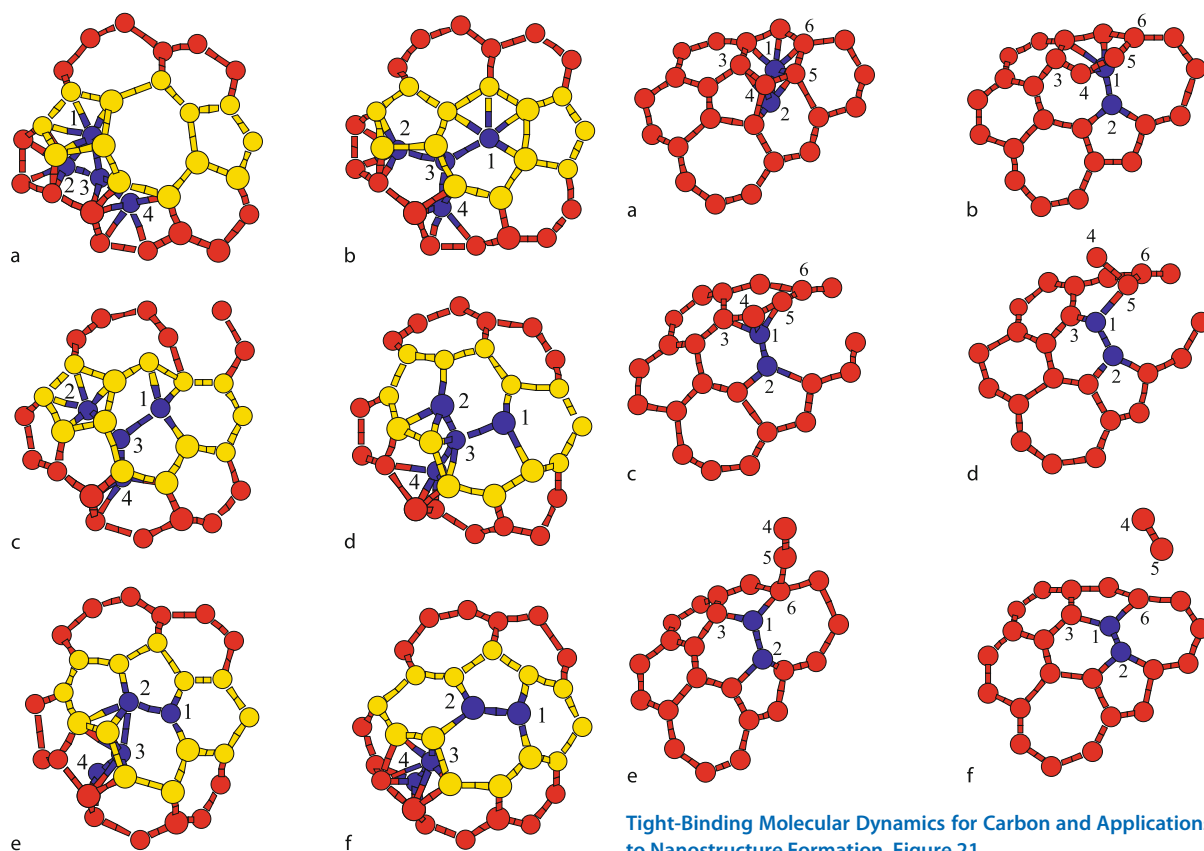


**Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 19**

Atomic processes of structural transformation of nanodiamond to capped nanotube by successive annealings. a 0 K (at time  $t = 0$  ps), b  $\sim 2,500$  K ( $t \approx 3$  ps), c  $\sim 2,500$  K ( $t \approx 19$  ps), d  $\sim 2,100$  K ( $t \approx 35$  ps), e  $\sim 1,900$  K ( $t \approx 50$  ps), f  $\sim 20$  sK ( $t \approx 120$  ps). Simulated annealings with temperatures up to 3,000 K are performed during the process e  $\rightarrow$  f. White color indicates atoms and bonds of 2 and less-fold coordination. Red and yellow colors indicate atoms and bonds of 3-fold coordination and 4-fold coordination, respectively. Green colors indicate atoms and bonds of 5 and higher-fold coordination. Note that two holes H1 and H2 are created in d which serve as exits for inner atoms to “flow-out” to the surface of the structure. (From Ref. [52])

the full vaporization, the system is cooled by decreasing the temperature from 2,500 K to 2,000 K in 10 ps (during the simulation time between 25 ps and 35 ps). The cluster is then further cooled down to a temperature of  $\sim 1,500$  K





#### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 20

Conversion of inner residual carbon atoms into surface by the 'direct adsorption' mechanism. The figure shows the part of the cluster marked by 'A' in Fig. 19e. Blue atoms indicate the inner atoms of the cluster at the initial configuration a. Red and yellow atoms indicate surface atoms. a Four inner atoms (labels 1, 2, 3, and 4) are seen to adhere to the surface of the nanotube from inside. b–c Atom '1' then moves to the center of a heptagonal ring, and breaks the bond at the boundary of the heptagonal and pentagonal ring. d A  $sp^2$  bonding configuration for atom "1" is formed and at the same time atom '2' also breaks the bond at the boundary of the heptagonal and the other pentagonal ring. e–f eventually, atoms '1' and '2' are combined and converted into surface atoms. (From Ref. [52])

in 20 ps when a stable cage structure is found to form. The two holes ('H1' and 'H2' in Fig. 19d), generated by the successive breaking of bonds among surface atoms play an important role in pumping the inner carbon atoms out onto the surface to form the graphitic layer. The process is referred to as the 'flow-out' mechanism by Lee et al. The 'flow-out' mechanism is a unique process in the transformation of a nanodiamond structure into a cage structure. This mechanism is distinguished from other transformation mechanisms observed on bulk diamond surfaces.

#### Tight-Binding Molecular Dynamics for Carbon and Applications to Nanostructure Formation, Figure 21

Conversion of residual inner carbon atoms into surface by the 'push out' mechanism. Blue and red colors indicate inner and surface atoms respectively at the initial configuration of a. a–b The inner atoms '1' and '2' are seen to move from one place to another, breaking their bonds with some surface atoms. c–d Atom '3' becomes a 4-fold coordinated atom by bonding to the inner atom '1', and then breaks the bond with atom '4' to keep its 3-fold coordination. d–e After the inner atom '1' forms a bond with atom '6', it breaks the bond with atom '5'. f Atom '6' also breaks the bond with atom '5' to keep its 3-fold coordination and the carbon dimer formed by atom '4' and '5' is evaporated and the inner atoms '1' and '2' finally convert into surface atoms. (From Ref. [52])

During the annealing process (stage e–f in Fig. 19), two other interesting atomic processes, namely the 'direct absorption' and 'push-out' mechanisms, have been identified from the simulation to play a crucial role in the conversion of the residual inner carbon atoms of the nanodiamond into the surface atoms of the nanotube. These mechanisms are illustrated in Figs. 20 and 21 respectively.

#### Future Directions

In this review, we have shown that tight-binding molecular dynamics is an accurate and efficient method for studying

the structures and properties of carbon fullerenes and carbon nanotubes. Tight-binding molecular dynamics simulation studies of various pathways of the fullerene and nanotube formations have provided useful insights into the formation mechanism of these nanostructures at the atomic scale.

Nevertheless, our understanding of the formation of carbon nanostructures is still very far from being completed. For example, it is a common practice in experiment to use small amount of transition metals (e. g., Fe) to catalyze the nucleation and growth of single-walled carbon nanotubes [15]. The atomic processes under such growth condition have not yet been well studied. Multi-walled carbon nanotubes and buckyonions are also not well studied even though such nanostructures are frequently seen in experiments. The bottleneck is the development of accurate and transferable tight-binding potentials. So far, there are no reliable potentials for carbon/transition metal interactions suitable for tight-binding molecular dynamics simulations. In addition, the tight-binding carbon potentials we discussed in this paper (i. e., the XWCH and EDTB potentials) do not describe the interaction between graphite layers accurately. They are therefore not suitable for studying multi-walled carbon nanotubes and buckyonions. Besides the carbon nanostructures, there is increasing interest in using tight-binding molecular dynamics to study other nanostructures (e. g. Si-based nanostructures and metal nanostructures). The success of such future studies depends on the development of more refined tight-binding potentials.

Very recently, the authors and co-worker have developed a method to extract environment-dependent minimal-basis-set orbitals from ab initio wavefunctions [55,56,57]. These orbitals give an exact description of the occupied electronic states and are highly localized on the individual atoms, making them ideal for use in an accurate tight-binding description of the system. This scheme is easily generalizable to systems involving different atomic species and will simplify the task of generating accurate tight-binding potentials for complex systems. We anticipate that more accurate and transferable tight-binding potentials for nanostructure simulations will be generated based on these highly localized environment-dependent minimal-basis-set orbitals.

## Acknowledgments

We would like to thank Dr. Wencai Lu for her help in preparing the figures. Ames Laboratory is operated for the US Department of Energy by Iowa State University under Contract No. DE-AC02-07CH11358. This work was

supported by the Director for Energy Research, Office of Basic Energy Sciences including a grant of computer time at the National Energy Research Supercomputing Center (NERSC) in Berkeley.

## Bibliography

1. Ballone P, Milani P (1990) *Phys Rev B* 42:3201
2. Bandow S, Takizawa M, Hirahara K, Yudasaka M, Iijima S (2001) *Chem Phys Lett* 337:48
3. Baskes MJ (1987) *Phys Rev Lett* 59:2666
4. Biswas R, Hamann DR (1985) *Phys Rev Lett* 55:2001
5. Car R, Parrinello M (1985) *Phys Rev Lett* 55:2471
6. Chang YK, Hsieh HH, Pong WF, Tsai MH, Chien FZ, Tseng PK, Chen LC, Wang TY, Chen KH, Bhusari DM (1999) *Phys Rev Lett* 82:5377
7. Chelikowsky JR (2000) *J Phys D* 33:R33
8. Chelikowsky JR, Louie SG (eds) (1996) *Quantum Theory of Real Materials*. Kluwer Press, Boston
9. Chelikowsky JR, Phillips JC, Kamal M, Strauss M (1989) *Phys Rev Lett* 62:292
10. Chico L, Crespi VH, Benedict LX, Louie SG, Cohen ML (1996) *Phys Rev Lett* 76:971
11. Cowley ER (1988) *Phys Rev Lett* 60:2379
12. Crespi VH, Cohen ML (1997) *Phys Rev Lett* 79:2093
13. Dodson BW (1987) *Phys Rev B* 35:2795
14. Dresselhaus MS, Dresselhaus G (1982) In: Cardona M, Guntherodt G (ed) *Light Scattering in Solids III*. Springer, Berlin, p 8
15. Dresselhaus MS, Dresselhaus G, Avouris P (eds) (2001) *Carbon Nanotubes: Synthesis, Structure, Properties, Applications*. Springer, Berlin
16. Goodwin L, Skinner AJ, Pettifor DG (1989) *Europhys Lett* 9:701
17. Greiner NR, Phillios DS, Johnson JD, Volk F (1988) *Nature* 333:440
18. Haerle R, Galli G, Baldereschi A (1999) *Appl Phys Lett* 75:1718
19. Haerle R, Riedo E, Pasquarello A, Baldereschi A (2001) *Phys Rev B* 65:045101
20. Hamada N, Sawada S-I, Oshiyama A (1992) *Phys Rev Lett* 68:579
21. Hashimoto A, Suenaga K, Gloter A, Urita K, Iijima S (2004) *Nature* 430:870
22. Hernandez E, Meunier V, Smith BW, Rurali R, Terrones H, Buongiorno M Nardelli, Terrones M, Luzzi DE, Charlierr J-C (2003) *Nanolett* 3:1037
23. Hohenberg P, Kohn W (1964) *Phys Rev* 136:B864
24. Iijima S (1991) *Nature* 354:56
25. Irle S, Zheng G, Elstner M, Morokuma K (2003) *Nanolett* 3:1657
26. Irle S, Zheng G, Elstner M, Morokuma K (2003) *Nanolett* 3:465
27. Irle S, Zheng G, Wang Z, Morokuma K (2006) *J Phys Chem. B* 110:14531
28. Lee I-H, Kim H, Lee J (2004) *J Chem Phys* 120:4672
29. Kawai T, Miyamoto Y, Sugino O, Koga Y (2002) *Phys Rev B* 66:033404
30. Kawai T, Miyamoto Y, Sugino O, Koga Y (2002) *Phys Rev Lett* 89:085901
31. Khan FS, Broughton JQ (1989) *Phys Rev B* 39:3688
32. Khor KE, Das S (1988) *Phys Rev B* 38:3318
33. Kilic C, Ciraci S, Gulseren O, Yildirim T (2000) *Phys Rev B* 62:R16345

34. Kim D-H, Chang KJ (2002) *Phys Rev B* 65:155402
35. Kim D-H, Sim H-S, Chang KJ (2001) *Phys Rev B* 64:115409
36. Kim Y-H, Lee I-H, Chang KJ, Lee S (2003) *Phys Rev Lett* 90:065501
37. Kohn W, Sham LJ (1965) *Phys Rev* 140:A1133
38. Krättschmer W, Lamb LD, Fostiropoulos K, Huffman DR (1990) *Nature* 347:354
39. Kresse G (1993) Ph.D Technische University of Wien
40. Kresse G, Furthmüller J (1996) *Comput Mater Sci* 6:15
41. Kresse G, Furthmüller J (1996) *Phys Rev B* 54:11 169
42. Kresse G, Hafner J (1993) *Phys Rev B* 47:558
43. Kresse G, Joubert J (1999) *Phys Rev B* 59:1758
44. Kroto HW, Heath JR, O'Brien SC, Curl RF, Smalley RE (1985) *Nature* 318:162
45. Kuznetsov VL, Chuvilin AL, Butenko YV, Mal'kov IY, Titov VM (1994) *Chem Phys Lett* 222:343
46. Laasonen K, Nieminen RM (1991) *J Phys Condens Matter* 2:1509
47. Laasonen K, Nieminen RM (1991) *ibid* 3:7455
48. Laszlo I (1998) *Europhys Lett* 44:741
49. Lee G-D, Wang CZ, Yoon E, Hwang N-M, Ho KM (2006) *Phys Rev B* 74:245411
50. Lee G-D, Wang CZ, Yoon E, Hwang N-M, Kim D-Y, Ho KM (2005) *Phys Rev Lett* 95:205501
51. Lee G-D, Wang CZ, Yu J, Yoon E, Hwang N-M, Ho K-M (2007) *Phys Rev B* 76:165413
52. Lee GD, Wang CZ, Yu J, Yoon E, Ho KM (2003) *Phys Rev Lett* 91:265701
53. Lewis RS, Anders E, Draine BT (1989) *Nature* 339:117
54. Lu JQ, Wu J, Duan WH, Liu F, Zhu BF, Gu BL (2003) *Phys Rev Lett* 90:156601
55. Lu WC, Wang CZ, Ruedenberg K, Ho KM (2004) *Phys Rev B* 70:041101
56. Lu WC, Wang CZ, Schmidt MW, Bytautas L, Ho KM, Ruedenberg K (2004) *J Chem Phys* 120:2629
57. Lu WC, Wang CZ, Schmidt MW, Bytautas L, Ho KM, Ruedenberg K (2004) *J Chem Phys* 120:2638
58. Luzzi DE, Smith BW (2000) *Carbon* 38:1751
59. Madelung O, Schulz M (eds) (1987) *Semiconductors: Intrinsic Properties of Group IV Elements and III-V, II-VI and I-VII Compounds*. Landolt-Börnstein New Series III/22a. Springer, Berlin
60. Madelung O, Schulz M, Weiss H (eds) (1982) *Semiconductors: Physics of Group IV Elements and III-V Compounds*. Landolt-Börnstein New Series III/17a. Springer, Berlin
61. Menon M, Andriotis AN, Srivastava D, Ponomareva I, Chernozatonskii LA (2003) *Phys Rev Lett* 91:144501
62. Menon M, Subbaswamy KR (1991) *Phys Rev Lett* 67:3487
63. Mintmire JW, Dunlap BI, White CT (1992) *Phys Rev Lett* 68:631
64. Morris JR, Wang CZ, Ho KM (1995) *Phys Rev B* 52:4138
65. Nikolaev P, Thess A, Rinzler AG, Colbert DT, Smalley RE (1997) *Chem Phys Lett* 266:422
66. Odom TW, Huang J-L, Kim P, Lieber CM (1998) *Nature* 391:62
67. Ogata S, Shibutani Y (2003) *Phys Rev B* 68:165409
68. Oh D-H, Lee YH (1998) *Phys Rev B* 58:7407
69. Ozaki T, Iwasa Y, Mitani T (2000) *Phys Rev Lett* 84:1712
70. Park C-J, Kim Y-H, Chang KJ (1999) *Phys Rev B* 60:10656
71. Porezag D, Frauenheim T, Kohler T, Seifert G, Kaschner R (1995) *Phys Rev B* 51:12947
72. Raty JY, Galli G, Bostedt C, Van Buuren TW, Terminello LJ (2003) *Phys Rev Lett* 90:037401
73. Saito R, Fujita M, Dresselhaus G, Dresselhaus MS (1992) *Appl Phys Lett* 60:2204
74. Sankey OF, Niklewski DJ (1989) *Phys Rev B* 40:3979
75. Slater JC, Koster GF (1954) *Phys Rev* 94:1498
76. Smith BW, Monthieux M, Luzzi DE (1998) *Nature* 396:323
77. Stillinger FH, Weber TA (1985) *Phys Rev B* 31:5262
78. Tang MS, Wang CZ, Chan CT, Ho KM (1996) *Phys Rev B* 53:979
79. Terrones M, Banhart F, Grobert N, Charlier J-C, Terrones H, Ajayan PM (2002) *Phys Rev Lett* 89:75505
80. Terrones M, Terrones H, Banhart F, Charlier J-C, Ajayan PM (2000) *Science* 288:1226
81. Tersoff J (1986) *Phys Rev Lett* 56:632
82. Tersoff J (1988) *Phys Rev B* 37:6991
83. Tersoff J (1988) *Phys Rev Lett* 61:2879
84. Tomita S, Fujii M, Hayashi S (2002) *Phys Rev B* 66:245424
85. Tomita S, Fujii M, Hayashi S, Yamamoto K (1999) *Chem Phys Lett* 305:225
86. Virkkunen R, Laasonen K, Nieminen RM (1991) *ibid* 3:7455
87. Wang CZ, Chan CT, Ho KM (1989) *Phys Rev B* 39:8592
88. Wang CZ, Chan CT, Ho KM (1992) *Phys Rev B* 46:9761
89. Wang CZ, Ho KM (1993) *Phys Rev Lett* 71:1184
90. Wang CZ, Ho KM (1994) *Phys Rev B* 50:12429
91. Wang CZ, Ho KM, Chan CT (1993) *Phys Rev B* 47:14835
92. Wang CZ, Ho KM, Chan CT (1993) *Phys Rev Lett* 70:611
93. Wang CZ, Ho KM, Shirk M, Molian P (2000) *Phys Rev Lett* 85:4092
94. Wang CZ, Xu CH, Chan CT, Ho KM (1992) *J Phys Chem* 96:3563
95. Wildoer JWG, Venema LC, Rinzler AG, Smalley RE, Dekker C (1998) *Nature* 391:59
96. Xu CH, Wang CZ, Chan CT, Ho KM (1992) *J Phys Condens Matter* 4:6047
97. Yang L, Han J (2000) *Phys Rev Lett* 85:154
98. Zhang BL, Wang CZ, Chan CT, Ho KM (1993) *J Phys Chem* 97:3134
99. Zhang BL, Wang CZ, Ho KM (1992) *Chem Phys Lett* 193:225
100. Zhang BL, Wang CZ, Ho KM (1992) *J Chem Phys* 96:7183
101. Zhang BL, Wang CZ, Ho KM, Xu CH, Chan CT (1993) *J Chem Phys* 97:5007(1992); 98:3095
102. Zhang BL, Xu CH, Wang CZ, Chan CT, Ho KM (1992) *Phys Rev B* 46:7333
103. Zhang P, Crespi VH (1999) *Phys Rev Lett* 83:1791
104. Zhao QZ, Nardelli MB, Bernholc J (2003) *Phys Rev B* 65:144105
105. Zheng G, Irle S, Morokuma K (2005) *J Chem Phys* 122:014708
106. Zheng G, Irle S, Elstner M, Morokuma K (2004) *J Phys Chem A* 108:3182

## Tiling Problem and Undecidability in Cellular Automata

JARKKO KARI

Department of Mathematics,  
University of Turku, Turku, Finland

### Article Outline

Glossary

Definition of the Subject

## Introduction

### The Tiling Problem and Its Variants

### Undecidability in Cellular Automata

### Future Directions

### Acknowledgments

### Bibliography

## Glossary

**Cellular automata (CA)** A  $d$ -dimensional cellular automaton consists of an infinite  $d$ -dimensional grid of cells, indexed by  $\mathbb{Z}^d$ . Each cell stores an element of a finite state set  $S$ . Configuration  $c: \mathbb{Z}^d \rightarrow S$  specifies the states of all cells. The set of all configurations is  $S^{\mathbb{Z}^d}$ . The neighborhood vector  $N = (\vec{n}_1, \vec{n}_2, \dots, \vec{n}_m)$  is a sequence of  $m$  distinct elements of  $\mathbb{Z}^d$  specifying the relative locations of the neighbors of the cells: A cell located at  $\vec{x} \in \mathbb{Z}^d$  has  $m$  neighbors, in positions  $\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \dots, \vec{x} + \vec{n}_m$ . Finally, the local update rule  $f: S^m \rightarrow S$  specifies the new state of a cell, based on the old states of its neighbors. In one step, configuration  $c$  is transformed into configuration  $e$  where, for all  $\vec{x} \in \mathbb{Z}^d$ ,

$$e(\vec{x}) = f[c(\vec{x} + \vec{n}_1), c(\vec{x} + \vec{n}_2), \dots, c(\vec{x} + \vec{n}_m)].$$

The mapping  $c \mapsto e$  is the global transition function, or the CA-function,  $G: S^{\mathbb{Z}^d} \rightarrow S^{\mathbb{Z}^d}$  specified by the CA  $\mathcal{A} = (d, S, N, f)$ .

**Tiles** A Wang tile is a unit square tile with colored edges. Tiles have an orientation, i.e. they may not be rotated or reflected. The colors give a local matching rule that specifies which tiles may be placed next to each other: Two adjacent tiles must have identical colors on the abutting edges. A Wang tile set consists of a finite number of Wang tiles.

A more general definition: a  $d$ -dimensional tile set is a quadruple  $\mathcal{T} = (d, T, N, R)$  where  $T$  is a finite set whose elements are called tiles,  $N = (\vec{n}_1, \vec{n}_2, \dots, \vec{n}_m)$  is a neighborhood vector of  $m$  distinct elements of  $\mathbb{Z}^d$  and  $R \subseteq T^m$  is a relation of allowed patterns. The neighborhood vector has the same interpretation as in the definition of cellular automata: it gives the relative locations of the neighbors of cells.

**Tiling** A covering of the plane using tiles. A valid Wang tiling by a Wang tile set  $T$  is an assignment  $t: \mathbb{Z}^2 \rightarrow T$  of tiles to cells such that the local matching rule is satisfied between all adjacent tiles. We say that  $T$  admits tiling  $t$ .

More general definition: A  $d$ -dimensional tiling using tile set  $\mathcal{T} = (d, T, N, R)$  is a mapping  $t: \mathbb{Z}^d \rightarrow T$ .

Tiling  $t$  is valid at cell  $\vec{x} \in \mathbb{Z}^d$  if

$$t(\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \dots, \vec{x} + \vec{n}_m) \in R.$$

Tiling  $t$  is called valid if it is valid at every cell  $\vec{x} \in \mathbb{Z}^d$ .

**Periodic tiling** A tiling that is invariant under some non-zero translation. A two-dimensional tiling is called totally periodic if it is invariant under two linearly independent translations. A totally periodic tiling is automatically periodic in horizontal and vertical directions, which means that it consists of a rectangular pattern that is repeated horizontally and vertically to fill the plane. A two-dimensional tile set that admits a valid periodic tiling automatically admits also totally periodic tiling.

More generally, a  $d$ -dimensional tiling  $t: \mathbb{Z}^d \rightarrow T$  is periodic with period  $\vec{p} \in \mathbb{Z}^d, \vec{p} \neq \vec{0}$ , if  $t(\vec{x}) = t(\vec{x} + \vec{p})$  for all  $\vec{x} \in \mathbb{Z}^d$ . It is totally periodic if it is periodic with  $d$  linearly independent periods  $\vec{p}_1, \vec{p}_2, \dots, \vec{p}_d$ . Note that when  $d > 2$  it is possible that a tile set admits a periodic tiling but does not admit any totally periodic tiling.

**Aperiodic tile set** A two-dimensional tile set that admits a valid tiling of the plane, but does not admit any valid periodic tilings. Smallest known aperiodic set of Wang tiles contain 13 tiles [3,13]

**Turing machine (TM)** Turing machines are computation devices commonly used to formally define the concept of an algorithm. They also provide us with the most basic undecidable decision problems. A Turing machine consists of a finite state control unit that moves along an infinite tape. The tape has symbols written in cells that are indexed by  $\mathbb{Z}$ . Depending on the state of the control unit and the symbol currently scanned on the tape the machine may overwrite the tape symbol, change the internal state and move along the tape one cell to the left or right. We formally define a Turing machine as a 6-tuple  $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$  where  $Q$  and  $\Gamma$  are finite sets (the state alphabet and the tape alphabet, respectively),  $q_0, q_h \in Q$  are the initial and the halting states, respectively,  $b \in \Gamma$  is the blank symbol and  $\delta: Q \times \Gamma \rightarrow Q \times \Gamma \times \{-1, 1\}$  is the transition function that specifies the moves of the machine. A configuration (or instantaneous description) of the machine is a triplet  $(q, i, t)$  where  $q \in Q$  is the current state,  $i \in \mathbb{Z}$  is the position of the machine on the tape and  $t: \mathbb{Z} \rightarrow \Gamma$  describes the content of the tape. In one time step configuration  $(q, i, t)$  becomes  $(q', i + d, t')$  if  $\delta(q, t(i)) = (q', \gamma, d)$  and  $t'(i) = \gamma$  and  $t'(j) = t(j)$  for all  $j \neq i$ . We denote



this move by

$$(q, i, t) \vdash (q', i + d, t').$$

The reflexive, transitive closure of  $\vdash$  is denoted by  $\vdash^*$ , that is,

$$(q, i, t) \vdash^* (q', i', t')$$

if and only if  $(q', i', t')$  can be reached from  $(q, i, t)$  by executing zero or more Turing machine moves.

**Decision problem** A decision problem is an algorithmic question with a yes/no -answer. The problem has an input (called the instance of the problem) and a well defined answer “yes” or “no” associated to each instance.

**The halting problem** TURING MACHINE HALTING ON BLANK TAPE is the decision problem whose input is a Turing machine  $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$  and the answer is positive if and only if the Turing machine eventually enters its halting state  $q_h$  when started in the initial state  $q_0$  on a totally blank tape, i. e. initially every tape location has the blank symbol  $b$ .

**(Un)decidability** Some decision problems can not be solved by any algorithm. Such problems are called undecidable. In contrast, decidable decision problems are solved by some algorithm. An example of an undecidable problem is the TURING MACHINE HALTING ON BLANK TAPE.

**Semi-algorithm** An algorithm-like procedure for a decision problem that correctly returns a positive answer on positive input instances, but on negative instances runs for ever without ever returning an answer.

**Semi-decidability** A decision problem is called semi-decidable if there is a semi-algorithm for it. For example, the decision problem TURING MACHINE HALTING ON BLANK TAPE is semi-decidable since one can simulate any given Turing machine step-by-step until (if ever) it halts.

**Recursive and recursively enumerable (re)** A formal language  $L$  is called recursive if it is decidable whether a given word belongs to  $L$ . The language is called recursively enumerable (re for short) if this membership problem is semi-decidable.

**The tiling problem** The decision problem that gets as input a tile set  $\mathcal{T}$ , and asks whether there exists a valid tiling by  $\mathcal{T}$ . The tiling problem was proved undecidable for Wang tiles by R. Berger [2]. Its complement (i. e., “Does there not exist a valid tiling?”) is semi-decidable.

**The tiling problem with a seed tile** The decision problem that gets as input a tile set  $\mathcal{T}$  and one tile  $s$ , and

asks whether  $\mathcal{T}$  admits a tiling that contains tile  $s$  at least once. The problem is undecidable for Wang tiles, but its complement is semi-decidable [24].

**The periodic tiling problem** The decision problem to determine if a given set of Wang tiles admits a periodic tiling. The problem is undecidable, but it is semi-decidable [7].

**The finite tiling problem** A decision problem where we are given a set  $T$  of Wang tiles and a specific blank tile  $B \in T$  whose all four edges are colored with the same color. A tiling  $t: \mathbb{Z} \rightarrow T$  is called finite if

$$\{\vec{x} \in \mathbb{Z}^2 \mid t(\vec{x}) \neq B\}$$

is a finite set. If  $t(\vec{x}) = B$  for all  $\vec{x} \in \mathbb{Z}^2$  then  $t$  is called trivial. The finite tiling problem asks whether there exist non-trivial valid finite tilings. The problem is undecidable but semi-decidable.

**Surjective cellular automata** A cellular automaton (CA) is called surjective if every configuration has a pre-image, that is, if its global transition function  $G: S^{\mathbb{Z}^d} \rightarrow S^{\mathbb{Z}^d}$  is surjective.

**Injective (reversible) CA** A CA is injective if every configuration has at most one pre-image, that is, the global transition function is one-to-one. It is well known that a CA is injective if and only if it is bijective (every configuration has a unique pre-image), which in turn is equivalent to reversibility (=there exists an inverse CA that traces the CA back in time.)

**Limit set** The limit set of a CA is its maximal attractor. In other words, it is the compact and translation invariant set

$$\bigcap_{i=0}^{\infty} G(S^{\mathbb{Z}^d})$$

where  $G$  is the global transition function and  $S$  is the state set.

**Nilpotent cellular automata** Nilpotent CA have trivial dynamics. A CA is called nilpotent if its limit set contains only one configuration. The unique element of the limit set is the quiescent configuration. This is equivalent to every initial configuration eventually becoming the quiescent configuration.

**Equicontinuous CA** Cellular automaton  $G$  is called equicontinuous if for every finite  $A \subseteq \mathbb{Z}^d$  there exists a finite  $B \subseteq \mathbb{Z}^d$  such that any two initial configurations that agree inside  $B$  will agree inside  $A$  for all subsequent steps. In other words,

$$\begin{aligned} \forall \vec{x} \in B: c(\vec{x}) = e(\vec{x}) \\ \implies \forall t \in \mathbb{N} \text{ and } \forall \vec{x} \in A: G^t(c)(\vec{x}) = G^t(e)(\vec{x}). \end{aligned}$$



This means that equicontinuous CA can be reliably simulated in finite windows. It is known that a CA is equicontinuous if and only if it is ultimately periodic [17]:  $\exists n, p \in \mathbb{N}: G^n = G^{n+p}$ . In this sense the dynamics of equicontinuous CA is trivial.

**Sensitive CA** Cellular automaton  $G$  is called sensitive to initial conditions if there exists a finite set  $B \subseteq \mathbb{Z}^d$  of cells such that for every configuration  $c$  and every finite set  $A \subseteq \mathbb{Z}$  of cells there exists a configuration  $e$  and time  $t \geq 0$  such that  $e(\vec{x}) = d(\vec{x})$  for all  $\vec{x} \in A$  but  $G^t(e)(\vec{x}) \neq G^t(c)(\vec{x})$  for some  $\vec{x} \in B$ . This means that arbitrarily distant modifications to any configuration  $c$  may propagate to a fixed observation window  $B$ .

**Topological entropy** The topological entropy  $h(G)$  of a CA  $G$  measures the complexity of its dynamics. For any finite  $A \subseteq \mathbb{Z}^d$  and positive integer  $n$  we define the equivalence relation  $\equiv_{A,n}$  among initial configurations as follows: For all  $c, e \in S^{\mathbb{Z}^d}$

$$\begin{aligned} c &\equiv_{A,n} e \\ \Leftrightarrow \forall \vec{x} \in A, 0 \leq t < n: G^t(c)(\vec{x}) &= G^t(e)(\vec{x}). \end{aligned}$$

In other words, two configurations are equivalent if we can not observe any difference in their orbits in region  $A$  within the first  $n$  time instances. Let us denote by  $N_G(A, n)$  the number of equivalence classes of  $\equiv_{A,n}$ . Then the topological entropy is

$$h(G) = \sup_A \lim_{n \rightarrow \infty} \frac{\log N_G(A, n)}{n}$$

where the supremum is over all finite  $A \subseteq \mathbb{Z}^d$ . The entropy always exists. In the one-dimensional case the entropy is always a finite, non-negative number. If  $d \geq 2$  then the entropy can also be infinite.

## Definition of the Subject

We consider the following algorithmic questions concerning cellular automata. All problems are decision problems, that is, the answer for each input instance is either yes or no. All problems considered are undecidable, i. e. no algorithm can solve them. We only consider problems whose undecidability is proved using a reduction from the tiling problem or its variant.

### INJECTIVITY

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  injective (i. e. reversible)?

There is an algorithm that solves INJECTIVITY for one-dimensional CA [1]. But the problem is undecidable among

two-dimensional CA [9,11]. The problem is semi-decidable in any dimension.

### SURJECTIVITY

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  surjective?

Also SURJECTIVITY is decidable among one-dimensional CA. The two-dimensional question is, however, undecidable [11]. The complement of the problem (i. e. non-surjectivity) is semi-decidable in any dimension.

### NILPOTENCY

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  nilpotent?

Nilpotency is undecidable even among one-dimensional CA [10]. It is undecidable even if  $A$  has a spreading state, i. e. a state  $q$  such that any cell whose neighborhood contains  $q$  becomes  $q$ . However, NILPOTENCY is semi-decidable in any dimension. Based on problem NILPOTENCY we can prove a Rice's theorem for Cellular Automata limit sets: any non-trivial decision problem concerning limit sets is undecidable [12].

### TOPOLOGICAL ENTROPY

**Input:** Cellular Automaton  $A$ .

**Question:** Is the topological entropy of  $A$  less than constant  $c > 0$ ?

Problem TOPOLOGICAL ENTROPY is undecidable for every constant  $c > 0$ , even in the one-dimensional case. This can be proved using a direct reduction from NILPOTENCY [8]. Also, direct reductions from NILPOTENCY prove the undecidability of the following two problems [5,14]:

### EQUICONTINUITY

**Input:** Cellular Automaton  $A$ .

**Question:** Is  $A$  equicontinuous?

### SENSITIVITY TO INITIAL CONDITIONS

**Input:** Cellular Automaton  $A$ .

**Question:** Is  $A$  sensitive to initial conditions?

## Introduction

Several decision problems concerning cellular automata are known to be undecidable, that is, no algorithm exists that solves them. Some undecidability results easily follow from the universal computation capabilities of cellular automata, while others require more elaborate proofs.

Reductions from the tiling problem and its variants turn out to be useful in proving various questions concerning CA undecidable. We consider the problems of determining if a given two-dimensional CA is surjective or injective, whether a one-dimensional CA is nilpotent or equicontinuous, and whether the topological entropy is less than some constant.

The fact that the tiling problem is closely related to cellular automata is not surprising considering their apparent similarity: both involve assignments of symbols over a finite alphabet onto integer lattice points. The difference is that tilings are static while cellular automata change the assignments dynamically according to the local rule. Undecidability of the tiling problem on the two-dimensional plane naturally leads to undecidability results concerning single step properties of two- and higher dimensional cellular automata. But also asymptotic properties of one-dimensional cellular automata can be related to tiling problems by viewing the space-time diagram of the CA as a tiling. This naturally leads to the definition of deterministic tile sets: Wang tiles where tiles are uniquely determined by some of their neighbors.

We start by discussing the tiling problem and its variants. We do not prove the undecidability of all the variants. Rather, literature references for the proofs are provided. We then define a particular tile set that has an interesting plane-filling property. This tile set is a useful tool in the reduction to prove that it is undecidable to tell whether a given two-dimensional CA is injective (reversible). We then provide reductions that show several questions concerning cellular automata undecidable.

### The Tiling Problem and Its Variants

In this section we discuss the tiling problem and several of its variants.

#### Definition of Tiles

For our purposes it is convenient to define tiles in a way that most closely resembles cellular automata. In the  $d$ -dimensional cellular space the cells are indexed by  $\mathbb{Z}^d$ . A neighborhood vector

$$N = (\vec{n}_1, \vec{n}_2, \dots, \vec{n}_m)$$

consists of  $m$  distinct elements  $\vec{n}_i \in \mathbb{Z}^d$ . Each  $\vec{n}_i$  specifies the relative location of a neighbor of each cell. More precisely, the  $i$ th neighbor of the cell in position  $\vec{x} \in \mathbb{Z}^d$  is located at  $\vec{x} + \vec{n}_i$ .

A tile set is a finite set  $T$  whose elements are called tiles. A local matching rule tells which patterns of tiles

are allowed in valid tilings. The matching rule is given as an  $m$ -ary relation  $R \subseteq T^m$  where  $m$  is the size of the neighborhood. Tilings are assignments

$$t: \mathbb{Z}^d \rightarrow T$$

of tiles into cells. Tiling  $t$  is valid at  $\vec{x} \in \mathbb{Z}^d$  if

$$t(\vec{x} + \vec{n}_1, \vec{x} + \vec{n}_2, \dots, \vec{x} + \vec{n}_m) \in R.$$

Tiling  $t$  is called valid if it is valid at every position  $\vec{x} \in \mathbb{Z}^d$ .

A convenient – and historically earlier – way of defining tiles is in terms of edge labelings. A Wang tile is a two-dimensional unit square with colored edges. The local matching rule is determined by these colors: A tiling is valid at position  $\vec{x} \in \mathbb{Z}^2$  iff each of the four edges of the tile in position  $\vec{x}$  have the same color as the abutting edge in the adjacent tile. Clearly this is a two-dimensional tile set with the neighborhood vector  $[(-1, 0), (1, 0), (0, -1), (0, 1)]$  and a particular way of defining the local relation  $R$ .

### Computations and Tilings

The basic observation in establishing undecidability results concerning tilings is the fact that valid tilings can be forced to contain a complete simulation of a computation by a given Turing machine. To any given Turing machine  $\mathcal{M} = (Q, \Gamma, \delta, q_0, q_h, b)$  we associate the Wang tiles shown in Fig. 1, and we call these tiles the *machine tiles* of  $\mathcal{M}$ . Note that in the illustrations, instead of colors, we use labeled arrows on the sides of the tiles. Two adjacent tiles match if and only if an arrow head meets an arrow tail with the same label. Such arrow representation can be converted into the usual coloring representation of Wang tiles by assigning to each arrow direction and label a unique color.

The machine tiles of  $\mathcal{M}$  contain the following tiles:

- (i) For every tape letter  $a \in \Gamma$  a *tape tile* of Fig. 1a,
- (ii) For every tape letter  $a \in \Gamma$  and every state  $q \in Q$  an *action tile* of Fig. 1b or c. Tile (b) is used if

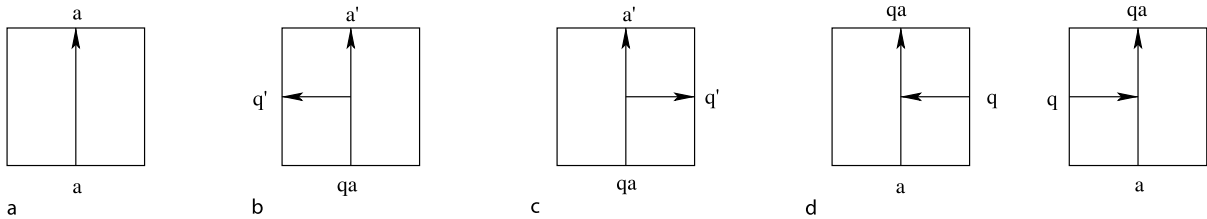
$$\delta(q, a) = (q', a', -1)$$

and tile (c) is used if

$$\delta(q, a) = (q', a', +1).$$

- (iii) For every tape letter  $a \in \Gamma$  and non-halting state  $q \in Q \setminus \{q_h\}$  two merging tiles shown in Fig. 1d.

The idea of the tiles is that a configuration of the Turing machine  $\mathcal{M}$  is represented as a row of tiles in such a way that the cell currently scanned by  $\mathcal{M}$  is represented by an



**Tiling Problem and Undecidability in Cellular Automata, Figure 1**  
Machine tiles associated to a Turing machine

action tile, its neighbor where the machine moves into has a merging tile and all other tiles on the row are tape tiles. If this row is part of a valid tiling then it is clear that the rows above must be similar representations of subsequent configurations in the Turing machine computation, until the machine halts.

The machine tiles above are the basic tiles associated to Turing machine  $\mathcal{M}$ . Additional tiles will be added depending on the actual variant of the tiling problem.

### The Tiling Problem

The tiling problem is the decision problem of determining if at least one valid tiling is admitted by the given set of tiles.

#### TILING PROBLEM

**Input:** Tile set  $\mathcal{T}$ .

**Question:** Does  $\mathcal{T}$  admit a valid tiling?

The tiling problem is easily seen decidable if the input is restricted to one-dimensional tile sets. It is classical result by R. Berger that the tiling problem of two-dimensional tiles is undecidable, even if the input consists of Wang tiles [2,21]:

**Theorem 1** TILING PROBLEM is undecidable for Wang tile sets  $\mathcal{T}$ . The complement problem (non-existence of valid tilings) is semi-decidable.

We do not prove this result here. The undecidability proofs in [2,21] are based on an explicit construction of an aperiodic tile set such that additional tiles implementing Turing machine simulations can be embedded in valid tilings. The aperiodic set is needed to force the presence of tiles that initiate Turing machine simulation in arbitrarily large regions.

Note that semi-decidability of the complement problem is apparent: a semi-algorithm simply tries to tile larger and larger regions until (if ever) a region is found that can not be properly tiled. Note also that a semi-algorithm ex-

ists for those tile sets that admit a valid, totally periodic tiling: All totally periodic tilings can be effectively enumerated and it is a simple matter to test each for validity of the tiling constraint. Combining the two semi-algorithms above yields a semi-algorithm that correctly identifies tile sets that (i) do not admit any valid tiling, or (ii) admit a valid periodic tiling. Only aperiodic tile sets fail to satisfy either (i) or (ii), so we see that the existence of aperiodic tile sets is implied by Theorem 1.

In the following sections we consider some variants of the tiling problem whose undecidability is easier to establish.

### Variants of the Tiling Problem

#### TILING PROBLEM WITH A SEED TILE

**Input:** Tile set  $\mathcal{T}$  and one tile  $s$ .

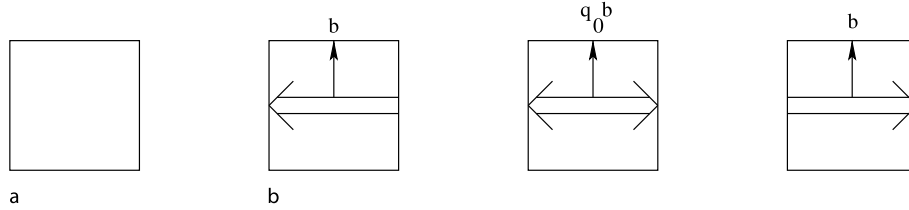
**Question:** Does  $\mathcal{T}$  admit a valid tiling such that tile  $s$  is used at least once?

The seeded version was shown undecidable by H. Wang [24]. We present the proof here because the proof is quite simple and shows the general idea of how Turing machine halting problem can be reduced to problems concerning tiles:

**Theorem 2** TILING PROBLEM WITH A SEED TILE is undecidable for Wang tile sets. The complement problem is semi-decidable.

*Proof* The semi-decidability of the complement problem follows from the following semi-algorithm: For  $r = 1, 2, 3, \dots$  try all tilings of the radius  $r$  square around the origin to see if there is a valid tiling of the square such that the origin contains the seed tile  $s$ . If for some  $r$  such a tiling is not found then halt and report that there is no tiling containing the seed tile.

Consider then undecidability. We reduce the decision problem TURING MACHINE HALTING ON BLANK TAPE, a problem that is well known to be undecidable. For any given Turing machine  $\mathcal{M}$  we can effectively construct a tile set and a seed tile in such a way that they form a positive in-



**Tiling Problem and Undecidability in Cellular Automata, Figure 2**  
**a the blank tile, and b three initialization tiles**

stance of TILING PROBLEM WITH A SEED TILE if and only if  $\mathcal{M}$  is a negative instance of TURING MACHINE HALTING ON BLANK TAPE. For the given Turing machine  $\mathcal{M}$  we construct the machine tiles of Fig. 1 as well as the four tiles shown in Fig. 2. These are the blank tile and three initialization tiles. They initialize all tape symbols to be equal to blank  $b$ , and the Turing machine to be in the initial state  $q_0$ . The middle initialization tile is chosen as the seed tile  $s$ .

Let us prove that a valid tiling containing a copy of the seed tile exists if and only if the Turing machine  $\mathcal{M}$  does not halt when started on the blank tape:

“ $\Leftarrow$ ”: Suppose that the Turing machine  $\mathcal{M}$  does not halt on the blank tape. Then a valid tiling exists where one horizontal row is formed with the initialization tiles, all tiles below this row are blank, and the rows above the initialization row contain consecutive configurations of the Turing machine.

“ $\Rightarrow$ ”: Suppose that a valid tiling containing the middle initialization tile exists. The seed tile forces its row to be formed by the initialization tiles, representing the initial configuration of the Turing machine on the blank tape. The machine tiles force the following horizontal rows above the seed row to contain the consecutive configurations of the Turing machine. There is no merge tile containing a halting state so the Turing machine does not halt – otherwise a valid tiling could not be formed.

Conclusion: Suppose we had an algorithm that solves TILING PROBLEM WITH A SEED TILE. Then we also have an algorithm (which simply constructs the tile set as above and determines if a tiling with seed tile exists) that solves TURING MACHINE HALTING ON BLANK TAPE. This contradicts the fact that this problem is known to be undecidable.  $\square$

In the following tiling problem variant we are given a Wang tile set  $T$  and specify one tile  $B \in T$  as the *blank tile*. The blank tile has all four sides colored by the same color. A *finite tiling* is a tiling where only a finite number of tiles are non-blank. A finite tiling where all tiles are blank is called *trivial*.

### FINITE TILING PROBLEM

**Instance:** A finite set  $T$  of Wang tiles and a blank tile  $B \in T$

**Problem:** Does there exist a valid finite tiling that is not trivial?

**Theorem 3** *The FINITE TILING PROBLEM is undecidable. It is semi-decidable while its complement is not semi-decidable.*

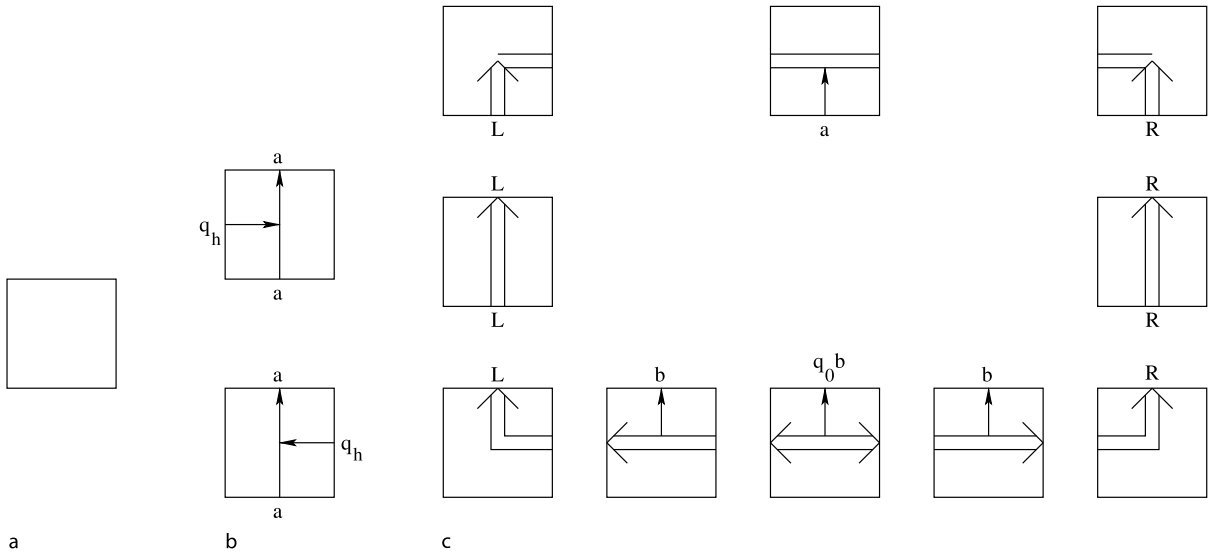
*Proof* For semi-decidability notice that we can try all valid tilings of larger and larger squares until we find a tiling of a square where all tiles on the boundary are blank, while some interior tile is different from the blank tile. If such a tiling is found then the semi-algorithm halts, indicating that a valid, finite, non-trivial tiling exists.

To prove the undecidability we reduce the problem TURING MACHINE HALTING ON BLANK TAPE. For any given Turing machine  $\mathcal{M}$  we construct the machine tiles of Fig. 1 as well as the blank tile, boundary tiles and the halting tiles shown in Fig. 3.

The halting tiles of Fig. 3b are constructed for all tape letters  $a \in \Gamma$  and the halting state  $q_h$ . The purpose of the halting tiles is to erase the Turing machine from the configuration once it halts. The lower border tiles of Fig. 3c initialize the configuration to consist of the blank tape symbol  $b$  and the initial state  $q_0$ . The top border tiles are made for every tape symbol  $a \in \Gamma$ . They allow the absorption of the configuration as long as the Turing machine has been erased. The border tiles on the sides are labeled with symbols  $L$  and  $R$  to identify the left and the right border of the computation area.

Let us prove that the tile set admits a valid, finite, non-trivial tiling if and only if the Turing machine halts on the empty tape.

“ $\Leftarrow$ ”: Suppose that the Turing machine halts on the blank tape. Then a tiling exists where the boundary tiles isolate a finite portion of the plane (a “board”) for the simulation of the Turing machine, the bottom tiles of the board initialize the Turing machine on the blank tape, and



**Tiling Problem and Undecidability in Cellular Automata, Figure 3**  
**a** the blank tile **B**, **b** halting tiles, and **c** border tiles

inside the board the Turing machine is simulated until it halts. After halting only tape tiles are used until they are absorbed by the topmost row of the board. If the board is made sufficiently large the entire computation fits inside the board, so the tiling is valid. All tiles outside the board are blank so the tiling is finite.

“ $\implies$ ”: Suppose then that a finite, non-trivial tiling exists. The only non-blank tiles with a blank bottom edge are the lower border tiles of Fig. 3c, so the tiling must contain a lower border tile. Horizontal neighbors of lower border tiles are lower border tiles, so we see that the only way to have a finite tiling is to have a contiguous lower border that ends at both sides in a corner tile where the border turns upwards. The vertical borders must again – due to the finiteness of the tiling – end at corners where the top border starts. All in all we see that the boundary tiles are forced to form a rectangular board.

The lower boundary of the board initializes the Turing machine configuration on the blank tape, and the rows above it are forced by the machine tiles to simulate consecutive configurations of the Turing machine. Because the Turing machine state symbol is not allowed to touch the side or the upper boundary of the board, the Turing machine must be erased by a halting tile, i. e. the Turing machine must halt.  $\square$

The third variation of the tiling problem we consider is the PERIODIC TILING PROBLEM where we ask whether a given set of tiles admits a valid periodic tiling.

#### PERIODIC TILING PROBLEM

**Input:** Tile set  $\mathcal{T}$ .

**Question:** Does  $\mathcal{T}$  admit a valid periodic tiling?

**Theorem 4** *The PERIODIC TILING PROBLEM is undecidable for Wang tile sets. It is semi-decidable while its complement is not semi-decidable.*

For a proof, see [7].

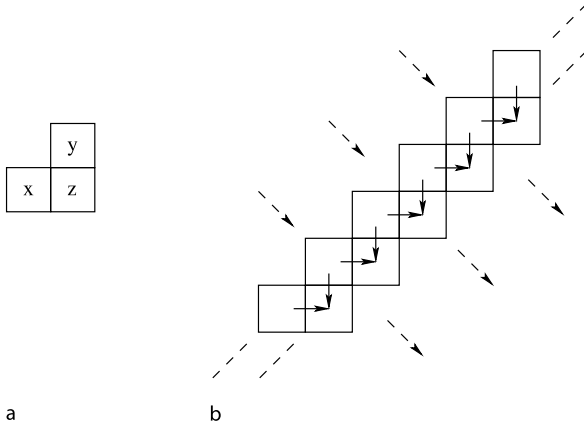
#### Deterministic Tiles

The tiling problem of one-dimensional tiles is decidable. However, tiles can provide undecidability results for one-dimensional CA when we use the trick that we view space-time diagrams as two-dimensional tilings. But not every tiling can be a space-time diagram of a CA: the tiling must be locally deterministic in the direction that corresponds to time. This leads to the consideration of determinism in Wang tiles.

Consider a set  $T$  of Wang tiles, i. e. squares with colored edges. We say that  $T$  is *NW-deterministic* if for all  $a, b \in T$ ,  $a \neq b$ , either the upper (=northern) edges of  $a$  and  $b$  or the left (=western) edges of  $a$  and  $b$  have different colors. See Fig. 4a for an illustration.

Consider now a valid tiling of the plane by NW-deterministic tiles. Each tile is uniquely determined by its left and upper neighbor. Then tiles on each diagonal in the NE-SW direction locally determine the tiles on the next diagonal below it. If we interpret these diagonals as con-





**Tiling Problem and Undecidability in Cellular Automata, Figure 4** NW-deterministic sets of Wang tiles: **a** there is at most one matching tile  $z$  for any  $x$  and  $y$ , **b** diagonals of NW-deterministic tilings interpreted as configurations of one-dimensional CA

figurations of a CA then there is a local rule such that valid tilings are space-time diagrams of the CA, see Fig. 4b.

We define analogously NE-, SW- and SE-deterministic tile sets. Finally, we call a tile set 4-way deterministic if it is deterministic in all four directions simultaneously.

The tiling problem is undecidable among NW-deterministic tile sets [10], even among 4-way deterministic tile sets [18]:

**Theorem 5** *The decision problem TILING PROBLEM is undecidable among 4-way deterministic sets of Wang tiles.*

As discussed at the end of Sect. “The Tiling Problem”, the theorem also means that 4-way deterministic aperiodic tile sets exist. In fact, the proof of Theorem 5 in [18] uses one such aperiodic set that was reported in [16].

### Plane Filling Directed Tiles

A  $d$ -dimensional *directed tile* is a tile that is associated a *follower vector*  $\vec{f} \in \mathbb{Z}^d$ . Let  $\mathcal{T} = (d, T, N, R)$  be a tile set, and let  $F: T \rightarrow \mathbb{Z}^d$  be a function that assigns tiles their follower vectors. We call  $\mathcal{D} = (d, T, N, R, F)$  a set of directed tiles. Let  $t \in T^{\mathbb{Z}^d}$  be an assignment of tiles to cells. For every  $\vec{p} \in \mathbb{Z}^d$  we call  $\vec{p} + F(t(\vec{p}))$  the follower of  $\vec{p}$  in  $t$ . In other words, the follower of  $\vec{p}$  is the cell whose position relative to  $\vec{p}$  is given by the follower vector of the tile in cell  $\vec{p}$ .

Sequence  $\vec{p}_1, \vec{p}_2, \dots, \vec{p}_k$  where all  $\vec{p}_i \in \mathbb{Z}^d$  is a (finite) *path* in  $t$  if

$$\vec{p}_{i+1} = \vec{p}_i + F(t(\vec{p}_i))$$

for all  $1 \leq i < k$ . In other words, a path is a sequence of cells such that the next cell is always the follower of the

previous cell. One-way infinite and two-way infinite paths are defined analogously.

In the following we only discuss the two-dimensional case ( $d = 2$ ) and the follower of each tile is one of the four adjacent positions:

$$F(a) \in \{(\pm 1, 0), (0, \pm 1)\} \quad \text{for all } a \in T.$$

In this case the follower is indicated in drawings as a horizontal or vertical arrow over the tile.

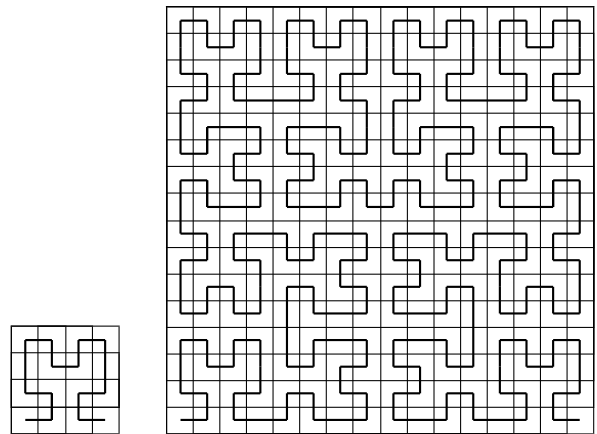
A set of two-dimensional directed tiles is said to have the *plane-filling property* if it satisfies the following two conditions:

- There exists  $t \in T^{\mathbb{Z}^2}$  and a one-way infinite path  $\vec{p}_1, \vec{p}_2, \vec{p}_3, \dots$  such that the tiling in  $t$  is valid at  $\vec{p}_i$  for all  $i = 1, 2, 3, \dots$
- For all  $t$  and  $\vec{p}_1, \vec{p}_2, \vec{p}_3, \dots$  as in (a), there are arbitrarily large  $n \times n$  squares of cells such that all cells of the squares are on the path.

Intuitively the plane-filling property means that the simple device that moves over tiling  $t$ , repeatedly verifies the correctness of the tiling at its present location and moves on to the follower, necessarily eventually either finds a tiling error or covers arbitrarily large squares. Note that the plane-filling property does not assume that the tiling  $t$  is correct everywhere: as long as it is correct along a path the path must snake through larger and larger squares.

Note that conditions (a) and (b) imply that the tile set is aperiodic. There exist tile sets that satisfy the plane filling-filling property, as proved in [11]:

**Theorem 6** *There exists a set of directed Wang tiles that has the plane-filling property.*



**Tiling Problem and Undecidability in Cellular Automata, Figure 5** Fractions of the plane-filling Hilbert curve through  $4 \times 4$  and  $16 \times 16$  squares

The proof of Theorem 6 in [11] constructs a set of Wang tiles such that the path that does not find any tiling errors is forced to follow the well known Hilbert-curve shown in Fig. 5.

### Undecidability in Cellular Automata

Let us begin with one-step properties of two-dimensional CA.

**Theorem 7** INJECTIVITY is undecidable among two-dimensional CA. It is semi-decidable in any dimension.

*Proof* The semi-decidability follows from the fact that injective CA have an inverse CA. One can effectively enumerate all CA and check them one-by-one until (if ever) the inverse CA is found.

Let us next prove INJECTIVITY undecidable by reducing the TILING PROBLEM into it. In the reduction a set  $D$  of directed tiles that has the plane filling property is used. The existence of such  $D$  was stated in Theorem 6.

Let  $T$  be a given set of Wang tiles that is an instance of the TILING PROBLEM. One can effectively construct a two-dimensional CA whose state set is

$$S = T \times D \times \{0, 1\}$$

and the local rule updates the bit component of a cell as follows:

- If either the  $T$ - or the  $D$ -components contain a tiling error at the cell then the state of the cell is not changed, but
- if the tilings according to both  $T$ - and  $D$ -components are valid at the cell then the bit of the follower cell (according to the direction in the  $D$ -component) is added to the present bit value modulo 2.

The tile components are not changed. Let us prove that this CA  $G$  is not injective if and only if  $T$  admits a valid tiling.

“ $\Leftarrow$ ”: Suppose a valid tiling exists. Construct two configurations  $c_0$  and  $c_1$  where the  $T$ - and  $D$ -components form the same valid tilings  $t \in T^{\mathbb{Z}^2}$  and  $d \in D^{\mathbb{Z}^2}$ , respectively. In  $c_0$  all bits are 0 while in  $c_1$  they are all 1. Since the tilings are everywhere valid, every cell performs modulo 2 addition of two bits, which means that every bit becomes 0. Hence  $G(c_0) = G(c_1) = c_0$ , and  $G$  is not injective.

“ $\Rightarrow$ ”: Suppose then that  $G$  is not injective. There are two different configurations  $c_0$  and  $c_1$  such that  $G(c_0) = G(c_1)$ . Tile components are not modified by the CA so they are identical in  $c_0$  and  $c_1$ . There is a cell  $\vec{p}_1$  such that  $c_0$  and  $c_1$

have different bits at cell  $\vec{p}_1$ . Since these bits become identical in the next configuration, the  $D$ -tiling must be correct at  $\vec{p}_1$  and  $c_0$  and  $c_1$  must have different bits in the follower position  $\vec{p}_2$ . We repeat the reasoning and obtain an infinite sequence of positions  $\vec{p}_1, \vec{p}_2, \vec{p}_3, \dots$  such that each  $\vec{p}_{i+1}$  is the follower of  $\vec{p}_i$ , and the  $D$  tiling is correct at each  $\vec{p}_i$ . It follows from the plane filling property of  $D$  that path  $\vec{p}_1, \vec{p}_2, \vec{p}_3, \dots$  covers arbitrarily large squares. Also the tiling according to the  $T$ -components must be valid at each cell of the path. Hence tile set  $T$  admits valid tilings of arbitrarily large squares, and therefore it admits a valid tiling of the entire plane.  $\square$

Analogously we can prove the undecidability of SURJECTIVITY. It is convenient to use the well known Garden-of-Eden theorem of Moore and Myhill to convert the surjectivity property into injectivity on finite configurations:

**Theorem 8 (Garden-of-Eden theorem)** A cellular automaton is non-surjective if and only if there are two distinct configurations that differ in a finite number of cells and that have the same successor [19,20].

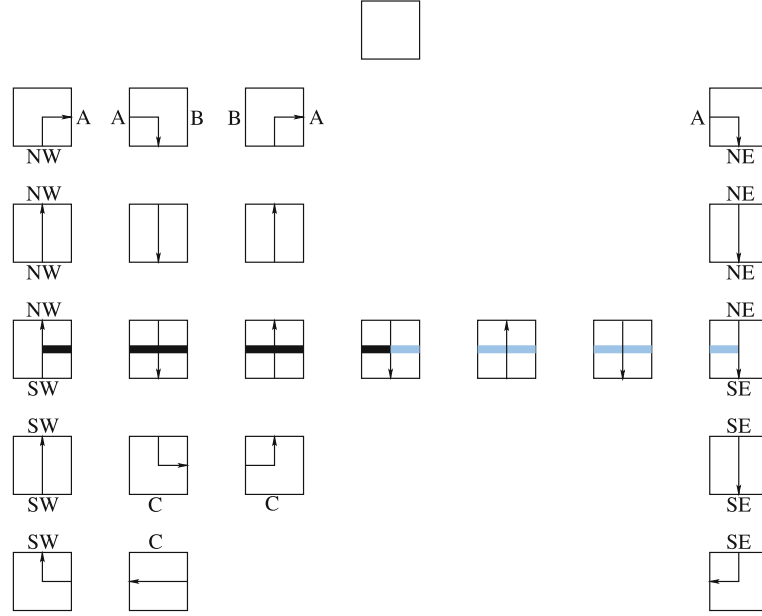
**Theorem 9** SURJECTIVITY is undecidable among two-dimensional CA. Its complement is semi-decidable in any dimension.

*Proof* A semi-algorithm for non-surjectivity enumerates all finite patterns one-by-one until a pattern is found that can not appear in  $G(c)$  for any configuration  $c$ .

To prove undecidability we reduce the FINITE TILING PROBLEM, using the set  $D$  of 23 directed tiles shown in Fig. 6. These directed tiles are used in an analogous way as in the proof of Theorem 7.

The topmost tile in Fig. 6 is called blank. All other tiles have a unique incoming and outgoing arrow. In valid tilings arrows and labels must match. The non-blank tiles are considered directed: the follower of a tile is the neighbor directed to by the outgoing arrow on the tile. Since each non-blank tile has exactly one incoming arrow, it is clear that if the tiling is valid at a tile then the tile is the follower of exactly one of its four neighbors.

The tile at the center in Fig. 6 where the dark and light thick horizontal lines meet is called the *cross*. It has a special role in the forthcoming proof. A *rectangular loop* is a valid tiling of a rectangle using tiles in  $D$  where the follower path forms a loop that visits every tile of the rectangle, and the outside border of the rectangle is colored blank. See Fig. 7 for an example of a rectangular loop through a rectangle of size  $12 \times 7$ . (The edge labels are not shown for the sake of clarity of the figure.) It is easy to see that a rectangular loop of size  $2n \times m$  exist for all  $n \geq 2$  and  $m \geq 3$ . Any tile in an even column in the interior of



**Tiling Problem and Undecidability in Cellular Automata, Figure 6**  
**Tiles used in the proof of the undecidability of SURJECTIVITY**

the rectangle can be made to contain the unique cross of the rectangular loop.

It is easy to see that the tile set  $D$  has the following property:

*Finite plane-filling property:* Let  $t \in D^{\mathbb{Z}^2}$  be a tiling, and  $\vec{p}_1, \vec{p}_2, \vec{p}_3, \dots$  a path in  $t$  such that the tiling  $t$  is valid at  $\vec{p}_i$  for all  $i = 1, 2, 3, \dots$ . If the path covers only a finite number of different cells then the cells on the path form a rectangular loop.

Let  $b$  and  $c$  be the blank and the cross of set  $D$ . For any given tile set  $T$  with blank tile  $B$  we construct the following two-dimensional cellular automaton. The state set  $S$  contains triplets

$$(d, t, x) \in D \times T \times \{0, 1\}$$

under the following constraints:

- If  $d = c$  then  $t \neq B$ , and
- if  $d = b$  or  $d$  is any tile containing label SW, SE, NW, NE, A, B or C, then  $t = B$ .

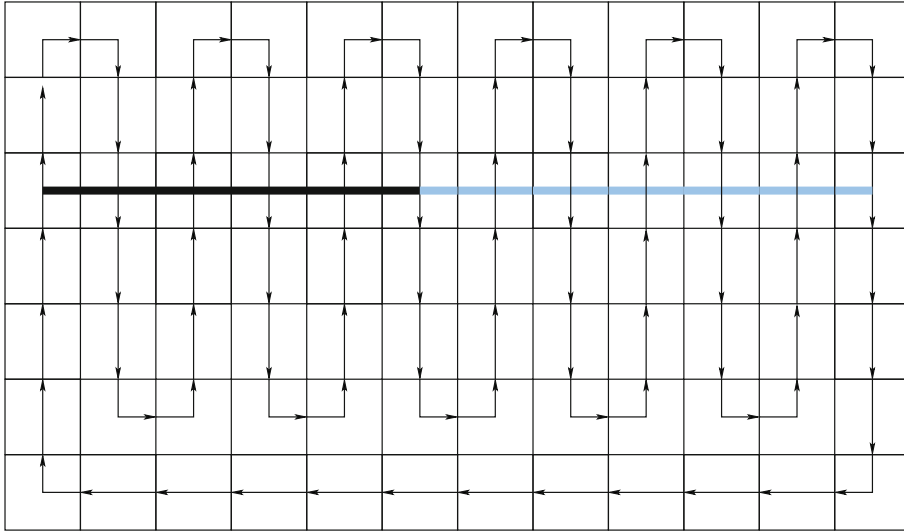
In other words, the cross must be associated with a non-blank tile in  $T$  while the blank of  $D$  as well as all the tiles on the boundary of a rectangular loop are forced to be associated with the blank tile of  $T$ . The triplet  $(b, B, 0)$  where both tile components are blank and the bit is 0 is the quiescent state of the CA. The local rule is as follows: Let  $(d, t, x)$  be the current state of a cell.

- If  $d = b$  then the state is not changed.
- If  $d \neq b$  then the cell verifies the validity of the tilings according to both  $D$  and  $T$  at the cell. If either tile component has a tiling error then the state is not changed. If both tilings are valid then the cell modifies its bit component by adding the bit of its follower modulo 2.

Let us prove that this CA is not surjective if and only if  $T$  admits a valid, finite, non-trivial tiling.

“ $\Leftarrow$ ”: Suppose a valid, finite, non-trivial tiling  $t \in T^{\mathbb{Z}^2}$  exists. Consider a configuration of the CA whose  $T$ -components form the valid tiling  $t$  and the  $D$ -components form a rectangular loop whose interior covers all non-blank elements of  $t$ . Tiles outside the rectangle are all blank and have bit 0. The cross can be positioned so that it is in the same cell as some non-blank tile in  $t$ . In such a configuration both  $T$  and  $D$  tilings are everywhere valid. The CA updates the bits of all tiles in the rectangular loop by performing modulo 2 addition with their followers, while the bits outside the rectangle remain 0. We get two different configurations that have the same image: In  $c_0$  all bits in the rectangle are equal to 0 while in  $c_1$  they are all equal to 1. The local rule updates the bits so that  $G(c_0) = G(c_1) = c_0$ . Configurations  $c_0$  and  $c_1$  only differ in a finite number of cells, so it follows from the Garden-of-Eden theorem that  $G$  is not surjective.

“ $\Rightarrow$ ”: Suppose then that the CA is not surjective. According to the Garden-of-Eden theorem there are two finitely



**Tiling Problem and Undecidability in Cellular Automata, Figure 7**  
A rectangular loop of size  $12 \times 7$

different configurations  $c_0$  and  $c_1$  such that  $G(c_0) = G(c_1)$ . Since only bit components of states are changed, the tilings in  $c_0$  and  $c_1$  according to  $D$ - and  $T$ -components of the states are identical. There is a cell  $\tilde{p}_1$  such that  $c_0$  and  $c_1$  have different bits at cell  $\tilde{p}_1$ . Since these bits become identical in the next configuration, the  $D$ -tiling must be correct at  $\tilde{p}_1$  and  $c_0$  and  $c_1$  must have different bits in the follower position  $\tilde{p}_2$ . We repeat the reasoning and obtain an infinite sequence of positions  $\tilde{p}_1, \tilde{p}_2, \tilde{p}_3, \dots$  such that each  $\tilde{p}_{i+1}$  is the follower of  $\tilde{p}_i$ , and the  $D$  tiling is correct at each  $\tilde{p}_i$ . Moreover,  $c_0$  and  $c_1$  have different bits in each position  $\tilde{p}_i$ . Because configurations  $c_0$  and  $c_1$  only differ in a finite number of cells, we see that the path can only contain a finite number of distinct cells. It follows then from the finite plane-filling property of  $D$  that the path must form a valid rectangular loop.

Also the tiling according to the  $T$ -components must be valid at each cell of the path. Because of the constraints on the allowed triplets, the  $T$ -components on the boundary of the rectangle are the blank  $B$ , while the cross in the interior contains a non-blank element of  $T$ . Hence there is a valid tiling of a rectangle according to  $T$  that contains a non-blank tile and has a blank boundary. This means that a finite, valid and non-trivial tiling is possible.  $\square$

### Undecidable Properties of One-Dimensional CA

Using deterministic Wang tiles and interpreting space-time diagrams as tilings one obtains undecidability results for long-term behavior of one-dimensional CA.

**Theorem 10** *NILPOTENCY is undecidable among one-dimensional CA. It is undecidable even among one-dimensional CA that have a spreading state  $q$ , i. e. a state that spreads to all neighbors. NILPOTENCY is semi-decidable in any dimension.*

*Proof* For semi-decidability notice that, for  $n = 1, 2, 3, \dots$ , we can effectively construct a cellular automaton whose global function is  $G^n$  and check whether the local rule of the CA maps everything into the same state. If that happens for some  $n$  then we halt and report that the CA is nilpotent.

To prove undecidability we reduce the TILING PROBLEM of NW-deterministic Wang tiles. Let  $T$  be a given NW-deterministic tile set. One can effectively construct a one-dimensional CA whose state set is  $S = T \cup \{q\}$  and the local rule turns a cell into the quiescent state  $q$  except in the case that the cell and its right neighbor are in states  $x, y \in T$ , respectively, and tile  $z \in T$  exists so that tiles  $xy, z$  match as in Fig. 4a. In this case  $z$  is the new state of the cell. Note that state  $q$  is a spreading state.

Let us prove that the CA is not nilpotent if and only if  $T$  admits a valid tiling.

“ $\Leftarrow$ ”: Suppose a valid tiling exists. If  $c \in T^{\mathbb{Z}}$  is a diagonal of this tiling then the configurations  $G^n(c)$  in its orbit are subsequent diagonals of the same tiling, for all  $n = 1, 2, \dots$ . This means that  $c$  never becomes quiescent, and the CA is not nilpotent.

“ $\Rightarrow$ ”: Suppose no valid tiling exists. Then there is number  $n$  such that no valid tiling of an  $n \times n$  square exists.

This means that for every initial configuration  $c \in S^{\mathbb{Z}}$  the configuration  $G^{2^n}(c)$  is quiescent: If it is not quiescent then a valid tiling of an  $n \times n$  square can be read from the space time diagram of configurations  $c, G(c), \dots, G^{2^n}(c)$ . We conclude that the CA is nilpotent.  $\square$

Undecidability of NILPOTENCY has some interesting corollaries. First, it implies that the topological entropy of a one-dimensional CA cannot be calculated, not even approximated [8].

**Theorem 11** TOPOLOGICAL ENTROPY is undecidable.

*Proof* Let us reduce NILPOTENCY. Let  $c > 0$  be any constant, and let  $n > 2^c$  be an integer. For any given one-dimensional CA  $G$  with state set  $S$  and a spreading state  $q \in S$  construct a new CA whose state set is  $S \times \{1, 2, \dots, n\}$ , and the local rule applies  $G$  in the first components of the states and shifts the second components one cell to the left. In addition, state  $(q, i)$  is turned into state  $(q, 1)$ .

If  $G$  is nilpotent then also the new CA is nilpotent and its topological entropy is 0. If  $G$  is not nilpotent then there is a configuration  $c \in S^{\mathbb{Z}}$  such that no cell ever turns into the spreading state  $q$ . But then the second components form a left shift over the alphabet  $\{1, 2, \dots, n\}$  so the topological entropy is at least  $\log_2 n > c$ .  $\square$

It also follows that is undecidable to determine if a given one-dimensional CA is ultimately periodic [5]:

**Theorem 12** EQUICONTINUITY is undecidable among one-dimensional CA.

*Proof* Among one-dimensional CA with a spreading state EQUICONTINUITY is equivalent to NILPOTENCY.  $\square$

**Theorem 13** SENSITIVITY TO INITIAL CONDITIONS is undecidable among one-dimensional CA.

*Proof* Originally the result was proved in [5] using an elaborate reduction of the Turing machine halting problem. However, undecidability of NILPOTENCY provides the result directly, as pointed out in [14]. Namely, a one-dimensional cellular automaton whose neighborhood vector contains only strictly positive numbers is either nilpotent or sensitive. Adding a constant to all elements of the neighborhood vector does not affect the nilpotency status of a CA. So for any given one-dimensional CA we proceed as follows: add a positive constant to the elements of the neighborhood vector so that they all become positive. The new CA is sensitive if and only if the original CA was not nilpotent. The result then follows from Theorem 10.  $\square$

As a final application of undecidability of NILPOTENCY consider other questions concerning the limit set (=maximal attractor) of one-dimensional CA. One can show

that NILPOTENCY can be reduced to any non-trivial question [12]. More precisely, let PROB be a decision problem that takes arbitrary one-dimensional CA as input. Suppose that PROB always has the same answer for any two CA that have the same limit set. Then we say that PROB is a decision problem concerning the limit sets of CA. We call PROB non-trivial if there exist both positive and negative instances.

**Theorem 14** Let PROB be any non-trivial decision problem concerning the limit sets of CA. Then PROB is undecidable [12].

## Other Undecidability Results

In the previous sections we only considered decision problems that have been proved undecidable using reductions from the tiling problem or its variant. There are many other decision problems that have been proved undecidable using other techniques. Below are a few, with literature references.

We call a CA  $G$  *periodic* if there is number  $n$  such  $G^n$  is the identity function. This is equivalent to saying that every configuration is periodic, that is, every configuration returns back to itself. Clearly a periodic CA is necessarily injective. In fact, periodic CA are exactly those CA that are injective and equicontinuous.

### PERIODICITY

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  periodic?

**Theorem 15** PERIODICITY is undecidable among one-dimensional CA [15].

A CA is called sensitive to initial conditions if there exists a finite set  $B \subseteq \mathbb{Z}^d$  of cells such that for every configuration  $c$  and every finite set  $A \subseteq \mathbb{Z}$  of cells there exists a configuration  $e$  and time  $t \geq 0$  such that  $e(\vec{x}) = d(\vec{x})$  for all  $\vec{x} \in A$  but  $G^t(e)(\vec{x}) \neq G^t(c)(\vec{x})$  for some  $\vec{x} \in B$ .

### SENSITIVITY TO INITIAL CONDITIONS

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  sensitive to initial conditions?

**Theorem 16** SENSITIVITY TO INITIAL CONDITIONS is undecidable among one-dimensional CA [5].

The following problems deal with dynamics on finite configuration. We hence suppose that the given CA has a quiescent state, i.e. a state  $q$  such that  $f(q, q, \dots, q) = q$  where  $f$  is the local update rule of the CA. A configuration  $c \in S^{\mathbb{Z}^d}$  is called finite (w.r.t.  $q$ ) if all but a finite number



of cells are in state  $q$ . Questions similar to NILPOTENCY and EQUICONTINUITY can be asked in the space of finite configurations:

#### NILPOTENCY ON FINITE CONFIGURATIONS

**Input:** Cellular Automaton  $A$  with a quiescent state

**Question:** Does every finite configuration evolve into the quiescent configuration?

#### EVENTUAL PERIODICITY ON FINITE CONFIGURATIONS

**Input:** Cellular Automaton  $A$  with a quiescent state

**Question:** Does every finite configuration evolve into a temporally periodic configuration?

**Theorem 17** NILPOTENCY ON FINITE CONFIGURATIONS and EVENTUAL PERIODICITY ON FINITE CONFIGURATIONS are undecidable for one-dimensional CA [4,23].

#### Future Directions

Several interesting and challenging open questions remain. In particular, the decidability statuses of the following decision problems concerning basic dynamical properties are unknown.

We call a CA  $G$  *periodic* if there is number  $n$  such  $G^n$  is the identity function. This is equivalent to saying that every configuration is periodic, that is, every configuration returns back to itself. Clearly a periodic CA is necessarily injective. In fact, periodic CA are exactly those CA that are injective and equicontinuous.

#### PERIODICITY

**Input:** Cellular Automaton  $A$

**Question:** Is  $A$  periodic?

The question is undecidable for two-dimensional CA (the construction in the proof of Theorem 7 shows it) but the decidability status is unknown for one-dimensional CA.

We call a one-dimensional CA  $G$  *positively expansive* if there exists a finite set  $A \subseteq \mathbb{Z}$  of cells such that for any two distinct configurations  $c$  and  $e$  there exists  $t \geq 0$  such that  $G^t(c)$  and  $G^t(e)$  differ in some cell in  $A$ . We call an injective CA  $G$  *expansive* for some finite  $A$  holds the following: for any two distinct configurations  $c$  and  $e$  there exists  $t \in \mathbb{Z}$  such that  $G^t(c)$  and  $G^t(e)$  differ in some cell in  $A$ . It is known that a two- and higher dimensional CA can not be expansive or positively expansive [6,22], so the following decision problems are only asked for one-dimensional CA:

#### POSITIVE EXPANSIVITY

**Input:** One-dimensional cellular Automaton  $A$

**Question:** Is  $A$  positively expansive?

#### EXPANSIVITY

**Input:** One-dimensional cellular Automaton  $A$

**Question:** Is  $A$  expansive?

The decidability status of both POSITIVE EXPANSIVITY and EXPANSIVITY is unknown.

#### Acknowledgments

Research supported by the Academy of Finland grant 211967.

#### Bibliography

##### Primary Literature

1. Amoroso S, Patt Y (1972) Decision Procedures for Surjectivity and Injectivity of Parallel Maps for Tessellation Structures. *J Comput Syst Sci* 6:448–464
2. Berger R (1966) The Undecidability of the Domino Problem. *Mem Am Math Soc* 66:1–72
3. Culik K II (1996) An aperiodic set of 13 Wang tiles. *Discret Math* 160: 245–251
4. Culik K II, Yu S (1988) Undecidability of CA Classification Schemes. *Complex Syst* 2:177–190
5. Durand B, Formenti E, Varouchas G (2003) On undecidability of equicontinuity classification for cellular automata. In: *Proceedings of Discrete Models for Complex Systems*, Lyon, France, 16–19 June 2003, pp. 117–128
6. Finelli M, Manzini G, Margara L (1998) Lyapunov Exponents versus Expansivity and Sensitivity in Cellular Automata. *J Complex* 14:210–233
7. Gurevich YS, Koryakov IO (1972) Remarks on Berger's paper on the domino problem. *J Sib Math J* 13:319–321
8. Hurd LP, Kari J, Culik K (1992) The topological Entropy of Cellular Automata is Uncomputable. *Ergodic Theor Dyn Syst* 12:255–265
9. Kari J (1990) Reversibility of 2D cellular automata is undecidable. *Physica D* 45:379–385
10. Kari J (1992) The nilpotency problem of one-dimensional cellular automata. *SIAM J Comput* 21:571–586
11. Kari J (1994) Reversibility and surjectivity problems of cellular automata. *J Comput Syst Sci* 48:149–182
12. Kari J (1994) Rice's theorem for the Limit Sets of Cellular Automata. *Theoret Comput Sci* 127:229–254
13. Kari J (1996) A small aperiodic set of Wang tiles. *Discret Math* 160:259–264
14. Kari J (2008) Undecidable properties on the dynamics of reversible one-dimensional cellular automata. In: *Proceedings of Journées Automates Cellulaires*, Uzès, France, 21–25 April 2008
15. Kari J, Ollinger N Periodicity and immortality in reversible computing. (in press)
16. Kari J, Papasoglu P (1999) Deterministic aperiodic tile sets. *J Geom Funct Anal* 9:353–369
17. Kurka P (1997) Languages, equicontinuity and attractors in cellular automata. *Ergod Theory Dyn Syst* 17:417–433
18. Lukkarila V (2007) The 4-way deterministic tiling problem is undecidable. (in press)

19. Moore EF (1962) Machine models of self-reproduction. In: *Proceedings of the Symposia in Applied Mathematics* 14:17–33
20. Myhill J (1963) The Converse to Moore's Garden-of-Eden Theorem. In: *Proceedings of the American Mathematical Society* 14:685–686
21. Robinson RM (1971) Undecidability and Nonperiodicity for Tilings of the plane. *Invent Math* 12:177–209
22. Shereshevsky MA (1993) Expansiveness, entropy and polynomial growth for groups acting on subshifts by automorphisms. *Indag Math* 4:203–210
23. Sutner K (1989) A note on the Culik-Yu classes. *Complex Syst* 3:107–115
24. Wang H (1961) Proving theorems by pattern recognition – II. *Bell Syst Techn J* 40:1–42

### Books and Reviews

- Codd EF (1968) *Cellular Automata*. Academic Press, New York
- Garzon M (1995) *Models of massive parallelism: analysis of cellular automata and neural networks*. Springer, New York
- Hedlund G (1969) Endomorphisms and automorphisms of shift dynamical systems. *Math Syst Theory* 3:320–375
- Kari J (2005) *Theory of cellular automata: a survey*. Theoret Comput Sci 334:3–33
- Toffoli T, Margolus N (1987) *Cellular Automata Machines*. MIT Press, Cambridge
- Wolfram S (ed) (1986) *Theory and Applications of Cellular Automata*. World Scientific Press, Singapore
- Wolfram S (2002) *A New Kind of Science*. Wolfram Media, Canada

## Tomography, Seismic

JOSE PUJOL

Dept. of Earth Sciences, The University of Memphis,  
Memphis, USA

### Article Outline

[Definition of the Subject](#)

[Introduction](#)

[Fundamentals of X-ray Computerized Tomography](#)

[Arrival-Time Seismic Tomography](#)

[Solution of Ill-Posed Linear Problems](#)

[Examples](#)

[Future Directions](#)

[Bibliography](#)

### Definition of the Subject

Seismic tomography refers to a number of techniques designed to investigate the interior of the earth using arrival times and/or waveforms from natural and artificial sources. The most common product of a tomographic study is a velocity model, although other parameters, such

as attenuation, are also studied. The importance of seismic tomography stems from two facts. One, it generally has higher resolution than that provided by other geophysical methods. Two, it provides information that (a) can help solve fundamental problems concerning the internal structure of the earth at a global scale, and (b) has been used in tectonic and seismic hazards studies at a local scale. Seismic tomography has also been applied to data collected in boreholes, but because of the high expenses associated with drilling, borehole tomography is relatively little used.

### Introduction

In the most general terms, seismic tomography problems are inverse problems, and before the word “tomography” entered the seismological literature the term inversion was used. This change occurred as a consequence of the revolution in medical imaging caused by computerized (or computed) X-ray tomography (CT) (also known as computer assisted or axial, tomography, CAT), introduced in the 1970s. The importance of CT on seismic tomography was that it provided efficient numerical techniques for the solution of systems of equations with extremely large number of unknowns ( $10^5$  or more), which allowed a great expansion of the seismological inverse problems that could be tackled. In addition, according to [1], changing the name inversion to tomography enhanced the “believability” of the inverse method.

Seismic tomography covers a wide range of scales, from the very small (e.g., borehole tomography, involving distances of a few kilometers at most) to the very large (i.e., whole earth tomography). Initially, seismic tomography involved body-wave arrival times, but later it was extended to include waveform and surface wave information. Arrival-time tomography is simpler and for this reason it is the most popular. Waveform tomography requires software for the computation of synthetic seismograms for comparison with the observed ones, which makes the whole inversion process more difficult both theoretically and practically. Surface wave tomography involves the determination of phase velocities, which is more complicated than picking arrival times. The three approaches, however, have one thing in common, namely, they all end up requiring the solution of a linear system of equations. This task might sound simple, but in practice it is not because the system generally does not have a unique solution, which means that solving it requires making decisions (either implicitly or explicitly) regarding the nature of the solution. In addition, given the complexity of wave propagation in heterogeneous media (such as the earth), the in-

verse problems solved by seismologists involve highly simplified models of the earth. These simplifications enter into the linear system to be solved, which adds another layer of uncertainty to the solution. This is in sharp contrast with X-ray tomography, which is comparatively unaffected by this kind of uncertainties. For this reason it might be argued that the name tomography does not do justice to the kind of problems that seismologists solve, and this fact should be born in mind by readers from other disciplines. Other differences between X-ray and seismic tomography, of more practical nature, are given below.

The goal of most of the seismic tomography work is to derive 3-D velocity models of portions of the earth. Currently, most of the research concentrates on two scales, global and local. At the global scale, the tomographic models generally have higher resolution than that provided by other geophysical methods, and for this reason it has the potential to provide constraints on the fate of the subducted slabs, on models of mantle convection, on petrological and geochemical models, on studies of the geomagnetic and gravity fields, on mineral physics, and on the core-mantle boundary, among others (see, e. g., [41,96,129,161]). Local tomography usually involves the simultaneous determination of a 3-D velocity model and the relocation of the seismic events used to determine the model. Traditionally, local velocity models have been used in structural and tectonic interpretations, but more recently they have become important in seismic hazard studies, as the ground motion caused by earthquakes are amplified by low-velocity materials, which increases the hazard in the case of large events (see, e. g., [160]). As discussed below, the standard practice of locating earthquakes with layered velocity models may lead to significant location errors when the lateral velocity variations are significant. Therefore, the simultaneous velocity-inversion and event relocation has the potential to produce improved event locations, which is also important in the context of seismic hazards studies.

## Fundamentals of X-ray Computerized Tomography

### Historical Overview

Computerized tomography began in the early 1970s with the introduction of the X-ray scanner developed by G. Hounsfield [65,66], who was able to combine data generated using X-ray techniques with computerized data acquisition and processing techniques. This resulted in the first commercially viable CT machine, with clinical applications presented in [6]. The history of the development of CT is extremely interesting and will be summarized here, but to put it into a broader perspective we will review

briefly the X-rays acquisition and processing techniques in use before CT was introduced.

Röntgen's discovery of X-rays in 1895 revolutionized the practice of medicine by allowing a view of the interior of the human body. However, an X-ray image is the 2-D projection of a 3-D body, which means that its interpretation is subject to ambiguity. This fact was noticed soon after the radiographic technique was introduced, and a number of researchers began to develop techniques to produce 3-D views (see, e. g., [155]). The goal was to generate X-ray images of thin slices of a patient. This was achieved by moving the film and the X-ray source in such a way that the objects in a particular plane were emphasized while others were blurred by the motion. This process is discussed in, e. g., [138]. Work on this goal began in 1914 in several European countries (Germany, France, Italy, Holland), and was motivated, at least in part, by World War I. Different approaches were tried and different names were assigned to some of them, but only one, tomography, remained. This word comes from the Greek word *tomos*, which means cut or slice, and was introduced by the Berliner physician G. Grossmann, whose tomograph was commercially available in 1935 [155,156].

Conventional tomography represented a considerable improvement over the original radiographic techniques, but the advent of computers opened new research avenues based on the digital processing of the X-ray images, or shadowgrams. A shadowgram is a photographic plate developed after it has been illuminated by X-rays that passed through an object and gives a measure of the absorptivity of the rays by the object, which in many materials is roughly proportional to its density [12]. The availability of computers allowed shadowgrams, as well as other images, to be scanned and digitized for further processing using Fourier transform techniques as well as numerical solution of matrix equations. This work will be referred to below, but it is important to realize that they were in place when CT was introduced, and were major contributors to the improvement of the quality of the early CT results.

One of the earliest papers on medical computerized tomography is by A. Cormack [26], who received the Nobel Prize in Physiology or Medicine in 1979 together with G. Hounsfield for their contributions to the development of the technique, which were carried out independently. Cormack was a South African physicist working at the University of Cape Town. In 1955 the physicist at the Cape Town hospital resigned and Cormack replaced him for six months in 1956. During this time he became interested in the determination of the absorption of X or gamma rays passing through an inhomogeneous medium. His motivation was medical, in the context of radiotherapy, which

required a good knowledge of the absorption coefficient of the bones and tissues of a patient. Cormack's approach was to consider a 2-D problem, as a 3-D one can be solved in terms of a succession of 2-D layers. The 2-D problem was well known and can be formulated as follows. A beam of monoenergetic rays of intensity  $I_o$  traverses a finite 2-D domain  $\mathcal{D}$  along a straight line  $L$ , and the intensity of the ray emerging from  $\mathcal{D}$  is  $I$ , given by

$$I = I_o \exp \int_L f(l) dl \quad (1)$$

where  $f$  is the absorption coefficient, which is a function of position within  $\mathcal{D}$ , and  $dl$  denotes a length element along  $L$ . Dividing both sides of Eq. (1) by  $I_o$  and taking the natural logarithm gives

$$g_L \equiv \ln \frac{I}{I_o} = \int_L f(l) dl. \quad (2)$$

The quantity  $g_L$  is known, and the question posed by Cormack was whether  $f$  could be computed using  $g_L$  determined for a number of lines. Cormack solved the problem in terms of a series expansion and demonstrated the feasibility of the method with two simple samples made of aluminum and either wood or Lucite [26,27]. His work, however, went essentially unnoticed until the publication of a paper [28] where the connection between the CT problem and the Radon transform was discussed. That paper in turn, was stimulated by Hounsfield's work, which became known in 1971 [29]. Cormack moved to Tufts University (United States) in 1957.

The actual implementation of the CT technique as a viable commercial enterprise is due to Hounsfield [65], who was an engineer working at the Central Research Laboratories of Electrical and Musical Industries, the English company that eventually became the well known musical records company EMI. As recounted by Hounsfield [67], after finding that a project he had been working on would not be commercially viable, he was given the opportunity to think about other areas of research that he thought might be more fruitful. This rare opportunity was probably the result of the considerable amount of money that the group *The Beatles* had brought to EMI [107]. One of the projects Hounsfield had been working on was connected with automatic pattern recognition, which in 1967 led to the idea of what eventually became the technique of computerized tomography. To materialize this idea Hounsfield built the CT machine and developed the numerical technique to process the data (see Subsect. "Iterative Solutions"). According to some, the second task may have been the more fundamental of the two [107].

We close this summary with two notes of historical interest. First, B. Korenblyum, S. Tetel'baum, and A. Tyutin worked on tomography at the Kiev Polytechnic Institute and published their results in obscure Russian journals in 1957 and 1958. These authors formulated the tomography problem using line integrals, solved it exactly in terms of the inverse Radon transform, and discussed a reconstruction algorithm. However, additional references on their work could not be found [11]. Second, W. Oldendorf, a neurologist at the University of California, Los Angeles, developed a scanning instrument, although he did not solve the problem mathematically [106]. Oldendorf patented his instrument in 1963 but was not able to attract companies interested in manufacturing a commercial version of it. Because of this contribution, Oldendorf was considered for the Nobel Prize together with Cormack and Hounsfield, but eventually was excluded. Possible reasons for the decision of the Nobel committee and the politics involved can be found in [19].

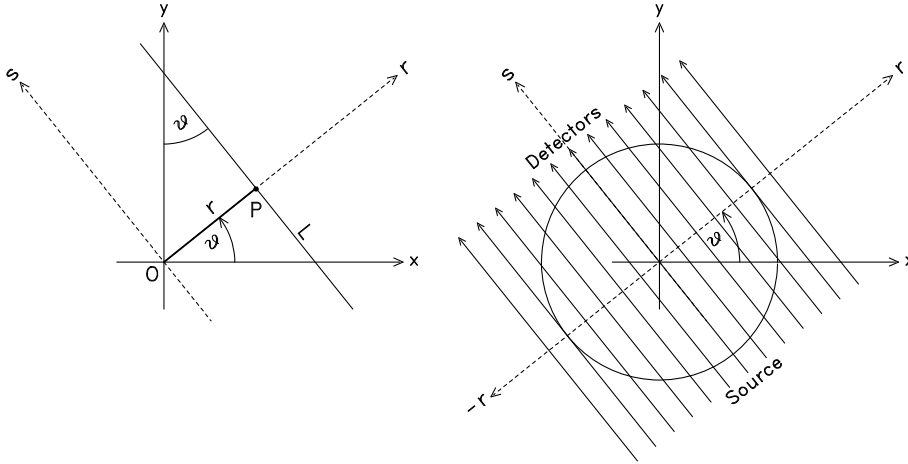
### Solving the CT Problem

The computerized tomography problem is to solve Eq. (2) when  $g_L$  has been determined for a large number of lines having different positions and directions. To make this problem more precise, the definition of projection will be introduced. First consider the following line integral of a function  $f(x, y)$

$$p(r, \vartheta) = \int_L f(x, y) dl \quad (3)$$

where the line  $L$  is defined in terms of its angle  $\vartheta$  with the  $y$  axis and its distance  $r$  to the origin (Fig. 1). Alternatively, it can be said that  $L$  passes through point  $P$  with polar coordinates  $(r \cos \vartheta, r \sin \vartheta)$  and is perpendicular to  $\overline{OP}$ . The collection of line integrals along parallel lines is known as a *parallel projection* of  $f(x, y)$ . Because this is the only type of projection considered here, the qualifier parallel will be dropped. In terms of Eq. (3), the projection of  $f$  is obtained by letting  $r$  be a variable while keeping the value of  $\vartheta$  fixed (i. e.,  $\vartheta$  must be considered a parameter). Note that some authors call a particular line integral a projection along that line (see, e. g., [132]).

Equation (3) was solved using two different approaches that broadly speaking can be referred to as analytical and iterative. The analytical approach followed different paths, but in each case the result was a closed-form solution, which had to be solved numerically. This approach is relevant to seismology for at least two reasons. First, it provides the theoretical underpinnings of the concept of backprojection, which has entered the seismic to-



Tomography, Seismic, Figure 1

**Left:** Geometry for the definition of projection.  $L$  is the projection line and  $r$  is the distance from the line to the origin. The axis  $s$  is parallel to  $L$ . **Right:** Geometry for parallel X-ray tomography. The circle bounds an object to be projected. Each line is equivalent to the line  $L$  on the left. There is a source of X-rays on one end of the lines and detectors on the opposite end

mography literature. Second, it is directly related to the concepts of slant-stack and Radon transform popular in the reflection seismology literature. The iterative approach was to write the CT problem in matrix form and to solve it iteratively. This approach was widely applied to the solution of seismological inverse problems. Because of their close relation to seismology, the two approaches will be discussed here.

**Analytical Solutions** The most popular solutions are based on the use of the Fourier transform, but before considering them we will briefly mention the work of Radon, the Bohemian mathematician that investigated Eq. (3) in a paper published in 1917. It was written in German, and an English translation (by R. Lohner) can be found in [35].

Radon considered the function

$$\begin{aligned} p(r, \theta) &= p(-r, \theta + \pi) \\ &= \int_{-\infty}^{\infty} f(r \cos \vartheta - s \sin \vartheta, r \sin \vartheta + s \cos \vartheta) ds \end{aligned} \quad (4)$$

(see Eq. (11) below; his symbols were somewhat different), and the mean value of  $p(r, \theta)$  for the lines tangent to a circle with center at a point  $Q = (x, y)$  and radius  $a$

$$\bar{p}_Q(a) = \frac{1}{2\pi} \int_0^{2\pi} p(x \cos \vartheta + y \sin \vartheta + a, \theta) d\theta \quad (5)$$

and proved that

$$f(Q) = -\frac{1}{\pi} \int_0^{\infty} \frac{d\bar{p}_Q(a)}{da} da = -\frac{1}{\pi} \int_0^{\infty} \frac{d\bar{p}_Q}{da} \frac{da}{a}, \quad (6)$$

where the last equality is given in [28]. A proof of Radon's result can be found in [62]. Equations (4) (or equivalent expressions) and (6) are known as the Radon and inverse Radon transform, respectively. Although Eq. (6) looks simple, its practical implementation is not [132] and probably for this reason it did not receive as much attention (after Cormack's 1973 paper, [28]) as the other methods developed to solve the problem. As a matter of historical interest we note that Radon's problem arises in a number of scientific fields and had been solved independently more than once in the early part of the twentieth century [30]. Yet, these results were not widely known and had to be derived again.

The analytical approach developed along two different lines and produced results formally different from that of Radon. One was pioneered by Cormack [26,27], but, as noted earlier, it went essentially unnoticed. A second line originated in radio astronomy [17,18], electron microscopy (e.g., [34,36]) and X-ray radiography (e.g., [12,124]), and was based on the use of the Fourier transform. This is the approach that will be taken here and its description will be based mainly on [20,76,88] and [132].

The early development of the CT technique was based on the use of a number of parallel projections determined for different values of  $\vartheta$  in Eq. (3). This is the case that will be analyzed here. To find the equation that represents  $L$  we will use that fact that its slope is equal to  $-1/\tan \vartheta$  and its intercept with the  $y$  axis is  $r/\sin \vartheta$ . Thus

$$y = -\frac{\cos \vartheta}{\sin \vartheta} x + \frac{r}{\sin \vartheta} \quad (7)$$



so that

$$x \cos \vartheta + y \sin \vartheta = r. \quad (8)$$

To tie this definition to the concept of *slant stack* in seismology we note that it is defined using Eq. (3) with  $f$  replaced by a function  $u(x, t)$ , where  $u$ ,  $x$ , and  $t$  represent seismic wave amplitude, distance, and time, respectively,  $y$  is replaced by  $t$ , and  $t$  is written in terms of the intercept and slope of the line  $L$  (generally indicated with  $\tau$  and  $p$ , so that  $t = px + \tau$ ) (see, e.g., [40,127]).

Now we will introduce a new coordinate system  $(r, s)$  obtained by a rotation of angle  $\vartheta$  of the  $(x, y)$  system (Fig. 1). In the  $(r, s)$  system the points on the line  $L$  have constant  $r$  and variable  $s$ . The two systems are related by the following transformations of coordinates

$$\begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} \cos \vartheta & \sin \vartheta \\ -\sin \vartheta & \cos \vartheta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \cos \vartheta + y \sin \vartheta \\ -x \sin \vartheta + y \cos \vartheta \end{pmatrix} \quad (9)$$

and

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos \vartheta & -\sin \vartheta \\ \sin \vartheta & \cos \vartheta \end{pmatrix} \begin{pmatrix} r \\ s \end{pmatrix} = \begin{pmatrix} r \cos \vartheta - s \sin \vartheta \\ r \sin \vartheta + s \cos \vartheta \end{pmatrix}. \quad (10)$$

Using Eq. (10), Eq. (3) can be rewritten as

$$\begin{aligned} p(r, \theta) &= \int_{-\infty}^{\infty} f(r \cos \vartheta - s \sin \vartheta, r \sin \vartheta + s \cos \vartheta) ds \\ &\equiv \int_{-\infty}^{\infty} \hat{f}_{\vartheta}(r, s) ds \end{aligned} \quad (11)$$

where  $\hat{f}_{\vartheta}$  represents the function  $f$  when it is written in terms of  $r$  and  $s$ , and the subscript  $\vartheta$  is used to emphasize that it enters in the computations as a parameter. In Eq. (11)  $r$  is allowed to vary continuously, although in practical applications (such as CT)  $r$  is a discrete variable, as sketched in Fig. 1. Also note that  $r$  is allowed to be negative, which means that  $0 \leq \vartheta < \pi$ , and that although  $|r|$  is allowed to extend to infinity, for functions defined over a finite domain in the  $(x, y)$  plane the projections for lines outside of the domain will be zero.

To proceed further we will work in the wavenumber domain. Let the one-dimensional Fourier transform (or 1-D F.T.) of  $p(r, \theta)$  with respect to  $r$  be

$$P(k, \theta) = \int_{-\infty}^{\infty} p(r, \theta) e^{-i2\pi kr} dr \quad (12)$$

where  $k$  indicates wavenumber (equivalent to the frequency in the time domain). Introducing Eq. (11) into this

expression and then going back to the  $(x, y)$  coordinate system gives

$$\begin{aligned} P(k, \theta) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \hat{f}_{\vartheta}(r, s) e^{-i2\pi kr} dr ds \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) e^{-i2\pi k(x \cos \vartheta + y \sin \vartheta)} dx dy \\ &= F_{\vartheta}(k_x, k_y) \end{aligned} \quad (13)$$

where  $r$  has been written using Eq. (9), the Jacobian of the coordinates transformation is equal to one,  $F_{\vartheta}$  is the two-dimensional Fourier transform of  $f$  and

$$k_x = k \cos \vartheta; \quad k_y = k \sin \vartheta. \quad (14)$$

Note that

$$k^2 = k_x^2 + k_y^2; \quad k_y/k_x = \tan \theta. \quad (15)$$

In summary,

$$P(k, \theta) = F_{\vartheta}(k_x, k_y). \quad (16)$$

In words, the 1-D F.T. of the projection  $p(r, \theta)$  of  $f(x, y)$  is equal to the 2-D F.T. of  $f(x, y)$ . This relation is known as the Fourier (central) slice theorem. This name comes from the fact that the 2-D F.T. is known in a slice (i.e., a line) through the origin in the  $(k_x, k_y)$  space, as Eq. (15) shows. Moreover, the angle of this line with the  $k_x$  axis is  $\vartheta$ , which is equal to the angle between the  $x$  and  $r$  axes. Equation (16) suggests one approach to the determination of  $f(x, y)$ . Find the 1-D F.T. of the projections for discrete values of  $\theta$  between 0 and  $\pi$  and then use the 2-D inverse F.T. to recover  $f(x, y)$  numerically. However, because the values of the 2-D F.T. will be defined on a polar coordinates grid, it must be interpolated to a Cartesian grid in the  $(k_x, k_y)$  space.

A different approach is as follows. Let  $F(k_x, k_y)$  be the 2-D F.T. of  $f(x, y)$ . Then,  $f(x, y)$  is given by the inverse F.T.

$$f(x, y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} F(k_x, k_y) e^{i2\pi(k_x x + k_y y)} dk_x dk_y. \quad (17)$$

In this expression  $k_x$  and  $k_y$  are defined on a Cartesian coordinate system, which will be converted to polar coordinates using Eq. (14). This gives

$$f(x, y) = \int_0^{\pi} \int_{-\infty}^{\infty} P(k, \vartheta) e^{i2\pi k(x \cos \vartheta + y \sin \vartheta)} |k| dk d\vartheta \quad (18)$$

where Eq. (16) was used and  $|k| d\vartheta dk$  is the Jacobian for the transformation of coordinates. Because the Jacobian relates the elements of area in the two coordinate systems, the absolute value of  $k$  is needed because the area is positive. Two comments are in order here. First,  $\vartheta$  is no longer a parameter; now it is used as a variable. Second, the polar coordinates have been defined using  $0 \leq \theta < \pi$  and  $-\infty < k < \infty$ . This convention is equivalent to the standard one (i.e.,  $0 \leq \theta < 2\pi$ ,  $0 \leq k < \infty$ .)

Now Eq. (18) will be rewritten as follows

$$f(x, y) = \int_0^\pi p^*(x \cos \vartheta + y \sin \vartheta, \vartheta) d\vartheta \quad (19)$$

where

$$p^*(r, \vartheta) = \int_{-\infty}^{\infty} P(k, \vartheta) |k| e^{i2\pi kr} dk. \quad (20)$$

The integral in Eq. (20) involves the product of two 1-D F.T.s,  $P(k, \theta)$  and  $|k|$ , so, in principle, one way to solve it would be to use the following property. Given two functions  $g(r)$  and  $h(r)$  with F.T.s  $G(k)$  and  $H(k)$ , then

$$\int_{-\infty}^{\infty} G(k) H(k) e^{i2\pi kr} dk = \int_{-\infty}^{\infty} g(\rho) h(r - \rho) d\rho \quad (21)$$

where the integral on the right-hand side is the convolution of  $g$  and  $h$  (see, e.g., [94]). A problem with this formulation is that  $|k|$  does not have an inverse F.T., as can be seen from the fact that it does not go to zero as  $k$  goes to infinity, which is one of the properties of the F.T. [149]. There are two ways to get around this problem. One is to solve Eq. (21) in terms of distributions. A basic yet rigorous introduction to distributions, which can be considered an extension of the concept of function, can be found in [120]. Let us rewrite Eq. (20) as follows

$$p^*(r, \theta) = \frac{1}{2\pi^2} \int_{-\infty}^{\infty} P(k, \theta) (i2\pi k) (-i\pi \operatorname{sgn} k) e^{i2\pi kr} dk \quad (22)$$

where

$$\operatorname{sgn} k = \begin{cases} 1 & k > 0 \\ -1 & k < 0. \end{cases} \quad (23)$$

Then, using Eq. (21) we can write

$$p^*(r, \theta) = \frac{1}{2\pi^2} \mathcal{F}^{-1} \{ i2\pi k P(k, \theta) \} * \mathcal{F}^{-1} \{ -i\pi \operatorname{sgn} k \} \quad (24)$$

where  $\mathcal{F}^{-1}$  represents the inverse F.T. of the function in braces and the  $*$  stands for convolution. The first inverse is just  $\partial p(r, \theta) / \partial r$  and the second inverse is equal to  $1/r$  [120]. Combining these two results gives

$$\begin{aligned} p^*(r, \theta) &= -\frac{1}{2\pi} \left( -\frac{1}{\pi r} * \frac{\partial p(r, \theta)}{\partial r} \right) \\ &\equiv -\frac{1}{2\pi} \mathcal{H} \left\{ \frac{\partial p(r, \theta)}{\partial r} \right\} \end{aligned} \quad (25)$$

where  $\mathcal{H}$  stands for the Hilbert transform of the function in braces [120]. The inverse expression for the slant stack can be derived using expressions similar to those presented above [25].

The importance of Eq. (25) is mostly theoretical, and from a practical point of view a different method of solution was found to be more useful. Before proceeding, however, note that Eq. (20) without  $|k|$  on the right-hand side corresponds to the 1-D inverse F.T. of  $P(k, \theta)$ , equal to  $p(r, \theta)$  (see Eq. (12)). Therefore,  $p^*$  is a *filtered projection*, with  $|k|$  the filter response. Clearly, the effect of the filter is to amplify the components of  $p$  corresponding to the higher wavenumbers. The approach used here is also based on the use of Eq. (21), with  $g(r) = p(r, \theta)$  (as before) and  $h(r)$  a function whose F.T. is an approximation to  $|k|$  in the sense that

$$H(k) \approx \begin{cases} |k|; & |k| < K \\ 0; & \text{elsewhere} \end{cases} \quad (26)$$

where  $K$  can be taken as the spatial Nyquist frequency  $k_N$  (equal to  $1/2a$ , where  $a$  is the spacing between projection lines). To avoid aliasing,  $P(k, \theta)$  must be zero for  $|k| > k_N$ . Clearly, there is some freedom in the selection of  $h(r)$ . For example, in [18] and [124]  $H(k)$  is defined by an expression similar to Eq. (26) with an equal sign in place of the less than sign. This work was followed by the introduction of an improved function [133]. For our purposes, however, what is important is not the particular choice of  $h$  but the fact that we can write the following approximation for  $f(x, y)$

$$\begin{aligned} f(x, y) &\approx \tilde{f}(x, y) \equiv \int_0^\pi \int_{-\infty}^{\infty} p(\rho, \theta) \\ &\quad h(x \cos \vartheta + y \sin \vartheta - \rho) d\rho d\theta. \end{aligned} \quad (27)$$

Finally, because the projections are determined for discrete and equispaced values of  $r$  and  $\theta$  we can write

$$\begin{aligned} \tilde{f}(x, y) &\approx \frac{\pi a}{N} \sum_{i=1}^N \sum_{j=-J}^J p(ja, \theta_i) \\ &\quad h(x \cos \vartheta_i + y \sin \vartheta_i - ja) \end{aligned} \quad (28)$$

where  $\theta_i = (i - 1)\pi/N$ ,  $N$  is fixed, and  $J$  is a finite limit that takes into account the finite size of the object (after [132]).

To end this discussion of the transform methods of solution we will go back to Eq. (19). The integration there is an example of the operation known as *backprojection*, to be indicated with  $\mathcal{B}$  and defined by

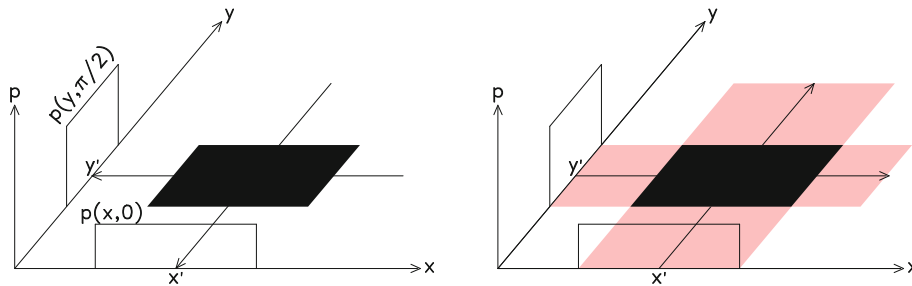
$$\mathcal{B}\{q\}(x, y) = \int_0^\pi q(x \cos \vartheta + y \sin \vartheta, \vartheta) d\vartheta \quad (29)$$

where  $q$  is an arbitrary function of  $r$  and  $\vartheta$ . An example of this operation is provided in Fig. 2a, which shows two projections of a rectangular object corresponding to  $\vartheta$  equal to 0 and  $\pi/2$ . The two projections, identified by  $p(x, 0)$  and  $p(y, \pi/2)$ , are nonzero for limited ranges of the  $y$  and  $x$  axes. To compute the backprojection let us consider a particular value  $y'$  of  $y$  and assign the value  $p(y', \pi/2)$  to every point in the line defined by the points  $(x, y')$ . The object may be part of a larger body, and for this reason the line will extend beyond the  $x$  values that bound the object. Repeating this operation for all the values  $y'$  and then proceeding in a similar way for the  $p(x, 0)$  projection produces an image such as that shown in Fig. 2b. The two backprojections overlap in the area occupied by the rectangular object, but also contribute values to other areas of the  $(x, y)$  plane, which is clearly not correct. The problem is not related to the use of only two projections; even when a larger number of projections is used, the contributions outside of the object do not disappear. The reason for that is that the determination of  $f(x, y)$  requires  $p^*$ , not  $p$ , as shown by Eq. (19) (see, e.g., [23]).

**Iterative Solutions** As was the case with the analytical solutions, the iterative solutions were developed before the advent of CT. The earliest one was the *algebraic reconstruction technique* (ART, [51]) and was developed for the

solution of electron microscopy problems as an alternative to analytical solutions based on the Fourier transform. These problems are represented by an equation similar to Eq. (2). The basic idea behind ART was the discretization of a plane object in terms of a square grid of points. The goal was to find the optical density  $\rho_{ij}$  at each point  $(i, j)$  of the grid. A ray of a projection at an angle  $\theta$  was defined as a band of width  $w$  across the plane at the same angle, and corresponds to each of our lines  $L$ . One possible choice for  $w$  is to make it equal to the grid spacing. The basis of ART is to start with an average value of the density (computed from the observations) and to use it to compute the projections for all the rays for all the angles. Then, two updating methods were used, one additive and the other multiplicative. In the original version of the additive method, for each ray the difference between observed and computed projections was used to update the density values in the cells along the ray. If an updated value of density became negative it was set to zero. This was done for all the rays, one ray at a time. This completed one iteration. Then a new one was started with the updated densities used as starting values. Originally, the difference between the observed and computed projections was divided equally among all the cells in the ray, but in later applications each cell received a weight representing its contribution to a given ray. The weight was used when the cell densities were updated (see, e.g., [50]). In the multiplicative version of ART the density value of a cell was updated by multiplying it by the ratio of computed to observed projections. Again, a weighting scheme was introduced later. A technique similar to additive ART was developed independently by Hounsfield and applied to his X-ray scanner (see, e.g., [20,133]).

A second technique, simultaneous iterative reconstruction technique (SIRT) was introduced in [48], which was highly critical of the performance of ART. SIRT also



Tomography, Seismic, Figure 2

Illustration of the concept of back projection. *Left*: projection of a rectangular box using two sets of parallel lines (two representative lines are shown). *Right*: backprojection of the two projections shown on the left. Note the change in the direction of the arrows. For each line, the value of its projection is assigned to all the points along the line and are added to the values that come from all the other projections. See text for details. After [20]

has additive and multiplicative versions, and the main difference with the ART counterparts is that at each iteration the density of each cell is updated using all the projections passing through that cell. Subsequent work showed that although ART is computationally more efficient than SIRT, it has the problem that is more affected by errors in the data (see, e.g., [20,39,63]). SIRT, on the other hand, has some undesirable properties, referred to in Subject. “Regularization Approach”.

### Arrival-Time Seismic Tomography

As noted earlier, this type of tomography is much simpler than surface wave and waveform tomography, and lends itself to a fairly self-contained discussion, to be presented here. The other two types of tomography will not be discussed here because they require a knowledge of seismic theory and data processing beyond the scope of this article.

This section is organized as follows. First, the differences between seismic and X-ray tomography will be discussed. The early applications of CT techniques to seismic problems ignored these differences, and although the results thus obtained opened the way to a new research approach in seismology, they may have been affected by several types of errors, the sources of which will become clear here. Second, tomography using local data will be addressed. In this case both the location of the events as well as the velocity model must be determined. In addition, a distinction between velocity and slowness tomography must be made. As noted earlier, before seismic tomography there was velocity inversion. Then, after the introduction of CT techniques the inversion parameter became slowness, not velocity. Therefore, the equations to be solved in the two cases are different, and because some of the velocity inversion programs are still in use, the two parameterizations, and their relationship, will be presented here. In addition, regardless of the parameterization used, it is necessary to decouple (or separate) the location part from the tomographic part of the problem, which requires introducing new analytical tools. Another topic related to local tomography is the computation of travel times, which will be briefly considered here. Finally, the case of tomography using teleseismic arrival times is analyzed.

### Comparison with X-ray Tomography

Seismic tomography differs from X-ray tomography in two fundamental ways because, first, the former is non-linear, and second, the locations of the seismic sources are generally unknown (i.e., when earthquakes are used). Another major difference is that in X-ray tomography there is control over position and number of sources and receivers,

which is not the case for seismic tomography. The first two differences will be considered in more detail.

The expression for travel time,  $t$ , along a ray is given by

$$t = \int_R \frac{1}{v} ds = \int_R u ds \quad (30)$$

where  $R$  denotes the raypath between two fixed points within an elastic medium with velocity  $v(\mathbf{x})$  that depends on position (indicated by the vector  $\mathbf{x}$ ),  $ds$  is a line element along the raypath, and  $u$  is the slowness, equal to the inverse of velocity

$$u(\mathbf{x}) = \frac{1}{v(\mathbf{x})}. \quad (31)$$

An advantage of writing  $t$  in terms of  $u$  instead of  $v$  is that in terms of  $u$ ,  $t$  has an expression similar to Eq. (2), which constitutes the basis of X-ray tomography. This similarity, however, is more apparent than real. To see that we will introduce the concept of linear operation. Given an operation  $\mathcal{O}$ , it is said to be linear if

$$\mathcal{O}(f + h) = \mathcal{O}(f) + \mathcal{O}(h) \quad (32)$$

and

$$\mathcal{O}(\alpha f) = \alpha \mathcal{O}(f) \quad (33)$$

where  $f$  and  $h$  are functions and  $\alpha$  is a scalar. A simple example of linear operation is the integration over a given interval

$$\int_a^b [f(x) + h(x)] dx = \int_a^b f(x) dx + \int_a^b h(x) dx. \quad (34)$$

In particular, the operation defined by Eq. (2) is linear because  $L$  is always a straight line that does not depend on  $f$ . Therefore, if we let  $f = f_1 + f_2$  we can write

$$\begin{aligned} g_L(f) &= \int_L [f_1(s) + f_2(s)] ds \\ &= \int_L f_1(s) ds + \int_L f_2(s) ds \\ &= g_L(f_1) + g_L(f_2). \end{aligned} \quad (35)$$

Now let us consider the travel time problem. Let  $R(u_1)$  and  $R(u_2)$  denote the raypaths between two fixed points within an elastic medium assuming that it had slownesses  $u_1$  and  $u_2 \neq u_1$  at two different times. This situation is generally not possible in the earth but can be simulated on a computer. The corresponding travel times are given by

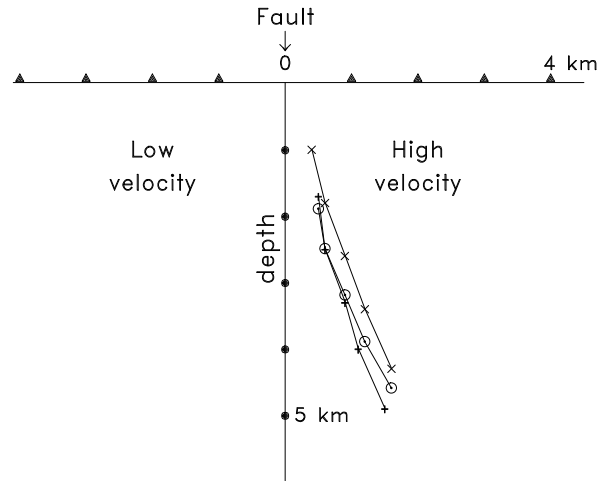
$$t_1 = \int_{R(u_1)} u_1(s) ds, \quad t_2 = \int_{R(u_2)} u_2(s) ds. \quad (36)$$

Note that the two raypaths will be different because  $u_1$  and  $u_2$  have been assumed to be different. The trivial case of  $u_1 = ku_2$ , where  $k$  is a constant, is ignored. Therefore, it will not be possible, in general, to write  $t_1 + t_2$  as an integral over a common path involving  $u_1 + u_2$ . Moreover, if the medium has slowness  $u_1 + u_2$  the raypath  $R(u_1 + u_2)$  between the two points will be different from  $R(u_1)$  and  $R(u_2)$  and, in general,

$$t_1 + t_2 \neq \int_{R(u_1+u_2)} [u_1(s) + u_2(s)] ds \quad (37)$$

which in turn means that the travel time problem cannot be expressed in terms of a linear operation. Therefore, the seismic tomography problem is nonlinear, and the CT solution techniques cannot be applied directly except when the 3-D velocity variations in the medium are small, a condition that severely limits its application to situations of seismic interest.

Another major difference between seismic tomography and CT is that the locations of the sources (mostly earthquakes) are generally not known. Therefore, the determination of a velocity model requires the simultaneous determination of the source locations. This is particularly important when the events are locally recorded. To see that we will consider a very simple example. Let us assume that a vertical fault separates rocks with low- and high-velocities, denoted  $v_1$  and  $v_2$ , respectively, and that earthquakes occur along a vertical line within the fault. Let  $v_1 = 3$  km/s and  $v_2 = 4$  km/s. The origin of the coordinate system will be the fault location at the surface and the event depths will be 1, 2, 3, 4, and 5 km. The earthquakes will be assumed to be recorded by a local network of eight stations on the surface. The theoretical arrival times computed for this geometry were used with a single-event location program and two constant velocity models with velocities equal to 3 and 3.5 km/s and with a 1-D model with five layers and velocities between 2.5 and 4 km/s. The locations obtained with these models (Fig. 3) are affected by significant errors, which cannot be inferred by the small root-mean-square errors, which range between 0.02 and 0.09 s, with the larger value corresponding to the deeper event. These results show that the effect of the lateral juxtaposition of high- and low-velocity rocks when the events are located with 1-D velocity models is a mislocation of the events in a direction away from the low-velocity zone. A well-documented example of this situation is provided by the aftershocks of the 1994  $M = 6.7$  Northridge, California, earthquake, which occurred within the sedimentary rocks of the San Fernando basin. These events and the velocity model determined using them are considered in



Tomography, Seismic, Figure 3

Simple example (2-D) of the effect of lateral velocity variations on earthquake location when those variations are ignored. A fault divides the medium into two quarter spaces with velocities equal to 3 km/s (left) and 4 km/s (right). The triangles and dots represent stations and hypocenters, respectively. Synthetic arrival times computed for these events, stations, and velocity model are used with a single-event location program and three 1-D velocity models. Two models have constant velocities, equal to 3 km/s and 3.5 km/s, and the other has five layers with velocities between 2.5 and 4 km/s. The locations determined using these models are indicated by pluses, crosses and circles

Subject. “P Wave Velocity Model for the San Fernando, California, Area”.

This example has a counterpart. What happens when mislocated events are kept at their erroneous locations and are used to determine a 2-D velocity model around the fault? Clearly, such a velocity model will be affected by the location errors and its reliability will be questionable. Examples showing the significant bias that these errors may have in the computed velocities are given in [54,145], and [146] for geometries similar to that in Fig. 3. In [146],  $v_1 = 5$  km/s and  $v_2 = 6$  km/s and there was only one earthquake on the fault. In this case the velocity determined by inversion was 5.44 km/s to the left of the fault and 5.56 km/s to the right of it. This example is clearly an oversimplification, but the result is useful because it shows that ignoring the mislocation introduced by lateral velocity variations will result in a velocity model showing variations smaller than the actual ones. For the case of teleseismic tomography, event mislocation may not be a serious problem except when earthquakes from subducting slabs are used, in which case the effect on the computed velocities may be important (see, e. g., [32,153]).

Another difference between seismic tomography and X-ray tomography is that in the later the position of the



sources and the receivers can be chosen so that the computed images have the desired resolution. In contrast, in seismic tomography there is no control over the position of the earthquakes, while the number and position of stations is dictated by financial considerations as well as practical constraints imposed by the nature of the terrain. For example, mountains frequently constitute a severe obstacle to station deployments. Even more serious is the presence of oceans, which cover two-thirds of the earth's surface and essentially prevent the routine deployment of permanent seismic networks similar to those on land. As a consequence, seismic tomographic models, particularly at a global scale, are affected to some extent by low-resolution problems.

Finally, we note that soon after X-ray tomography began, several competing methods of solution were proposed (see Subsect. "Solving the CT Problem"), and to compare their performances synthetic data generated for a model of a cross section of the skull were used [133]. With this approach it was possible to detect artifacts in some of the solutions and to investigate the effect of noise in the data. In contrast, seismic tomography methods were rarely subject to a similar validation analysis in spite of the approximations and simplifying assumptions made. These differences stem, in part, on the complicated nature of wave propagation in the earth, which is difficult to replicate in a computer on account of theoretical difficulties and computational cost. On the other hand, the propagation of X-rays is relatively simple and easy to simulate on a computer. In addition, medical tomography, in general, is well funded, while seismic tomography is only poorly funded [84].

### Local Velocity Tomography

The simultaneous determination of the locations of a group of earthquakes and a 3-D velocity model can be considered an extension of the standard method of earthquake location. Usually, the velocity structure is modeled in terms of constant-velocity blocks or by velocity values on a grid. The basic ideas were introduced by Aki and Lee [2] and Crosson [33], although the latter only considered layered velocity models. Because this inverse problem is nonlinear in both the earthquake location and the velocity determination, it is solved by linearizing it about initial estimates of the locations and origin times of the events and model velocities. With this approach the original problem is replaced by other in which the unknowns are adjustments to the initial estimates. Then these adjustments are added to the initial estimates, the corresponding values are used as new initial estimates, and the pro-

cess is repeated iteratively until some stopping criterion is met. This process will be described mathematically in the following.

Let us use subscripts  $i$  and  $j$  to identify the stations and earthquakes used, respectively, and let  $M$  and  $N_j$  be the number of events and the number of stations with arrival times for the  $j$ th event, respectively. The blocks or grid points used to parameterize the velocity structure constitute a 3-D array, but for computational purposes each block or point will be identified with a single subscript, say  $k$ . This requires establishing an ordering scheme that assigns a single index to a triplet of indices. For example, we may use the following scheme:  $(1, 1, 1) \rightarrow 1$ ,  $(2, 1, 1) \rightarrow 2$ ,  $(n, 1, 1) \rightarrow n$ ,  $(n, 2, 1) \rightarrow n + 1$ , and so on. Let  $K$  represent the total number of velocity parameters. The linearized problem can be written as

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \sum_k \frac{\partial t}{\partial v_k} dv_k \right) \quad (38)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where

$$r_{ij} = T_{ij}^{\text{obs}} - T_{ij}^{\text{comp}} \equiv T_{ij}^{\text{obs}} - (T_j + t_{ij}) \quad (39)$$

is the arrival time residual for the  $j$ th earthquake and the  $i$ th station. The subscript  $k$  labels the blocks or grid points associated with the raypath from the  $j$ th hypocenter to the  $i$ th station. Therefore,  $k$  is a function of  $i$  and  $j$ , but this dependence is left implicit to simplify the notation. Also note that the values of  $k$  associated with a given ray do not have any particular ordering. The meaning of the other variables is as follows:

$dT_j, (dx_j, dy_j, dz_j), dv_k$ : adjustments to origin time, hypocentral coordinates, and velocities, respectively,  
 $T_{ij}^{\text{obs}}$ : observed arrival time,  
 $T_{ij}^{\text{comp}}$ : computed arrival time,  
 $T_j$ : origin time,  
 $t_{ij}$ : computed travel time,  
 $t, v$ : travel time and velocity,  
 $w_{ij}$ : quality weight.

In Eqs. (38) and (39) the only unknowns are the adjustments. All the other quantities are computed using initial estimates of the velocities ( $v_k$ ) and the origin time ( $T_j$ ) and hypocentral coordinates ( $x_j, y_j, z_j$ ) of each earthquake. The expressions for the derivatives can be found in [85].

### Local Slowness Tomography

As noted earlier, the travel time problem cannot be expressed in terms of a linear operation on slowness, and for this reason the problem is linearized assuming that the difference between the initial and actual velocity models is small. Let us go back to Eq. (36) with  $u_2 = u_1 + du$ ,  $du \ll u_1$ . Then the difference  $dt$  in travel times is given by

$$dt = t_2 - t_1 = \int_{R(u_1+du)} [u_1(s) + du(s)] ds - \int_{R(u_1)} u_1(s) ds \approx \int_{R(u_1)} du(s) ds. \quad (40)$$

Here we have invoked *Fermat's principle*, which allows replacing  $R(u_1 + du)$  with  $R(u_1)$ . Recall that this principle states that raypaths are paths of stationary travel time (see, e.g., [3,120]), which in turn means that the variation of travel time along a raypath is zero. Two points must be noted here, however. First, the principle applies to small variations in raypaths, which in turn requires a small  $du$ , a condition which may not be valid in reality. This question was tested in the context of the propagation of rays through subducting slabs [31] and it was found that arrival times at teleseismic distances calculated using Fermat's principle were the same or later than those determined using exact ray tracing. The difference between exact and approximate times depended on event depth, with the larger errors corresponding to intermediate-depth events. A consequence of these errors was an underestimation of the velocity anomaly. Second, Fermat's principle applies to curves with the same end points, but in practice the end point of  $R(u_1)$  corresponding to the event location is not the true location because it has been determined with an incorrect velocity model. Therefore,  $R(u_1)$  is not necessarily close to the true ray path. In practice, however, the approximations involved in the use of Fermat's principle in local tomography improve as the iterations proceed, and can essentially be ignored.

For completeness, the relation between small variations  $dv$  in velocity and slowness will be considered. Because travel time is equal to distance divided by velocity, here we are interested in  $1/(v + dv)$ , with  $dv \ll v$ , which can be approximated as follows

$$\frac{1}{v + dv} = \frac{1}{v} \left( 1 + \frac{dv}{v} \right)^{-1} \approx \frac{1}{v} - \frac{1}{v^2} dv \equiv u + du. \quad (41)$$

Therefore, it does not make any theoretical difference whether the inverse problem is formulated in terms of slowness or velocity as long as  $du$  and  $dv$  are such that the approximations introduced are valid.

Now let us go back to the joint hypocentral location and determination of a 3-D slowness model. The only difference with the expression for velocity tomography (Eq. (38)) is in the last term, which is replaced by the right-hand side of Eq. (40). This gives

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \int_{R_{ij}} du ds \right) \quad (42)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where  $R_{ij}$  indicates the raypath between the corresponding station-event pair. To proceed further the integral must be discretized, which can be done using a block model. After that Eq. (42) becomes

$$w_{ij}r_{ij} = w_{ij} \left( dT_j + \frac{\partial t}{\partial x} dx_j + \frac{\partial t}{\partial y} dy_j + \frac{\partial t}{\partial z} dz_j + \sum_k l_{ijk} du_k \right) \quad (43)$$

$$i = 1, \dots, N_j \quad j = 1, \dots, M$$

where  $l_{ijk}$  and  $du_k$  are the length of the ray in the  $k$ th block and the corresponding slowness perturbation.

Note that although the last terms in Eqs. (38) and (43) are different, in principle, the velocity and slowness models derived using the two formulations should satisfy  $u(\mathbf{x}) = 1/v(\mathbf{x})$ . Let us consider this question for the block parameterization. To compute the derivative of travel time with respect to velocity we will use the following approximation

$$\frac{\partial t_{ij}}{\partial v_k} \approx \frac{\partial t_{ijk}}{\partial v_k} \quad (44)$$

[85], where the subscripts in  $t$  are used to identify the rays and blocks involved. The meaning of this relation is that the change in travel time along a ray path due to a change in the velocity of a given block is approximately equal to the change in travel time in that block. Note that

$$t_{ijk} = \frac{l_{ijk}}{v_k} \quad (45)$$

which allows us to write

$$\frac{\partial t_{ij}}{\partial v_k} dv_k \approx l_{ijk} \left( -\frac{1}{v_k^2} dv_k \right) \equiv l_{ijk} du_k. \quad (46)$$

Using this result we see that Eqs. (38) and (43) are identical, which means that the two formulations should give the

same results as long as the block structure is the same in the two cases and the approximations introduced are valid. Of course, the numerical implementation of the software corresponding to the two approaches should be the same.

### Decoupling of the Earthquake Location and Tomography Problems

Equation (38) can be written in matrix form as:

$$\mathbf{W}_j \mathbf{A}_j d\mathbf{x}_j + \mathbf{W}_j \mathbf{B}_j d\mathbf{v} = \mathbf{W}_j \mathbf{r}_j; \quad j = 1, M \quad (47)$$

where  $\mathbf{r}_j$  is the vector of residuals  $r_{ij}$ ,  $\mathbf{A}_j$  is an  $N_j \times 4$  matrix of partial derivatives of time with respect to origin time and hypocentral coordinates,  $d\mathbf{x}_j$  is the vector of origin time and hypocenter adjustments,  $\mathbf{W}_j$  is an  $N_j \times N_j$  matrix of weights,  $\mathbf{B}_j$  is an  $N_j \times K$  matrix of partial derivatives of travel time with respect to velocities, and  $d\mathbf{v}$  is a vector of  $K$  velocity adjustments. The matrix  $\mathbf{B}_j$  has zero entries for the blocks not traversed by any ray. The matrix  $\mathbf{W}_j$  has only one nonzero entry per row and column and is not necessarily diagonal because the ordering of stations for different earthquakes may not be always the same.

For Eq. (43) the matrix form is

$$\mathbf{W}_j \mathbf{A}_j d\mathbf{x}_j + \mathbf{W}_j \mathbf{C}_j d\mathbf{u} = \mathbf{W}_j \mathbf{r}_j; \quad j = 1, M \quad (48)$$

with  $\mathbf{C}_j$  an  $N_j \times K$  matrix whose entries are raypath lengths in individual blocks and  $d\mathbf{u}$  is a vector of  $K$  slowness adjustments. All the other quantities are as in Eq. (47).

Clearly, Eqs. (47) and (48) are formally equivalent, which means that for the following discussion on how to solve them it is unnecessary to distinguish between velocity and slowness. For this reason, from now on we will refer exclusively to Eq. (47), with the understanding that the results below apply to Eq. (48) as well.

Equation (47) represents  $M$  systems of equations coupled through the common vector  $d\mathbf{v}$ . Therefore, these systems could be combined, in principle, into a single system, but because its size could become computationally unmanageable, it is necessary to decouple the earthquake location part from the inversion part. A very efficient approach to do that is based on the singular value decomposition (SVD) of a matrix [110] and is known as the method of parameter separation. The following presentation is based on [119]. The SVD of an arbitrary  $n \times m$  matrix  $\mathbf{G}$  is given by

$$\mathbf{G} = \mathbf{U} \mathbf{\Lambda} \mathbf{V}^T \quad (49)$$

(see, e. g., [3,46]) where the superscript T indicates matrix transposition,  $\mathbf{U}$  and  $\mathbf{V}$  are  $n \times n$  and  $m \times m$  matrices with

columns given by the eigenvectors of  $\mathbf{G}\mathbf{G}^T$  and  $\mathbf{G}^T\mathbf{G}$ , respectively, and  $\mathbf{\Lambda}$  is an  $n \times m$  matrix with diagonal elements equal to the singular values of  $\mathbf{G}$  and off-diagonal elements equal to zero. The singular values are positive and equal to the square roots of the eigenvalues of  $\mathbf{G}\mathbf{G}^T$  and  $\mathbf{G}^T\mathbf{G}$ . The number of nonzero singular values, say  $p$ , cannot exceed the minimum of  $m$  and  $n$ . Matrices  $\mathbf{U}$  and  $\mathbf{V}$  are orthogonal, i. e.,

$$\mathbf{U}\mathbf{U}^T = \mathbf{I}_n, \quad \mathbf{V}\mathbf{V}^T = \mathbf{I}_m, \quad (50)$$

where  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Let us assume that the singular values are sorted in nonincreasing order (i. e., largest first). Then, matrices  $\mathbf{U}$  and  $\mathbf{V}$  can be partitioned as follows:

$$\mathbf{U} = (\mathbf{U}_p \quad \mathbf{U}_0), \quad \mathbf{V} = (\mathbf{V}_p \quad \mathbf{V}_0) \quad (51)$$

where the subscripts  $p$  and  $0$  indicate that the columns of the matrices come from the eigenvectors corresponding to the nonzero and zero singular values, respectively. Partitioned matrices are discussed in, e. g., [103]. Matrix  $\mathbf{A}$  can be partitioned in a similar way. Then, writing Eq. (49) in terms of these partitioned matrices gives

$$\mathbf{G} = (\mathbf{U}_p \quad \mathbf{U}_0) \begin{pmatrix} \mathbf{A}_p & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \begin{pmatrix} \mathbf{V}_p^T \\ \mathbf{V}_0^T \end{pmatrix} = \mathbf{U}_p \mathbf{A}_p \mathbf{V}_p^T \quad (52)$$

where  $\mathbf{A}_p$  is the  $p \times p$  diagonal matrix having as elements the  $p$  nonzero singular values and  $\mathbf{O}$  represents zero matrices of appropriate sizes. Using this result and

$$\mathbf{U}_0^T \mathbf{U}_p = \mathbf{O} \quad (53)$$

which is a consequence of the fact that the columns of  $\mathbf{U}$  are orthogonal to each other, we get the following important result

$$\mathbf{U}_0^T \mathbf{G} = \mathbf{U}_0^T \mathbf{U}_p \mathbf{A}_p \mathbf{V}_p^T = \mathbf{O}. \quad (54)$$

To apply Eq. (54) to our problem we will use

$$\begin{aligned} \mathbf{G} &= \mathbf{W}_j \mathbf{A}_j = \mathbf{U}_j \mathbf{A}_j \mathbf{V}_j^T \\ &= (\mathbf{U}_{jp} \quad \mathbf{U}_{j0}) \mathbf{A}_j (\mathbf{V}_{jp} \quad \mathbf{V}_{j0})^T = \mathbf{U}_{jp} \mathbf{A}_{jp} \mathbf{V}_{jp}^T. \end{aligned} \quad (55)$$

Because  $\mathbf{W}_j \mathbf{A}_j$  has four columns and at least as many rows,  $p \leq 4$ . If  $p < 4$ , this means that one or more of the columns of the matrix is a linear combination of the others and the location of the corresponding event cannot be determined uniquely. Events with  $p < 4$  should not be used in the inversion. Now multiply Eq. (47) by  $\mathbf{U}_{j0}^T$  on the left and use the equivalent of Eq. (54) for the  $j$ th event. This gives

$$\mathbf{U}_{j0}^T \mathbf{W}_j \mathbf{B}_j d\mathbf{v} = \mathbf{U}_{j0}^T \mathbf{W}_j \mathbf{r}_j; \quad j = 1, M. \quad (56)$$

Because  $\mathbf{A}_j$  does not appear in Eq. (56), we have been able to decouple the velocity parameters from the earthquake parameters. Note, however, that this equation depends on  $\mathbf{A}_j$  implicitly via  $\mathbf{U}_{j0}$ , which means that location errors will translate into errors in  $\mathbf{U}_{j0}$  and, thus, in  $d\mathbf{v}$ .

To simplify the following discussion Eq. (56) will be rewritten as

$$\underbrace{\mathbf{B}'_j}_{(N_j-4) \times K} \underbrace{d\mathbf{v}}_{K \times 1} = \underbrace{\mathbf{r}'_j}_{(N_j-4) \times 1}; \quad j = 1, M \quad (57)$$

with obvious expressions for the primed quantities. The sizes of the matrix and vectors involved are also indicated (assuming  $p = 4$ ). Equation (57) represents  $M$  systems of equations with the common unknown vector  $d\mathbf{v}$  and can be written in compact form in terms of a partitioned matrix and vector

$$\mathcal{B} d\mathbf{v} = \boldsymbol{\rho} \quad (58)$$

where

$$\mathcal{B} = \left( \mathbf{B}'_1{}^T \mathbf{B}'_2{}^T \dots \mathbf{B}'_M{}^T \right)^T \quad (59)$$

and

$$\boldsymbol{\rho} = \left( \mathbf{r}'_1{}^T \mathbf{r}'_2{}^T \dots \mathbf{r}'_M{}^T \right)^T. \quad (60)$$

Let  $N$  be the largest of the  $N_j$ . Then, matrix  $\mathcal{B}$  has  $K$  columns and up to  $M \times (N - 4)$  rows, which means that it may become very large. For example, there may be thousands of earthquakes, thousands of velocity parameters, and tens of stations, which means that how to solve Eq. (60) becomes an issue. If the size of the matrix and the computer resources allow it, a solution based on the use of the SVD would be convenient (see Sect. “[Solution of Ill-Posed Linear Problems](#)”). If this approach is not feasible, one can solve Eq. (60) by least squares, which requires solving

$$\mathcal{B}^T \mathcal{B} d\mathbf{v} = \mathcal{B}^T \boldsymbol{\rho} \quad (61)$$

which on account of Eqs. (59) and (60) becomes

$$\left( \sum_{j=1}^M \mathbf{B}'_j{}^T \mathbf{B}'_j \right) d\mathbf{v} = \sum_{j=1}^M \mathbf{B}'_j{}^T \mathbf{r}'_j. \quad (62)$$

This approach however, may have two problems. One is that the resulting matrix may still be too large for the computer facilities available and the other is the possibility of numerical loss of precision due to the matrix multiplications. This second problem can be alleviated through the

use of double precision, although at the expense of longer computer times. For these reasons, for large tomographic problems the ensuing linear systems are solved using iterative matrix solvers (see Subsect. “[Regularization Approach](#)”). Also note that Eq. (62) is not the result of the “accumulation” of individual equations  $\mathbf{B}'_j{}^T \mathbf{B}'_j d\mathbf{v} = \mathbf{B}'_j{}^T \mathbf{r}'_j$ , as it is sometimes stated.

Once Eq. (58) has been solved, the earthquakes must be relocated. To do that Eq. (47) will be used, but to make the analysis more general it will be assumed that both  $P$  and  $S$  wave arrivals are available. Because the  $P$  wave arrivals are independent of the  $S$  wave velocity, and vice versa, we can write two pairs of equations similar to Eqs. (47) and (56), and to solve for  $d\mathbf{v}^P$  and  $d\mathbf{v}^S$ , where the superscripts identify the type of arrivals. After that is done we can write

$$\mathbf{W}_j^P \mathbf{A}_j^P d\mathbf{x}_j = \mathbf{W}_j^P \mathbf{r}_j^P - \mathbf{W}_j^P \mathbf{B}_j^P d\mathbf{v}^P; \quad j = 1, M \quad (63)$$

and

$$\mathbf{W}_j^S \mathbf{A}_j^S d\mathbf{x}_j = \mathbf{W}_j^S \mathbf{r}_j^S - \mathbf{W}_j^S \mathbf{B}_j^S d\mathbf{v}^S; \quad j = 1, M. \quad (64)$$

These two equations are coupled through the common vector  $d\mathbf{x}_j$  and can be written as a single equation as follows

$$\begin{pmatrix} \mathbf{W}_j^P \mathbf{A}_j^P \\ \mathbf{W}_j^S \mathbf{A}_j^S \end{pmatrix} d\mathbf{x}_j = \begin{pmatrix} \mathbf{W}_j^P \mathbf{r}_j^P - \mathbf{W}_j^P \mathbf{B}_j^P d\mathbf{v}^P \\ \mathbf{W}_j^S \mathbf{r}_j^S - \mathbf{W}_j^S \mathbf{B}_j^S d\mathbf{v}^S \end{pmatrix}; \quad j = 1, M. \quad (65)$$

[122]. The matrix on the left-hand side of this equation is small ( $2N \times 4$  at most) and can be solved using any of the standard methods. Once each  $d\mathbf{x}_j$  has been found, it is used to get new initial estimates of origin time and hypocentral estimates (equal to  $T_j + dT_j$ ,  $x_j + dx_j$  and so on). After updating the velocities (i.e., they become  $v_k + dv_k$ ) a new iteration is started. This iterative process is repeated until some stopping criterion is met. An early criterion [144] was based on the use the statistical  $F$  test. With this test it is possible to determine if the decrease in the sum of arrival time residuals squared from a given iteration to the next is statistically significant ([80] and references therein). A simpler, yet effective approach is to stop when the root-mean square of all the travel-time residuals reaches the expected value of the error in the data.

### Computation of Local Travel Times

The iterative determination of a 3-D velocity model and simultaneous determination of earthquake locations requires the availability of software for the computation of

travel times in that type of models. The theoretical solution to this problem using ray theory was well known (see, e. g., [22,44,85,112], and references therein) at the time the seismic tomographic method began to be developed, but because exact ray tracing is a computationally time-consuming task, in most of the earlier tomographic codes approximate ray tracing methods were used. One of them was introduced by Thurber [144], and was based on the use of arcs of circles to approximate the ray paths connecting the source and station. Although the method was fast, its accuracy was questionable. For example, in [42] the weights of arrivals with epicentral distances between 20 and 45 km were decreased and arrivals with distances larger than 45 km were not used. Because this approximate method was popular, the readers of the earlier literature should be aware of the limitations of the method, which can result in errors in the computed velocities and in the locations determined with them.

A significant improvement to Thurber's [144] method was introduced by Um and Thurber [152], who developed a method based on the perturbation of an initial raypath between the source and the station using a ray-theoretical equation. Ray tracing methods that find the raypath between two points by iterative perturbation of an initial estimate are known as *bending methods* (see, e. g., [75] and references therein). The Um and Thurber method is based on an approximate computation of the perturbations, and for this reason it is commonly referred to as a *pseudo-bending method*. However, one of the disadvantages of this method, noted by the authors, is that it is not appropriate for use in media with velocity discontinuities and where nearly constant velocities are present. This method is included in a popular tomographic package, and these limitations should be taken into account when considering the results obtained using it. Pseudo-bending solutions that do not have the limitations of the Um and Thurber method exist [99,115], but their applications to seismic tomography appear to be limited. The Um and Thurber's method was used in combination with Snell's law to account for velocity discontinuities at the Conrad and Moho boundaries and the upper boundary of subducting plates [163], but these boundaries must be known from other studies.

An early example of tomography with exact ray tracing can be found in [61]. In this case the ray tracing problem was solved by Runge–Kutta integration of six simultaneous differential equations that can be derived from the eikonal equation. This approach corresponds to the *shooting method*, which requires the specification of the two angles at the source that define a ray. These two angles must be changed until the end point is within a specified distance from the station. This approach does not introduce

any approximations (beyond those inherent to ray theory), but, as already noted, is more time consuming than the approximate methods. The software based on this approach was later improved [14] by incorporation of the method of parameter separation described in Subsect. “[Decoupling of the Earthquake Location and Tomography Problems](#)”. This software was applied to the 1989,  $M = 7.1$ , Loma Prieta, California, mainshock-aftershock sequence [116]. Another example of tomography with exact ray tracing [79] is based on the division of the earth into constant-velocity blocks and ray tracing based on the use of Snell's law.

A different kind of approach to the computation of travel times, which does not require ray tracing, was introduced by Vidale [157,158]. In his method, the velocity model consists of a set of values assigned to points in a regular equispaced 2-D or 3-D grid and travel times are computed using a finite-difference approximation to the partial derivatives of travel times with respect to the spatial coordinates and plane or circular wavefront approximations. It must be noted, however, that the eikonal equation is not solved using the classical finite-difference method, which is not a simple task and is seldomly used [21]. The output of Vidale's method is a series of wavefronts of minimum travel times. This method is more time consuming than the approximate methods referred to above and has some problems when large velocity contrasts are present. Moreover, the approximations involved may not be adequate in the vicinity of the source, where the wavefronts are highly curved [21]. Vidale's method has been implemented in software for earthquake location in media with 3-D variations [102].

Vidale's [157,158] approach motivated a method developed by Podvin and Lecomte [113], who used Huygens's principle for the computation of travel times. The application of the principle is equivalent to the propagation of local wavefronts, which is done as follows. Let us assume that the four vertices of one of the sides of a unit cube in the grid have known arrival times. These points can be combined into four groups of three adjacent points. Each group defines a plane wavefront, which is used to compute the arrival time at each of the other four corners of the same cube. The selection of the appropriate local wavefront follows a set of pre-established rules. In addition, each point within the grid is the common vertex of eight unit cubes around it. These cubes form a larger cube with the grid point at its center and each side contributes sixteen local wavefronts. Therefore, for each grid point within the grid, ninety-six wavefronts must be considered. The advantage of this approach is that it accounts for the existence of transmitted, diffracted, and head waves and performs well even in the presence of large velocity



contrasts. The computations can be carried out in parallel in a multiprocessor machine or sequentially when only one processor is used. In the latter case the method becomes computationally very time consuming, but its implementation in a tomographic software package [15] has resulted in highly detailed velocity models. Examples are given in Sect. “Examples”. Podvin and Lecomte’s software is becoming very popular and for this reason it is worth noting that the software has a minor flaw that results in time differences for rays moving equal distances to the left and right of the source [151]. For example, for a 0.5 km grid size and source-receiver offsets up to 30 km the difference is between about 0.03 and 0.04 s, with the larger values for offsets less than 10 km. Decreasing the grid size reduces the error.

The two wavefront methods referred to above do not compute raypaths directly. However, in seismic tomography they are needed to associate arrival time residuals with specific velocity grid points. In Podvin and Lecomte’s software, rays are traced from the receiver to the source following a direction opposite to the time gradient, with the endpoint of the ray at most a distance  $h$  from the source, where  $h$  is the grid spacing. Recall that in isotropic media rays are defined as curves whose tangents are everywhere perpendicular to a wavefront (see, e. g., [69,120]).

There is yet another approach to the computation of travel times, namely, the shortest-path method, which is based on concepts borrowed from network theory. In this approach the velocity model is also based on a grid of points and connections are established between close points. Each connection is given a weight, equal to the travel time between them. The shortest path between any two points is that along the connections for which the sum of the weights is smallest. This path is an approximation to the seismic raypath. This method was introduced in seismology in [101] and was further developed in [98]. Additional references can be found in, e. g., [9,21], and [147]. Applications to seismic tomography and 3-D earthquake location can be found in [7,8] and [100], respectively.

To end this section we note that the methods considered here assume a Cartesian coordinate system, so that the curvature of the earth is ignored, i. e., the earth is assumed to be flat. This assumption, however, has limitations (see, e. g., [125,134]), which should be taken into account when dealing with epicentral distances that exceed about 150–200 km. This range is provided as guide only. For a given epicentral distance, the error in travel time generally increases with hypocentral depth. For example, for a depth of 200 km and an epicentral distance of  $2^\circ$  the error is about 0.25 s [125]. When the errors introduced by ignoring the earth’s curvature become important (e. g.,

larger than the arrival-time picking errors) it is advisable to introduce a cut-off epicentral distance beyond which the arrival times are either not used or weighted down (see, e. g. [78] and Subsect. “*P* and *S* Wave Velocity Models for Taiwan”).

### Teleseismic Tomography

The data most commonly used in teleseismic tomographic studies are *P* wave arrivals for the investigation of the mantle and *PcP* and *PKP* arrivals for the investigation of the core-mantle boundary (see, e. g. [96]), although other arrivals have also been used. A major source for the data is the International Seismological Center, which has collected millions of arrivals contributed by hundreds of stations around the world. Although it is possible to perform a simultaneous event location and velocity determination following a formulation similar to that described for local tomography (see, e. g. [137]), the most common approach is to keep the event locations fixed. This is the approach to be describe here.

The use of teleseismic data to determine a 3-D velocity model of the crust and upper mantle underneath a seismic array was introduced by Aki et al. [4]. They divided the portion of the earth being investigated into homogeneous horizontal layers subdivided into blocks, computed the theoretical arrival times for earthquakes having published locations, formed arrival time residuals, and solved for slowness perturbations in each of the blocks with respect to a model with constant slowness in each layer. A simplified version of the arguments presented in [4] follows. Let  $u_k$  and  $u_k^o$  be the actual and reference slowness of block  $k$  and let  $m_k$  be the fractional slowness perturbation, equal to

$$m_k = \frac{u_k - u_k^o}{u_k^o} \equiv \frac{du_k}{u_k^o} \quad (66)$$

where  $du_k$  is equivalent to the  $du$  in Eq. (40) and is assumed to be much smaller than  $u_k$ . Solving for  $u_k$  we get

$$u_k = u_k^o(1 + m_k). \quad (67)$$

Given a particular earthquake-station pair, the actual travel time in the  $k$ th block traveled by the corresponding ray is given by

$$t_k = l_k u_k = l_k u_k^o(1 + m_k) \approx t_k^o + t_k^o m_k \quad (68)$$

where  $l_k$  is the actual length of the ray in the  $k$ th block and the approximate sign arises because the actual path and the path in the reference medium are assumed to be close on account of Fermat’s principle. Then, the difference in

travel time introduced by the slowness perturbation  $m_k$  is given by

$$dt_k = t_k - t_k^o \approx t_k^o m_k. \quad (69)$$

Introducing the right-hand side of Eq. (66) in this expression gives

$$t_k \approx \frac{t_k^o}{u_k^o} du_k = l_k^o du_k \quad (70)$$

where  $l_k^o$  is the raypath length in the  $k$ th block for the reference model. The right-hand side of this expression is equivalent to  $l_{ijk} du_k$  in Eq. (43).

Let us consider the arrival time residual that arises when using teleseismic data and the earthquake locations used are those determined with the reference slowness model. Using the notation introduced earlier we can write

$$r_{ij} = \delta t_j + \sum_k l_{ijk}^o du_k \quad (71)$$

where the subscripts  $i$  and  $j$  denote station and event, and the sum is over all the blocks. A block not traversed by any ray will have the corresponding  $l_{ijk}^o$  equal to zero. The term  $\delta t_j$  includes the contributions to the residuals not accounted for by the second term. For example,  $\delta t_j$  includes errors in event origin time and location, picking errors, possible errors in the reference slowness model outside of the volume investigated, and possible errors in the model parameterization. This  $\delta t_j$  is unknown, but because it is assumed to be effectively constant for a given earthquake, it can be eliminated by averaging  $r_{ij}$  over all the stations that recorded the  $j$ th event and subtracting the result from Eq. (71). The average residual is given by

$$\overline{r_{ij}} = \frac{1}{N_j} \sum_{i=1}^{N_j} r_{ij} = \delta t_j + \sum_k \overline{l_{ijk}^o} du_k \quad (72)$$

where the overbar indicates average, as defined by the first equality. Next, subtracting Eq. (72) from Eq. (71) gives the relative residual

$$r'_{ij} = r_{ij} - \overline{r_{ij}} = \sum_k (l_{ijk}^o - \overline{l_{ijk}^o}) du_k. \quad (73)$$

A similar result can be found in [3]. This equation shows that the slowness perturbations  $du_k$  are obtained by solving a linear system, as expected, and that the main difference with Eq. (43) is the presence of source terms in the latter (aside from a weight factor). Once Eq. (73) is solved, the perturbations are added to the reference slownesses and the tomographic problem is solved. It is important

to note however, that the resulting model is not well resolved vertically. In fact, the effect of a uniform perturbation over a layer cannot be distinguished from a change in  $\delta t_j$ . This problem is intrinsic to the method and cannot be removed [3]. Examples showing the capabilities and limitations of the method can be found in [45]. A potentially severe limitation is the effect introduced by the presence of large sedimentary basins, which usually have significantly smaller velocities than the surrounding rocks. These basins may introduce travel time anomalies as large as the upper mantle residuals [93]. Clearly, the tomographic results will be incorrect if the arrival times are not corrected for large crustal effects, which requires information derived independently (see, e. g., [93,131]).

### Solution of Ill-Posed Linear Problems

Regardless of the type of arrivals considered, the tomographic problem reduces to the solution of equations of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad (74)$$

where  $\mathbf{A}$  is a known  $m \times n$  matrix with information about the model,  $\mathbf{x}$  is an  $n$  vector of unknowns, and  $\mathbf{b}$  is a known  $m$  vector. Solving this problem requires consideration of the question of the existence of a solution, but before proceeding it is convenient to introduce the following definition, due to the French mathematician Hadamard (see, e. g., [47,53], and references therein). The problem represented by Eq. (74) is said to be *well posed* if it has a solution, it is unique, and depends continuously on  $\mathbf{b}$ . The latest condition means that small changes in  $\mathbf{b}$  cause small changes in  $\mathbf{x}$ . When any of these conditions is not satisfied the problem is said to be *ill posed*. For a unique solution of Eq. (74) to exist a necessary (but not sufficient) condition is that  $m \geq n$ . If  $m > n$ , the system is said to be *overdetermined*. It must be noted, however that this condition (to be assumed here) does not imply that a solution exists, or that it is unique (if it exists). For example, if some of the rows of  $\mathbf{A}$  constitute a linearly dependent set (e. g., some of them are linear combinations of other rows) a solution may exist but it will not be unique. On the other hand, a solution may not exist at all. In this case the system is said to be *inconsistent*, a condition that in seismic tomography arises because of errors in both the data and the model. The magnitude of the data errors depend on arrival type. Teleseismic data generally have larger errors than local data, while local  $S$  wave arrivals have larger errors than the corresponding  $P$  arrivals because they are more difficult to identify. Model errors arise because the parameterization of the velocity (or slowness) field does not reproduce

the actual variations appropriately. Although a more faithful representation of the earth would be advantageous, an obvious problem is that the data may not be adequate to resolve the parameters of a more detailed model. Clearly, if the block size or grid spacing used to parameterize the velocity (or slowness) field are too small with respect to the interstation spacing, and the spatial distribution of the seismic sources is not favorable, then a large number of blocks may not be traversed by any ray, and matrix  $\mathbf{A}$  will be very large with a large number of zero entries (i. e., the matrix is said to be *sparse*). For a given set of arrival times, an increase in the number of unknown parameters results in a relative decrease in the number of constraints, and in the resolution of the solution (see below).

A very powerful tool for the solution of the system (74) is provided by the Moore–Penrose generalized inverse  $\mathbf{A}^\dagger$  of  $\mathbf{A}$ . Let  $\mathbf{A}$  have  $p$  nonzero singular values, sorted by decreasing order. Then, using the SVD (see Eq. (49))

$$\mathbf{A}^\dagger = \mathbf{V}\mathbf{A}^\dagger\mathbf{U}^T = \mathbf{V}_p\mathbf{A}_p^{-1}\mathbf{U}_p^T \quad (75)$$

where  $(\mathbf{A}^\dagger)_{ij}$  is equal to  $1/\lambda_i$  if  $\lambda_i \neq 0$  and zero otherwise [3,111]. The second equality follows from the product of partitioned matrices. The generalized inverse solution is given by

$$\mathbf{x}^\dagger = \mathbf{A}^\dagger\mathbf{b} \quad (76)$$

which has the property that is a minimum-length solution (see, e. g., [3]). A more general solution is of the form

$$\mathbf{x} = \mathbf{x}^\dagger + \mathbf{z} \quad (77)$$

where

$$\mathbf{A}\mathbf{z} = \mathbf{0}. \quad (78)$$

The vectors  $\mathbf{z}$  that satisfy Eq. (78) constitute the *null space* of  $\mathbf{A}$ . The column vectors of  $\mathbf{A}$  corresponding to the zero singular value satisfy an equation similar to Eq. (78) and constitute a base for the null space of  $\mathbf{A}$ . Therefore,  $\mathbf{z}$  is a linear combination of the columns of the matrix  $\mathbf{V}_0$ .

The definition of  $\mathbf{A}^\dagger$  is based on a sharp distinction between zero and nonzero singular values. In practice, however, this clear-cut situation does not occur, with some of them nonzero yet much smaller than the largest one. In this case the solution of Eq. (76) may be strongly affected by errors in the data. This question is made more precise when the condition number,  $\kappa$ , of  $\mathbf{A}$  is introduced

$$\kappa = \frac{\lambda_{\text{largest}}}{\lambda_{\text{smallest}}} \geq 1. \quad (79)$$

Then, a perturbation  $\mathbf{db}$  (such as errors) in the data introduces a perturbation  $\mathbf{dx}$  in the solution that satisfies

$$\frac{|\mathbf{dx}|}{|\mathbf{x} + \mathbf{dx}|} \leq \kappa \frac{|\mathbf{db}|}{|\mathbf{b} + \mathbf{db}|} \quad (80)$$

(e. g., [46,49]). Therefore, when  $\kappa$  is large a small change in the data may cause a significant change in the solution. If  $\kappa$  is large the matrix is said to be *ill-conditioned*; otherwise it is said to be *well-conditioned*. Although the terms “large” and “small” are obviously vague, the idea is not. For example, for a given data set and two velocity models leading to matrices  $\mathbf{A}$  having condition numbers that differ by a factor of say 10, the matrix with the larger  $\kappa$  may lead to a solution more affected by the errors in the data.

In the following, ill-conditioned problems will be considered a form of ill-posed problems and the distinction between the two will no longer be made. The question that must be addressed now is how to solve this type of problems. Two approaches are available, one based on the so-called regularization of the problem and the other based on the Bayesian statistics. The two approaches are discussed in detail below.

### Regularization Approach

As noted above, tomographic problems are likely to be ill-posed, and to get a solution it is necessary to recourse to mathematical techniques that turn them well-posed. This process is known as regularization (see, e. g., [53,55,80,148,164]). The reader must be aware, however, that there is a price to be paid; namely one ends up with a family of solutions and to objectively select the most appropriate among them may not be easy or even possible. For a system with a matrix for which the computation of its singular value decomposition is practically feasible, the generalized inverse solution has several advantages, but for most tomographic problems the size of the matrices involved is so large that this option is not practical and will not be pursued here. Instead, we will discuss techniques that lead to systems that can be solved with computationally efficient methods.

Before proceeding we will introduce the following definitions. A square symmetric matrix  $\mathbf{C}$  is said to be *positive semidefinite* if

$$\mathbf{y}^T\mathbf{C}\mathbf{y} \geq 0; \quad \mathbf{y} \neq \mathbf{0}. \quad (81)$$

The matrix  $\mathbf{C}$  is *positive definite* if the  $\geq$  sign in Eq. (81) is replaced by  $>$ .

Let us summarize some of the properties of these matrices (see, e. g., [121]).

- (1) Let  $\mathbf{v}_i$  be an eigenvector of  $\mathbf{C}$  and  $\lambda_i$  its corresponding eigenvalue. Then

$$\mathbf{v}_i^T \mathbf{C} \mathbf{v}_i = \lambda_i \mathbf{v}_i^T \mathbf{v}_i = \lambda_i |\mathbf{v}_i|^2. \quad (82)$$

Because  $|\mathbf{v}_i| > 0$ ,  $\lambda_i \geq 0$  if  $\mathbf{C}$  is positive semidefinite, and  $\lambda_i > 0$  if  $\mathbf{C}$  is positive definite. In the second case the inverse of  $\mathbf{C}$  exists because  $\mathbf{C} = \mathbf{U} \mathbf{A} \mathbf{U}^T$ , with  $\mathbf{U}$  and  $\mathbf{A}$  similar to those in Eq. (49) and Eq. (50) (see, e.g., [103]), and  $\mathbf{C}^{-1} = \mathbf{U} \mathbf{A}^{-1} \mathbf{U}^T$ .

- (2) Any matrix of the form  $\mathbf{B}^T \mathbf{B}$  is either positive definite or semidefinite. Consider

$$\mathbf{y}^T (\mathbf{B}^T \mathbf{B}) \mathbf{y} = (\mathbf{B} \mathbf{y})^T \mathbf{B} \mathbf{y} = |\mathbf{B} \mathbf{y}|^2 \geq 0; \quad \mathbf{y} \neq \mathbf{0}. \quad (83)$$

Therefore,  $\mathbf{B}^T \mathbf{B}$  is at least positive semidefinite. In addition, if the inverse of this matrix exists, all of its eigenvalues will be positive and  $\mathbf{B}^T \mathbf{B}$  will be positive definite. If the inverse does not exist, the matrix will be positive semidefinite.

- (3) Let  $\mathbf{C}$  be a diagonal matrix with positive diagonal elements. Then  $\mathbf{C}$  is positive definite because

$$\mathbf{y}^T \mathbf{C} \mathbf{y} = \sum_i c_i y_i^2 > 0; \quad \mathbf{y} \neq \mathbf{0}, \quad c_i = (\mathbf{C})_{ii} > 0. \quad (84)$$

- (4) Let matrices  $\mathbf{B}$  and  $\mathbf{P}$  be positive semidefinite and definite, respectively, and let  $\lambda^2$  be a scalar. Then  $\mathbf{B} + \lambda^2 \mathbf{P}$  is positive definite because

$$\mathbf{y}^T (\mathbf{B} + \lambda^2 \mathbf{P}) \mathbf{y} = \mathbf{y}^T \mathbf{B} \mathbf{y} + \lambda^2 \mathbf{y}^T \mathbf{P} \mathbf{y} > 0; \quad \mathbf{y} \neq \mathbf{0}. \quad (85)$$

These results will be used to show that the typical methods of solution of Eq. (74) when  $\mathbf{A}^{-1}$  does not exist or is ill-conditioned are based on the solution of a new well-posed problem.

The simplest regularization approach is to constrain (in some sense) the length of the solution vector  $\mathbf{x}$ . Let us consider the minimization of the function

$$\tilde{S}(\mathbf{x}) = w |\mathbf{A} \mathbf{x} - \mathbf{b}|^2 + \mathbf{x}^T \mathbf{D} \mathbf{x} \equiv w S(\mathbf{x}) + Q(\mathbf{x}) \quad (86)$$

with respect to  $\mathbf{x}$ . Here  $w$  is a positive weighting factor,  $\mathbf{D}$  is a symmetric positive definite matrix, and  $S$  and  $Q$  are defined by the identity. This problem was solved by Levenberg [87] and a similar one by Marquardt (who used  $\mathbf{D} = \mathbf{I}$ ) [92] in the context of non-linear problems. The minimization of  $\tilde{S}$  leads to the following equation

$$(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D}) \mathbf{x} = \mathbf{A}^T \mathbf{b}; \quad \lambda^2 = \frac{1}{w} \quad (87)$$

(see, e.g., [121]). A similar equation was introduced by Tihonov [148], who investigated the regularization of ill-

posed linear problems in the context of integral equations. The  $\mathbf{x}$  that solves Eq. (87) is known as the *damped least-squares estimator*. When  $\lambda^2 = 0$ , Eq. (87) corresponds to the *ordinary* least-squares solution.

Note that Eq. (87) is a special case of Eq. (85) with  $\mathbf{B} = \mathbf{A}^T \mathbf{A}$  and  $\mathbf{P} = \mathbf{D}$ , which means that there is a value of  $\lambda^2$  that makes the matrix on the left side of Eq. (87) well conditioned. Therefore, for a given  $\mathbf{D}$ , Eq. (87) will have a family of solutions, which will depend on  $\lambda^2$ . A problem with this equation, however, is that it may be too large for tomographic problems. For this reason, Eq. (87) will be derived by consideration of the following system

$$\begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \quad (88)$$

This new system will be solved using ordinary least squares, which requires solving

$$(\mathbf{A}^T \quad \lambda \mathbf{D}^{1/2}) \begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{D}^{1/2} \end{pmatrix} \mathbf{x} = (\mathbf{A}^T \quad \lambda \mathbf{D}^{1/2}) \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix} \quad (89)$$

where  $\mathbf{D}^{1/2}$  is known as the square root of  $\mathbf{D}$  (see, e.g., [121]). After performing the matrix multiplications we obtain Eq. (87), which shows that the minimization of  $\tilde{S}(\mathbf{x})$  is equivalent to solving Eq. (74) under the constraint

$$\lambda \mathbf{D}^{1/2} \mathbf{x} = \mathbf{0}. \quad (90)$$

Clearly, the  $\lambda$  in this equation is unnecessary, but it is introduced to indicate that it is a weighting factor. We also note that in many applications  $\mathbf{D} = \mathbf{I}$ .

Let us mention some of the properties of the damped least-squares solution in the context of nonlinear linearized inverse problems. This type of problems require an initial estimate of the solution, say  $\mathbf{x}_0$ . If this estimate is poorly chosen, it may lead to a non-convergent iterative process, and for this reason Levenberg [87] decided to minimize  $\tilde{S}$ . Moreover, when dealing with linearized problems,  $S$  is the linearized version of the true residuals. Let  $s$  represent the sum of residuals whose linearization leads to  $S$ . A question investigated by Levenberg [87] regards the relations between  $\tilde{S}$ ,  $S$  and  $s$ , and between the damped and ordinary least-squares solutions, indicated by  $\mathbf{x}_\lambda$  and  $\mathbf{x}^{\text{ls}}$ , respectively. Some of his results are as follows:

$$S(\mathbf{x}_\lambda) < S(\mathbf{x}_0) \quad (91)$$

$$Q(\mathbf{x}_\lambda) < Q(\mathbf{x}^{\text{ls}}) \quad (92)$$

and

$$s(\mathbf{x}_\lambda) < s(\mathbf{x}_0) \quad \text{for some } \lambda^2 \quad (93)$$

as long as  $\mathbf{x}_0$  is not a stationary point of  $s$  [87,121]. Equations (91) and (92) show that it is possible to simultaneously minimize  $S$  and  $Q$  for all  $\lambda^2$ , while Eq. (93) shows that there are values of  $\lambda^2$  that minimize  $s$ , which is the quantity of actual interest to us. Although none of these results is obvious, they are implicitly assumed when solving linearized problems. In addition, if  $\lambda^2$  is large enough, Eq. (87) becomes

$$\mathbf{x} \approx \frac{1}{\lambda^2} \mathbf{D}^{-1} \mathbf{A}^T \mathbf{b} \quad (94)$$

with  $\mathbf{x}$  going to zero as  $\lambda^2$  goes to infinity. Therefore, when solving linearized nonlinear problems  $\lambda^2$  should be relatively large in the early iterations, so that  $\mathbf{x}$  is small enough to assure convergence to a solution. Then, as the iterations proceed the value of  $\lambda^2$  should be decreased gradually until it reaches its regularization value.

Another regularization approach is to limit the “roughness” of the solution by constraining its Laplacian to be equal to zero ([86] and references therein). This constraint can be introduced by writing the Laplacian using a finite-difference approximation

$$6x_{i,j,k} - (x_{i-1,j,k} + x_{i+1,j,k} + x_{i,j-1,k} + x_{i,j+1,k} + x_{i,j,k-1} + x_{i,j,k+1}) = 0 \quad (95)$$

where  $x_{i,j,k}$  is the value of  $\mathbf{x}$  at a grid point (identified by the three indices) and  $x_{i\pm 1,j,k}$ ,  $x_{i,j\pm 1,k}$ , and  $x_{i,j,k\pm 1}$  are the values of  $\mathbf{x}$  at adjacent points. This 3-D constraint is used in [15]. Equation (95) must be translated into a matrix form, which will depend on the number of grid points in the three dimensions. Let  $\mathbf{L}$  indicate this matrix. Then, the constrained system becomes

$$\begin{pmatrix} \mathbf{A} \\ \lambda \mathbf{L} \end{pmatrix} \mathbf{x} = \begin{pmatrix} \mathbf{b} \\ \mathbf{0} \end{pmatrix}. \quad (96)$$

To solve this new system by ordinary least squares we must write an expression similar to Eq. (89), which in turn gives

$$(\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{L}^T \mathbf{L}) \mathbf{x} = \mathbf{A}^T \mathbf{b}. \quad (97)$$

In general,  $\mathbf{L}$  is positive definite (see, e. g., [5,114]) and so is  $\mathbf{L}^T \mathbf{L}$ , which means that this approach also leads to a well-conditioned system.

In practice, Eqs. (88) and (96) can be solved using methods that compute the solution iteratively, such as the ART, SIRT, and LSQR methods. As noted earlier, ART is affected by noise in the data, while SIRT has the disadvantage that it introduces an unwanted scaling of the problem. This question will not be discussed here, but it should be

noted that the term SIRT does not refer to a single technique; rather it refers to a family of iterative methods based on ideas similar to those described in [48]. The earlier application to geophysical problems is described in [39], but the algorithm used there was equivalent to the solution of a problem that was a scaled version of the original one, i. e., an equation similar to Eq. (58) is multiplied by a certain diagonal matrix which is not controlled by the user [73]. A similar problem affects the original SIRT formulation [63]. A general discussion of this question can be found in [154]. The LSQR method [108] was introduced in seismology by Nolet ([104] and references therein), does not have the problems that affect ART and SIRT, and is very well suited for the tomographic method because the matrices involved are highly sparse, which helps speed the computations considerably. For these reasons, the LSQR method is used widely used in seismic tomography (see, e. g., [15,86,163]).

So far we have not discussed how the parameter  $\lambda^2$  should be chosen. This question does not have a definite answer. In their discussion of the damped least-squares method, Lawson and Hanson [83] plotted the norm (i. e. length) of the solution vector versus the norm of the vector of residuals obtained for different values of the damping parameter and noted (for specific examples) that the resulting curve had an approximate L shape. The optimal value of the damping parameter was that corresponding to the corner of the curve at which it goes from nearly vertical to nearly horizontal. The idea here is to minimize both the residual and the length of the solution. This approach is discussed further in [55] and [56], where the logarithms of the two norms, rather than the norms themselves, are used. A somewhat related approach is based on the plot of the data variance versus the solution variance [42]. A recent example of choice of damping parameter motivated by this latter approach is provided in [78].

We end this section with the definition of *resolution*. Regardless of how Eq. (74) is solved, we can write

$$\hat{\mathbf{x}} = \hat{\mathbf{A}} \mathbf{b} \quad (98)$$

where  $\hat{\mathbf{A}}$  can be considered the inverse, in some sense, of  $\mathbf{A}$ . Now write  $\mathbf{b}$  using Eq. (74). This gives

$$\hat{\mathbf{x}} = \hat{\mathbf{A}} \mathbf{A} \mathbf{x} \equiv \mathbf{R} \mathbf{x} \quad (99)$$

where  $\mathbf{R}$  is known as the resolution matrix (see, e. g., [3]) and is defined by the identity. If  $\mathbf{R} = \mathbf{I}$ ,  $\hat{\mathbf{x}} = \mathbf{x}$ . If  $\mathbf{A}$  were known perfectly well, this result would imply that the solution vector  $\hat{\mathbf{x}}$  is exactly equal to the true solution. In practice, in seismic tomography  $\mathbf{A}$  is only poorly known, and



even if  $\mathbf{R} = \mathbf{I}$  there is no reason to say that  $\hat{\mathbf{x}}$  represents the true solution, as sometimes stated. To see that, consider an extreme case. Let us suppose that a velocity model consists of just one block. In this case  $\mathbf{A}$  will have only one column and will be well conditioned, which means that the corresponding Eq. (74) can be solved by least squares. The resulting resolution matrix will be equal to  $\mathbf{I}$ , but it is obvious that the computed solution will not correspond to the true velocity of the earth (except in the unlikely case that the velocity is a constant). If the number of blocks is gradually increased to make the model more realistic, at some point  $\mathbf{A}$  will become ill-posed and a regularized solution must be introduced. In this case from Eq. (87) we get

$$\hat{\mathbf{A}} = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D})^{-1} \mathbf{A}^T \quad (100)$$

and

$$\mathbf{R} = (\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D})^{-1} \mathbf{A}^T \mathbf{A} = \frac{\mathbf{A}^T \mathbf{A}}{\mathbf{A}^T \mathbf{A} + \lambda^2 \mathbf{D}} \neq \mathbf{I} \quad \text{for } \lambda^2 \neq 0. \quad (101)$$

Here  $\mathbf{R}$  was written as a ratio for visual purposes only. Note that  $\mathbf{R}$  is not equal to  $\mathbf{I}$  when  $\lambda^2 = 0$  unless  $(\mathbf{A}^T \mathbf{A})^{-1}$  exists. On the other hand, from Eq. (85) we know that the inverse of the matrix in parentheses always exists because  $\mathbf{D}$  is assumed to be positive definite.

Now consider the implications of a resolution matrix not equal to the identity. Using Eq. (98) we can write the  $i$ th component of the solution vector as

$$\hat{x}_i = \sum_k R_{ik} x_k, \quad (102)$$

so that  $\hat{x}_i$  is equal to a linear combination of some of the components of  $\mathbf{x}$ , not to  $x_i$  (as would be the case if  $\mathbf{R} = \mathbf{I}$ ). This result can be interpreted as follows. As the number of parameters in a velocity model is increased (e. g., the number of blocks is increased), it is expected that it will become a more realistic representation of the actual velocity variations in the earth. However, the available data do not allow the determination of individual values of the parameters; only average values can be computed. We may summarize this situation by saying that in geophysics, as in economics, “there is no such thing as a free lunch”. Note that  $\mathbf{R}$  does not depend on the data. Rather, it is controlled by the distribution of stations and seismic sources. Seismologists have some control on the former (within financial and logistic limits), but not on the latter, which places a strong constraint on the resolution that can be achieved in any given case. The Bayesian approach discussed below is an attempt to circumvent this intrinsic limit.

The computation of the resolution matrix when the number of velocity parameters is very large is computationally too expensive, and in such cases the resolution is estimated using synthetic arrival times using velocity models with simple patterns. An early approach [52] was based on the use of several high velocity areas extending through the whole model depth embedded in a constant velocity medium. The synthetic data were computed using the data raypaths and inverted as the actual data. Comparison of the model obtained by inversion with the model used to generate the synthetic data gives an idea of the resolving capabilities of the distribution of sources and receivers used. In a similar approach, synthetic data were generated for a checkerboard model with alternating positive and negative velocities in three dimensions [72]. In another variation only two blocks of the model had velocities different from those of the rest of the blocks [70]. The checkerboard approach has become very popular, but in principle, other synthetic data sets can be used (see Sect. “Examples”).

Finally, let us note that in addition to its mathematical definition, the term resolution is generally used (i. e., not only in seismology) to indicate level of detail. For example, given two images of the same physical object acquired using different instruments, the one showing more detail is said to have more resolution. The same idea applies to velocity models, and it is not uncommon to see references in the literature to high- or low-resolution models. These terms are useful in a qualitative way, and may not be directly related to what one would call high or low in a mathematical sense. In fact,  $\mathbf{R}$  can be made closer (and even equal) to  $\mathbf{I}$  by decreasing the number of parameters (see, e. g., [74]), which by itself will decrease the level of detail in the model. For example, if a number of blocks in a model is replaced by a single block occupying the same volume, the velocity in the larger block will be the average (in some sense) of the velocities in the smaller blocks. As a consequence, if the small-block velocities are significantly different, the model based on the larger blocks would have lost detail and, thus, will have less resolution, although mathematically it may be larger.

### Bayesian Approach

This approach is based on probability considerations, and to discuss it we need a few basic definitions and results, based on [130]. Let  $X$  and  $Y$  be random variables and  $x$  and  $y$  be elements of a discrete set of real values. Also let  $P_X(x)$  be the probability that the event “ $X = x$ ” occurs and  $P_{X,Y}(x, y)$  be the probability that the event “ $X = x$  and  $Y = y$ ” occurs. Then the *conditional probability* that

the event “ $Y = y$ ” occurs given that the event “ $X = x$ ” has occurred is given by

$$P_{Y|X}(y|x) = \frac{P_{X,Y}(x, y)}{P_X(x)}. \quad (103)$$

Similarly,

$$P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}. \quad (104)$$

These two equations can be combined by elimination of the common factor, giving

$$P_{X|Y}(x|y) = \frac{P_{Y|X}(y|x)P_X(x)}{P_Y(y)}. \quad (105)$$

This result is known as *Bayes' theorem* (or rule).

When the random variables can take on values defined over a continuous interval, which is the case of interest to us, the following definitions are introduced.

(a) *Probability distribution function*  $F_X(x)$ :

$$F_X(x) = P(-\infty < X \leq x) \quad (106)$$

where  $P$  indicates the probability of the event in parentheses.

(b) *Probability density function*  $p_X(x)$ :

$$p_X(x) = \frac{dF_X(x)}{dx}. \quad (107)$$

From Eq. (107) it follows that

$$F_X(x) = \int_{-\infty}^x p_X(a) da. \quad (108)$$

A well known example of probability density function (or pdf, for short) is the Gaussian pdf, introduced below.

The definitions given above are for a single random function. When dealing with a number of them it is convenient to introduce a vector notation. For the case of  $n$  random variables  $X_1, X_2, \dots, X_n$ , we have the following two definitions

$$p_{\mathbf{X}}(\mathbf{x}) = p_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n) \quad (109)$$

and

$$F_{\mathbf{X}}(\mathbf{x}) = \int_{-\infty}^{\mathbf{x}} p_{\mathbf{X}}(\mathbf{a}) d\mathbf{a} \quad (110)$$

where the subscript in  $p_{\mathbf{X}}$  and  $F_{\mathbf{X}}$  is the vector  $\mathbf{X} = (X_1 X_2 \dots X_n)^T$ , the integral symbol represents an  $n$ -fold integral, and  $d\mathbf{a} = da_1 da_2 \dots da_n$ . Then

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \dots \partial x_n} \quad (111)$$

and

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} &= F_{\mathbf{X}}(\mathbf{x} + d\mathbf{x}) - F_{\mathbf{X}}(\mathbf{x}) \\ &= P(\mathbf{x} < \mathbf{X} \leq \mathbf{x} + d\mathbf{x}) \end{aligned} \quad (112)$$

where  $d\mathbf{x}$  is the vector with  $i$ th component given by  $dx_i$ .

For vector random variables Bayes' theorem becomes

$$p_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) = \frac{p_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) p_{\mathbf{X}}(\mathbf{x})}{p_{\mathbf{Y}}(\mathbf{y})}. \quad (113)$$

In the following we will adopt the common practice of dropping the subscripts of the probability density functions.

The application of Bayes' theorem to inverse problems is based on the following interpretation of the theorem. Before an experiment is conducted, a random variable  $\mathbf{X}$  has a given a priori pdf  $p(\mathbf{x})$ . During the experiment data  $\mathbf{y}$  are collected, and as a result  $\mathbf{X}$  has an a posteriori pdf given by  $p(\mathbf{x}|\mathbf{y})$  (see, e. g. p. 36 in [10]). The pdf  $p(\mathbf{y}|\mathbf{x})$  represents the probability that an experiment would have produced the result  $\mathbf{y}$  if the value of the variable had been  $\mathbf{x}$ . To apply Bayes' theorem to our problem, we need to rewrite Eq. (74) adding a random error vector  $\mathbf{e}$

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \mathbf{e} \quad (114)$$

and to assume that  $\mathbf{x}$  and  $\mathbf{b}$  are random variables. In this formulation the data are represented by  $\mathbf{b}$  (which plays the role of  $\mathbf{y}$ ), and  $p(\mathbf{b}|\mathbf{x})$  for a given  $\mathbf{x}$  is the probability of the error  $\mathbf{e}$  [71]. The denominator in Eq. (113) does not depend on  $\mathbf{x}$  and can be ignored. The question that remains is the selection of the form of the probability density functions on the right side of the equation. The standard choice is a *Gaussian* pdf, given, in general, by

$$\begin{aligned} p(\mathbf{x}) &= (2\pi)^{-n/2} |\mathbf{C}_{\mathbf{x}}|^{-1/2} \\ &\cdot \exp \left[ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}) \right] \end{aligned} \quad (115)$$

where  $n$  is as in Eq. (109),  $\mathbf{C}_{\mathbf{x}}$  and  $\boldsymbol{\mu}_{\mathbf{x}}$  are the covariance matrix and mean of  $\mathbf{x}$ , respectively, and the vertical bars indicate determinant. Introducing this expression and a similar one for  $p(\mathbf{e})$  in Eq. (113), and then taking logarithms on both sides of the resulting equation gives

$$\begin{aligned} \ln[p(\mathbf{x}|\mathbf{b})] &= -\frac{1}{2} [(m+n) \ln(2\pi) + \ln |\mathbf{C}_{\mathbf{e}}| + \ln |\mathbf{C}_{\mathbf{x}}| + S] \\ &\quad - \ln(p(\mathbf{y})) \end{aligned} \quad (116)$$

where  $m$  is the number of components of  $\mathbf{b}$  and

$$S = (\mathbf{e} - \boldsymbol{\mu}_{\mathbf{e}})^T \mathbf{C}_{\mathbf{e}}^{-1} (\mathbf{e} - \boldsymbol{\mu}_{\mathbf{e}}) + (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}})^T \mathbf{C}_{\mathbf{x}}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{\mathbf{x}}). \quad (117)$$

Now we will make the additional assumption that  $\mu_e = \mathbf{0}$  and will write  $\mathbf{e} = \mathbf{Ax} - \mathbf{b}$  (using Eq. (114)). Under these conditions Eq. (117) becomes

$$S = (\mathbf{Ax} - \mathbf{b})^T \mathbf{C}_e^{-1} (\mathbf{Ax} - \mathbf{b}) + (\mathbf{x} - \mu_x)^T \mathbf{C}_x^{-1} (\mathbf{x} - \mu_x). \quad (118)$$

The last step is to find the value of  $\mathbf{x}$  that maximizes  $\ln p(\mathbf{x}|\mathbf{b})$  (and thus,  $p(\mathbf{x}|\mathbf{b})$ ) and, therefore, minimizes  $S$ . Taking the derivative of  $S$  with respect to  $\mathbf{x}$  and setting it equal to zero leads to the *maximum a posteriori estimator*  $\mathbf{x}_{\text{MAP}}$ , given by

$$\mathbf{x}_{\text{MAP}} = \mathbf{P} (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{b} + \mathbf{C}_x^{-1} \mu_x) \quad (119)$$

where

$$\mathbf{P} = (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1}. \quad (120)$$

An alternative expression for  $\mathbf{x}_{\text{MAP}}$  is

$$\mathbf{x}_{\text{MAP}} = \mu_x + \mathbf{PA}^T \mathbf{C}_e^{-1} (\mathbf{b} - \mathbf{A}\mu_x). \quad (121)$$

These results are based on [13]. An expression similar to Eq. (119) is given in [130]. The Bayesian approach presented here was introduced in the context of X-ray tomography under the assumption that  $\mu_x = \mathbf{0}$  [71]. An application can be found in [64]. In geophysics a similar approach was introduced in [141] and [142].

Equation (121) is important for two reasons. First,  $\mathbf{x}_{\text{MAP}}$  can be interpreted as the sum of two terms: one corresponding to the prior information about the value of  $\mathbf{x}$  (given by  $\mu_x$ ) and another one that contains the new information provided by the data collected in the experiment. Second, using Eq. (121) and the fact that the mean values of  $\mathbf{x}$  and  $\mathbf{e}$  are  $\mu_x$  and  $\mathbf{0}$ , respectively, it can be shown that the expected value of  $\mathbf{x}_{\text{MAP}}$  is  $\mu_x$ . Therefore, the estimator is *biased* (i. e., its expected value is not  $\mathbf{x}$ ). The significance of this result is that the expected value of the estimator is independent of the data, which is not a desirable feature. Finally, we also note that the covariance matrix of the difference  $\mathbf{x}_{\text{MAP}} - \mathbf{x}$  is equal to the matrix  $\mathbf{P}$  [13].

Now let us consider two special cases. First, no a priori information is used. In this case  $S$  in Eq. (118) is equal to the first term on the right hand side and its minimization leads to the *generalized least-squares estimator*, given by

$$\hat{\mathbf{x}} = (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A})^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{b} \quad (122)$$

(if the inverse exists). The *ordinary least-squares estimator* is obtained from Eq. (122) with  $\mathbf{C}_e = \sigma_e^2 \mathbf{I}$ , where  $\sigma_e$  is a measure of the standard deviation of the errors. The estimator  $\hat{\mathbf{x}}$  has the properties that it is unbiased and minimizes the determinant of its covariance matrix, and does not require a Gaussian pdf (see, e. g., [10]).

Second, let us assume that the a priori information is  $\mu_x = \mathbf{0}$  and  $\mathbf{C}_x \neq \mathbf{0}$ . Then, using Eqs. (119) and (120) we get

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= (\mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{A} + \mathbf{C}_x^{-1})^{-1} \mathbf{A}^T \mathbf{C}_e^{-1} \mathbf{b} \\ &= \mathbf{C}_x \mathbf{A}^T (\mathbf{A} \mathbf{C}_x \mathbf{A}^T + \mathbf{C}_e)^{-1} \mathbf{b}. \end{aligned} \quad (123)$$

A derivation of the second equality is given in [13]. This result agrees with similar results derived in [47] and [74] using different approaches. An alternative derivation can be found in [3]. If we further simplify the problem by letting  $\mathbf{C}_e = \sigma_e^2 \mathbf{I}$  and  $\mathbf{C}_x = \sigma_x^2 \mathbf{I}$ , the first equality in Eq. (123) gives

$$\begin{aligned} \mathbf{x}_{\text{MAP}} &= \left( \frac{1}{\sigma_e^2} \mathbf{A}^T \mathbf{A} + \frac{1}{\sigma_x^2} \mathbf{I} \right)^{-1} \frac{1}{\sigma_e^2} \mathbf{A}^T \mathbf{b} \\ &= \left( \mathbf{A}^T \mathbf{A} + \frac{\sigma_e^2}{\sigma_x^2} \mathbf{I} \right)^{-1} \mathbf{A}^T \mathbf{b}. \end{aligned} \quad (124)$$

This particular form of  $\mathbf{x}_{\text{MAP}}$  has been referred to as the *stochastic inverse* [2,4]. If we now put  $\lambda = \sigma_e/\sigma_x$  we see that this solution is formally similar to the damped least-squares solution obtained from Eq. (87) with  $\mathbf{D} = \mathbf{I}$ .

Another important result concerning  $\mathbf{x}_{\text{MAP}}$ , as given in Eq. (119), is that it can be obtained by simply treating the prior information as a constraint, i. e.,  $\mathbf{x} = \mu_x$ , which then can be added to Eq. (114) in a way similar to what was done in Eqs. (88) and (96), and then finding the generalized least-squares estimator, defined by Eq. (122) [74,136]. When proceeding in this way there is no need to make the Gaussianity assumption, although without it (or some other assumption) it is not possible to derive statistical properties of the estimator.

Finally, we note that although the Bayesian approach seems to be a good way to handle the problem of low resolution that affects many seismic tomography problems, it has some serious drawbacks that should not be overlooked. As noted earlier, the Bayesian solution is biased by the prior information, and if this information is incorrect, the solution will be affected by some amount of error that cannot to be quantified objectively. The assumption that the velocity adjustments can be described by a Gaussian pdf (or any other simple pdf) is introduced by mathematical convenience, not because it is likely to represent the actual 3-D variations of the velocity adjustments within the earth. This applies, particularly, to crustal areas within complicated tectonic settings. Finally, the covariance matrix is also needed, but because it is essentially unknowable, it is frequently assumed that it is diagonal. Again, this is an assumption based on convenience, rather than on scientific facts. A critique of the Bayesian approach as applied

to geophysical inverse problems can be found in [109]. For a somewhat different perspective see [104]. A good overview of the Bayesian method can be found in [57]. An example of the problems that erroneous prior information may introduce is provided in Subsect. “Effect of Inaccurate Prior Information on the Location of Aftershocks”.

## Examples

Here we present two examples of local tomography, one from the San Fernando basin in southern California and one from Taiwan, and also give an example of the problems that can be created when erroneous prior information is used. The tomography software used was written by H. Benz [15] and solves for earthquake locations and slownesses using the formulation described in Subsect. “Local Slowness Tomography” – Subsect. “Decoupling of the Earthquake Location and Tomography Problems”. The slowness model is parameterized in terms of constant velocity blocks and the computation of arrival times is performed using the software of Podvin and Lecomte [113]. As noted in Subsect. “Computation of Local Travel Times”, this software has become popular because of its ability to handle sharp velocity variations accurately. As a consequence, the resulting velocity models show high resolution (i. e. a high level of detail). For the examples presented here, none of the other existing models show such resolution. Additional results derived using Benz’s software can be found in, e. g., [95,105] and [159]. The original software relocated the events with a constant  $v_p/v_s$  ratio but was modified according to Eq. (65). To regularize the solution Eq. (96) is used.

### ***P* Wave Velocity Model for the San Fernando, California, Area**

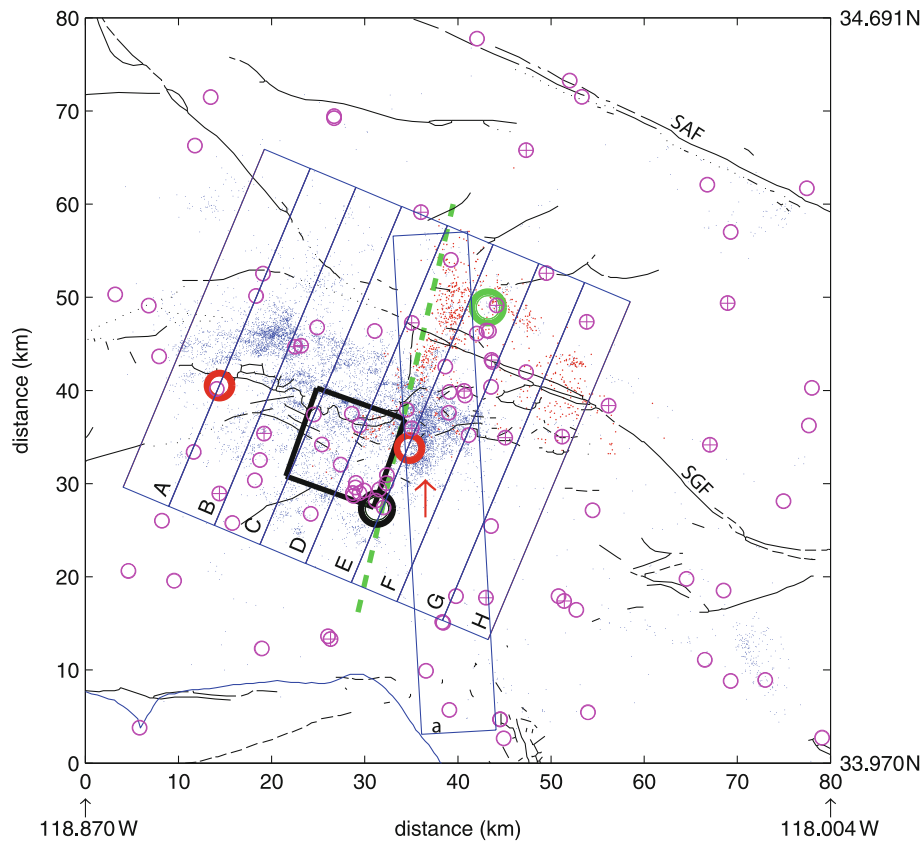
The Los Angeles, California, region is one of the most seismically active in the United States, and because of its extremely large population (over ten million), a large earthquake there may be catastrophic in terms of human and/or economic costs. A major source of seismic hazard there is the presence of large sedimentary basins, which have much lower velocity than their host rocks. As a consequence, they can amplify significantly the ground motion caused by earthquakes. For example, a study published in 1989 [160] showed that the ground motion in the San Fernando and Los Angeles basins can be amplified by factors of about three and four with respect to rock sites, and that the 3-D velocity models available were not capable of generating this amplification. To address this and other problems, the Southern California Earthquake Center (SCEC) supported the development of reference of 3-D *P* and *S*

wave velocity models for the major basins of southern California [91]. These models were constructed using depth and age of sediments data compiled as part of oil and water exploration studies and geological studies. Empirical relations between these two types of data were used to estimate *P* wave velocities. Additional empirical relations between *P* wave velocities, density, and Poisson’s ratio were used to calculate *S* wave velocity. For depths less than 300 m, the *P* and *S* wave velocities were constrained using borehole velocity data. Deep borehole information was used for calibration purposes. For the San Fernando basin four boreholes were available, but the deepest one was only one 3.5 km deep. Therefore, below that depth the model does not have hard constraints. For the rocks outside of the basins an existing 3-D tomographic velocity model [58] was used. This model assigns velocities to points on a grid, which has a 15 km x 15 km horizontal spacing at depths of 1, 4, 6, 10, 15, 17, 22, 31 and 33 km.

The *P* wave velocities in the SCEC model for the Los Angeles basin were compared to sonic log velocities in [140]. This comparison shows that the standard deviation of the velocity differences is about 440 m/s, which is up to 20% of the model velocities. An underestimation of the velocities near the center of the basin and an overestimation near the border was also observed. An additional comparison [139] produced results that are consistent with those in [140].

The 1994,  $M = 6.7$ , Northridge, California, earthquake occurred on a previously unknown blind thrust fault in the San Fernando valley, to the NW of Los Angeles. This was the costliest earthquake in the United States (about US\$ 20,000 million), although the number of deaths was small (58, [143]), thanks to building code provisions and the fact that the event occurred early in the morning (4:30 AM local time). Because of its importance, this earthquake has been extensively studied, but more than ten years after its occurrence, a number of important questions still remain partially or totally unanswered (see, e. g., [123]). One of them is the exact nature of the 3-D velocity variations in the area, as the existing velocity models (see, e. g., [59,97,118,162]) have low resolution.

The velocity model issue has been addressed in [123]. The Benz’s tomography software was applied to events in the Northridge mainshock-aftershock sequence and to aftershocks of the nearby 1971,  $M = 6.6$ , San Fernando earthquake. The data used were 192,421 *P* wave first arrivals from 12,656 events recorded during 1981–2000 by 81 permanent and temporary stations and 799 aftershocks of the San Fernando earthquake recorded by a portable network of 20 stations (Fig. 4). The velocity model covers a volume with a surface area of 80 km by 80 km and a depth



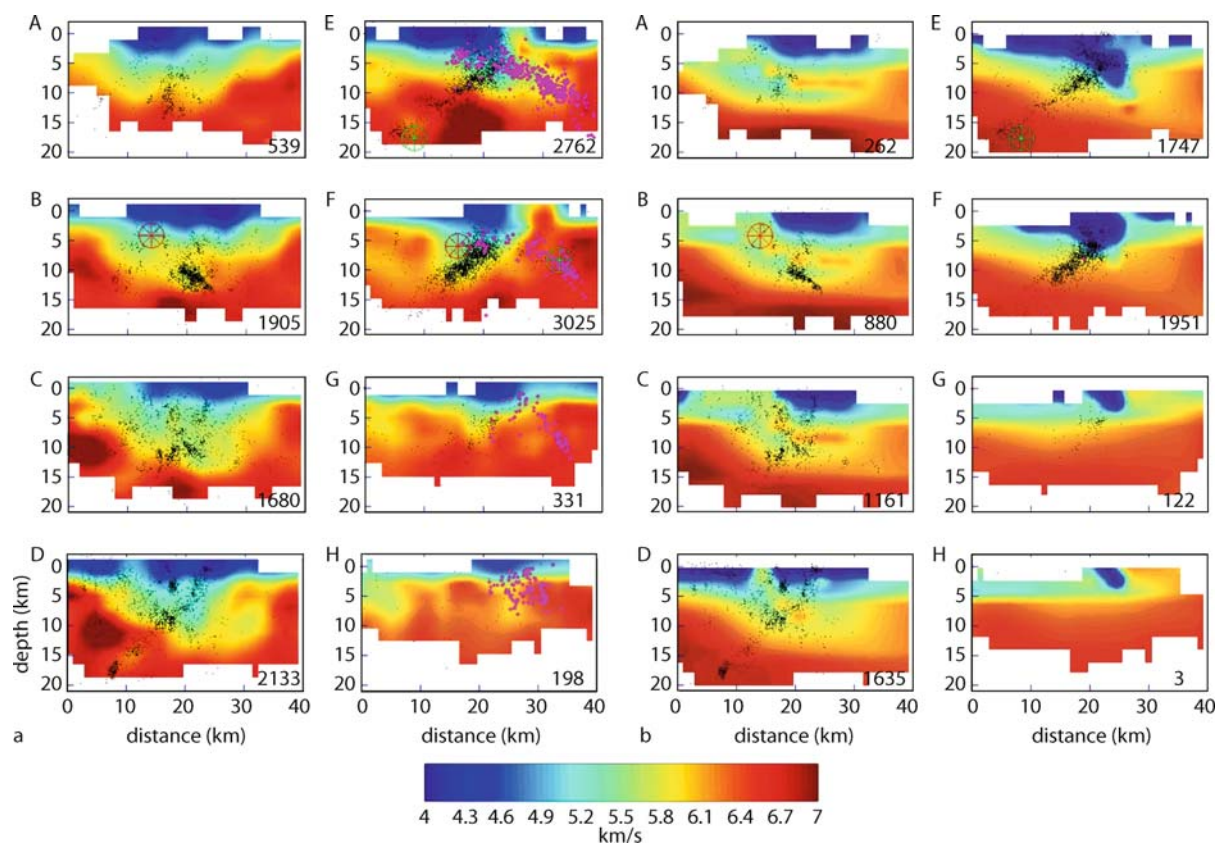
**Tomography, Seismic, Figure 4**

Area around the San Fernando valley, California, for which a 3-D  $P$  wave velocity model has been determined. *Blue dots* indicate the epicenters of 12,656 seismic events recorded during 1981–2000 (mostly aftershocks of the 1994,  $M = 6.7$ , Northridge earthquake). *Red dots* indicate epicenters of 799 aftershocks of the 1971,  $M = 6.6$ , San Fernando earthquake recorded during February–April 1971. The locations shown here were determined as part of the velocity determination and earthquake relocation. The large *black* and *green circles* indicate the epicenters of the Northridge and San Fernando mainshocks, respectively, the latter provided by the U.S. Geological Survey (U.S.G.S.). *Magenta open* and *crossed circles* indicate the stations that recorded the Northridge and San Fernando events, respectively. The large *red circles* indicate the epicenters of the two largest Northridge aftershocks (the location of the eastern one is from the U.S.G.S. and is poorly constrained). *Black lines* indicate faults. The San Andreas and San Gabriel faults are labeled SAF and SGF. The bold box is the projection of the Northridge earthquake fault plane determined in [68] using geodetic data. The events and velocities in boxes A through H are shown in cross section form in Fig. 5. Most of the San Fernando events are to the east of the *green dashed line* while the Northridge mainshock rupture occurred to the west of that line. The prominent band of Northridge seismicity in a roughly N-S direction (identified by the red arrow in box F) did not occur within the area that ruptured during the main shock. The events and velocity within the box labeled a are shown in cross section form in Fig. 5. From [123]

of 24 km and was divided into cubic blocks with sides of 2 km. For the computation of travel times cubic blocks with sides of 1 km were used. The initial locations were computed using a standard 1-D velocity model with three layers having thicknesses of 5.5, 10.5 and 16 km, and corresponding velocities of 5.5, 6.3, and 6.7 km/s. For the inversion the initial velocity model was the 1-D model described in [60] with a small modification (the velocity in the upper 2 km was reduced from 4.8 to 4.5 km/s). A cross section of the model is shown in a subsequent figure. The number of

iterations was ten, and the value of  $\lambda$  in Eq. (96) ranged between 64 (first four iterations) and 20 (last two iterations) with intermediate values in between. Because some of the stations are in the mountains, a depth of  $-2$  km was used as a reference depth for the model. The root-mean square residual for all the events was 0.17 s and 0.07 s for the first and last iterations. Representative velocity cross sections (Fig. 5) show that the resulting model has much more detail than any of the other published models. To make sure that this detail was not artificial the initial model was





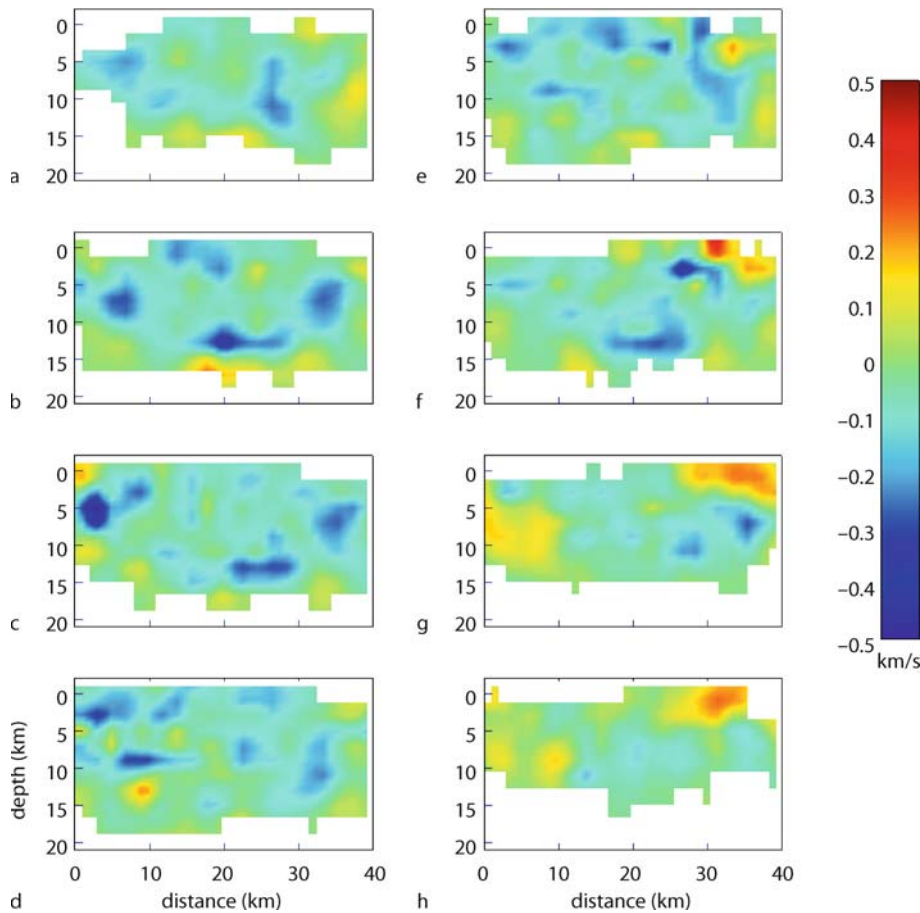
Tomography, Seismic, Figure 5

**a** Depth cross sections for the 3-D  $P$  wave velocity model determined by inversion and the events in the boxes A through H in Fig. 4. The width of the cross sections is 5.3 km. The letters are on the southern ends of the cross sections. The velocities are assigned to the centers of the model blocks and are interpolated along planes passing through the centers of the cross sections. *Black and magenta dots* indicate Northridge and San Fernando aftershocks. The large circled asterisks in cross sections E, B and F indicate the Northridge mainshock, its two largest aftershocks and the San Fernando mainshock (see Fig. 4). The number in the right lower corner of each cross section denotes the number of events. Only the velocity blocks covered by a combined ray length of 0.1 km or more are shown. Note the correlation between seismicity and velocity. The events between about 10 and 20 km in D and E are within high-velocity, basement rocks, and form narrow and well-defined lineations. These events span about 10 km horizontally and basically define the width of the fault that slipped during the main shock. The Northridge aftershocks in F correspond mostly to those indicated by the *red arrow* in Fig. 4. Most of these events are shallower than about 14 km and form a band of seismicity within and near the edge of the basin. The north-dipping events in B below about 10 km probably occurred on the Santa Susana fault. **b** Corresponding cross sections for the SCEC 3-D velocity model [91] and the events that occurred during 1994. From [123]

changed and different data subsets were used, with the result that the most important aspects of the model were robust. In addition, realistic synthetic arrival times were generated with the velocity model and event locations from the last iteration. This involved using the stations that had recorded the observed data (for each event) and applying the original weights. Then the synthetic arrival times were inverted as the actual data (i. e., the same initial locations and velocity model were used). The corresponding velocity model is very close to the input model (Fig. 6). A few areas have velocity differences of up to  $\pm 0.5$  km/s, but this

does not detract from an overall good agreement. Regarding the hypocentral locations, the average difference between the true and computed values is 0.15 km in epicenter and 0.23 km in depth.

The inversion model is also supported by two additional pieces of evidence. One is the SCEC model, which was sampled at the centers of the blocks in the inversion model. However, because the SCEC model has the surface as zero depth and the elevation in the area ranges between 0 and 1.6 km, a direct depth comparison is not strictly possible and for comparison purposes it was re-



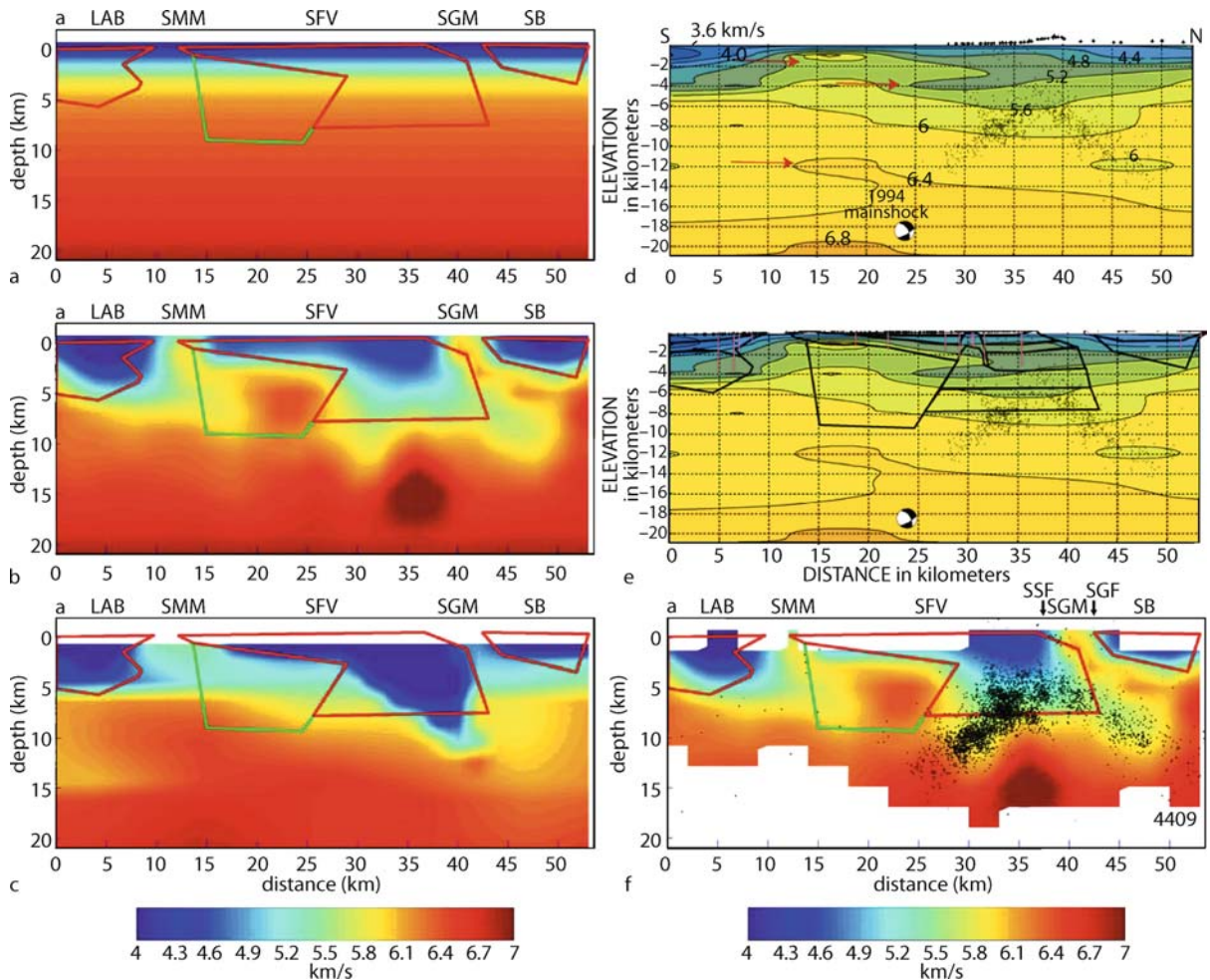
**Tomography, Seismic, Figure 6**

**a** Depth cross sections of the difference between the 3-D model shown in Fig. 5a and the 3-D model determined by inversion of synthetic arrival times generated using the model in Fig. 5a and the event locations determined during the inversion. See text for further details. From [123]

ferred to a base elevation of  $-0.6$  km. As Fig. 5 shows, the inversion model has the main features of the SCEC model, and although there are some obvious differences, some of them come from deficiencies in the SCEC model, as discussed next. A second piece of evidence is the excellent agreement (Fig. 7) with a roughly 55 km long density model derived from the analysis of gravity data [82]. In particular, the low-velocity areas in the model correlate with the Los Angeles, San Fernando and Soledad basins, while high velocities underlie the Santa Monica and San Gabriel mountains. In contrast, the SCEC model does not match the density model equally well. Figure 7 also shows another tomographic model [59], computed for a grid with a  $10 \text{ km} \times 10 \text{ km}$  horizontal spacing at depths of 1, 4, 6, 8, 12, 16 and 20 km. Clearly, this model has a poor resolution and may be affected by significant artifacts, which may be a direct consequence of the large horizontal dimensions of

the blocks. Therefore, for a more direct comparison the inversion with Benz's software was repeated with blocks having horizontal sides of 10 km. As Fig. 8 shows, the corresponding results closely resembles those in Fig. 7b, which seems to indicate that the lack of resolution of the model in Fig. 7d may be due to either the software or the inversion parameters used, or both.

The results of the event relocation are also important. There are two types of differences between the initial and final locations. One is quasi-systematic, with the initial locations on average 1.0 km to the east and 0.5 km to the south of the inversion locations, and 1.5 km deeper. These epicentral differences are a consequence of the low velocities within the basin, which bias the event locations as noted earlier (see Fig. 3). These results are in agreement with a similar shift found using the joint hypocentral determination (JHD) technique [118], and have not been re-



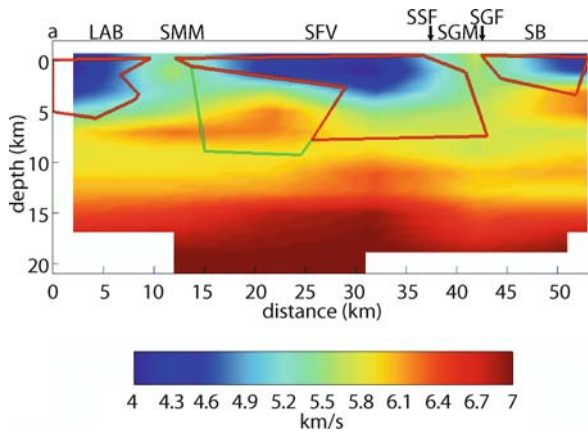
**Tomography, Seismic, Figure 7**

Velocity depth cross section along the center of box a in Fig. 5 for several velocity models. **a** Initial model. **b** Inversion model. All the blocks, regardless of the ray coverage, are shown. The polygons represent the bodies (simplified) used in [82] to match a gravity profile. The *red lines* bound materials with densities ranging between 2.00 and 2.55 g/cc. The area bounded by *red and green lines* corresponds to low-density basement (2.65 g/cc). Elsewhere in the figure the density is 2.71 g/cc. Note the excellent agreement between the extent of the low velocities in the Los Angeles (LAB), San Fernando (SFV) and Soledad (SB) basins and the low-density bodies, as well as the presence of high velocities in the areas of the Santa Monica (SMM) and the San Gabriel (SGM) mountains. **c** SCEC 3-D velocity model. This model does not fit the density model as well as the inversion model. **d** and **e** the 3-D velocity model described in [59] and original density model (from [82]). The contour labels in **d** indicate velocities in km/s. The *red arrows* indicate possible artifacts in the model. *Crosses* indicate the locations of gravity stations. **f** Inversion velocity model for blocks having ray coverage of 0.1 km or larger. The fact that the initial model is one-dimensional indicates that the 3-D velocity variations seen in **a** for the blocks not shown here were determined in earlier iterations. Also shown are the events within box a in Fig. 4. The box width is 8 km. Note that the Northridge and San Fernando aftershocks are underlain by a wedge of basement, with the seismicity occurring where rocks with lower velocities are present. The *arrows* labeled SSF and SGF indicate the positions of the Santa Susana and San Gabriel faults. From [123]

ported in other published tomographic inversion papers. A second difference between the initial and final locations is a considerable reduction in the epicentral scatter seen in the single-event locations. On average, there is a 1.5 km difference between the single-event and inversion epicentral locations.

The combination of a high-resolution velocity model and improved event locations has important tectonic implications. For example, Fig. 5a shows that the most of the seismicity occurred within the sedimentary rocks of the basin. There is little activity within the basement, mostly confined to an area around the fault plane of





**Tomography, Seismic, Figure 8**

Similar to Fig. 7f for a velocity model with blocks having the two horizontal sides and the vertical side 10 km and 2 km long respectively. Comparison with Fig. 7b shows that even though the block size is horizontally much larger, the major features of the model are recovered

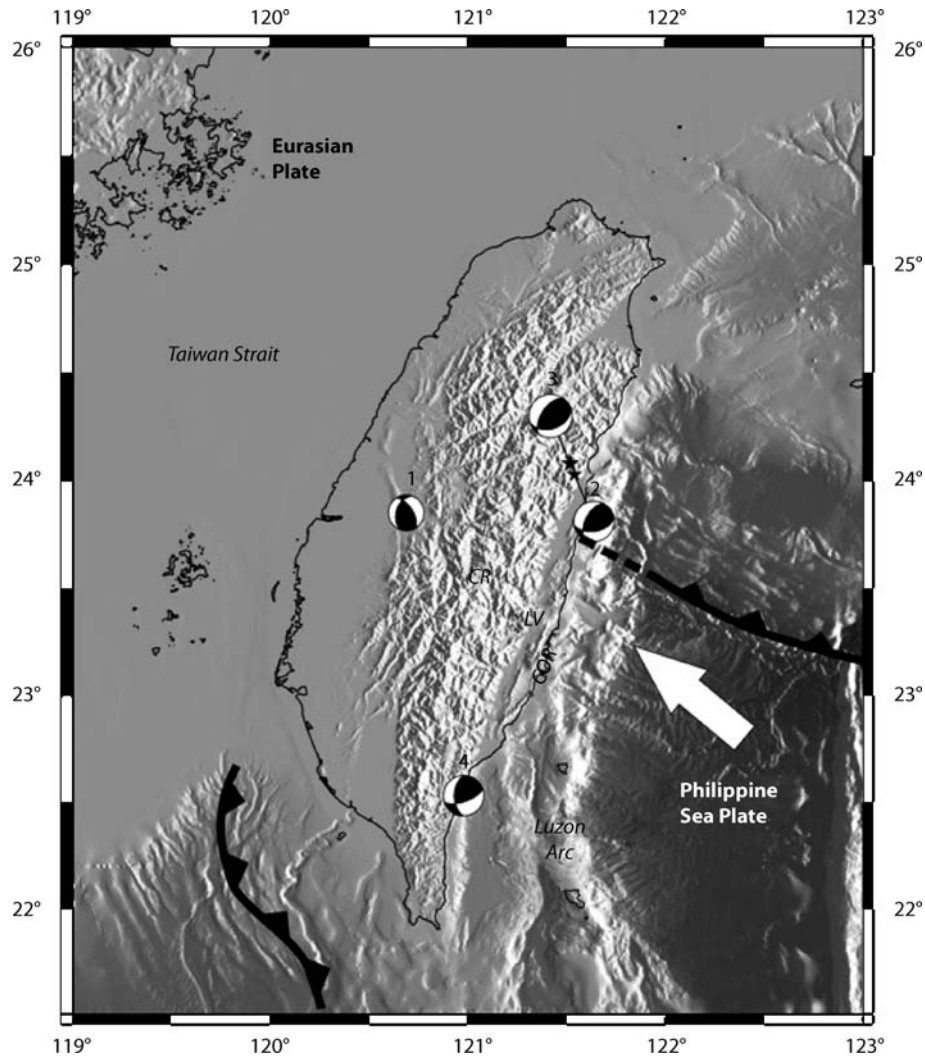
the Northridge earthquake. The relation between the Northridge earthquake aftershocks to the east of the main shock and San Fernando aftershocks has also been clarified; they occur on the flanks of a common high velocity block (Fig. 7f). Other significant results, not discussed here, can be found in [123].

### **P and S Wave Velocity Models for Taiwan**

The seismic hazard in Taiwan is also very high and for this reason a large network of seismic stations covers the island (Fig. 9). In addition, the collision and subduction processes involving the Eurasian and Philippine plates in the Taiwan region have resulted in very strong ongoing tectonic and orogenic activities. As a consequence, Taiwan is the focus of numerous studies, past and present, aimed at getting a better understanding of the nature of the deformation processes there. Some of those studies involve the determination of the crustal and upper mantle velocity structure, which has been investigated by a number of researchers (see, e.g., [90,126,128]). However, although the resulting information has been highly valuable, the resolution of the velocity models was relatively low. This situation was improved by the work described in [77], which resulted in *P* and *S* wave velocity models with much higher resolutions. The main results will be described here (see also [78]). As noted above, the inversion was carried out using Benz's software. The block size was 8 km × 8 km (horizontally) and 2 km in depth for velocity and 2 km × 2 km × 2 km for travel time computation.

The dataset included 69,758 *P* wave arrivals and 42,733 *S* wave arrivals from 6285 events distributed as uniformly as possible recorded by 78 permanent stations as well as 18,502 and 10,789 *P* and *S* wave arrivals, respectively, from 1298 events recorded during two 30-station portable deployments (Fig. 10). The initial and final values of  $\lambda$  in Eq. (96) for the *P* and *S* wave inversions were 128 and 50 and 128 and 20, respectively. The number of iterations was 20 and the initial and final average root mean square residuals were 0.58 s and 0.15 s and 0.67 s and 0.21 s for the *P* and *S* waves, respectively. To avoid exceeding the limitations of the flat-earth approximation (see Subject, "Computation of Local Travel Times"), station-event pairs with epicentral distances greater than 140 km were not used in the inversion. The number of events used in the inversion was less than 5% of the total number of events recorded. Once the velocity models were determined, they were used to relocate all the locatable events using a 3-D location program based on the tomographic software [24].

Representative cross-sectional views of the computed 3-D *P*- and *S*-wave velocity models are shown in Fig. 11. As in the previous example, these models show higher resolution than other published models. The resolution of these models was investigated with a checkerboard test, which shows that most of the original velocity pattern is well recovered to depths of 25 to 30 km for most blocks (Figs. 12, 13). Below that resolution decreases. Further evidence for the reliability of the velocity models comes from the comparison of observed station corrections and those determined using synthetic data generated with those models. This approach was introduced in [116] and [118]. The station corrections were computed using the JHD technique (see, e.g., [116,119]). The JHD corrections carry information on the lateral velocity under an array and are usually much larger than the corresponding station residuals. For example, in Taiwan they can be up to about  $\pm 1$  s and  $\pm 2$  s for *P* and *S* waves (Fig. 14), with the positive corrections corresponding to low-velocity areas (such as sedimentary basins) and the negative corrections associated with high-velocity areas. In general, if the observed and synthetic station corrections do not agree well with each other, it can be stated with confidence that the velocity model used to generate the synthetic data cannot be correct. On the other hand, a good agreement indicates that the inversion model is able to reproduce, at least, the actual velocity variations in an average sense along raypaths. For the Taiwan 3-D velocity model the agreement between the actual and synthetic *P* and *S* wave JHD corrections for two subsets of events is good (see Fig. 14 for an example), which gives confidence to the overall quality of the model.



**Tomography, Seismic, Figure 9**

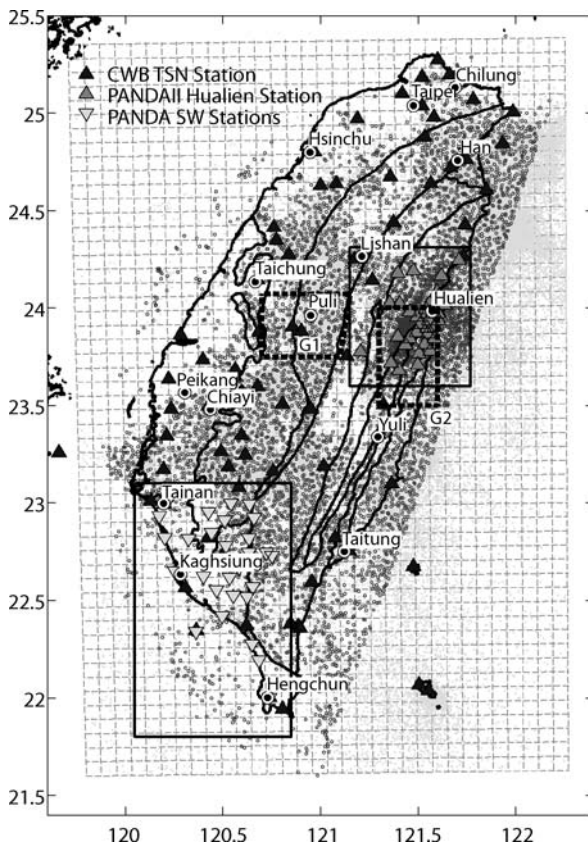
Tectonic setting of Taiwan and surrounding area. The Philippine Sea plate both subducts beneath, and collides with, the Eurasian plate, while the South China Sea sub-plate subducts beneath the Philippine Sea plate in southern Taiwan. The Longitudinal Valley (LV) is the suture zone corresponding to the collision and separates the Coastal Range (COR, part of the Philippine Sea plate) from the Central Mountain Range (CR). The beach ball symbols numbered 2–4 correspond to earthquakes with abnormal  $P_n$  waves recorded by stations along the eastern coast. Number 1 corresponds to an earthquake that does not show those  $P_n$  waves. See text for details.

From [89]

Additional evidence in support of the  $P$  wave model is provided by the analysis of anomalous  $P_n$  waves recorded by stations along the collision suture zone in eastern Taiwan and generated by shallow eastern Taiwan events [89]. These waves can be observed at stations with epicentral distances as small as 60 km. This critical distance is much smaller than that for stations elsewhere in Taiwan and indicates the presence of an elevated Moho (i.e., a thinner crust) along the suture zone. Figure 15 shows travel times for three eastern earthquakes showing the normal

and anomalous  $P_n$  waves and the corresponding Moho depths (having 36–38 km and 22–23 km depth ranges, respectively). In contrast, an earthquake in western Taiwan only shows the normal  $P_n$  waves. The presence of thinner crust along the suture zone agrees well with the 3-D velocity described here [77,78], as Fig. 16 shows. Because this zone is on the edge of the network, the resolution, although still acceptable, is not as good as under the network and the details of the model are not well resolved. However, an inversion of synthetic data similar to that de-





**Tomography, Seismic, Figure 10**

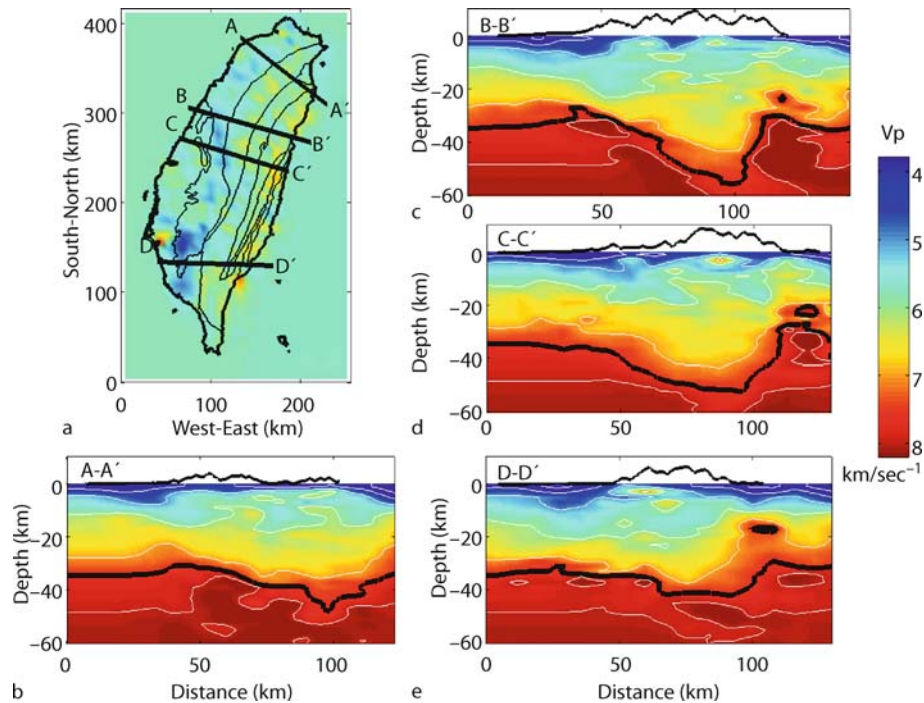
Seismic stations and events used in the 3-D tomographic study of Taiwan. *Solid triangles* indicate stations of the Taiwan Seismic Network (TSN), operated by the Central Weather Bureau (CWB). *Grey and inverted triangles* indicate stations of a portable network (PANDA) deployed in the Hualien and Pingtung areas (*solid rectangles*), respectively. The *small circles* correspond to the epicenters of the events used in the 3-D tomography. The *grey background* corresponds to the epicenters determined by the CWB between 1991 and 2002. The events within the *dashed rectangles* (labeled G1 and G2) were used with the joint hypocentral determination (JHD) technique. The 8 km  $\times$  8 km grid of the 3-D velocity is shown by the *dashed lines*. From [78]

scribed for the Northridge data showed that the main features of the velocity model on the eastern side are well recovered [77] and confirms that the profiles in Figs. 11 and 16 are representative of the actual velocity variations.

### Effect of Inaccurate Prior Information on the Location of Aftershocks

The relocation of the mainshock and aftershocks of the 1989,  $M = 7.1$ , Loma Prieta, California, earthquake, offers a good example of the pitfalls that inaccurate prior information may introduce. This sequence occurred along

the San Andres fault and was located using two 1-D  $P$  wave velocity models, one for stations to the southwest of the fault and one for stations to the northeast of it [37]. This division was based on differences in surface geology across the fault. The velocity of the NE model was up to about 0.1 km/s lower in the upper 1 km, and between 0.2 and 0.5 km/s higher between 1 and 9 km depth, with respect to the SE model. Below that depth the velocity differences did not exceed 0.1 km/s (Fig. 17). The two models were derived using a 1-D velocity inversion program that also included the computation of station corrections. According to [37], these models reflected the presence of elevated basement to the NE of the fault and Tertiary and younger sediments above 9 km depth to the SW of the fault. Note that although the prior information was not used as a quantitative constraint, it was essential to establish the boundary between the two velocity models and was implicitly used to justify the main features of the models. The seismicity was located with a single-event location program and these two velocity models and the corresponding station corrections. The mainshock fault plane inferred from the seismicity, however, had a considerable discrepancy with the fault plane determined using geodetic data, with the former consistently to the northeast of the latter by as much as 3 km. A possible explanation for this discrepancy was the presence of lateral variations in the values of the elastic constants in the vicinity of the fault zone, which were not taken into account when the geodetic fault plane was determined [43]. A more likely explanation, however, is that the discrepancy was the result of a systematic mislocation of the events introduced by the use of incorrect velocity models. This problem was identified in [117], where the geodetic fault plane was compared to the event locations computed with the joint hypocentral determination (JHD) technique and a single velocity model, equal to the average of the two models described above [116]. With these new locations, the discrepancy was greatly reduced except in the southern end of the rupture zone, where the difference was about 1.2 m. This residual discrepancy is probably due to a significant lateral velocity contrast there. Subsequently, the analysis described in [37] was repeated with two modified velocity models [38], with the result that the new locations became similar to those in [116]. Interestingly, the new NE and SW velocity models in [38] are significantly different from their earlier counterparts (Fig. 17), with higher velocities to the SW of the San Andreas fault, rather than to the NE. Thus, these new models contradict the geological arguments used as supporting evidence for the earlier models, which in turn means that the prior information was not quite correct. Also worth noting is the fact that



**Tomography, Seismic, Figure 11**

Cross-sectional views of the Taiwan 3-D  $P$  wave velocity model. The 7.8 km/s contour line (*black line*) marks the depth to the Moho (approximately). The cross section locations are shown in **a**. The *bold lines* below 0 depth indicate the 7.8 km/s contour line. The *bold lines* above 0 depth indicate elevation (with a vertical exaggeration of 3). From [78]

the locations determined using a 3-D velocity model also showed the discrepancy with the geodetic fault plane (see, e. g., [43]). Clearly, this result casts doubts on the reliability of the 3-D model used.

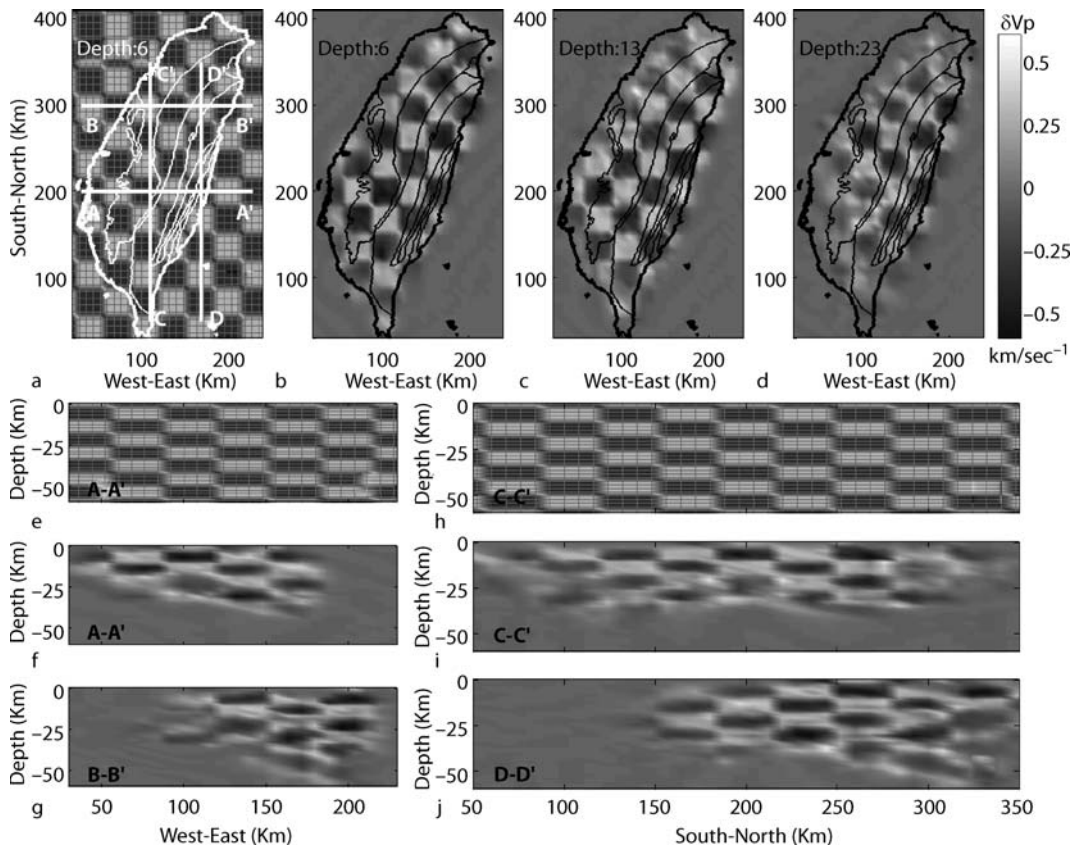
### Future Directions

Seismic tomography began about thirty years ago [2,4,33] and since then the field has grown steadily. However, because of the intrinsic difficulties of wave propagation in the earth and its computer simulation, the ill-posedness of the inverse problem that must be solved (particularly at the global scale), and very limited funding, progress has been slow. Fortunately, in spite of these obstacles the point has been reached where it is possible to establish with some certainty which features of the internal composition of the earth are well resolved and which ones are not, and what needs to be done to improve the existing knowledge (see, e. g. [16,41,129,135,150]). It is clear that progress will come from several fronts and will be fueled by the steady decline in the price of computers and their increased power as well as the availability of relatively inexpensive PC clusters. This development will allow the generation of more realistic synthetic seismograms (see,

e. g. [81]) and the computation of more accurate inverse solutions as well as the computation of resolution and covariance matrices. The integration of tomographic, geodynamic, and mineral physics models will lead, hopefully, to a better understanding of the earth's interior and the processes therein. However, because oceans cover two-thirds of the earth's surface, it will be necessary to make progress in the deployment of ocean-bottom seismographs before these goals can be achieved fully.

### Bibliography

1. Aki K (1993) Overview. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 1–8
2. Aki K, Lee W (1976) Determination of three-dimensional velocity anomalies under a seismic array using first  $P$  arrival times from local earthquakes, 1. A homogeneous initial model. *J Geophys Res* 81:4381–4399
3. Aki K, Richards P (1980) *Quantitative seismology*, vol 2. Freeman, San Francisco
4. Aki K, Christofferson A, Husebye E (1977) Determination of the three-dimensional seismic structure of the lithosphere. *J Geophys Res* 82:277–296
5. Allen M, Isaacson E (1998) *Numerical analysis for applied science*. Wiley, New York

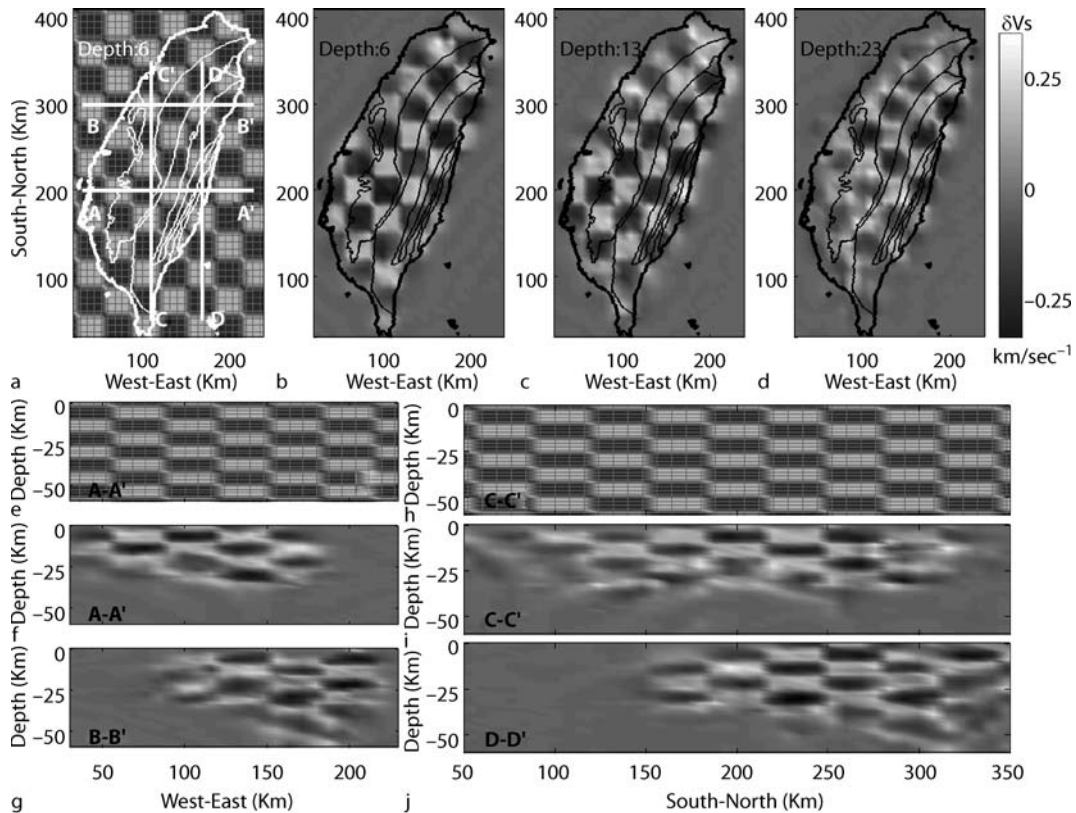


### Tomography, Seismic, Figure 12

Results of the checkerboard test for the Taiwan  $P$  wave velocity model. The checkerboard pattern is shown in a, e and h. The velocity variations across the block boundaries is 0.6 km/s. Synthetic arrival times were generated with this model and the earthquake and station locations used in the inversion of the actual data. The results of the inversion of the synthetic data are shown in map view for different depths in b–d. Cross sectional views are shown in f, g, i, and j. From [78]

- Ambrose J (1973) Computerized transverse axial scanning (tomography): Part 2. Clin Appl Br J Radiol 46:1023–1047
- Bai C-Y, Greenhalgh S (2005) 3-D multi-step travel time tomography: imaging the local, deep velocity structure of Rabaul volcano, Papua New Guinea. Phys Earth Planet Inter 151:259–275
- Bai C-Y, Greenhalgh S (2006) 3D local earthquake hypocenter determination with an irregular shortest-path method. Bull Seism Soc Am 96:2257–2268
- Bai C-Y, Greenhalgh S, Zhou B (2007) 3D ray tracing using a modified shortest-path method. Geophysics 72(4):T27–T36
- Bard Y (1974) Nonlinear parameter estimation. Academic Press, New York
- Barret H, Hawkins W, Joy M (1983) Historical note on computed tomography. Radiology 147:172
- Bates R, Peters T (1971) Towards improvements in tomography. NZ J Sci 14:883–896
- Beck J, Arnold K (1977) Parameter estimation in engineering and science. Wiley, New York
- Benz H, Smith R (1984) Simultaneous inversion for lateral velocity variations and hypocenters in the Yellowstone region using earthquake and refraction data. J Geophys Res 89:1208–1220
- Benz H, Chouet B, Dawson P, Lahr J, Page R, Hole J (1996) Three-dimensional  $P$  and  $S$  wave velocity structure of Redoubt Volcano, Alaska. J Geophys Res 101:8111–8128
- Boschi L, Ampuero J-P, Peter D, Mai P, Soldati G, Giardini D (2007) Petascale computing and resolution in global seismic tomography. Phys Earth Planet Inter 163:245–250
- Bracewell R (1956) Strip integration in radio astronomy. Aust J Phys 9:198–217
- Bracewell R, Riddle A (1967) Inversion of fan-beam scans in radio astronomy. J Astrophys 150:427–434
- Broad W (1980) Riddle of the Nobel debate. Science 207:37–38
- Brooks R, Di Chiro G (1976) Principles of computer assisted tomography (CAT) in radiographic and radioisotopic imaging. Phys Med Biol 5:689–732
- Červený V (2001) Seismic ray theory. Cambridge University Press, Cambridge
- Červený V, Molotkov I, Pšenčík I (1977) Ray method in seismology. Charles University, Prague

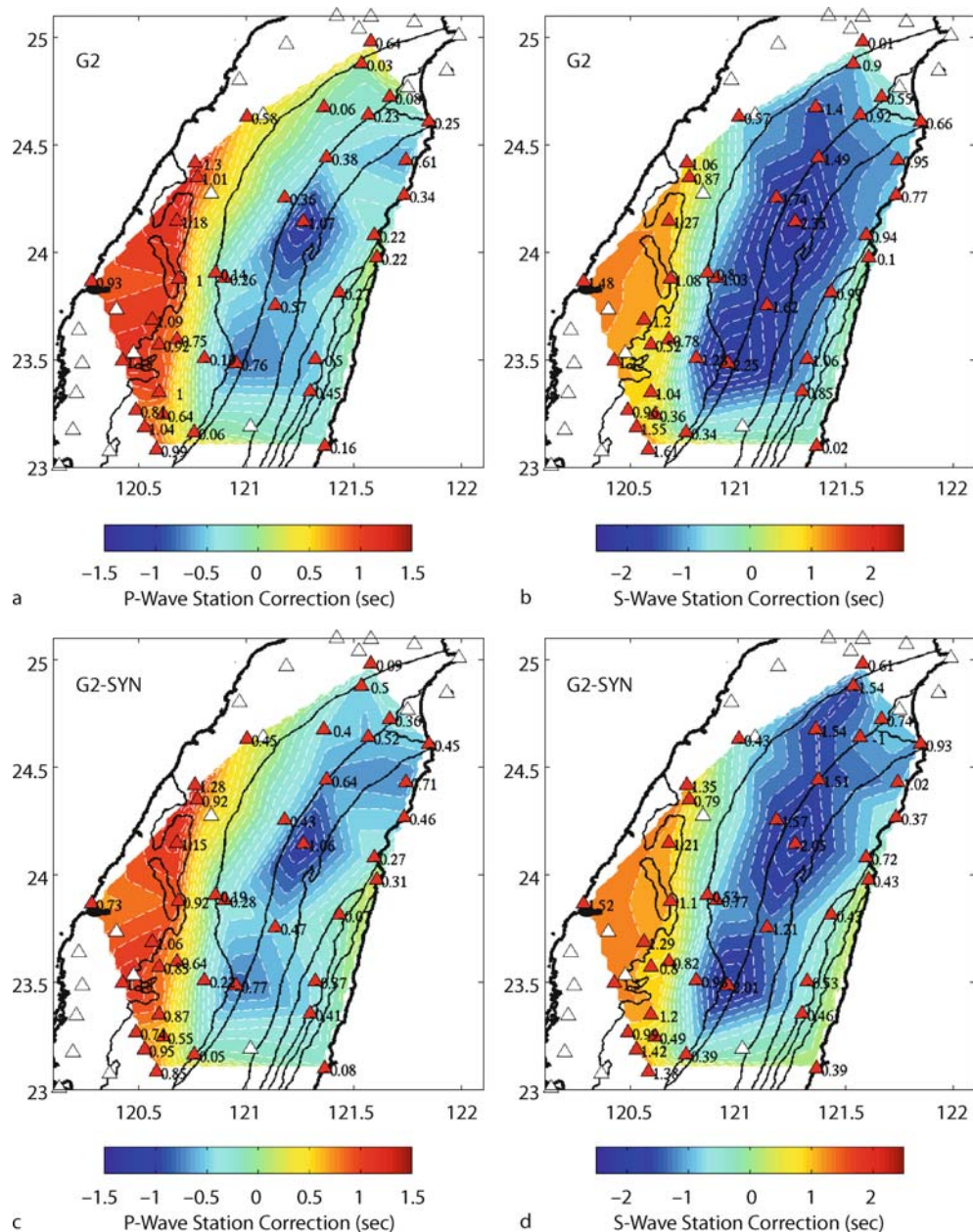




**Tomography, Seismic, Figure 13**

Similar to Fig. 12 for the S wave velocity model. The velocity variations across box boundaries is 0.346 km/s. From [78]

23. Chapman C (1987) The Radon transform and seismic tomography. In: Nolet G (ed) Seismic tomography. Reidel, Dordrecht, pp 25–47
24. Chen H, Chiu J-M, Pujol J, Kim K, Chen K-C, Huang B-S, Yeh Y-H, Chiu S-C (2006) A simple algorithm for local earthquake location using 3D  $V_P$  and  $V_S$  models: test examples in the central United States and in central eastern Taiwan. Bull Seis Soc Am 96:288–305
25. Claerbout J (1985) Imaging the Earth's Interior. Blackwell Scientific Publications, Boston
26. Cormack A (1963) Representation of a function by its line integrals, with some radiological applications. J Appl Phys 34:2722–2727
27. Cormack A (1964) Representation of a function by its line integrals, with some radiological applications, II. J Appl Phys 35:2908–2913
28. Cormack A (1973) Reconstruction of densities from their projections, with applications in radiological physics. Phys Med Biol 18:195–207
29. Cormack A (1980) Recollections of my work with computer assisted tomography. Mol Cell Biochem 32:57–61
30. Cormack A (1982) Computed tomography: some history and recent developments. Proc Symp Appl Math 27:35–42
31. Creager K (1984) Geometry, velocity structure, and penetration depths of descending slabs in the western Pacific. Ph D dissertation, University of California, San Diego
32. Creager K, Boyd T (1992) Effects of earthquake mislocation on estimates of velocity structure. Phys Earth Planet Inter 75: 63–76
33. Crosson R (1976) Crustal structure modeling of earthquake data. 1. Simultaneous least squares estimation of hypocenter and velocity parameters. J Geophys Res 81:3036–3046
34. Crowther R, DeRosier D, Klug A (1970) The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. Proc R Soc Lond Ser A 317:319–340
35. Deans S (1983) The Radon transform and some of its applications. Wiley, New York
36. DeRosier D, Klug A (1968) Reconstruction of three dimensional structures from electron micrographs. Nature 217: 130–134
37. Dietz L, Ellsworth W (1990) The October 17, 1989 Loma Prieta, California, earthquake and its aftershocks: Geometry of the sequence from high-resolution locations. Geophys Res Lett 17:1417–1420
38. Dietz L, Ellsworth W (1997) Aftershocks of the Loma Prieta earthquake and their tectonic implications. In: P Reasenber (ed) The Loma Prieta, California, earthquake of October 17, 1989 – Aftershocks and postseismic effects. US Geol Surv Prof Pap 1550-D, D5-D47
39. Dines K, Lytle R (1979) Computerized geophysical tomography. Proc Inst Electr Electron Eng 67:1065–1073

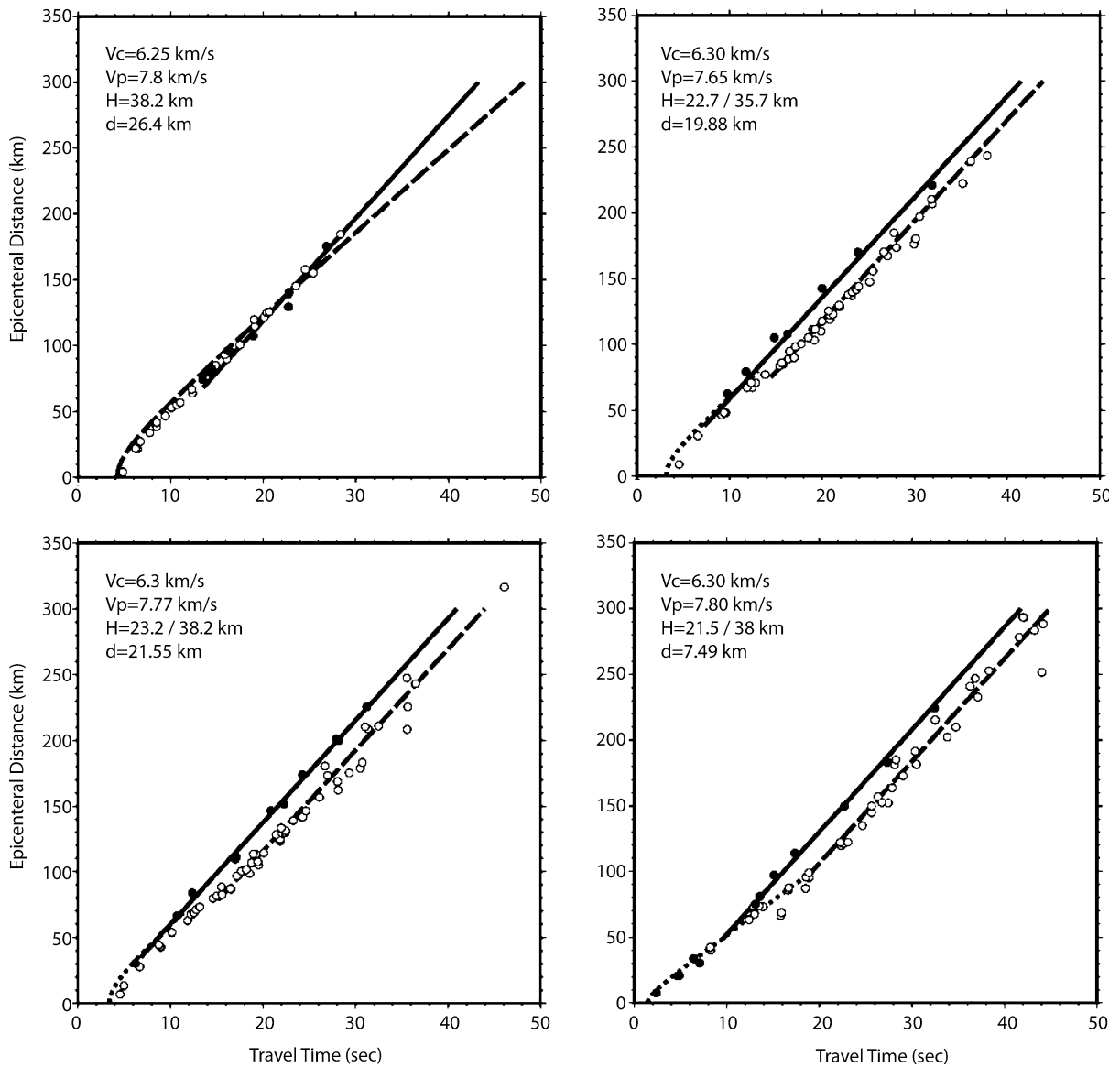


#### Tomography, Seismic, Figure 14

Top: P and S wave station corrections determined using the joint hypocentral determination technique (JHD) and the data in box G2 in Fig. 10. Bottom: JHD station corrections obtained using synthetic data generated with the 3-D velocity model and locations determined by tomographic inversion. From [78]

40. Durrani T, Bisset D (1984) The Radon transform and its properties. *Geophysics* 49:1180–1187; Errata, 1985, 50:884–886
41. Dziewonski A (2003) Global seismic tomography: What we really can say and what we make up. *Geol Soc Am Penrose Conference, Plume IV: Beyond the Plume Hypothesis*, Abstracts (available at: [www.mantleplumes.org/Penrose/PenPDFAbstracts/Dziewonski\\_Adam\\_abs.pdf](http://www.mantleplumes.org/Penrose/PenPDFAbstracts/Dziewonski_Adam_abs.pdf))
42. Eberhart-Phillips D (1986) Three-dimensional velocity structure in northern California Coast Ranges from inversion of local earthquake arrival times. *Bull Seism Soc Am* 76: 1025–1052
43. Eberhart-Phillips D, Stuart W (1992) Material heterogeneity simplifies the picture: Loma Prieta. *Bull Seism Soc Am* 82:1964–1968

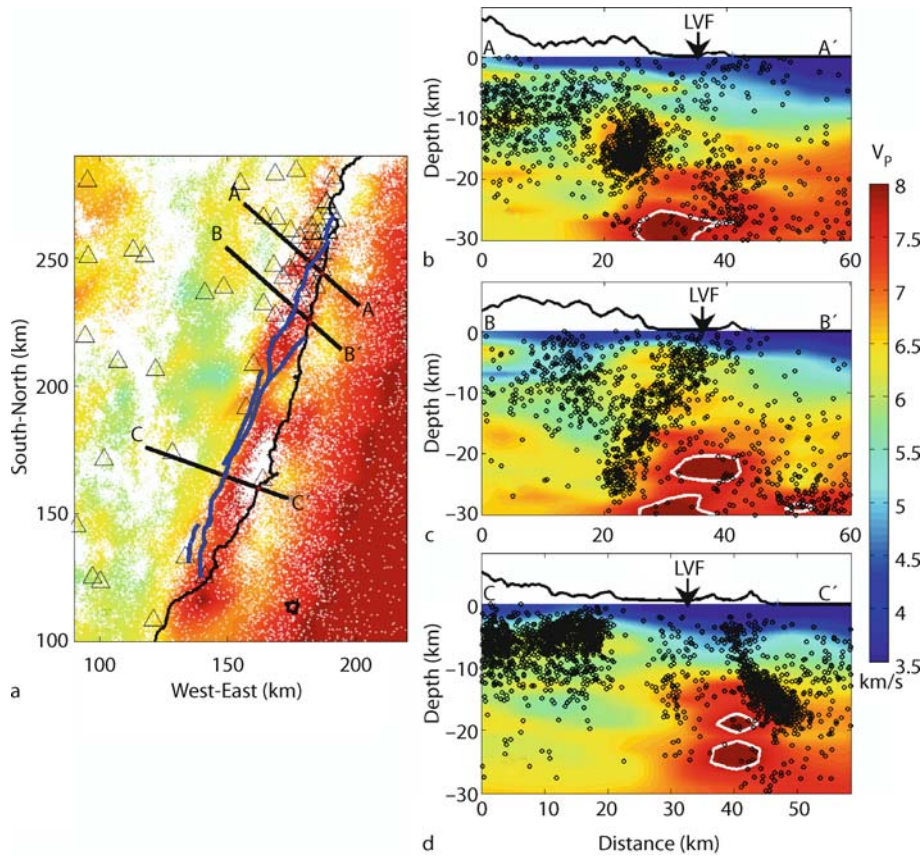




**Tomography, Seismic, Figure 15**

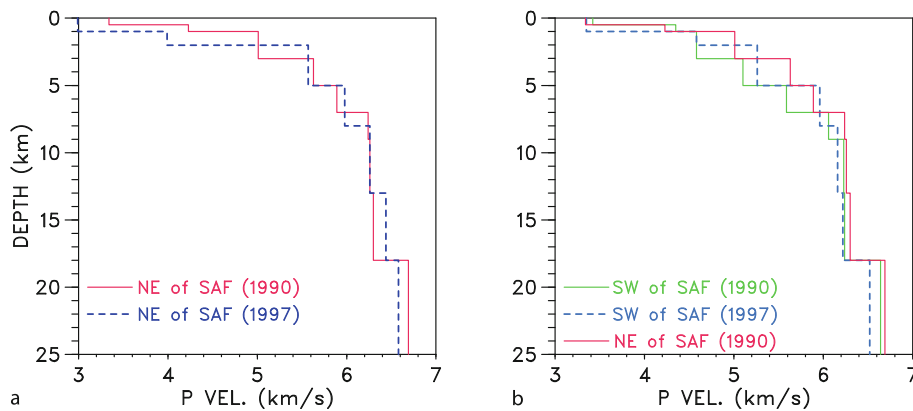
Epicentral distance versus travel time plots for the four earthquakes in Fig. 9. The plot in the *upper left corner* corresponds to the event numbered 1. The *solid* and *dashed lines* are best-fit lines corresponding to the anomalous and normal *Pn* waves. For events 2–4 (along Taiwan's east coast) the anomalous *Pn* waves are seen in the eastern stations. Vc: crustal velocity; Vp: mantle velocity; H: depth(s) to the Moho; d: event depth. From [89]

44. Eliseevnin V (1965) Analysis of rays propagating in an inhomogeneous medium. *Sov Phys Acoust* 10:242–245
45. Evans J, Achauer U (1993) Teleseismic velocity tomography using the ACH method: theory and application to continental-scale studies. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 319–360
46. Forsythe G, Malcolm M, Moler C (1977) *Computer methods for mathematical computations*. Prentice-Hall, Englewood Cliffs
47. Franklin J (1970) Well-posed extensions of ill-posed linear problems. *J Math Anal Appl* 31:682–716
48. Gilbert P (1972) Iterative methods for the three-dimensional reconstruction of an object from projections. *J Theor Biol* 36:105–117
49. Gill P, Murray W, Wright M (1981) *Practical optimization*. Academic Press, London
50. Gordon R (1974) A tutorial on ART. *Inst Electr Electron Eng Trans Nucl Sci* NS-21:78–93



**Tomography, Seismic, Figure 16**

Left: Map view of the Taiwan  $P$  wave velocity model at a depth of 20 km (from [78]). The coordinate system is as in Fig. 11a. Blue lines: surface trace of active faults. White dots: epicentral locations. Right: Cross-sectional views of  $P$  wave velocity across the Longitudinal Valley (LVF) and seismicity. The locations of the cross sections are shown on the left. Note the elevated high-velocity oceanic upper mantle along the entire collision suture zone. The 7.8 km/s contour is indicated by the white line around the dark-red areas. From [89]



**Tomography, Seismic, Figure 17**

$P$  wave velocity models used in [37] and [38] to locate the 1989,  $M = 7.1$ , Loma Prieta, California, earthquake and its aftershocks. Left: Models for stations to the NE of the San Andreas fault. Right: Models for stations to the SW of the San Andreas fault. For a comparison, the NE 1990 model is also shown. After [119]

51. Gordon R, Bender R, Herman G (1970) Algebraic reconstruction techniques (ART) for three-dimensional electron microscopy and X-ray photography. *J Theor Biol* 29:471–481
52. Grand S (1987) Tomographic inversion for shear velocity beneath the North American plate. *J Geophys Res* 92:14065–14090
53. Groetsch C (1993) Inverse problems in the mathematical sciences. Vieweg, Braunschweig
54. Gubbins D (1981) Source location in laterally varying media. In: Husebye E, Mykkeltveit S (eds) Identification of seismic sources – Earthquake or underground explosion. Reidel, Dordrecht, pp 543–573
55. Hansen P (1992) Analysis of discrete ill-posed problems by means of the L-curve. *SIAM Rev* 34:561–580
56. Hansen P (1994) Regularization Tools: A Matlab package for analysis and solution of discrete ill-posed problems. *Num Algorithms* 6:1–35; (Software available at: <http://www2.imm.dtu.dk/~pch/Regutools/>)
57. Hanson K (1987) Bayesian and related methods in image reconstruction from incomplete data. In: Stark H (ed) Image recovery: theory and applications. Academic, Orlando, pp 79–125
58. Hauksson E (2000) Crustal structure and seismicity distribution adjacent to the Pacific and North America plate boundary in southern California. *J Geophys Res* 105:13875–13903
59. Hauksson E, Haase J (1997) Three-dimensional  $V_P$  and  $V_P/V_S$  velocity models of the Los Angeles basin and central Transverse Ranges, California. *J Geophys Res* 102:5423–5453
60. Hauksson E, Jones L, Hutton K (1995) The 1994 Northridge earthquake sequence in California: seismological and tectonic aspects. *J Geophys Res* 100:12335–12355
61. Hawley B, Zandt G, Smith R (1981) Simultaneous inversion for hypocenters and lateral velocity variations: an iterative solution with a layered model. *J Geophys Res* 86:7073–7086
62. Herman G (1980) Image reconstruction from projections. Academic Press, New York
63. Herman G, Lent A (1976) Iterative reconstruction algorithms. *Comput Biol Med* 6:273–294
64. Herman G, Hurwitz H, Lent A, Lung H-P (1979) On the Bayesian approach to image reconstruction. *Inform Contr* 42:60–71
65. Hounsfield G (1973) Computerized transverse axial scanning (tomography): Part 1. Description of system. *J Br Radiol* 46:1016–1022
66. Hounsfield G (1979) Computed medical imaging, Nobel Lecture. (available at: [nobelprize.org/nobel\\_prizes/medicine/laureates/1979/hounsfield-lecture.pdf](http://nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-lecture.pdf))
67. Hounsfield G (1980) Autobiography. In: Wilhelm O (ed) The Nobel Prizes 1979. The Nobel Foundation, Stockholm (available at: [nobelprize.org/nobel\\_prizes/medicine/laureates/1979/hounsfield-autobio.html](http://nobelprize.org/nobel_prizes/medicine/laureates/1979/hounsfield-autobio.html))
68. Hudnut K et al (1996) Co-seismic displacements of the 1994 Northridge, California, earthquake. *Bull Seism Soc Am* 86(1B):S19–S36
69. Hudson J (1980) The excitation and propagation of elastic waves. Cambridge University Press, Cambridge
70. Humphreys E, Clayton R (1988) Adaptation of back projection tomography to seismic travel times problems. *J Geophys Res* 93:1073–1085
71. Hurwitz H (1975) Entropy reduction in Bayesian analysis of measurements. *Phys Rev A* 12:698–706
72. Inoue H, Fukao Y, Tanabe K, Ogata Y (1990) Whole mantle P-wave travel time tomography. *Phys Earth Planet Inter* 59:294–328
73. Ivansson S (1983) Remark on an earlier proposed iterative tomographic algorithm. *J Geophys R Astr Soc* 75:855–860
74. Jackson D (1979) The use of a priori data to resolve non-uniqueness in linear inversion. *J Geophys R Astr Soc* 57:137–157
75. Julian B, Gubbins D (1977) Three-dimensional seismic ray tracing. *J Geophys* 43:95–113
76. Kak A, Slaney M (1988) Principles of computerized tomographic imaging. Inst Electr Electron Eng Press, New York
77. Kim K-H (2003) Subsurface structure, seismicity patterns, and their implication to tectonic evolution in Taiwan. Ph D dissertation, University of Memphis, Memphis
78. Kim K-H, Chiu J-M, Pujol J, Chen K-C, Huang B-S, Yeh Y-H, Shen P (2005) Three-dimensional  $V_P$  and  $V_S$  structural models associated with the active subduction and collision tectonics in the Taiwan region. *J Geophys Int* 162:204–220
79. Koch M (1985) Non-linear inversion of local seismic travel times for the simultaneous determination of 3D-velocity structure and hypocenters – application to the seismic zone Vrancea. *J Geophys* 56:160–173
80. Koch M (1993) Simultaneous inversion for 3-D crustal structure and hypocenters including direct, refracted and reflected phases – I Development, validation and optimal regularization of the method. *J Geophys Int* 112:385–412
81. Komatitsch D, Tsuboi S, Tromp J (2005) The spectral-element method in seismology. In: Levander A, Nolet G (eds) Seismic earth: array analysis of broadband seismograms. Geophysical Monograph Series, vol 157. Am Geophys Union, Washington DC, pp 205–227
82. Langenheim V, Griscom A, Jachens R, Hildenbrand T (2000) Preliminary potential-field constraints on the geometry of the San Fernando basin, southern California. US Geol Survey Open-File Report 00–219
83. Lawson C, Hanson R (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs
84. Lee W, Pereyra V (1993) Mathematical introduction to seismic tomography. In: Iyer H, Hirahara K (eds) Seismic tomography. Chapman, London, pp 9–22
85. Lee W, Stewart S (1981) Principles and applications of microearthquake networks. Academic Press, New York
86. Lees J, Crosson R (1989) Tomographic inversion for three-dimensional velocity structure at Mount St. Helens using earthquake data. *J Geophys Res* 94:5716–5728
87. Levenberg K (1944) A method for the solution of certain nonlinear problems in least squares. *Quart Appl Math* 2:164–168
88. Lewitt R (1983) Reconstruction algorithms: transform methods. *Proc Inst Electr Electron Eng* 71:390–408
89. Liang W-T, Chiu J-M, Kim K (2007) Anomalous Pn waves observed in eastern Taiwan: implications of a thin crust and elevated oceanic upper mantle beneath the active collision-zone suture. *Bull Seism Soc Am* 97:1370–1377
90. Ma K-F, Wang J-H, Zhao D (1996) Three-dimensional seismic velocity structure of the crust and uppermost mantle beneath Taiwan. *J Phys Earth* 44:85–105
91. Magistrale H, Day S, Clayton R, Graves R (2000) The SCEC southern California reference three-dimensional seismic velocity model version 2. *Bull Seism Soc Am* 90(6B):S65–S76

92. Marquardt D (1963) An algorithm for least-squares estimation of nonlinear parameters. *J Soc Ind Appl Math* 11:431–441
93. Martin M, Ritter J, CALIXTO Working Group (2005) High-resolution teleseismic body-wave tomography beneath SE Romania – I. Implications for three-dimensional versus one-dimensional crustal correction strategies with a new crustal velocity model. *J Geophys Int* 162:448–460
94. Meskó A (1984) Digital filtering: applications in geophysical exploration for oil. Wiley, New York
95. Monna S, Filippi L, Beranzoli L, Favali P (2003) Rock properties of the upper-crust in Central Apennines (Italy) derived from high-resolution 3-D tomography. *Geophys Res Lett* 30(61): 1–4 doi:10.1029/2002GL016780
96. Morelli A (1993) Teleseismic tomography: core-mantle boundary. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 163–189
97. Mori J, Wald D, Wesson R (1995) Overlapping fault planes of the (1971) San Fernando and 1994 Northridge, California earthquakes. *Geophys Res Lett* 22:1033–1036
98. Moser T (1991) Shortest path calculation of seismic rays. *Geophysics* 56:59–67
99. Moser T, Nolet G, Snieder R (1992) Ray bending revisited. *Bull Seismol Soc Am* 82:259–288
100. Moser T, Van Eck T, Nolet G (1992) Hypocenter determination in strongly heterogeneous earth models using the shortest path method. *J Geophys Res* 97:6563–6572
101. Nakanishi I, Yamaguchi K (1986) A numerical experiment on nonlinear image reconstruction from first-arrival times for two-dimensional island arc structure. *J Phys Earth* 34:195–201
102. Nelson G, Vidale J (1990) Earthquake locations by 3-D finite-difference travel times. *Bull Seism Soc Am* 80:395–410
103. Noble B, Daniel J (1977) *Applied linear algebra*. Prentice-Hall, Englewood Cliffs
104. Nolet G (1993) Solving large linearized tomographic problems. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 227–247
105. Okubo P, Benz H, Chouet B (1997) Imaging the crustal magma sources beneath Mauna Loa and Kilauea Volcanoes, Hawaii. *Geology* 25:867–870
106. Oldendorf W (1961) Isolated flying spot detection of radio density discontinuities – Displaying the internal structural pattern of a complex object. *IRE Trans Biomed Elec BME-8*: 68–72
107. Oransky I (2004) Obituary. Sir Godfrey N Hounsfield. *Lancet* 364:1032
108. Paige C, Saunders M (1982) LSQR: An algorithm for sparse linear equations and sparse least square problems. *ACM Trans Math Softw* 8:43–71
109. Parker R (1994) *Geophysical inverse theory*. Princeton University Press, Princeton
110. Pavlis G, Booker J (1980) The mixed discrete-continuous inverse problem: application to the simultaneous determination of earthquake hypocenters and velocity structure. *J Geophys Res* 85:4801–4810
111. Penrose R (1955) A generalized inverse for matrices. *Proc Camb Phil Soc* 51:406–413
112. Pereyra V, Lee W, Keller H (1980) Solving two-point seismic-ray tracing problems in a heterogeneous medium. *Bull Seism Soc Am* 70:79–99
113. Podvin P, Lecomte I (1991) Finite difference computation of traveltimes in very contrasted velocity models: a massively parallel approach and its associated tools. *J Geophys Int* 105:271–284
114. Press W, Teukolsky S, Vetterling W, Flannery B (1992) *Numerical Recipes*. Cambridge University Press, Cambridge
115. Prothero W, Taylor W, Eickemeyer J (1988) A fast, two-point, three-dimensional raytracing algorithm using a simple step search method. *Bull Seism Soc Am* 78:1190–1198
116. Pujol J (1995) Application of the JHD technique to the Loma Prieta, California, mainshock-aftershock sequence and implications for earthquake location. *Bull Seism Soc Am* 85: 129–150
117. Pujol J (1996) Comment on: “The 1989 Loma Prieta earthquake imaged from inversion of geodetic data” by Thora Árnadóttir and Paul Segall. *J Geophys Res* 101:20133–20136
118. Pujol J (1996) An integrated 3D velocity inversion – joint hypocentral determination relocation analysis of events in the Northridge area. *Bull Seism Soc Am* 86(1B):S138–S155
119. Pujol J (2000) Joint event location – The JHD technique and applications to data from local seismic networks. In: Thurber C, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer, Dordrecht, pp 163–204
120. Pujol J (2003) *Elastic wave propagation and generation in seismology*. Cambridge University Press, Cambridge
121. Pujol J (2007) The solution of nonlinear inverse problems and the Levenberg-Marquardt method. *Geophysics* 72(4): W1–W16
122. Pujol J et al (1989) 3-D P- and S-wave velocity structure of the Andean foreland in San Juan, Argentina, from local earthquakes. *Eos Trans Am Geoph Union* 70(43):1213
123. Pujol J, Mueller K, Peng S, Chitupolu V (2006) High-resolution 3D P-wave velocity model for the East Ventura–San Fernando basin, California, and relocation of events in the Northridge and San Fernando aftershock sequences. *Bull Seism Soc Am* 96:2269–2280
124. Ramachandran G, Lakshminarayanan A (1971) Three-dimensional reconstruction from radiographs and electron micrographs: application of convolutions instead of Fourier transforms. *Proc Natl Acad Sci USA* 68:2236–2240
125. Ratchkovsky N, Pujol J, Biswas N (1997) Relocation of earthquakes in the Cook Inlet area, south central Alaska, using the joint hypocenter determination method. *Bull Seism Soc Am* 87:620–636
126. Rau R-J, Wu F (1995) Tomographic imaging of lithospheric structures under Taiwan. *Earth Planet Lett* 133:517–532
127. Robinson E (1982) Spectral approach to geophysical inversion by Lorentz, Fourier, and Radon transforms. *Proc Inst Electr Electron Eng* 70:1039–1054
128. Roecker S, Yeh Y, Tsai Y (1987) Three-dimensional P and S wave velocity structures beneath Taiwan: deep structure beneath an arc-continent collision. *J Geophys Res* 92:10547–10570
129. Romanowicz B (2003) Global mantle tomography: Progress status in the past 10 years. *Annu Rev Earth Planet Sci* 31: 303–328
130. Sage A, Melsa J (1971) *Estimation theory with applications to communications and control*. McGraw-Hill, New York
131. Sandoval S, Kissling E, Ansorge J, Svekopalapko Seismic Tomography Working Group (2003) High-resolution body wave tomography beneath the Svekopalapko array: I, A priori three-dimensional crustal model and associated traveltime effects on teleseismic wave fronts. *J Geophys Int* 153:75–87



132. Shepp L, Kruskal J (1978) Computerized tomography: the new medical X-ray technology. *Am Math Mon* 85:420–439
133. Shepp L, Logan B (1974) The Fourier reconstruction of a head section. *Inst Electr Electron Eng Trans Nucl Sci NS-21*:21–43
134. Snoke J, Lahr J (2001) Locating earthquakes: at what distances can the Earth no longer be treated as flat? *Seism Res Lett* 72:538–541
135. Soldati G, Boschi L, Piersanti A (2006) Global seismic tomography and modern parallel computers. *Ann Geophys* 49: 977–986
136. Sorenson H (1980) Parameter estimation: principles and problems. Dekker, New York
137. Spakman W (1993) Iterative strategies for non-linear travel time tomography using global earthquake data. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 190–226
138. Stanton L (1969) Basic medical radiation physics. Appleton-Century-Crofts, New York
139. Stewart J, Choi Y, Graves R, Shaw J (2005) Uncertainty of southern California basin depth parameters. *Bull Seism Soc Am* 95:1988–1993
140. Süss M, Shaw J (2003) *P* wave seismic velocity structure derived from sonic logs and industry reflection data in the Los Angeles basin, California. *J Geophys Res* 108(13):1–18 [doi:10.1029/2001JB001628](https://doi.org/10.1029/2001JB001628)
141. Tarantola A, Valette B (1982) Inverse problems = quest for information. *J Geophys* 50:159–170
142. Tarantola A, Valette B (1982) Generalized nonlinear inverse problems solved using the least squares criterion. *Rev Geophys Space Phys* 20:219–232
143. Teng T-L, Aki K (1996) Preface to the 1994 Northridge earthquake special issue. *Bull Seism Soc Am* 86(1B):S1–S2
144. Thurber C (1983) Earthquake locations and three-dimensional crustal structure in the Coyote Lake area, central California. *J Geophys Res* 88:8226–8236
145. Thurber C (1992) Hypocenter-velocity structure coupling in local earthquake tomography. *Phys Earth Planet Inter* 75: 55–62
146. Thurber C (1993) Local earthquake tomography: velocities and  $V_P/V_S$  – theory. In: Iyer H, Hirahara K (eds) *Seismic tomography*. Chapman, London, pp 563–583
147. Thurber C, Kissling E (2000) Advances in travel-time calculations for three-dimensional structures. In: Thurber C, Rabinowitz N (eds) *Advances in seismic event location*. Kluwer, Dordrecht, pp 71–99
148. Tikhonov A (1963) Solution of incorrectly formulated problems and the regularization method. *Sov Math* 4:1035–1038; (Note: a more common transliteration of this author's Russian name is Tikhonov.)
149. Titchmarsh W (1948) Introduction to the theory of Fourier integrals. Oxford University Press, Oxford
150. Trampert J, Van der Hilst R (2005) Towards a quantitative interpretation of global seismic tomography. In: Van der Hilst R, Bass J, Matas J, Trampert J (eds) *Earth's deep mantle: structure, composition, and evolution*. Geophysical Monograph Series, vol 160. Am Geophys Union, Washington DC, pp 47–62
151. Tryggvason A, Bergman B (2006) A traveltimes reciprocity discrepancy in the Podvin & Lecomte time3d finite difference algorithm. *J Geophys Int* 165:432–435
152. Um J, Thurber C (1987) A fast algorithm for two-point seismic ray tracing. *Bull Seism Soc Am* 77:972–986
153. Van der Hilst R, Engdahl E (1992) Step-wise relocation of ISC earthquake hypocenters for linearized tomographic imaging of slab structure. *Phys Earth Planet Inter* 75:39–53
154. Van der Sluis A, Van der Vorst H (1987) Numerical solution of large, sparse linear algebraic systems arising from tomographic problems. In: Nolet G (ed) *Seismic tomography*. D Reidel, Dordrecht, pp 49–83
155. Vidale R (2002) In search for the third dimension: from radiostereoscopy to three-dimensional imaging. *JBR-BTR* 85:266–270
156. Van Tiggelen R, Pouders E (2003) Ultrasound and computed tomography: spin-offs of the World Wars. *JBR-BTR* 86: 235–241
157. Vidale J (1988) Finite-difference calculation of travel times. *Bull Seism Soc Am* 78:2062–2076
158. Vidale J (1990) Finite-difference calculation of travel times in three dimensions. *Geophysics* 55:521–526
159. Villaseñor A, Benz H, Filippi L, De Luca G, Scarpa R, Patanè G, Vinciguerra S (1998) Three-dimensional *P*-wave velocity structure of Mt. Etna, Italy. *Geophys Res Lett* 25: 1975–1978
160. Wald D, Graves R (1998) The seismic response of the Los Angeles basin, California. *Bull Seism Soc Am* 88:337–356
161. Woodhouse J, Dziewonski A (1989) Seismic modelling of the Earth's large-scale three-dimensional structure. *Phil Trans Roy Soc Lond A* 328:291–308
162. Zhao D, Kanamori H (1995) The 1994 Northridge earthquake: 3-D crustal structure in the rupture zone and its relation to the aftershock locations and mechanisms. *Geophys Res Lett* 22:763–766
163. Zhao D, Hasegawa A, Horiuchi S (1992) Tomographic imaging of *P* and *S* wave velocity structure beneath northeastern Japan. *J Geophys Res* 97:19909–19928
164. Zhdanov M (2002) Geophysical inverse theory and regularization problems. Elsevier, Amsterdam

## Topological Complexity of Molecules

DUŠANKA JANEŽIČ<sup>1</sup>, ANTE MILIČEVIĆ<sup>2</sup>,  
SONJA NIKOLIĆ<sup>3</sup>, NENAD TRINAJSTIĆ<sup>3</sup>

<sup>1</sup> National Institute of Chemistry, Ljubljana, Slovenia

<sup>2</sup> The Institute of Medical Research and Occupational Health, Zagreb, Croatia

<sup>3</sup> The Rugjer Bošković Institute, Zagreb, Croatia

### Article Outline

Glossary

Definition of the Subject

Introduction

The Hierarchical Approaches to Molecular Complexity

Criteria for Topological Complexity Indices

Examples of Topological Complexity Indices



## The Usefulness of the Reviewed Topological Complexity Indices to Assess the Complexity of Molecular Graphs

### Future Directions

### Bibliography

## Glossary

**Complexity** The word ‘complexity’ is made up from the Latin roots – ‘com’ meaning ‘together’ and ‘plectere’ meaning ‘to plait’. Complexity is a difficult concept to define. One way to define it is as follows. A system is complex if it consists of a number of components interacting with each other in many different ways, so that these interactions sometimes lead to unexpected collective (emergent) properties. Hence, increasing complexity is associated with an increasing number of components in a system and with increasing versatility of their interactions.

**Complexity (descriptor) index** A complexity (descriptor) index is a number that is used to assess the complexity of a system.

**Graph** A graph, sometimes called a non-directed graph and, usually denoted by  $G$ , is a mathematical object which consists of two non-empty sets: one set, denoted usually by  $V$ , is a set of elements called vertices and the other, usually denoted by  $E$ , is a set of unordered pairs of distinct elements of  $V$  called edges. Thus,  $G = (V, E)$ . In directed graphs,  $E$  is a set of ordered pairs of elements of  $V$ . In multigraphs more than one edge can join two vertices. A graph  $G$  is connected if every pair of vertices is joined by a path. If there is no path between two vertices in  $G$ , then  $G$  is the disconnected graph having two or more components. A graph  $G$  is a planar graph if it can be drawn in a plane in such a way that no two edges intersect. A dual  $G^*$  of a planar graph  $G$  can be constructed in this way: Place a vertex in each region of  $G$ , including the exterior region, and if two regions have an edge  $e$  in common, join the corresponding vertices by an edge  $e^*$  crossing only  $e$ . The inner dual is a subgraph of dual which does not contain the vertex corresponding to the exterior region. A graph is complete if each pair of its vertices is adjacent.

**Graph invariant** An invariant of a graph  $G$  is a number associated with  $G$  which has the same value for any graph isomorphic to  $G$ .

**Graph-theoretical distance** The length of the shortest path between two vertices in a graph  $G$  is the graph-theoretical distance.

**Laplacian matrix** The Laplacian matrix  $L$  of a graph  $G$  is a real symmetric matrix whose diagonal elements are

the vertex-degrees of  $G$  and off-diagonal elements are  $-1$  if vertices are connected in  $G$ , otherwise zero.

**Molecular graph** A molecular graph is a 2-dimensional representation of a molecule and is generated by replacing atoms and bonds with vertices and edges, respectively. All molecular graphs in this article are connected graphs.

**Path** A path is a sequence of edges, each edge sharing one vertex with the sequence-adjacent edges and sharing no vertices with any other edge. The length of a path is the number of edges it contains.

**Subgraph** A subgraph of a graph  $G$  is a graph with all of its vertices and edges in  $G$ . A spanning subgraph is a subgraph containing all the vertices of  $G$ .

**Topological complexity** Topological complexity is the complexity of graphs and is determined completely by the particular adjacency of a graph's vertices – the issues of metrics and geometry are not relevant.

**Tree** A tree is an acyclic graph – one which has no cycles. A spanning tree of a graph  $G$  is a connected, acyclic subgraph containing all the vertices of  $G$ .

**Vertex-adjacency matrix** The vertex-adjacency matrix  $A$  is a 0-1 matrix representing graph. Entry 1 denotes adjacent vertices, all other entries are zero.

**Walk** A walk in a graph  $G$  is an alternating sequence of vertices and edges of  $G$ , such that each edge  $e$  begins and ends with the vertices immediately preceding and following  $e$  in the sequence. The length of a walk is the number of edges it contains. The number of walks of length  $l$  beginning at vertex  $i$  and ending at vertex  $j$  is given by the  $i, j$ -element of the  $l$ th power of the vertex-adjacency matrix, whilst the number of the self-returning walks of length  $l$  is given by the  $i, i$ -element of the  $l$ th power of the vertex-adjacency matrix.

## Definition of the Subject

The concept of complexity has intrigued people from the beginning of history, but only in our times have attempts to quantify it begun to appear with a new science emerging – the science of complexity. Manifestation of complexity can be found everywhere in nature and life [1] and different levels of complexity are encountered in arts, humanities and sciences [2].

Like many concepts in chemistry, the concept of complexity appears to be a fuzzy, but useful concept [3]. The fuzziness of this concept has not prevented chemists from attempting to quantify it [4].

Here we are concerned with topological complexity of molecules. It should be stated at the outset that there are

different levels of complexity regarding the structure of molecules [5], i. e., elemental or compositional or one-dimensional complexity (which is determined by the partition of a graph's vertices into classes of different type), topological or two-dimensional complexity and dynamical or three-dimensional complexity (a complexity feature of fluctuating molecules).

The topological complexity index will be abbreviated as TCI, and this symbol will be used throughout the article.

## Introduction

Initial work on the topological complexity of molecules was based on use of information theory in studies of the complexity of living systems and the most of the pioneering papers appeared in the *Bulletin of Mathematical Biophysics* (this journal changed the title in 1972 to the *Bulletin of Mathematical Biology*). The idea behind this approach was reductionistic: the complexity of living systems is largely determined by the complexity of the constituent organic molecules. More recently, Bonchev and Buck [6] carried out a comparative analysis of the topological structure of molecules and biological networks and pointed out both similarities and differences.

The first attempt to quantify the complexity of biological molecules *via* an index (descriptor) was carried out in 1953 by Dancoff and Quastler [7] within the framework provided by the Shannon information theory [8]. The Dancoff–Quastler complexity index was based on the information of the atomic composition of a molecule. Morowitz [9] in 1955 produced another complexity index that combined the information on the atomic composition with that on type and multiplicity of bonds in the molecule. In the same year, Rashevsky [10] pointed out that there are molecules with the same atomic composition and the same atomic valencies but differ in the way in which atoms are connected, that is, they differ by their topology. Rashevsky called his complexity measure as the topological information content – this was the first structural complexity measure devised for molecules. He was also first to use graphs to compute the information content of organic molecules. Karreman [11] immediately pointed out that in living processes chemical reactions between the molecules rather than the molecules themselves are responsible for the complexity of living systems. Hence, he proposed the topological information content of reactions to be used in assessing the complexity in living systems. Karreman did not use the term graph, but graphs appear in his paper under the term the topology of a molecule. A year later (1956), Trucco [12,13] re-

formulated the Rashevsky complexity index in terms of the automorphism group of the molecule, since the topology of a molecule is not always sufficient. He used graph-theoretical and group-theoretical arguments in his work. Mowshowitz [14] in 1968 used Rashevsky's and Trucco's ideas to propose structural information content as an index of the relative complexity of graphs. His paper is based on concepts from graph theory, group theory and combinatorics.

The first to study the topological complexity of graphs without using information theory was Minoli [15]. Bonchev and Trinajstić [16], however, again employed information-theoretical arguments to derive the branching index. Though they did not specify that their index is a complexity index, their approach was soon after extended by Bertz [17] who introduced an information theoretic complexity index that became known as the Bertz (complexity) index [18]. These three initial papers and Bertz's two subsequent papers [19,20] stimulated a number of authors to search for a qualified index that will account for the complexity of molecules; many following the path set by Minoli, deriving complexity indices from consideration of graphs representing molecules [4] without depending upon information theory.

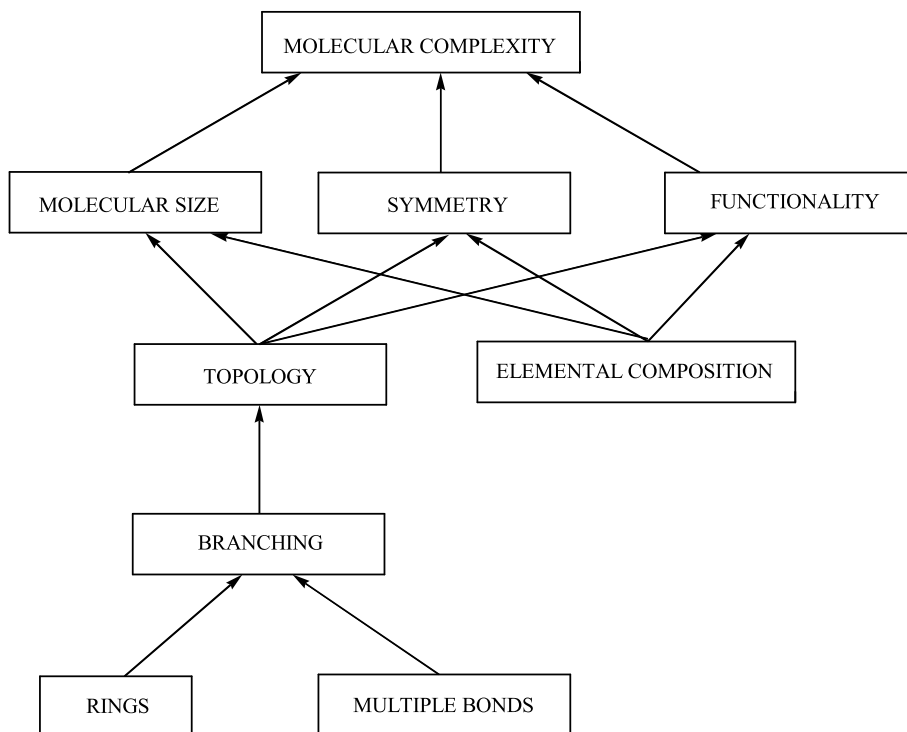
In the text that follows, first we present several hierarchical approaches to molecular complexity, then various criteria for TCIs and a selection of TCIs. The last part of the text contains a comparison between the selected TCIs for trees, cycles, isomeric trees on six vertices and planar polycyclic graphs on four vertices, thoughts on the possible development of the subject and a number of references pertinent to the material discussed.

## The Hierarchical Approaches to Molecular Complexity

The first hierarchical approach to molecular complexity was proposed by Bertz in 1983 [19]. His approach, consisting of four hierarchical levels of molecular complexity, is presented in Fig. 1.

As seen from this figure, the Bertz approach contains both topological (branching, rings, multiple bonds) and non-topological (molecular size, symmetry, functionality, elemental composition) structural features. However, the Bertz approach has really only three hierarchical levels of molecular complexity since branched structures are not more complex than rings and multiple bonds and cannot be derived from them.

Four years later (1987), Bonchev and Polansky [21] proposed the hierarchical organization of the total complexity of chemical systems consisting of five levels (see



Topological Complexity of Molecules, Figure 1

A hierarchical organization of molecular complexity as proposed by Bertz [19]

Fig. 2) and then the hierarchical levels of topological complexity of chemical systems.

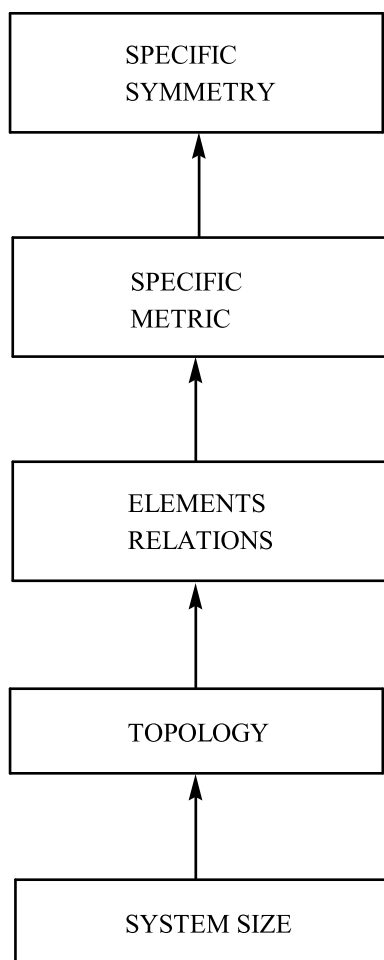
The basic level in this scheme is the size of a chemical system and is followed by its topology. Topology discriminates between systems of the same size. The physical nature of the system is placed on the third level. The physical nature of the system elements and pairwise relations which in chemistry stand for the elemental composition and bond types, discriminate between the systems of the same size and topology. If systems remain in the same complexity class after the first three levels have been considered, the next level to be employed is the specific metric of a system, e.g., the geometric characteristic of a system. The last level is specific symmetry, e.g., the point groups of a system symmetry, which takes care of uniformity or diversity of element distribution. The mathematical model based on the Bonchev–Polansky hierarchical concept of total complexity, denoted by  $C_{\text{total complexity}}$ , is a multicomponent vector whose components are the hierarchical levels:

$$C_{\text{total complexity}} = (C_{\text{size}}, C_{\text{topology}}, C_{\text{physical nature}}, C_{\text{metric}}, C_{\text{symmetry}}) \quad (1)$$

The Bonchev–Polansky proposal for the topological complexity of chemical systems was slightly condensed in 1997 by Bonchev and Seitz [5] who used graph-theoretical terminology. The difference between the Bonchev–Polansky and the Bonchev–Seitz proposals is exclusion of symmetry from the latter proposal. The Bonchev–Seitz proposal in modified form is shown in Fig. 3.

The hierarchical levels of topological complexity start with the system connectedness which allows differentiation between connected, disconnected and non-planar graphs. The next level, the system adjacency, consists of three classes: directed graphs, non-directed graphs and multigraphs. The connectivity patterns are placed in the third level. Four such patterns were considered: linearity, bridging, branching and cyclicity. In the fourth level metric properties of graphs, e.g., distances, paths, walks are placed in one branch, and the other contains subgraphs of increasing size and number. The mathematical model based on the Bonchev–Seitz hierarchical concept of complexity can be summarized as:

$$C_{\text{topological complexity}} = (C_{\text{connectedness}}, C_{\text{adjacency}}, C_{\text{connectivity pattern}}, C_{\text{metric}}, C_{\text{subgraphs}}) \quad (2)$$



**Topological Complexity of Molecules, Figure 2**

A hierarchical organization of total complexity of chemical systems as proposed by Bonchev and Polansky [21]

### Criteria for Topological Complexity Indices

TCIs arose from the need to compare molecules according to their structural complexity. However, to be useful, TCIs must fulfill certain criteria. Several authors proposed criteria that TCIs should satisfy. A first proposal is due to Minoli [15] in 1975. He proposed the following four criteria for the TCI:

1. TCI should monotonically increase with the number of vertices of the graph.
2. TCI should monotonically increase with the number of edges of the graph.
3. TCI should reflect the degree of connectedness of the graph.
4. TCI should satisfy one's intuition regarding the complexity of objects by assigning a higher value to a graph which looks complicated and vice versa.

Twelve years (1987) after Minoli, Bonchev and Polansky [21] proposed the following nine criteria for the TCI:

1. TCI should be independent of the nature of the systems.
2. TCI should be specified within a unique theoretical concept.
3. TCI should take into account different complexity levels and their hierarchy.
4. TCI should exhibit a stronger dependence on the relations between the elements of a system than on the number of elements.
5. TCI should increase monotonically with the number of different complexity features.
6. TCI should agree with the intuitive ideas of complexity.
7. TCI should differentiate non-isomorphic systems.
8. TCI should not be too sophisticated.
9. TCI should be applicable to practical purposes.

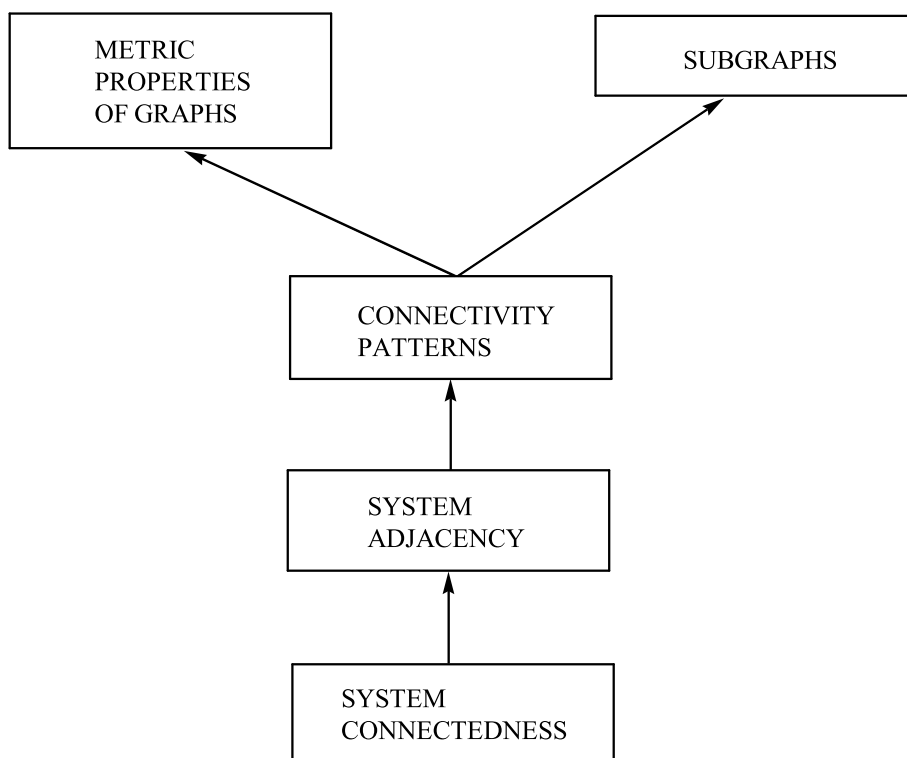
Eleven years later (1998), Bertz and Wright [22] analyzed the Minoli and the Bonchev–Polansky criteria and rejected intuition as a criterion which had been included in both sets of criteria. The following five criteria can be extracted from their work:

1. TCI should increase monotonically with factors that increase complexity as embodied in homologous series.
2. TCI should increase with increasing branching and cyclicity.
3. TCI should increase with the number of multiple edges.
4. TCI should take into account symmetry as a simplifying feature.
5. TCI should be sensitive to heteroatoms.

Intuition should not, however, be treated lightly, since many results achieved in chemistry as well as in other natural sciences were based on guidelines provided by the researcher's intuition.

Two years after Bertz and Wright, Rücker and Rücker [23] discussed the following desirable properties of the TCI:

1. TCI should increase with increasing graph size.
2. TCI should increase with increasing branching.
3. TCI should increase with increasing cyclicity.
4. TCI should increase with increasing number of multiple edges.
5. TCI should increase with increasing number of weighted vertices.
6. TCI should increase with decreasing symmetry.



Topological Complexity of Molecules, Figure 3

A hierarchical organization of topological complexity of chemical systems as proposed by Bonchev and Seitz [5]

All of the above criteria point out that useful TCI should increase with the structural characteristics such as the graph size, branching, cyclicity, the number of weighted vertices and multiple edges but decrease with symmetry. These criteria will be used here to judge the applicability of TCIs.

### Examples of Topological Complexity Indices

In this section, beginning with the Minoli index [15], we present chronologically several TCIs that were specifically designed to study the complexity of graphs and molecules. It should be noted that we use the term “graph” to mean both mathematical graphs and molecular graphs.

#### The Minoli Index

In 1975, Minoli proposed an index with aim to assess the combinatorial complexity of graphs. The Minoli index MI is defined as:

$$MI = [VE/(V + E)] \sum_l P_l, \quad (3)$$

where  $V$  and  $E$  are the number of vertices and edges, respectively, in a graph  $G$ , while  $P_l$  is the number of paths of length  $l$ . The Minoli index is a highly degenerate TCI – it does not distinguish between isomeric graphs. In order to increase the discriminatory ability of MI, Bonchev [24] replaced the sum of the number of paths with the total length of all paths  $L_p$ . This variant of the Minoli index, called the Minoli–Bonchev index MBI, is defined as:

$$MBI = [VE/(V + E)] L_p \quad (4)$$

where  $L_p$  is computed by:

$$L_p = \sum_l l P_l. \quad (5)$$

#### The Bonchev–Trinajstić Index

Two years after Minoli, Bonchev and Trinajstić [16] proposed an information-theoretic index based on the graph-theoretical distances. The Bonchev–Trinajstić index BTI is defined as:

$$BTI = n \log_2 n - \sum_l n_l \log_2 n_l \quad (6)$$



where  $n$  is the total number of distances in a graph  $G$ ,  $n_l$  is the number of graph-theoretical distances of length  $l$  and  $n = \sum_l n_l$ . The logarithm is taken at basis 2 for measuring the information content in bits. BTI was originally used to study molecular branching and was only later used to study the complexity of graphs and molecules.

### The Bertz Index

Another information-theoretic index was proposed by Bertz [17]. The Bertz index BI is defined as:

$$BI = 2n \log_2 n - \sum_i n_i \log_2 n_i \quad (7)$$

where  $n$  is the number of pairs of adjacent edges in a graph  $G$  and  $n_i$  is the number of pairs of adjacent edges in the  $i$ th class by symmetry. Equation (7) represents a pragmatic modification of Eq. (6). Bertz added to Eq. (7) the extra term  $n \log_2 n$  to prevent  $BI = 0$  when all pairs of adjacent edges in  $G$  are equivalent. The first part of Eq. (7) takes into account structural characteristics of  $G$ , such as size, branching and cyclicity, and the second part deals with the symmetry of  $G$  in terms of equivalent pairs of adjacent edges. Bertz introduced his index for use in synthesis planning and analysis of synthetic strategies [19,20,22,25,26,27].

### The Number of Spanning Trees

The first authors who used spanning trees explicitly as TCIs were Bonchev, Kamenski and Temkin in 1987 [28] in their work on the complexity of the linear mechanisms of chemical reactions. These authors also suggested an unwieldy method for computing spanning trees. However, in 1983, Gutman, Mallion and Essam [29] produced an elegant method for computing spanning trees of labeled graphs. These authors stated in passing "... the number of spanning trees of a labeled molecular-graph, sometimes referred to as the complexity of the graph", but they did not explore this use of the spanning trees.

The number of spanning trees of a planar graph  $G$ ,  $t(G)$ , based on the matrix-tree theorem, is given by:

$$t(G) = \det \mathbf{L}^* \quad (8)$$

where  $\mathbf{L}^*$  is the following difference matrix:

$$\mathbf{L}^* = \mathbf{A}^* - \mathbf{A}^* \quad (9)$$

$\mathbf{A}^*$  is the vertex-adjacency matrix of an inner dual  $G^*$  of  $G$  and  $\mathbf{A}^*$  is a diagonal matrix of  $G^*$  with elements equal to the size of each cycle making up the polycyclic graph.

Matrix  $\mathbf{L}^*$  may also be regarded as a Laplacian matrix of the vertex-weighted inner dual  $G^*$ , the weights of vertices in  $G^*$  being equal to the sizes of cycles in the parent graph  $G$ . The spanning trees can also be computed using eigenvalues  $\lambda_i$  of the Laplacian matrix of  $G$  [30,31]:

$$t(G) = (1/V) \prod_{i=2}^V \lambda_i(\mathbf{L}) \quad (10)$$

It should be noted that the smallest eigenvalue of the Laplacian matrix  $\lambda_1$  is always zero, thus the multiplication of eigenvalues in Eq. (10) starts with  $\lambda_2$ . In  $d$ -regular graphs, that is graphs in which every vertex has the same degree  $d$ , the eigenvalues of the Laplacian matrix  $\lambda_i(\mathbf{L})$  and the vertex-adjacency matrix  $\lambda_i(\mathbf{A})$  are related by:

$$\lambda_i(\mathbf{L}) = d - \lambda_i(\mathbf{A}) \quad (11)$$

In fullerenes whose graphs are three regular graphs ( $d=3$ ), Eq. (10) transforms by means of Eq. (11) into [32]:

$$t(G) = (1/V) \prod_{i=2}^V [3 - \lambda_i(\mathbf{A})] \quad (12)$$

This formula was used by Fowler [33] to study the complexity of isomeric fullerenes.

### Indices Based on Connected Subgraphs

Bertz and Sommer [34] and Bertz and Wright [22] advocated the use of the number of kinds of connected subgraphs  $N_s$  and the total number of connected subgraphs  $N_t$  of a graph  $G$  as TCIs. The set of all connected subgraphs starts with vertices and is followed by edges, 3-vertex subgraphs, etc. and ends with  $G$  itself, since in formal graph theory  $G$  is its own subgraph. The generation and computation of the subgraph-counting numbers  $N_s$  and  $N_t$  is a rather difficult problem. But, Rücker and Rücker [35,36] solved it by designing an efficient computer program for generating and counting all connected subgraphs of  $G$ . These authors named their program BERTZ in recognition of Bertz's work on the complexity of molecules and reactions.

Independently, Bonchev [37] also used  $N_t$  to evaluate the complexity of molecules and in the same paper he also proposed two novel TCIs based on the connected subgraphs. One of these, he called simply the topological complexity index, denoted by TC. TC is equal to the sum of vertex degrees in all connected subgraphs (or the subgraph total adjacency) of  $G$ :

$$TC = \sum_s \sum_i d_i(s) \quad (13)$$

where the first sum is over all connected subgraphs  $s$  in  $G$ , the second sum is over all vertices in a subgraph  $i$  and  $d_i$  is the degree of vertices in  $i$ . It should be noted that in computing TC the vertex-degrees in subgraphs are taken as they are in  $G$ . If the vertex-degrees are taken as they are in each subgraph, the corresponding index is denoted by TC1. Thus, TC is always greater than TC1, except for a single-vertex graph, where  $TC = TC1 = 0$ . TC and TC1 were the first overall TCIs.

Bonchev later used other graph invariants to generate overall indices, such as the overall Wiener index [38], based on the Wiener indices [39] of subgraphs and overall Zagreb indices [40], based on the Zagreb indices [41,42].

### The Randić Complexity Index

The Randić complexity index RCI [43] is a symmetry-dependent complexity index based on the concept of augmented vertex-degree [44,45]. The augmented degree of a vertex  $i$  in a graph  $G$  is the sum of its degree and degrees of all other vertices  $j$  with weight  $1/2^{l(i,j)}$  depending on their distance  $l$  from the vertex  $i$ . Then the RCI is the sum of all augmented degrees of vertices that are not symmetrically equivalent:

$$RCI = \sum_{i=1}^{N^*} \sum_{j=1}^N d_j / 2^{l(i,j)} \quad (14)$$

where  $N^*$  is the number of vertices that are not symmetrically equivalent and  $N$  is the number of all vertices in a graph  $G$ . The first sum is over only the symmetrically non-equivalent vertices and the second sum is over all vertices in  $G$ .

### The Total Walk Count

Rücker and Rücker [23,46] proposed the total walk count (twc) as a measure of the complexity of graphs and molecules. The twc's can be computed from atomic walk count sums (awcs's), molecular walk counts (mwc's) and the vertex-adjacency matrices of (molecular) structures considered.

The atomic walk count of order  $l$  of atom  $i$   $(awc)_l(i)$  is the number of all possible walks of length  $l$  which start at the vertex  $i$  and end at any vertex  $j$ :

$$(awc)_l(i) = \sum_{j=1}^V (A^l)_{ij} . \quad (15)$$

The molecular walk count of order  $l$   $(mwc)_l$  is equal to the sum of all atomic walk counts of order  $l$ :

$$(mwc)_l = \sum_{i=1}^V (awc)_l(i) . \quad (16)$$

Finally, the twc is equal to the sum of all molecular walk counts for  $l$  from 1 to  $V - 1$ :

$$twc = \sum_{l=1}^{V-1} (mwc)_l . \quad (17)$$

There is another way to compute the total walk count by utilizing a relationship between the Morgan extended connectivities MEC [47] and powers of the vertex-adjacency matrix. This relationship was first noticed by Razinger [48]; Rücker and Rücker [46] proved that MECs and  $(awc)_l$ 's are identical. Thus, the MEC of vertex  $i$  of the  $l$ th order can be computed by the iterative summation of the  $(l - 1)$ th order contributions of the neighbors of  $i$  and is equal to  $(awc)_l(i)$ . Consequently, the calculation of twc for reasonably sized (molecular) structures requires only a sequence of addition steps. Rücker and Rücker [46] termed this procedure the Morgan summation procedure, and based on this procedure they prepared the program MORGAN, so named as a tribute to Morgan's work, which is now almost forgotten.

The twc index does not take care of the graph's symmetry. In order to do that the twc needs to be modified by symmetry. The symmetry-modified twc is called the walk complexity (wxc) [23,49]. It can be obtained from the atomic walk count sum awcs. The awcs of atom  $i$  can be computed by summing up the  $(awc)_l(i)$  over all  $l$  from 1 to  $V - 1$ :

$$awcs(i) = \sum_{l=1}^{V-1} (awc)_l(i) . \quad (18)$$

Then, the wxc is given by the following:

$$wxc = \sum_i awcs(i) \quad (19)$$

where  $i$  is the first integer from each set of integers denoting vertices in an equivalence class by symmetry.

### The Usefulness of the Reviewed Topological Complexity Indices to Assess the Complexity of Molecular Graphs

Comparison of the reviewed TCIs is carried out on four sets of simple molecular graphs, that is, on linear trees

representing carbon skeletons of  $n$ -alkanes containing from two to eight carbon atoms, cycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms, trees on six vertices representing carbon skeletons of isomeric hexanes and three planar polycyclic graphs on four vertices representing carbon skeletons of cyclobutane, bicyclo[1.1.0]butane and tetrahedrane. The idea behind this choice of graphs is that if the TCIs fail to assess the complexity of simple graphs they certainly would not be able to handle graphs with much more complicated structures.

### Complexity of Linear Trees

We considered linear trees from two to eight vertices to study the behavior of TCIs with the size increase. Linear trees are convenient to use because the structural factor strongly influencing their complexity is their size. However, there is another (silent) structural feature present in trees that cannot be avoided – the symmetry. The symmetry is identified through the number of vertices not equivalent by symmetry. It will influence the symmetry-dependent TCIs, but not dramatically. The studied linear trees are depicted in Fig. 4 and their TCIs are given in Table 1.

From Table 1, it is evident that all of considered TCIs increase with increasing linear alkane tree size. The reason for such behavior is due to the regularity of the linear alkane tree structures that allows most of the indices considered to be defined trivially in terms of the number of vertices.

### MI and MB Indices

$$MI = V^2(V-1)^2/(4V-2) \quad (20)$$

$$MB = V^2(V+1)(V-1)^2/6(2V-1). \quad (21)$$

### BT and BI Indices

$$BT = \binom{V}{2} \log_2 \binom{V}{2} - \sum_{i=2}^{V-1} i \log_2 i \quad (22)$$

$$BI = \begin{cases} 2(V-2) \log_2(V-2) - (V-2) & \text{if } V = \text{even} \\ 2(V-2) \log_2(V-2) - (V-3) & \text{if } V = \text{odd} \end{cases} \quad (23)$$

### $t(G)$ Index

The  $t(G)$  index is not given because for all acyclic structures; it is always equal to 1, that is, the spanning tree of a tree is identical to the tree itself.

### $N_s$ and $N_t$ Indices

$$N_s = V \quad (24)$$

$$N_t = V(V+1)/2. \quad (25)$$

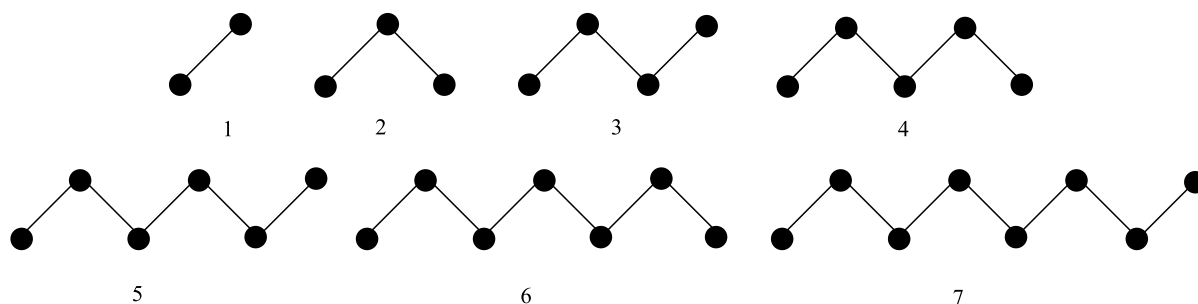
### TC and TC1 Indices

$$TC = V(V-1)(V+4)/3 \quad (26)$$

$$TC1 = V(V^2-1)/3. \quad (27)$$

### RCI Index

$$RCI = \begin{cases} 3V-6+6/2^V & \text{if } V = \text{even} \\ 3V-3-3(2^{(V-1)/2}-1)/2^{V-1} & \text{if } V = \text{odd} \end{cases} \quad (28)$$



Topological Complexity of Molecules, Figure 4

Linear trees representing carbon skeletons of  $n$ -alkanes from two to eight carbon atoms

Topological Complexity of Molecules, Table 1

Topological complexity indices of smaller linear trees representing carbon skeletons of linear alkanes from two to eight carbon atoms

Linear alkane tree <sup>a</sup>	Topological complexity index <sup>b</sup>										
	MI	MB	BT	BI	N <sub>s</sub>	N <sub>t</sub>	TC	TC1	RCI	twc	wcx
1	0.7	0.7	0	–	2	3	4	2	1.5	2	1
2	3.6	4.8	2.8	0.0	3	6	14	8	5.3	10	7
3	10.3	17.1	8.8	2.0	4	10	32	20	6.4	32	16
4	22.2	44.4	18.5	7.5	5	15	60	40	11.4	88	56
5	40.9	95.5	32.2	12.0	6	21	100	70	12.1	222	111
6	67.8	180.9	50.4	19.2	7	28	154	112	17.7	536	320
7	104.5	313.6	73.1	25.0	8	36	224	168	18.0	1254	617

<sup>a</sup>Considered linear alkane trees, given in Fig. 4<sup>b</sup>MI = Minoli index; MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index, BI = Bertz index; N<sub>s</sub>, N<sub>t</sub> = the number of kinds and the total number of connected sub-graphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity**twc Index**

$$\text{twc} = \text{mwc}_V - 2V + \sum_{i=0}^{V-1} s_i \quad (29)$$

where

$$\text{mwc}_l = 2^l V - \sum_{i=0}^{l-1} 2^{l-1-i} s_i \quad (l \leq V) \quad (30)$$

and

$$s_i = \begin{cases} 2 \binom{i}{i/2} & \text{if } i = \text{even} \\ 2 \binom{i}{(i+1)/2} & \text{if } i = \text{odd} \end{cases} \quad (31)$$

**wcx Index**The closed form of the wcx index is available only for  $V = \text{even}$ :

$$\text{wcx} = \left( \text{mwc}_V - 2V + \sum_{i=0}^{V-1} s_i \right) / 2. \quad (32)$$

**Complexity of Monocycles** The next set of graphs considered is the set of lower monocycles from three to eight vertices. In their case, besides the cycle size, a more visible role is played by symmetry. The studied monocycles are depicted in Fig. 5 and their TCIs are given in Table 2.

All considered TCIs increase with the size of the cycle, thus indicating the larger the cycle the greater complexity

of the cycle. The regular structures of the monocycles allows one to express the employed TCIs simply in terms of the number of vertices.

**MI and MB Indices**

$$\text{MI} = [V^2(V-1)]/2 \quad (33)$$

$$\text{MB} = V^3(V-1)/4. \quad (34)$$

**BT and BI Indices**

$$\text{BT} = \begin{cases} \binom{V}{2} \log_2(V-1) - V(V-2)/2 & \text{if } V = \text{even} \\ \binom{V}{2} \log_2[(V-1)/2] & \text{if } V = \text{odd} \end{cases} \quad (35)$$

$$\text{BI} = V \log_2 V. \quad (36)$$

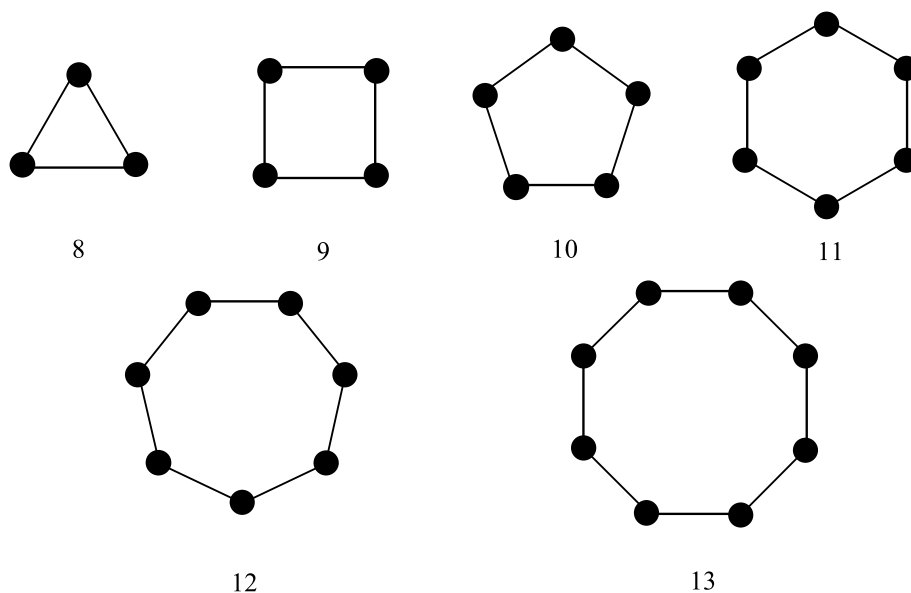
**t(G) Index**

$$t(G) = V. \quad (37)$$

**N<sub>s</sub> and N<sub>t</sub> Indices**

$$N_s = V + 1 \quad (38)$$

$$N_t = V^2 + 1. \quad (39)$$



Topological Complexity of Molecules, Figure 5

Cycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms

Topological Complexity of Molecules, Table 2

Topological complexity indices of smaller monocycles representing carbon skeletons of cycloalkanes from three to eight carbon atoms

Cycle <sup>a</sup>	Topological complexity index <sup>b</sup>											
	MI	MB	BT	BI	$t(G)$	$N_s$	$N_t$	TC	TC1	RCI	twc	wcx
8	9	13.5	0	4.8	3	4	10	42	24	4.00	18	6
9	24	48.0	5.5	8.0	4	5	17	88	56	4.50	56	14
10	50	125.0	10.0	11.6	5	6	26	160	110	5.00	150	30
11	90	270.0	22.8	15.5	6	7	37	264	192	5.25	372	62
12	147	514.5	33.3	19.7	7	8	50	406	308	5.50	882	126
13	224	896.0	54.6	24.0	8	9	65	592	464	5.63	2032	254

<sup>a</sup>Considered monocycles, given in Fig. 5

<sup>b</sup>MI = Minoli index; MB = Minoli-Bonchev index; BT = Bonchev-Trinajstić index; BI = Bertz index;  $t(G)$  = the number of spanning trees;  $N_s$ ,  $N_t$  = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wx = the walk complexity

#### TC and TC1 Indices

$$TC = V[V(V + 1) + 2] \quad (40)$$

$$TC1 = V[V(V - 1) + 2] . \quad (41)$$

#### twc and wx Indices

$$twc = 2V(2^{V-1} - 1) \quad (43)$$

$$wx = 2(2^{V-1} - 1) . \quad (44)$$

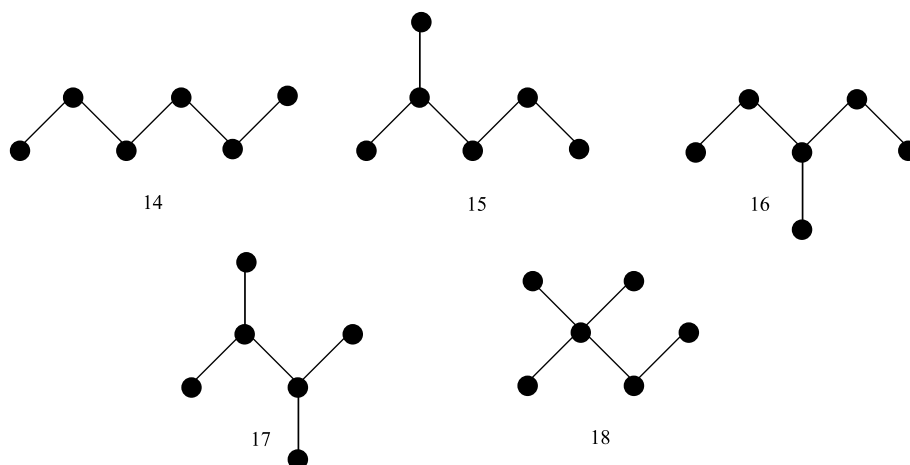
#### RCI Index

$$RCI = \begin{cases} 6(2^{V/2} - 1)/2^{V/2} & \text{if } V = \text{even} \\ 6 - 1/2^{(V-5)/2} & \text{if } V = \text{odd} . \end{cases} \quad (42)$$

#### Branching and Complexity

To illustrate the impact of branching on the complexity, we selected isomeric trees on six vertices representing the carbon skeletons of hexanes. It should also be pointed out





Topological Complexity of Molecules, Figure 6

Hexane trees representing carbon skeletons of isomeric hexanes

that, as before, the impact of branching cannot be isolated from the impact of symmetry on the complexity. Branching is intuitively identified through the appearance of vertices with degrees higher than two – trees containing only vertices with degrees one and two are not branched. The symmetry impact will be strongly reflected on the symmetry-dependent TCIs. The isomeric hexane trees are depicted in Fig. 6 and their TCIs are given in Table 3. The Minoli index is not included since it possesses the same value (40.9) for all hexane trees.

As can be seen from Table 3, four of the symmetry-independent TCIs ( $N_t$ , TC, TC1, twc) indicate that the increasing branching of a hexane tree *via* the number of branched vertices increases the tree complexity as one would expect intuitively. The symmetry-dependent TCIs

(MB, BT) produced the anti-intuitive prediction, that is, increasing branching decreases the tree complexity. However, if we consider  $-MB$  and  $-BT$ , then these indices agree in their prediction with the above four symmetry-independent TCIs. The four remaining TCIs (BI,  $N_s$ , RCI, wcx) behave erratically, but two of them (BI, wcx) agree at least in predicting that **14** will be the least and **18** the most complex structure among the five isomeric hexane trees.  $N_s$  and RCI predict **14** to be the least complex structure, but the complexity-ordering of the remaining four isomeric trees is dubious.

Topological Complexity of Molecules, Table 3

Trees on six vertices representing carbon skeletons of isomeric hexanes

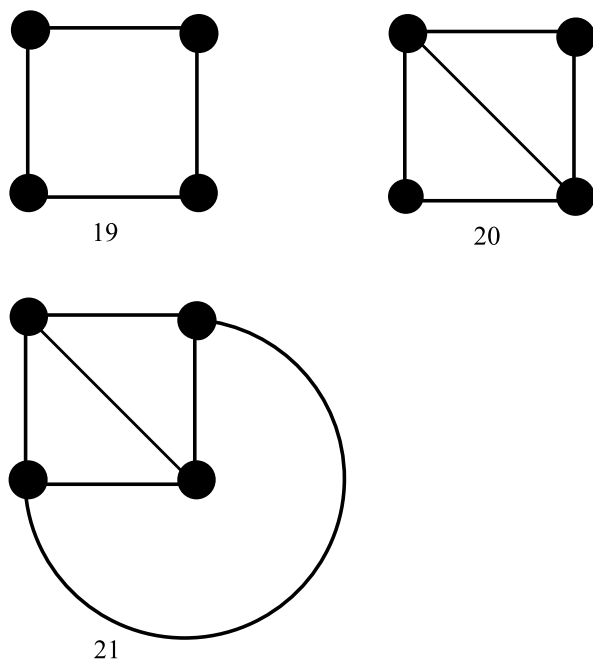
Hexane tree <sup>a</sup>	Topological complexity index <sup>b</sup>									
	MB	BT	BI	$N_s$	$N_t$	TC	TC1	RCI	twc	wcx
14	95.5	32.2	12.0	6	21	100	70	12.1	222	111
15	87.3	28.6	21.2	8	24	127	88	21.9	268	234
16	84.5	27.4	19.2	8	25	136	94	17.8	284	201
17	79.1	23.5	21.0	7	28	164	112	9.8	330	124
18	76.4	22.6	29.8	8	30	181	122	19.5	370	272

<sup>a</sup>Hexane trees are given in Fig. 6

<sup>b</sup>MB = Minoli–Bonchev index; BT = Bonchev–Trinajstić index; BI = Bertz index;  $N_s$ ,  $N_t$  = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

## Cyclicity and Complexity

Cyclicity can be considered either in terms of closing a cycle or increasing the number of cycles. If we consider the former case, then we can compare TCIs in Tables 1 and 2. Comparison between the linear trees and monocycles with the same number of vertices shows that seven TCIs (MI, MB,  $N_s$ ,  $N_t$ , TC, TC1, twc) always possess greater values for monocycles than for linear alkanes, thus indicating that the monocyclic structures are more complex than linear structures. This agrees with conclusion reached by Bertz and Zamfirescu [50] that  $C_V$  is more complex than  $P_V$  for all  $V$ , where  $C_V$  and  $P_V$  denote monocycles and linear trees, respectively, with  $V$  vertices. However, the four symmetry-dominated TCIs (BT, BI, RCI, wcx) violate this. This is expected since monocycles are more symmetry-rich than linear alkanes. However, it should be noted that whilst BT, RCI and wcx are always smaller for monocycles than for the corresponding linear trees, the BI index ex-



Topological Complexity of Molecules, Figure 7  
Planar cyclic graphs on four vertices

hibits such behavior only for linear trees and monocycles with eight or more vertices.

As stated above cyclicity also manifests itself through the increasing number of cycles. To illustrate this, we considered three planar cyclic graphs on four vertices depicting carbon skeletons of well-known organic molecules: cyclobutane (one cycle), bicyclo[1.1.0]butane (two cycles) and tetrahedrane (three cycles). These three cyclic graphs are depicted in Fig. 7 and their TCIs are given in Table 4.

Cyclicity ordering as implied by consideration of cyclic graphs *via* the numbers of cycles is followed by ten TCIs

(MI, MB, BI,  $t(G)$ ,  $N_s$ ,  $N_t$ , TC, TC1, twc, wcx). BT predicts the inverse order, but its prediction could be amended if  $-BT$  is considered. Only in the case of RCI, a strong symmetry-dependent TCI predicts counter-intuitively that **20** is more complex than **21** since **21** is a symmetry-richer structure than **20**.

### Summary

We applied 12 TCIs on four sets of simple molecular graphs. They exhibit the following characteristics:

- (i) MI increases with increasing size of linear trees and monocycles and with increasing cyclicity. It is not applicable to isomeric structures.
- (ii) MB and BT increase with increasing size of linear trees and monocycles and decrease with increasing branching. However, if one considers  $-MB$  and  $-BT$ , then these indices increase with increasing branching. MB increases, but BT decreases with increasing cyclicity. Again if one considers  $-BT$ , it increases with increasing cyclicity.
- (iii)  $t(G)$  can be used to assess the complexity only of graphs containing cycles.
- (iv)  $N_s$  increases with increasing size of linear trees and monocycles and with increasing cyclicity, but fails in the case of branching where it predicts only the least branched structure.
- (v)  $N_t$  increases with increasing size of linear trees and monocycles and with increasing branching and cyclicity. This is a good TCI, but the procedure of getting  $N_t$  increases exponentially with problem size.
- (vi) TC and TC1 behave similarly to  $N_t$ , as expected, since they depend on the knowledge of all connected subgraphs. Thus, they suffer from the same computational problem as does  $N_t$ . However, for small to

Topological Complexity of Molecules, Table 4

Planar cyclic graphs on four vertices representing carbon skeletons of cyclobutane, bicyclo[1.1.0]butane and tetrahedrane

Cyclic graph <sup>a</sup>	Topological complexity index <sup>b</sup>											
	MI	MB	BT	BI	$t(G)$	$N_s$	$N_t$	TC	TC1	RCI	twc	wcx
19	24.0	48.0	5.5	8.0	4	5	17	88	56	4.5	56	14
20	42.2	86.7	3.9	36.0	8	9	33	254	152	12.0	102	51
21	72.0	158.4	0.0	43.0	16	10	64	648	372	7.5	156	39

<sup>a</sup>Planar cyclic graphs, given in Fig. 7

<sup>b</sup>MI = Minoli index; MB = Minoli-Bonchev index; BT = Bonchev-Trinajstić index, BI = Bertz index;  $t(G)$  = the number of spanning trees;  $N_s$ ,  $N_t$  = the number of kinds and the total number of connected subgraphs; TC, TC1 = Bonchev indices; RCI = Randić complexity index; twc = the total walk count; wcx = the walk complexity

medium size molecular graphs all these three indices appear to be useful TCIs.

- (vii) The *twc* increases with increasing size of linear trees and monocycles and with increasing branching and cyclicity. Therefore, the *twc* is also a useful TCI regarding the graphs considered.
- (viii) BI increases with increasing size of linear trees and monocycles, but decreases with increasing cyclicity. With respect to branching, BI predicts only the least and the most branched structures. This index also possesses a strongly symmetry-dependent component and hence it does not produce a consistent ordering between monocycles and the corresponding linear trees.
- (ix) RCI and *wcx* suffer from their overemphasis on the role of symmetry. Therefore, the use of these two TCIs and BI require explicit input of symmetry information *via* a reliable approach for symmetry recognition e. g., [51,52,53].

### Future Directions

Assessment of complexity apparently requires complex approaches. Perhaps a way to proceed is to consider that complexity is not a strictly numerical quantity. Some authors [50,54,55] already pointed out that the complexity is a partially ordered quantity. The Hasse diagram reflecting the partial order of indices appears to be a very useful device with which to appraise the topological complexity of (molecular) graphs. One way to construct the Hasse diagram for a set of molecular graphs is as follows: The graph **A** in a diagram is placed above graph **B** if all TCIs considered have a smaller value for **A** than for **B**. Besides, two graphs **A** and **B** are directly linked by an edge downward from **A** to **B** if and only if no third graph is placed by this partial ordering between them.

### Bibliography

#### Primary Literature

1. Waldrop MM (1992) Complexity: The emerging science at the edge of order and chaos. Touchstone, New York
2. Rouvray DH (2003) An introduction to complexity. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry. Taylor & Francis, London, pp 1–27
3. Rouvray DH (1997) Are the concepts in chemistry all fuzzy? In: Rouvray DH (ed) Concepts in chemistry: A contemporary challenge. Research Studies Press, New York, pp 1–15
4. Nikolić S, Trinajstić N, Tolić IM, Rücker G, Rücker C (2003) On molecular complexity indices. In: Bonchev D, Rouvray DH (eds) Complexity in chemistry. Taylor & Francis, London, pp 29–89
5. Bonchev D, Seitz WA (1997) The concept of complexity in chemistry. In: Rouvray DH (ed) Concepts in chemistry: A contemporary challenge. Research Studies Press, Taunton, pp 353–381
6. Bonchev D, Buck GA (2007) From molecular to biological structure and back. J Chem Inf Model 47:909–917
7. Dancoff SM, Quastler H (1953) The information content and error rate in living things. In: Quastler H (ed) Essays on the Use of Information Theory in Biology. University of Illinois Press, Urbana
8. Shannon C, Weaver W (1949) Mathematical theory of communications. University of Illinois Press, Urbana
9. Morowitz H (1955) Some order-disorder considerations in living systems. Bull Math Biophys 17:81–86
10. Rashevsky N (1955) Life, information theory and topology. Bull Math Biophys 17:229–235
11. Karreman G (1955) Topological information content and chemical reactions. Bull Math Biophys 17:279–285
12. Trucco E (1956) A note on the information content of graphs. Bull Math Biophys 18:129–135
13. Trucco E (1956) On the information content of graphs: Compound symbols. Different states for each point. Bull Math Biophys 18:237–253
14. Mowshovitz A (1968) Entropy and the complexity of graphs: I. An index of the relative complexity of a graph. Bull Math Biophys 30:175–204
15. Minoli D (1975) Combinatorial graph complexity. Atti della Accademia Nazionale dei Lincei – Classe di Scienze fisiche, matematiche e naturali (Serie 8) 59:651–661
16. Bonchev D, Trinajstić N (1977) Information theory, distance matrix and molecular branching. J Chem Phys 67: 4517–4533
17. Bertz SH (1983) The first general index of molecular complexity. J Am Chem Soc 103:3599–3601
18. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley, Weinheim, pp 300
19. Bertz SH (1983) A mathematical model of complexity. In: King RB (ed) Chemical applications of topology and graph theory. Elsevier, Amsterdam, pp 206–221
20. Bertz SH (1983) On the complexity of graphs and molecules. Bull Math Biol 45:849–855
21. Bonchev D, Polansky OE (1987) On the topological complexity of chemical systems. In: King RB, Rouvray DH (eds) Graph Theory and Topology in Chemistry. Elsevier, Amsterdam, pp 125–158
22. 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 35:32–48
23. Rücker G, Rücker C (2000) Walk counts, labirinthicity and complexity of acyclic and cyclic graphs and molecules. J Chem Inf Comput Sci 40:99–106
24. Bonchev D (1990) Problems of computing molecular complexity. In: Rouvray DH (ed) Computational Chemical Graph Theory. Nova Science Publishers, NY, pp 33–63
25. Bertz SH (1982) Convergence, molecular complexity and synthetic analysis. J Am Chem Soc 104:5801–5803
26. Bertz SH (2003) Complexity of molecules and their synthesis. In: Bonchev D, Rouvray DH (eds) Complexity in Chemistry. Taylor & Francis, London, pp 91–156
27. Bertz SH, Rücker C (2004) In search of simplification: the use of topological complexity indices to guide retrosynthetic analysis. Croat Chem Acta 77:221–235

28. Bonchev D, Kamenski D, Temkin ON (1987) Complexity index of the linear reaction mechanisms of chemical reactions. *J Math Chem* 1:345–388
29. Gutman I, Mallion RG, Essam JW (1983) Counting the spanning trees of a labelled molecular-graph. *Mol Phys* 50:859–877
30. Mohar B (1989) Laplacian matrices of graphs. In: Graovac A (ed) *MATH/CHEM/COMP 1989*. Elsevier, Amsterdam, pp 1–8
31. Trinajstić N, Babić D, Plavšić, Amić D, Mihalić Z (1994) The Laplacian matrix in chemistry. *J Chem Inf Comput Sci* 34: 368–376
32. Mihalić Z, Trinajstić N (1994) On the number of spanning trees in fullerenes. *Fullerene Sci Technol* 2:89–95
33. Fowler PW (2003) Complexity, spanning trees and relative energies of fullerene isomers. *MATCH Commun Math Comput Chem* 48:87–96
34. Bertz SH, Sommer TJ (1997) Rigorous mathematical approaches to strategic bonds and synthetic analysis based on conceptually simple new complexity indices. *Chem Com* 2409–2410
35. Rücker G, Rücker C (2001) On finding non-isomorphic connected subgraphs and distinct molecular substructures. *J Chem Inf Comput Sci* 41:314–320; erratum 825
36. 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
37. Bonchev D (1997) Novel indices for the topological complexity of molecules. *SAR QSAR Environ Res* 7:23–43
38. Bonchev D (2001) The Overall Wiener index – A new tool for characterizing molecular topology. *J Chem Inf Comput Sci* 41:582–592
39. Wiener H (1947) Structural determination of paraffin boiling points. *J Am Chem Soc* 69:17–20
40. Bonchev D, Trinajstić N (2001) Overall molecular descriptors. 3. Overall Zagreb indices. *Sar & QSAR Environ Res* 12:213–235
41. Gutman I, Trinajstić N (1972) Graph theory and molecular orbitals. III. Total  $\pi$ -Electron Energy of Alternant Hydrocarbons. *Chem Phys Lett* 17:535–538
42. Gutman I, Ružić B, Trinajstić N, Wilcox Jr CF (1975) Graph theory and molecular orbitals. XII. Acyclic polyenes. *J Chem Phys* 62:3399–3405
43. Randić M (2001) On complexity of transitive graphs representing degenerate rearrangements. *Croat Chem Acta* 74:683–705
44. Randić M, Plavšić D (2002) On the concept of molecular complexity. *Croat Chem Acta* 75:107–116
45. Randić M, Plavšić D (2003) On characterization of molecular complexity. *Int J Quantum Chem* 91:20–31
46. Rücker G, Rücker C (1993) Counts of all walks as atomic and molecular descriptors. *J Chem Inf Comput Sci* 33:683–695
47. Morgan HL (1965) The generation of a unique machine description for chemical structures – A technique developed at Chemical Abstracts Service. *J Chem Doc* 5:107–113
48. Razinger M (1982) Extended connectivity of chemical graphs. *Theor Chim Acta* 61:581–586
49. Gutman I, Rücker C, Rücker G (2001) On walks in molecular graphs. *J Chem Inf Comput Sci* 41:739–745
50. Nikolić S, Tolić IM, Trinajstić N, Baučić I (2000) On the Zagreb indices as complexity indices. *Croat Chem Acta* 73:909–921; see Note 28 and remarks by DJ Klein on complexity partial ordering
51. Rücker G, Rücker C (1990) Computer perception of constitutional (topological) symmetry: TOPSYM, a fast algorithm for partitioning atoms and pairwise relations among atoms into equivalence classes. *J Chem Inf Comput Sci* 30:187–191
52. Rücker G, Rücker C (1991) On using the adjacency matrix power method for perception of symmetry and for isomorphism testing of highly intricate graphs. *J Chem Inf Comput Sci* 31:123–126
53. Rücker G, Rücker C (1991) Isocodal and isospectral points, edges and pairs in graphs and how to cope with them in computerized symmetry recognition. *J Chem Inf Comput Sci* 31:422–427
54. Bertz SH, Zamfirescu C (2000) New complexity indices based on edge covers. *MATCH Commun Math Comput Chem* 42:39–70
55. Rajtmajer SM, Miličević A, Trinajstić N, Randić M, Vukičević D (2006) On the complexity of Archimedean solids. *J Math Chem* 39:119–132

## Books and Reviews

- Cvetković DM, Doob M, Sachs H (1995) *Spectra of graphs – Theory and application*, 3rd edn. Barth, Heidelberg
- Devillers J, Balaban AT (eds) (1999) *Topological descriptors and related descriptors in QSAR and QSPR*. Gordon & Breach, Amsterdam
- Harary F (1971) *Graph theory*, 2nd print. Addison-Wesley, Reading
- Janežič D, Miličević A, Nikolić S, Trinajstić N (2007) *Graph-Theoretical Matrices in Chemistry*. University of Kragujevac, Kragujevac
- Trinajstić N (1992) *Chemical graph theory*, 2nd edn. CRC, Boca Raton
- Wilson RJ (1972) *Introduction to graph theory*. Oliver & Boyd, Oxford

## Topological Dynamics

ETHAN AKIN

Mathematics Department, The City College,  
New York City, USA

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction and History  
 Dynamic Relations, Invariant Sets  
 and Lyapunov Functions  
 Attractors and Chain Recurrence  
 Chaos and Equicontinuity  
 Minimality and Multiple Recurrence  
 Future Directions  
 Cross References  
 Bibliography

## Glossary

**Attractor** An attractor is an invariant subset for a dynamical system such that points sufficiently close to the set

remain close and approach the set in the limit as time tends to infinity.

**Equilibrium** An equilibrium, or a fixed point, is a point which remains at rest for all time.

**Invariant set** A subset is invariant if the orbit of each point of the set remains in the set at all times, both positive and negative. The set is + invariant, or forward invariant, if the forward orbit of each such point remains in the set.

**Lyapunov function** A continuous, real-valued function on the state space is a Lyapunov function when it is non-decreasing on each orbit as time moves forward.

**Orbit** The orbit of an initial position is the set of points through which the system moves as time varies positively and negatively through all values. The forward orbit is the subset associated with positive times.

**Recurrence** A point is recurrent if it is in its own future. Different concepts of recurrence are obtained from different interpretations of this vague description.

**Repellor** A repellor is an attractor for the reverse system obtained by changing the sign of the time variable.

**Transitivity** A system is transitive if every point is in the future of every other point. Periodicity, minimality, topological transitivity and chain transitivity are different concepts of transitivity obtained from different interpretations of this vague description.

## Definition of the Subject

A *dynamical system* is model for the motion of a system through time. The time variable is either discrete, varying over the integers, or continuous, taking real values. The systems considered in topological dynamics are primarily deterministic, rather than stochastic, so the the future states of the system are functions of the past. As the name suggests, topological dynamics concentrates on those aspects of dynamical systems theory which can be grasped by using topological methods.

## Introduction and History

The many branches of dynamical systems theory are outgrowths of the study of differential equations and their applications to physics, especially celestial mechanics. The classical subject of ordinary differential equations remains active as a topic in analysis, see, e. g. the older texts Codrington and Levinson [28] and Hartman [39] as well as Murdock [50]. One can observe the distinct fields of differentiable dynamics, measurable dynamics (that is, ergodic theory) and topological dynamics all emerging in the work of Poincaré on the Three Body Problem (for a history, see Barrow-Green [23]).

The transition from the differential equations to the dynamical systems viewpoint can be seen in the two parts of the great book Nemitskii and Stepanov [52]. We can illustrate the difference by considering the *initial value problem* in ordinary differential equations:

$$\begin{aligned}\frac{dx}{dt} &= \xi(x) \\ x(0) &= p.\end{aligned}\tag{1}$$

Here  $x$  is a vector variable in a Euclidean space  $X = \mathbb{R}^n$  or in a manifold  $X$ , and the initial point  $p$  lies in  $X$ . The infinitesimal change  $\xi(x)$  is thought of as a vector attached to the point  $x$  so that  $\xi$  is a vector field on  $X$ .

The associated *solution path* is the function  $\phi$  such that as time  $t$  varies,  $x = \phi(t, p)$  moves in  $X$  according to the above equation and with  $p = \phi(0, p)$  so that  $p$  is associated with the initial time  $t = 0$ . The solution is a curve in the space  $X$  along which  $x$  moves beginning at the point  $p$ . A theorem of differential equations asserts that the function  $\phi$  exists and is unique, given mild smoothness conditions, e. g. Lipschitz conditions, on the function  $\xi$ .

Because the equation is *autonomous*, i. e.  $\xi$  may vary with  $x$ , but is assumed independent of  $t$ , the solutions satisfy the following *semigroup identities*, sometimes also called the *Kolmogorov equations*:

$$\phi(t, \phi(s, p)) = \phi(t + s, p).\tag{2}$$

Suppose we solve Eq. (1), beginning at  $p$ , and after  $s$  units of time, we arrive at  $q = \phi(s, p)$ . If we again solve the equation, beginning now at  $q$ , then the identity (2) says that we continue to move along the old curve at the same speed. Thus, after  $t$  units of time we are where we would have been on the old solution at the same time,  $t + s$  units after time 0.

The initial point  $p$  is a parameter here. For each solution path it remains constant, the fixed base point of the path. The solution path based at  $p$  is also called the *orbit* of  $p$  when we want to emphasize the role of the initial point. It follows from the semigroup identities that distinct solution curves, regarded as subsets of  $X$ , do not intersect and so  $X$  is subdivided, *foliated*, by these curves. Changing  $p$  may shift us from one curve to another, but the motion given by the original differential equation is always on one of these curves.

The gestalt switch to the dynamical systems viewpoint occurs when we reverse the emphasis between  $t$  and  $p$ . Above we thought of  $p$  as a fixed parameter and  $t$  as the time variable along the solution path. Instead, we now think of the initial point, relabeled  $x$ , as our variable and the time value as the parameter.



For each fixed  $t$  value we define the time- $t$  map  $\phi^t: X \rightarrow X$  by  $\phi^t(x) = \phi(t, x)$ . For each point  $x \in X$  we ask whether it has moved in  $t$  units of time. The function  $\phi: T \times X \rightarrow X$  is called the *flow* of the system and the semigroup identities can be rewritten:

$$\phi^t \circ \phi^s = \phi^{t+s} \text{ for all } t, s \in T. \quad (3)$$

These simply say that the association  $t \mapsto \phi^t$  is a group homomorphism from the additive group  $T$  of real numbers to the automorphism group of  $X$ . In particular, observe that the time-0 map  $\phi^0$  is the identity map  $1_X$ .

We originally obtained the flow  $\phi$  by solving a differential equation. This requires differentiable structure on the underlying space which is why we specified that  $X$  be a Euclidean space or a manifold. The automorphism group is then the group of *diffeomorphisms* on  $X$ .

For the subject of topological dynamics we begin with a flow, a continuous map  $\phi$  subject to the condition (3). In modern parlance a flow is just a continuous group action of the group  $T$  of additive reals on the topological space  $X$ . The automorphism group is the group of *homeomorphisms* on  $X$ . If we replace the group of reals by letting  $T$  be the group of integers then the action is entirely determined by the generator  $f =_{\text{def}} \phi^1$ , the time 1 homeomorphism with  $\phi^n$  obtained by iterating  $f$   $n$  times if  $n$  is positive and iterating the inverse  $f^{-1}|n|$  times if  $n$  is negative.

Above I mentioned the requirement that the vector field  $\xi$  be smooth in order that the flow function  $\phi$  be defined. However, I neglected to point out that in the absence of some sort of boundedness condition the solution path might go to infinity in a finite time. In such a case the function  $\phi$  would not be defined on the entire domain  $T \times X$  but only on some open subset containing  $\{0\} \times X$ . The problems related to this issue are handled in the general theory by assuming that the space  $X$  is compact. Of course, Euclidean space is only locally compact. In applying the general theory to systems on a noncompact space one usually restricts to some compact invariant subset or else compactifies, i. e. embeds the system in one on a larger, compact space. Already in Nemytskii and Stepanov [52] much attention is devoted to conditions so that a solution path has a compact closure, see also Bahtia and Szego [27].

As topological dynamics has matured the theory has been extended to cover the action of more general topological groups  $T$ , usually countable and discrete, or locally compact, or *Polish* (admits a complete, separable metric). This was already emphasized in the first treatise which explicitly concerned topological dynamics, Gottschalk and Hedlund [38].

In differentiable dynamics we return to the case where the space  $X$  is a manifold and the flow is smooth. The breadth and depth of the results then obtained make it much more than a subfield of topological dynamics, see, for example, Katok and Hasselblatt [45]. For measurable dynamics we weaken the assumption of continuity to mere measurability but assume that the space carries a measure invariant with respect to the flow, see, for example, Petersen [54] and Rudolph [56]. The measurable and topological theories are especially closely linked with a number of parallel results, see Glasner [36] as well as Alpern and Prasad [12]. The reader should also take note of Oxtoby [53], a beautiful little book which explicitly describes this parallelism using a great variety of applications.

The current relationship between topological dynamics and dynamical systems theory in general is best understood by analogy with that between point-set, or general, topology and analysis.

General topology proper, even excluding algebraic topology and homotopy theory, is a large specialty with a rich history and considerable current research (for some representative surveys see the Russian Encyclopedia volumes [13,14,15]). But much of this work is little known to nonspecialists. On the other hand, the fundamentals of point-set topology are part of the foundation upon which modern analysis is built. Compactness was a rather new idea when it was used by Jesse Douglas in his solution of the Plateau Problem, Douglas [32] (see Almgren [11]). Nowadays continuity, compactness and connectedness are in the vocabulary of every analyst. In addition, there recur unexpected applications of hitherto specialized topics. For example, indecomposable continua, examined by Bing and his students, see Bing [25], are now widely recognized and used in the study of strange attractors, see Brown [26].

Similarly, the area of topological dynamics proper is large and some of the more technical results have found application. We will touch on some of these in the end, but for the most part this article will concentrate on those basic aspects which provide a foundation for dynamical systems theory in general. We will focus on chain recurrence and the theory of attractors, following the exposition of Akin [1] and Akin, Hurley and Kennedy [8].

To describe what we want to look for, let us begin with the simplest qualitative situation: the differential equation model (1) where the vector field  $\xi$  is the gradient of some smooth real-valued potential function  $U$  on  $X$ . Think of  $X$  as the Euclidean plane and the graph of  $U$  as a surface in space over  $X$ . The motion in  $X$  can be visualized on the surface above. On the surface it is always upward, perpendicular to the contour curves of constant height. For simplicity we will assume that  $U$  has isolated critical points.

These critical points: local maxima, minima and saddles, are *equilibria* for the system, points at which the gradient field  $\xi$  vanishes. We observe two kinds of behavior. The orbit of a critical point is constant, resting at equilibrium. The other kind exhibits what engineers call *transient* behavior. A non-equilibrium solution path moves asymptotically toward a critical point (or towards infinity). As it approaches its limit, the motion slows, becoming imperceptible, indistinguishable from rest at the limit point. The set of points whose orbits tend to a particular critical point  $e$  is called the *stable set* for  $e$ .

Each local maximum  $e$  is an *attractor* or *sink*. The stable set for  $e$  is an open set containing  $e$  which is called the *domain of attraction* for  $e$ . Such a state is called *asymptotically stable* illustrated by the rest state of a cone on its base.

The local minima are *repellers* or *sources* which are attractors for the system with time reversed. Solution paths near a repeller move away from it. The stable set for a repeller  $e$  consists of  $e$  alone. Consider a cone balanced on its point.

A saddle point between two local maxima is like the highest point of a pass between two mountains. Separating the domains of attraction for the two peaks are solution paths which have limit the saddle point equilibrium.

There is a kind of knife, used for cutting bread dough, which has a semi-circular blade. The saddle point equilibrium is a state like this knife balanced on the midpoint of its blade. A slight perturbation will cause it to fall down on one side or the other. But if you start the knife balanced elsewhere along its blade there remains the possibility – not achievable in practice – of its rolling back along the blade toward balance at the midpoint equilibrium. Notice that these are first-order systems with no momentum. Imagine everything going on in thick, clear molasses. As this model illustrates, it is usually true that the stable set of a saddle point is a lower dimensional set in  $X$  and the union of the domains of attraction of the local maxima is a dense open subset of  $X$ .

There do exist examples where the stable set of a saddle has nonempty interior. This is a pathology which we will hope to exclude by imposing various conditions. For example, the potential function  $U$  is called a *Morse function* when all of its critical points are nondegenerate. That is, the *Hessian* matrix of second partials is nonsingular at each critical point. For the gradient system of a Morse function each equilibrium is of a type called *hyperbolic*. For the saddle points of such a system the stable sets are manifolds of lower dimension.

From the cone example, we omitted what physicists call *neutral stability*, the cone resting on its side. From our point of view this is another sort of pathology: an infinite,

connected set of equilibria. Each of these equilibria is *stable* but not asymptotically stable. If we perturb the cone by lifting its point and turning it a bit, then it drops back toward an equilibrium near to but not necessarily identical with the original state (Remember, no momentum).

We obtain a similar classification into sinks, saddles, etc. and complementary transient behavior when we remove the assumption that the system comes from the gradient of  $U$  and retain only the condition that  $U$  increases along nonequilibrium solution paths. The function  $U$  is then called a *strict Lyapunov function* for the system. Instead of the steepest path ascent of a mountain goat, we may observe the spiraling upward of a car on a mountain road.

However, these gradient-like systems are too simple to represent a typical dynamical system. Lacking is the general behavior complementary to transience, namely *recurrence*. A point is recurrent – in some sense – if it is “in its own future” – in the appropriate sense. Beyond equilibrium the simplest kind of recurrence occurs on a periodic orbit. A periodic orbit returns infinitely often to each point on the orbit. As we will see, there are increasingly broad concepts of recurrence obtained by extending the notion of the “future” of a point. Clearly, a real-valued function cannot be strictly increasing along a periodic orbit. When we consider Lyapunov functions in general we will see that they remain constant along the orbit of each recurrent point.

Nonetheless, the picture we are looking for can be related to the gradient landscape by replacing the critical points by blobs of various sizes. Each blob is a closed, invariant set of a special type. A subset  $A$  is an *invariant set* for the system when  $A$  contains the entire orbit of each of its points. The special condition on each blob  $A$  which replaces an equilibrium is a kind of *transitivity*. This means that if  $p$  and  $q$  are points of  $A$  then each point is in the “future” of the other and so, in particular, each point is recurrent in the appropriate sense. Here again it remains to provide a meaning – or actually several different meanings – for this vague notion of “future”.

In this more general situation the transient orbits need not converge to a point. Instead, each accumulates on a closed subset of one of these blobs. If there are only finitely many of the blobs then there is a classification of them as attractor, repeller, or saddle analogous to the description for equilibria in the gradient system.

Within each blob the motion may be quite complicated. It is in attempting to describe such motions that the concept of *chaos* arises.

For some applied fields this sort of thinking is relatively new. When I learned population genetics – admit-

tedly that was over thirty years ago – most of the analysis consisted of identifying and classifying the equilibria, tricky enough in several variables. This is perfectly appropriate for gradient-like systems and is a good first step in any case. However, it has become apparent that more complicated recurrence may occur and so requires attention.

While most applications use differential equations and the associated flows, it is more convenient to develop the discrete time theory and then to derive from it the results for flows. In what follows we will describe the results for a *cascade*, a homeomorphism  $f$  on a compact metric space  $X$  with the dynamics introduced by iteration. Focusing on this case, we will not discuss further real flows or noninvertible functions, and we will omit as well the extensions to noncompact state spaces and to compact, nonmetrizable spaces.

### Dynamic Relations, Invariant Sets and Lyapunov Functions

It is convenient to assume that our state spaces are nonempty and metrizable. It is essential to assume that they are compact. Recall that if  $A$  is a subset of a compact metric space  $X$  then  $A$  is closed if and only if it is compact. Also, the continuous image of a compact set is compact and so for continuous maps between compact metric spaces the image of a closed set is closed. Perhaps less familiar are the following important results:

**Proposition 1** *Let  $X$  be a compact metric space and  $\{A_n\}$  be a decreasing sequence of closed subsets of  $X$  with intersection  $A$ .*

- (a) *If  $U$  is an open subset of  $X$  with  $A \subseteq U$  then for sufficiently large  $n$ ,  $A_n \subseteq U$ .*
- (b) *If  $A_n$  is nonempty for every  $n$ , then the intersection  $A$  is nonempty.*
- (c) *If  $h: X \rightarrow Y$  is a continuous map with  $Y$  a metric space then*

$$\bigcap_n h(A_n) = h(A). \quad (4)$$

*Proof* (a): We are assuming  $A_{n+1} \subseteq A_n$  for all  $n$  and  $A = \bigcap_n A_n$ . The complementary open sets  $V_n = X \setminus A_n$  are increasing and, together with  $U$ , they cover  $X$ . By compactness,  $\{U, V_1, \dots, V_N\}$  covers  $X$  for some  $N$  and so  $\{U, V_N\}$  suffice to cover  $X$ . Hence,  $A_N$  is a subset of  $U$  as  $A_n$  for any  $n \geq N$ .

(b): If  $A$  is empty then  $U = \emptyset$  is an open set containing  $A$  and so by (a),  $A_n$  is empty for sufficiently large  $n$ .

(c): Since  $A \subseteq A_n$  for all  $n$ , it is clear that  $h(A)$  is contained in  $\bigcap_n h(A_n)$ . On the other hand, if  $y$  is a point of

the latter intersection then  $\{h^{-1}(y) \cap A_n\}$  is a decreasing sequence of nonempty compact sets. By (b) the intersection  $h^{-1}(y) \cap A$  is nonempty.  $\square$

If  $\{B_n\}$  is any sequence of closed subsets of  $X$  then we define the *lim sup* :

$$\limsup_n B_n =_{\text{def}} \bigcap_n \overline{\bigcup_{k \geq n} B_k} \quad (5)$$

where  $\overline{Q}$  denotes the closure of  $Q$ . It follows from (b) above that the lim sup of a sequence of nonempty sets is nonempty. It is easy to check that

$$\overline{\bigcup_n B_n} = \left( \bigcup_n B_n \right) \cup \left( \limsup_n B_n \right). \quad (6)$$

We want to study a homeomorphism on a space. By compactness this is just a bijective (= one-to-one and onto) continuous function from the space to itself. It will be convenient to use the more general language of relations.

A function  $f: X \rightarrow Y$  is usually described as a rule associating to every point  $x$  in  $X$  a unique point  $y = f(x)$  in  $Y$ . In set theory the function  $f$  is defined to be the set of ordered pairs  $\{(x, f(x)): x \in X\}$ . Thus, the function  $f$  is a subset of the product  $X \times Y$ . It is what other people call the graph of the function. We will use this language so that, for example, the identity map  $1_X$  on  $X$  is the diagonal subset  $\{(x, x): x \in X\}$ . The notation is extended by defining a *relation from  $X$  to  $Y$* , written  $F: X \rightarrow Y$ , to be an arbitrary subset of  $X \times Y$ . Then  $F(x) = \{y: (x, y) \in F\}$ . Thus, a relation is a function exactly when the set  $F(x)$  contains a single point for every  $x \in X$ . In the function case, we will use the symbol  $F(x)$  for both the set and the single point it contains, the latter being the usual meaning of  $F(x)$ .

As they are arbitrary subsets of  $X \times Y$  we can perform set operations like union, intersection, closure and interior on relations. In addition, for  $F: X \rightarrow Y$  we define the *inverse*  $F^{-1}: Y \rightarrow X$  by

$$F^{-1} =_{\text{def}} \{(y, x): (x, y) \in F\}. \quad (7)$$

If  $A \subseteq X$  then its *image* is

$$\begin{aligned} F(A) &=_{\text{def}} \{y: (x, y) \in F \text{ for some } x \in A\} \\ &= \bigcup_{x \in A} F(x) = \pi_2((A \times Y) \cap F), \end{aligned} \quad (8)$$

where  $\pi_2: X \times Y \rightarrow Y$  is the projection to the second coordinate.

If  $G: Y \rightarrow Z$  is another relation then the *composition*  $G \circ F: X \rightarrow Z$  is the relation given by

$$\begin{aligned} G \circ F &=_{\text{def}} \{(x, z): \text{there exists } y \in Y \\ &\text{such that } (x, y) \in F \text{ and } (y, z) \in G\} \\ &= \pi_{13}((X \times G) \cap (F \times Z)), \end{aligned} \quad (9)$$

where  $\pi_{13}: X \times Y \times Z \rightarrow X \times Z$  is the projection map. This generalizes composition of functions and, as with functions, composition is associative. Clearly,  $(G \circ F)^{-1} = F^{-1} \circ G^{-1}$ .

We call  $F$  a *closed relation* when it is a closed subset of  $X \times Y$ . Clearly, the inverse of a closed relation is closed and by compactness, the composition of closed relations is closed. If  $A$  is a closed subset of  $X$  and  $F$  is a closed relation then the image  $F(A)$  is a closed subset of  $Y$ . For relations being closed is analogous to being continuous for functions. In fact, a function is continuous if and only if, regarded as a relation, it is closed. This is another application of compactness.

If  $Y = X$ , so that  $F: X \rightarrow X$ , then we call  $F$  a *relation on  $X$* . For a positive integer  $n$  we define  $F^n$  to be the  $n$ -fold composition of  $F$  with  $F^0 =_{\text{def}} 1_X$  and  $F^{-n} =_{\text{def}} (F^{-1})^n = (F^n)^{-1}$ . This is well-defined because composition is associative. Clearly,  $F^m \circ F^n = F^{m+n}$  when  $m$  and  $n$  have the same sign, i. e. when  $mn \geq 0$ . On the other hand, the equations  $F \circ F^{-1} = F^{-1} \circ F = 1_X = F^0$  all hold if and only if the relation  $F$  is a bijective function.

The utility of this relation-speak, once one gets used to it, is that it allows us to extend to this more general situation our intuitions about a function as a way of moving from input here to output there. For example, if  $\epsilon \geq 0$  then we can use the metric  $d$  on  $X$  to define the relations on  $X$

$$\begin{aligned} V_\epsilon &=_{\text{def}} \{(x, y): d(x, y) < \epsilon\}, \\ \bar{V}_\epsilon &=_{\text{def}} \{(x, y): d(x, y) \leq \epsilon\}. \end{aligned} \quad (10)$$

Thus,  $V_\epsilon(x)$  and  $\bar{V}_\epsilon(x)$  are the open ball and the closed ball centered at  $x$  with radius  $\epsilon$ . We can think of these relations as ways of moving from a point  $x$  to a nearby point.

Each  $\bar{V}_\epsilon$  is a closed, symmetric and reflexive relation. The triangle inequality is equivalent to the inclusion  $\bar{V}_\epsilon \circ \bar{V}_\delta \subseteq \bar{V}_{\epsilon+\delta}$ .

In general, a relation  $F$  on  $X$  is *reflexive* if  $1_X \subseteq F$ , *symmetric* if  $F^{-1} = F$  and *transitive* if  $F \circ F \subseteq F$ .

Now we apply this relation notation to a homeomorphism  $f$  on  $X$ .

For  $x \in X$  the *orbit sequence* of  $x$  is the bi-infinite sequence  $\{\dots, f^{-2}(x), f^{-1}(x), x, f(x), f^2(x), \dots\}$ . This is just the discrete time analogue of the solution path discussed in the [Introduction](#). We are thinking of it as a se-

quence and so as a function from the set  $\mathbb{Z}$  of discrete times to the state space  $X$  with parameter the initial point  $x$ .

Now we define the *orbit relation*

$$\mathcal{O}f =_{\text{def}} \bigcup_{n=1}^{\infty} f^n. \quad (11)$$

Thus,  $\mathcal{O}f(x) = \{f(x), f^2(x), \dots\}$  is a set, not a sequence, consisting of the states which follow the initial point  $x$  in time. Notice that for reasons which will be clear when we consider the cyclic sets below, we begin the union with  $n = 1$  rather than  $n = 0$  and so the initial point  $x$  itself need not be included in  $\mathcal{O}f(x)$ .

If with think of the point  $f(x)$  as the immediate temporal successor of  $x$  then  $\mathcal{O}f(x)$  is the set of points which occur on the orbit of  $x$  at some positive time. This is the first – and simplest – interpretation of the “future” of  $x$  with respect to the dynamical system obtained by iterating  $f$ .

It is convenient to extend this notion by including the limit points of the positive orbit sequence. For  $x \in X$  define

$$\begin{aligned} \omega f(x) &= \limsup_n \{f^n(x)\}, \\ \mathcal{R}f(x) &=_{\text{def}} \overline{\mathcal{O}f(x)} = \mathcal{O}f(x) \cup \omega f(x). \end{aligned} \quad (12)$$

Thus, from  $f$  we have defined the orbit relation  $\mathcal{O}f$  and the *orbit closure relation*  $\mathcal{R}f$  with  $\mathcal{R}f = \mathcal{O}f \cup \omega f$ .

While  $\mathcal{R}f(x)$  is closed for each  $x$ , the relation  $\mathcal{R}f$  itself is usually not closed. As was mentioned above, among relations the closed relations are the analogues of continuous functions. We obtain closed relations by defining

$$\begin{aligned} \Omega f &= \limsup_n f^n, \\ \mathcal{N}f(x) &=_{\text{def}} \overline{\mathcal{O}f(x)} = \mathcal{O}f(x) \cup \Omega f(x). \end{aligned} \quad (13)$$

Here we are taking the closure in  $X \times X$  and so we obtain closed relations. The relation  $\mathcal{N}f$ , defined by Auslander et al. [18,19,20] (see also Ura [63,64]), is called the *prolongation* of  $f$ .  $\mathcal{N}f(x)$  is our next, broader, notion of the “future” of  $x$ .

To compare all these suppose that  $x, y \in X$ . Then  $y \in \mathcal{O}f(x)$  when there exists a positive integer  $n$  such that  $y = f^n(x)$  while  $y \in \omega f(x)$  if there is a sequence of positive integers  $n_i \rightarrow \infty$  such that  $f^{n_i}(x) \rightarrow y$ . On the other hand,  $y \in \Omega f(x)$  iff there are sequences  $x_i \rightarrow x$  and  $n_i \rightarrow \infty$  such that  $f^{n_i}(x_i) \rightarrow y$ . Thus,  $y \in \mathcal{R}f(x)$  if for every  $\epsilon > 0$  we can run along the orbit of  $x$  and at some positive time make a smaller than  $\epsilon$  jump to  $y$ . Similarly,  $y \in \mathcal{N}f(x)$  if for every  $\epsilon > 0$  we can make an initial  $\epsilon$

small jump to a point  $x_1$ , run along the orbit of  $x_1$  and at some positive time make an  $\epsilon$  small jump to  $y$ .

The relations  $\mathcal{O}f$  and  $\mathcal{R}f$  are transitive but usually not closed. In general, when we pass to the closure, obtaining  $\mathcal{N}f$ , we lose transitivity. When we consider Lyapunov functions we will see why it is natural to want both of these properties, closure and transitivity.

The intersection of any collection of closed, transitive relations is a closed, transitive relation. Notice that  $X \times X$  is such a relation. Thus, we obtain  $\mathcal{G}f$ , the smallest closed, transitive relation which contains  $f$  by intersecting:

$$\mathcal{G}f =_{\text{def}} \bigcap \{Q \subseteq X \times X : \overline{Q} = Q \text{ and } f, Q \circ Q \subseteq Q\}. \quad (14)$$

There is an alternative procedure, due to Conley, see [29], which constructs a closed, transitive relation, generally larger than  $\mathcal{G}f$ , in a simple and direct fashion.

A *chain* or 0-chain is a finite or infinite sequence  $\{x_n\}$  such that  $x_{n+1} = f(x_n)$  along the way, i. e. a piece of the orbit sequence. Given  $\epsilon \geq 0$  an  $\epsilon$ -chain is a finite or infinite sequence  $\{x_n\}$  such that each  $x_{n+1}$  at most  $\epsilon$  distance away from the point  $f(x_n)$ , i. e.  $x_{n+1} \in \tilde{V}_\epsilon(f(x_n))$ . If the chain has at least two terms, but only finitely many, then the first and last terms are called the beginning and the end of the chain. The number of terms minus 1 is then called the *length* of the chain. We say that  $x$  *chains to*  $y$ , written  $y \in Cf(x)$  if for every  $\epsilon > 0$  there is an  $\epsilon$  chain which begins at  $x$  and ends at  $y$ . That is,

$$Cf =_{\text{def}} \bigcap_{\epsilon > 0} \mathcal{O}(\tilde{V}_\epsilon \circ f). \quad (15)$$

Compare this with (12) and (13). If  $y \in \mathcal{R}f(x)$  then we can get to  $y$  by moving along the orbit of  $x$  and then taking an arbitrarily small jump at the end. If  $y \in \mathcal{N}f(x)$  then we are allowed a small jump at the beginning as well as the end. Finally, if  $y \in Cf(x)$  then we are allowed a small jump at each iterative step.

The chain relation is of great importance for applications. Suppose we are computing the orbit of a point on a computer. At each step there is usually some round-off error. Thus, what we take to be an orbit sequence is in reality an  $\epsilon$  chain for some positive  $\epsilon$ . It follows that, in general, what we can expect to compute directly about  $f$  is only that level of information which is contained in  $Cf$ .

As the intersection of transitive relations,  $Cf$  is transitive. It is not hard to show directly that the chain relation  $Cf$  is also closed. Since  $x_1 = x$ ,  $x_2 = f(x)$  is a 0-chain beginning at  $x$  and ending at  $f(x)$ , we have  $f \subseteq Cf$ . It follows that  $\mathcal{G}f \subseteq Cf$ .

This inclusion may be strict. The identity map  $f = 1_X$  is already a closed equivalence relation. Hence,  $\mathcal{G}1_X = 1_X$ . On the other hand, for any  $\epsilon > 0$  the relation  $\mathcal{O}V_\epsilon$  is an open equivalence relation, and so each equivalence class is clopen (= closed and open). If  $X$  is connected then the entire space is a single equivalence class. Since this is true for every positive  $\epsilon$  we have

$$X \text{ connected} \implies C1_X = X \times X. \quad (16)$$

Since  $Cf$  is transitive, the composites  $\{(Cf)^n\}$  form a decreasing sequence of closed transitive relations. We denote by  $\Omega Cf$  the intersection of this sequence. That is,

$$\Omega Cf =_{\text{def}} \bigcap_{n=0}^{\infty} (Cf)^n. \quad (17)$$

One can show that  $y \in \Omega Cf(x)$  if and only if for every  $\epsilon > 0$  and positive integer  $N$  there is an  $\epsilon$  chain of length greater than  $N$  which begins at  $x$  and ends at  $y$ . In addition, the following identity holds (compare (12) and (13)):

$$Cf = \mathcal{O}f \cup \Omega Cf. \quad (18)$$

Thus, built upon  $f$  we have a tower of relations:

$$f \subseteq \mathcal{O}f \subseteq \mathcal{R}f \subseteq \mathcal{N}f \subseteq \mathcal{G}f \subseteq Cf. \quad (19)$$

These are the relations which capture the successively broader notions of the “future” of an input  $x$ .

A useful identity which holds for  $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$  and  $C$  is

$$\mathcal{A}f = f \cup (\mathcal{A}f) \circ f = f \cup f \circ (\mathcal{A}f). \quad (20)$$

These are easy to check directly for all but  $\mathcal{G}$ . For that one, observe that  $f \cup (\mathcal{G}f) \circ f$  and the other composite are closed and transitive and so each contains  $\mathcal{G}f$ .

It is also easy to show that for  $\mathcal{A} = \mathcal{O}, \mathcal{N}, \mathcal{G}$  and  $C$ :  $\mathcal{A}(f^{-1}) = (\mathcal{A}f)^{-1}$  and so we can omit the parentheses in these cases. The analogue for  $\mathcal{R}$  is usually false and we define

$$\alpha f =_{\text{def}} \omega(f^{-1}). \quad (21)$$

Thus,  $\alpha f(x)$  is the set of limit points of the negative time orbit sequence  $\{x, f^{-1}(x), f^{-2}(x), \dots\}$  and it is usually not true that  $\alpha f$  equals  $(\omega f)^{-1}$ .

For  $\Theta = \alpha, \omega, \Omega$  and  $\Omega C$  it is true that

$$\Theta f = f \circ \Theta f = f^{-1} \circ \Theta f = \Theta f \circ f = \Theta f \circ f^{-1}. \quad (22)$$

Now we are ready to consider the variety of recurrence concepts. Recall that a point  $x$  is recurrent – in some



sense – if it lies in its own “future”. Thus, for any relation  $F$  on  $X$  we define the *cyclic set*

$$|F| =_{\text{def}} \{x: (x, x) \in F\}. \quad (23)$$

Clearly, if  $F$  is a closed relation then  $|F|$  is a closed subset of  $X$ .

A point  $x$  lies in  $|f|$  when  $x = f(x)$  and so  $|f|$  is the set of *fixed points* for  $f$ , while  $x \in |\mathcal{O}f|$  when  $x = f^n(x)$  for some positive integer  $n$  and so  $|\mathcal{O}f|$  is the set of *periodic points*. It is easy to check that every periodic point is contained in  $|\omega f|$  and so  $|\omega f| = |\mathcal{R}f|$ . These are called *recurrent points* or sometimes the *positive recurrent points* to distinguish them from  $|\alpha f|$ , the set of *negative recurrent points*. Similarly, we have  $|\Omega f| = |\mathcal{N}f|$ , called the set of *nonwandering points*. The points of  $|\mathcal{G}f|$  are called *generalized nonwandering* and those of  $|Cf| = |\Omega Cf|$  are called *chain recurrent*. The set of periodic points and the sets of recurrent points need not be closed. The rest, associated with closed relations, are closed subsets.

For an illustration of these ideas, observe that for  $x, y, z \in X$

$$y, z \in \omega f(x) \implies z \in \Omega f(y). \quad (24)$$

Just hop from  $y$  to a nearby point on the orbit of  $x$ , moving arbitrarily far along the orbit, you repeatedly arrive nearby  $z$  and then can hop to it. In particular, with  $y = z$  we see that every point  $y$  of  $\omega f(x)$  is non-wandering. However, the points of  $\omega f(x)$  need not be recurrent. That is, while  $y \in \Omega f(y)$  for all  $y \in \omega f(x)$  it need not be true that  $y \in \omega f(y)$ . In particular,  $\mathcal{R}f(y)$  can be a proper subset of  $\mathcal{N}f(y)$ .

On the other hand, it is true that for most points  $x$  in  $X$  the orbit closure  $\mathcal{R}f(x)$  is equal to the prolongation set  $\mathcal{N}f(x)$ . Recall that a subset of a complete metric space is called *residual* when it is the countable intersection of dense, open subsets. By the Baire Category Theorem a residual subset is dense.

**Theorem 1** *If  $f$  is a homeomorphism on  $X$  then  $\{x \in X: \omega f(x) = \Omega f(x)\} = \{x \in X: \mathcal{R}f(x) = \mathcal{N}f(x)\}$  is a residual subset of  $X$ .*

In particular, if every point is nonwandering, i.e.  $x \in \Omega f(x)$  for all  $x$ , then the set of recurrent points is residual in  $X$ . However, if the set of nonwandering points is a proper subset of  $X$ , it need not be true that most of these points are recurrent. The closure of  $|\omega f|$  can be a proper subset of the closed set  $|\Omega f|$ .

From recurrence we turn to the notion of invariance.

Let  $F$  be a relation on  $X$  and  $A$  be a closed subset of  $X$ . We call  $A$  *+ invariant* for  $F$  if  $F(A) \subseteq A$  and *invariant* for

$F$  if  $F(A) = A$ . For a homeomorphism  $f$  on  $X$ , the set  $A$  is invariant for  $f$  if and only if it is + invariant for  $f$  and for  $f^{-1}$ . For example, (20) implies that for  $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$  and  $C$  each of the sets  $\mathcal{A}f(x)$  is + invariant for  $f$  and (22) implies that for  $\Theta = \alpha, \omega, \Omega$  and  $\Omega C$  each of the sets  $\Theta f(x)$  is invariant for  $f$ . Finally, for  $\mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}$  and  $C$  each of the cyclic sets  $|\mathcal{A}f|$  is invariant for  $f$ .

If  $A$  is a nonempty, closed invariant subset for a homeomorphism  $f$  then the restriction  $f|_A$  is a homeomorphism on  $A$  and we call this dynamical system the *subsystem* determined by  $A$ . In general, if  $F$  is a relation on  $X$  and  $A$  is any subset of  $X$  then we call the relation  $F \cap (A \times A)$  on  $A$  the *restriction* of  $F$  to  $A$ .

For a homeomorphism  $f$  the families of + invariant subsets and of invariant subsets are each closed under the operations of closure and interior and under arbitrary unions and intersections. If  $A$  is + invariant then the sequence  $\{f^n(A)\}$  is decreasing and the intersection is  $f$  invariant. Furthermore, this intersection contains every other  $f$  invariant subset of  $A$  and so is the maximum invariant subset of  $A$ .

If  $A$  is + invariant for  $f$  then it is + invariant for  $\mathcal{O}f$ . If, in addition,  $A$  is closed then it is + invariant for  $\mathcal{R}f$ . However, + invariance with respect to the later relations in the tower (19) are successively stronger conditions, and the relations of (19) provide convenient tools for studying these conditions.

We call a closed + invariant subset  $A$  a *stable* subset, or a *Lyapunov stable* subset, if it has a neighborhood basis of + invariant neighborhoods. That is, if  $G$  is open and  $A \subseteq G$  then there exists a + invariant open set  $U$  such that  $A \subseteq U \subseteq G$ .

**Theorem 2** *A closed subset  $A$  is + invariant for  $\mathcal{N}f$  if and only if it is a stable + invariant set for  $f$ .*

*Proof* If  $G$  is an open set which contains  $A$  and  $\mathcal{N}f(A) \subseteq A$  then  $U = \{x: \mathcal{N}f(x) \subseteq G\}$  is an open set which contains  $A$  and which is + invariant by (20). The reverse implication is easy to check directly.  $\square$

Invariance with respect to  $\mathcal{G}f$  is characterized by using Lyapunov functions, which generalize the strict Lyapunov functions described in the introduction.

For a closed relation  $F$  on  $X$ , a Lyapunov function  $L$  for  $F$  is a continuous, real-valued function on  $X$  such that

$$(x, y) \in F \implies L(x) \leq L(y), \quad (25)$$

(some authors, e.g. Lyapunov, use the reverse inequality).

For any continuous, real-valued function  $L$  on  $X$  the set  $\{(x, y): L(x) \leq L(y)\}$  is a closed, transitive relation on  $X$ . To say that  $L$  is a Lyapunov function for  $F$  is exactly to say that this relation contains  $F$ . It follows that if

$L$  is a Lyapunov function for a homeomorphism  $f$  then it is automatically a Lyapunov function for the closed, transitive relation  $\mathcal{G}f$ . That  $L$  be a Lyapunov function for  $Cf$  is usually a stronger condition. For example, any continuous, real-valued function is a Lyapunov function for  $1_X$ , but by (16) if  $X$  is connected then constant functions are the only Lyapunov functions for  $C1_X$ .

The following result is a dynamic analogue of Urysohn's Lemma in general topology and it has a similar proof.

**Theorem 3** *A closed subset  $A$  is  $+$  invariant for  $\mathcal{G}f$  if and only if there exists a Lyapunov function  $L: X \rightarrow [0, 1]$  for  $f$  such that  $A = L^{-1}(1)$ .*

A Lyapunov function for a homeomorphism  $f$  is non-decreasing on each orbit sequence  $x, f(x), f^2(x), \dots$ . Suppose that  $x$  is a periodic point, i.e.  $x \in |\mathcal{O}f|$ . Then  $x = f^n(x)$  for some positive integer  $n$ , and it follows that  $L$  must be constant on the orbit of  $x$ .

Now suppose, more generally, that  $x$  is a generalized recurrent point, i.e.  $x \in |\mathcal{G}f|$ . By (20)  $(f(x), x) \in \mathcal{G}f$  and so  $L(f(x)) = L(x)$  whenever  $L$  is a Lyapunov function for  $f$  (and hence for  $\mathcal{G}f$ ). Recall that the set  $|\mathcal{G}f|$  of generalized recurrent points is  $f$  invariant. It follows that  $f^n(x) \in |\mathcal{G}f|$  for every integer  $n$  and so  $L(f^{n+1}(x)) = L(f^n(x))$  for all  $n$ . Thus, a Lyapunov function for a homeomorphism  $f$  is constant on the orbit of each generalized recurrent point  $x$ .

Similarly, if  $x$  is a chain recurrent point, i.e.  $x \in |Cf|$ , then  $L(f(x)) = L(x)$  whenever  $L$  is a Lyapunov function for  $Cf$  and so a Lyapunov function for  $Cf$  is constant on the orbit of each chain recurrent point.

Of fundamental importance is the observation that one can construct Lyapunov functions for  $f$  (and for  $Cf$ ) which are increasing on all orbit sequences which are not generalized recurrent (respectively, chain recurrent).

**Theorem 4** *For a homeomorphism  $f$  on a compact metric space  $X$  there exist continuous functions  $L_1, L_2: X \rightarrow [0, 1]$  such that  $L_1$  is a Lyapunov function for  $f$ , and hence for  $\mathcal{G}f$ , and  $L_2$  is a Lyapunov function for  $Cf$  and, in addition,*

$$\begin{aligned} x \in |\mathcal{G}f| &\iff L_1(x) = L_1(f(x)), \\ x \in |Cf| &\iff L_2(x) = L_2(f(x)). \end{aligned} \quad (26)$$

Lyapunov functions which satisfy the conditions of (26) are called *complete Lyapunov functions* for  $f$  and for  $Cf$ , respectively.

If  $L$  is a Lyapunov function for  $f$  then we define the set of *critical points* for  $L$

$$|L| =_{\text{def}} \{x \in X: L(x) = L(f(x))\}. \quad (27)$$

This language is a bit abusive because here criticality describes a relationship between  $L$  and  $f$ . It does not depend only upon  $L$ . However, we adopt this language to compare the general situation with the simpler strict Lyapunov function case in the introduction. Similarly, we call  $L(|L|) \subseteq \mathbb{R}$  the set of *critical values* for  $L$ . The complementary points of  $X$  and  $\mathbb{R}$  respectively are called *regular points* and *regular values* for  $L$ . Thus, the generalized recurrent points are always critical points for a Lyapunov function  $L$  and for a complete Lyapunov function these are the only critical points.

For invariance with respect to  $Cf$  we turn to the study of attractors.

## Attractors and Chain Recurrence

For a homeomorphism  $f$  on  $X$  we say that a closed set  $U$  is *inward* if  $f(U)$  is contained in  $U^\circ$ , the interior of  $U$ . By compactness this implies that  $\tilde{V}_\epsilon \circ f(U) \subseteq U$  for some  $\epsilon > 0$ . That is,  $U$  is  $+$  invariant for the relation  $\tilde{V}_\epsilon \circ f$ . Hence, any  $\epsilon$  chain for  $f$  which begins in  $U$  remains in  $U$ . It follows that an inward set for  $f$  is  $Cf$   $+$  invariant.

For example, assume that  $L$  is a Lyapunov function for  $f$ . Then for any  $s \in \mathbb{R}$  the closed set  $U_s = \{x: L(x) \geq s\}$  is  $+$  invariant for  $f$ . Suppose now that  $s$  is a regular value for  $L$ . This means that for all  $x$  such that  $L(x) = s$  we have  $L(f(x)) > L(x) = s$ . On the other hand, for the remaining points  $x$  of  $U_s$  we have  $L(f(x)) \geq L(x) > s$ . Thus,  $f(U_s)$  is contained in the open set  $\{x: L(x) > s\} \subseteq U_s$  and so  $U_s$  is inward. It easily follows that  $L$  is a Lyapunov function for  $Cf$  if the set of critical values is nowhere dense.

If  $U$  is inward for  $f$  then we define  $A = \bigcap_{n=0}^{\infty} \{f^n(U)\}$  to be the *attractor* associated with  $U$ . Since an inward set  $U$  is  $+$  invariant for  $f$ , the sequence  $\{f^n(U)\}$  is decreasing. In fact, for  $U$  inward and  $n, m \in \mathbb{Z}$  we have

$$n > m \implies f^n(U) \subseteq f^m(U)^\circ. \quad (28)$$

The associated attractor is the maximum  $f$  invariant subset of  $U$  and  $\{f^n(U): n \in \mathbb{Z}\}$  is a sequence of inward neighborhoods of  $A$ , forming a neighborhood basis for the set  $A$ . That is, if  $G$  is any open which contains  $A$  then by Proposition 1(a),  $f^n(U) \subseteq G$  for sufficiently large  $n$ .

For example, the entire space  $X$  is an inward set and is its own associated attractor. In general, a set  $A$  is inward and equal to its own attractor if and only if  $A$  is a clopen,  $f$  invariant set.

The power of the attractor idea comes from the equivalence of a number of descriptions of different apparent

strength. We use the “weak” ones to test for an attractor and then apply the “strong” conditions. The following collects these alternative descriptions.

**Theorem 5** *Let  $f$  be a homeomorphism on  $X$  and  $A$  be a closed  $f$  invariant subset of  $X$ . The following conditions are equivalent.*

- (i)  $A$  is an attractor. That is, there exists an inward set  $U$  such that  $\bigcap_{n=0}^{\infty} f^n(U) = A$ .
- (ii) There exists a neighborhood  $G$  of  $A$  such that  $\bigcap_{n=0}^{\infty} f^n(G) \subseteq A$ .
- (iii)  $A$  is  $\mathcal{N}f$  + invariant and the set  $\{x: \omega f(x) \subseteq A\}$  is a neighborhood of  $A$ .
- (iv) The set  $\{x: \Omega f(x) \subseteq A\}$  is a neighborhood of  $A$ .
- (v) The set  $\{x: \Omega Cf(x) \subseteq A\}$  is a neighborhood of  $A$ .
- (vi)  $A$  is  $\mathcal{G}f$  + invariant and the set  $A \cap |\mathcal{G}f|$  is clopen in the relative topology of the closed set  $|\mathcal{G}f|$ .
- (vii)  $A$  is  $Cf$  + invariant and the set  $A \cap |Cf|$  is clopen in the relative topology of the closed set  $|Cf|$ .

Applying Theorem 2 to condition (iii) of Theorem 5 we see that if  $A$  is a stable set for  $f$  and, in addition, the orbit of every point in some neighborhood of  $x$  tends asymptotically toward  $A$  then  $A$  is an attractor. The latter condition alone does not suffice, although the strengthening in (iv) is sufficient. For example, the homeomorphism of  $[0, 1]$  defined by  $t \mapsto t^2$  has  $\{0\}$  as an attractor. If we identify the two fixed points  $0, 1 \in [0, 1]$  by mapping  $t$  to  $z = e^{2\pi i t}$  then we obtain a homeomorphism  $f$  on the unit circle  $X$ . For every  $z \in X$ , we have  $\omega f(z) = \{1\}$ , the unique fixed point. However,  $\{1\}$  is not an attractor for  $f$ . In fact,  $\Omega f(1) = X$ .

The class of attractors is closed under finite union and finite intersection. Using infinite intersections we can characterize  $Cf$  invariance.

**Theorem 6** *Let  $f$  be a homeomorphism on  $X$  and  $A$  be a closed  $f$  invariant subset of  $X$ . The following conditions are equivalent. When they hold we call  $A$  a quasi-attractor for  $f$ .*

- (i)  $A$  is  $Cf$  + invariant.
- (ii)  $A$  is the intersection of a (possibly infinite) set of attractors.
- (iii) The set of inward neighborhoods of  $A$  form a basis for the neighborhood system of  $A$ .

From Theorem 2 again it follows that a quasi-attractor is stable.

If  $A$  is a closed invariant set for a homeomorphism  $f$  then  $A$  is called an *isolated invariant set* if it is the maxi-

mum invariant subset of some neighborhood  $U$  of  $A$ . That is,

$$A = \bigcap_{n=-\infty}^{+\infty} f^n(U). \quad (29)$$

In that case,  $U$  is called an *isolating neighborhood* for  $A$ . Notice that if  $U$  is a closed isolating neighborhood for  $A$  and the positive orbit of  $x$  remains in  $U$ , i. e.  $f^n(x) \in U$  for  $n = 0, 1, \dots$  then  $\omega f(x) \subseteq A$  because  $\omega f(x)$  is an invariant subset of  $U$ .

Since an attractor is the maximum invariant subset of some inward set, it follows that an attractor is isolated. Conversely, by condition (iii) of Theorem 5 an invariant set is an attractor precisely when it is isolated and stable. In particular, a quasi-attractor is an attractor exactly when it is an isolated invariant set.

An attractor for  $f^{-1}$  is called a *repellor* for  $f$ . If  $U$  is an inward set for  $f$  then  $X \setminus \overline{U} = X \setminus (U^\circ)$  is an inward set for  $f^{-1}$  and  $B = \bigcap_{n=0}^{\infty} f^{-n}(X \setminus \overline{U})$  is the associated repellor for  $f$ . We call  $B$  the repellor *dual* to the attractor  $A = \bigcap_{n=0}^{\infty} f^n(U)$ . Recall that an  $f$  invariant set is  $f^{-1}$  invariant. In particular, attractors and repellors are both  $f$  and  $f^{-1}$  invariant. The open set  $\bigcup_{n=0}^{\infty} f^{-n}(U) = X \setminus B$  is called the *domain of attraction* for  $A$ . The name comes from the implication:

$$\begin{aligned} x \in X \setminus B &\implies \omega(x) \subseteq \Omega Cf(x) \subseteq A, \\ x \in X \setminus A &\implies \alpha(x) \subseteq \Omega Cf^{-1}(x) \subseteq B. \end{aligned} \quad (30)$$

For example, the entire space  $X$  is both an attractor and a repellor with dual  $\emptyset$ .

If  $A$  is an attractor then we call the set of chain recurrent points in  $A$ , i. e.  $A \cap |Cf|$ , the *trace* of the attractor  $A$ . An attractor is determined by its trace via the equation

$$Cf(A \cap |Cf|) = A. \quad (31)$$

The trace of an attractor is a clopen subset of  $|Cf|$  by part (vii) of Theorem 5. Conversely, suppose  $A_0$  is a subset of  $|Cf|$  which is + invariant for the restriction of  $Cf$  to  $|Cf|$ . That is,  $x \in A_0$  and  $y \in Cf(x) \cap |Cf|$  implies  $y \in A_0$ . If  $A_0$  is clopen in  $|Cf|$ , then  $Cf(A_0)$  is the attractor with trace  $A_0$  and  $Cf^{-1}(|Cf| \setminus A_0)$  is the dual repellor. By Theorem 6 if  $A_0$  is merely closed then  $Cf(A_0)$  is a quasi-attractor.

With the relative topology the set  $|Cf| = |Cf^{-1}|$  of chain recurrent points is a compact metric space and so has only countably many clopen subsets (Every clopen subset is a finite union of members of a countable basis for the topology). It follows that, while there are often

uncountably many inward sets, there are only countably many attractors.

When restricted to  $|Cf| = |Cf^{-1}|$  the closed relation  $Cf \cap Cf^{-1}$  is reflexive as well as symmetric and transitive. The individual equivalence classes are closed  $f$  invariant subsets of  $X$  called the *chain components* of  $f$ , or the *basic sets* of  $f$ . These chain components are the analogues of the individual critical points in gradient case described in the introduction.

Any two points of a chain component are related by  $Cf$ . This is a type of transitivity condition. As with recurrence, there are several – increasingly broad – notions of dynamic transitivity and these can be associated with the relations of (19). First, we consider when the entire system  $f$  on  $X$  is transitive in the some way. Then we say that a closed  $f$  invariant subset  $A$  is a transitive subset in this way, when the subsystem  $f|_A$  on  $A$  is transitive in the appropriate way.

A group action on a set is called transitive when one can move from any element of the set to any other by some element of the group. That is, the entire set is a single orbit of the group action. Notice that this use of the word is unrelated to transitivity of a relation. Recall that a relation  $F$  on a set  $X$  is transitive when  $F \circ F \subseteq F$ . We are now considering when it happens that any two points of  $X$  are related by  $F$ . This just says that  $F = X \times X$  which we will call the *total relation* on  $X$ .

First, what does it mean to say that  $\mathcal{O}f$  is total, that is to say all of  $X$  lies in a single orbit of  $f$ ? First, compactness implies that  $X$  is then finite and so the homeomorphism  $f$  is a permutation of the finite set  $X$ . Such permutation is a product of disjoint cycles and  $\mathcal{O}f = X \times X$  exactly when all of  $X$  consists of a single cycle. The associated invariant subsets are the periodic orbits of  $f$ , including the fixed points.

Next, we consider when the orbit closure relation  $\mathcal{R}f$  is total, that is, when every point is in the orbit closure of every other.

**Theorem 7** *Let  $f$  be a homeomorphism on  $X$ . The following conditions are equivalent and when they hold we call  $f$  minimal.*

- (i) *For all  $x \in X$ ,  $\mathcal{O}f(x)$  is dense in  $X$ .*
- (ii) *For all  $x \in X$ ,  $\mathcal{R}f(x) = X$ , i. e.  $\mathcal{R}f = X \times X$ .*
- (iii) *For all  $x \in X$ ,  $\omega f(x) = X$ , i. e.  $\omega f = X \times X$ .*
- (iv)  *$X$  is the only nonempty, closed  $f +$  invariant subset of  $X$ .*
- (v)  *$X$  is the only nonempty, closed  $f$  invariant subset of  $X$ .*

Recall our convention that the state space of a dynamical system is nonempty, although we do allow the empty set as

an invariant subset. For example, the empty set is the repeller/attractor dual to the attractor/repeller which is the entire space.

Thus, a closed  $f$  invariant subset  $A$  of  $X$  is *minimal* when it is nonempty but contains no nonempty, proper  $f$  invariant subset. From compactness it follows via the usual Zorn's Lemma argument that every nonempty, closed  $f +$  invariant subset of  $X$  contains a minimal, nonempty, closed  $f$  invariant subset.

**Theorem 8** *Let  $f$  be a homeomorphism on  $X$ . The following conditions are equivalent and when they hold we call  $f$  topologically transitive.*

- (i) *For some  $x \in X$ ,  $\mathcal{O}f(x)$  is dense in  $X$ , i. e.  $\mathcal{R}f(x) = X$ .*
- (ii) *For all  $x \in X$ ,  $\mathcal{N}f(x) = X$ , i. e.  $\mathcal{N}f = X \times X$ .*
- (iii) *For all  $x \in X$ ,  $\Omega f(x) = X$ , i. e.  $\Omega f = X \times X$ .*
- (iv)  *$X$  is the only closed  $f +$  invariant subset with a nonempty interior.*

If  $f$  is topologically transitive then the set  $\{x: \omega f(x) = X\}$  is residual, i. e. it is the countable intersection of dense, open subsets of  $X$ .

Every point in a minimal system has a dense orbit. If  $f$  is topologically transitive then the set of *transitive points* for  $f$ ,

$$Trans_f =_{\text{def}} \{x: \omega f(x) = X\} = \{x: \mathcal{R}f(x) = X\}, \quad (32)$$

is dense by the Theorem 8. However, its complement is either empty, which is the minimal case, or else is dense as well. Also, there is a usually rich variety of invariant subsets. It may happen, for instance, that the set of periodic points is dense. Most well-studied examples of chaotic dynamical systems are non-minimal, topologically transitive systems. In fact, Devaney used the conjunction of topological transitivity and density of periodic points in an infinite system as a definition of chaos.

The broadest notion of transitivity is associated with the chain relation  $Cf$ .

**Theorem 9** *Let  $f$  be a homeomorphism on  $X$ . The following conditions are equivalent and when they hold we call  $f$  chain transitive.*

- (i) *For all  $x \in X$ ,  $Cf(x) = X$ , i. e.  $Cf = X \times X$ .*
- (ii) *For all  $x \in X$ ,  $\Omega Cf(x) = X$ , i. e.  $\Omega Cf = X \times X$ .*
- (iii)  *$X$  is the only nonempty inward set.*
- (iv)  *$X$  is the only nonempty attractor.*

Before proceeding further, we pause to observe that each of these three concepts is the same for  $f$  and for its inverse.



Because the  $f$  invariant subsets are the same as the  $f^{-1}$  invariant subsets, we see that  $f^{-1}$  is minimal when  $f$  is. Since  $\mathcal{N}(f^{-1}) = (\mathcal{N}f)^{-1}$  and  $C(f^{-1}) = (Cf)^{-1}$ , it follows as well that  $f^{-1}$  is topologically transitive or chain transitive when  $f$  satisfies the corresponding property.

There are many naturally occurring chain transitive subsets. For every  $x \in X$  the limit sets  $\alpha f(x)$  and  $\omega f(x)$  are chain transitive subsets as are each of the chain components. Notice that if  $x$  and  $y$  are points of some chain component  $B$  then by definition of the equivalence relation  $Cf \cap Cf^{-1}$ ,  $x$  and  $y$  can be connected by  $\epsilon$  chains in  $X$  for any positive  $\epsilon$ . To say that  $B$  is a chain transitive subset is to make the stronger statement that they can be connected via  $\epsilon$  chains which remain in  $B$ .

Contrast this positive result with the trap set by implication (24) which suggests that  $B = \omega f(x)$  is a topologically transitive subset. However, (24), which says  $B \times B \subseteq \mathcal{N}f$ , does not imply that the restriction  $f|B$  is topologically transitive, i. e.  $B \times B = \mathcal{N}(f|B)$ . It may not be possible to get from near  $y$  to near  $z$  without a hop which takes you outside  $B$ .

In fact, if  $f$  is any chain transitive homeomorphism on a space  $X$  then it is possible to embed  $X$  as a closed subset of a space  $Y$  and extend  $f$  to a homeomorphism  $g$  on  $Y$  in such a way that  $X = \omega g(y)$  for some point  $y \in Y$ .

Any chain transitive subset for  $f$  is contained in a unique chain component of  $|Cf|$ . In fact the chain components are precisely the maximal chain transitive subsets. In particular, each subset  $\alpha f(x)$  and  $\omega f(x)$  is contained in some chain component of  $f$ .

By using (16) one can show that each connected component of the closed subset  $|Cf|$  is contained in some chain component as well. It follows that the space of chain components, that is, the space of  $Cf \cap Cf^{-1}$  equivalence classes with the quotient topology from  $|Cf|$ , is zero-dimensional. So this space is either countable or is the union of a Cantor set with a countable set.

An individual chain component  $B$  is called *isolated* when it is a clopen subset of the chain recurrent set  $|Cf|$  or, equivalently, if it is an isolated point in the space of chain components. Thus,  $B$  is isolated when it has a neighborhood  $U$  in  $X$  such that  $U \cap |Cf| = B$ . It can be proved that  $B$  is an isolated chain component precisely when it is an isolated invariant set, i. e. it admits a neighborhood  $U$  satisfying (29).

The individual chain components generalize the role played by the critical points in the gradient case. They are the blobs we described at the end of the introduction. We complete the analogy by identifying which chain components are like the relative maxima and minima among the critical points.

**Theorem 10** *Let  $B$  be a nonempty, closed,  $f$  invariant set for a homeomorphism  $f$  on  $X$ . The following conditions are equivalent and when they hold we call  $B$  a terminal chain component.*

- (i)  $B$  is a chain transitive quasi-attractor.
- (ii)  $B$  is a  $Cf +$  invariant, chain transitive subset of  $X$ .
- (iii)  $B$  is a  $Cf +$  invariant chain component.
- (iv) In the set of nonempty, closed  $Cf +$  invariant subsets of  $X$  the set  $B$  is an element which is minimal with respect to inclusion.

*In particular,  $B$  is a chain transitive attractor if and only if it is an isolated, terminal chain component.*

From condition (iv) and Zorn's Lemma it follows that every nonempty, closed  $Cf +$  invariant subset of  $X$  contains a terminal chain component. In particular,  $\Omega Cf(x)$  contains a terminal chain component for every  $x \in X$ .

Clearly, if there are only finitely many chain components then each terminal chain component is isolated and so is an attractor.

**Corollary 1** *Let  $f$  be a homeomorphism on  $X$  and let  $B$  be a nonempty, closed,  $Cf +$  invariant subset of  $X$ . If for some  $x \in X$ ,  $B \subseteq \omega f(x)$  then  $B$  is a terminal chain component and  $B = \omega f(x)$ .*

*Proof*  $B$  contains some terminal chain component  $B_1$  and  $\omega f(x)$  is contained in some chain component  $B_2$ . Thus,  $B_1 \subseteq B \subseteq \omega f(x) \subseteq B_2$ . Since distinct chain components are disjoint,  $B_1 = B_2$ .  $\square$

Because  $C(f^{-1}) = (Cf)^{-1}$ , the chain components for  $f$  and  $f^{-1}$  are the same. A chain component is called an *initial chain component* for  $f$  when it is a terminal chain component for  $f^{-1}$ .

Let  $L$  be a Lyapunov function for  $Cf$ .  $L$  is constant on each chain component and Theorem 4 says that  $L$  can be constructed to be strictly increasing on the orbit of every point of  $X \setminus |Cf|$ . That is, for such a complete Lyapunov function the set  $|L|$  of critical points equals  $|Cf|$ . In that case, the local maxima occur at terminal chain components and, as a partial converse, an isolated terminal chain component is a local maximum for any complete  $Cf$  Lyapunov function.

**Theorem 11** *For a homeomorphism  $f$  on  $X$  let  $L$  be a  $Cf$  Lyapunov function and let  $B$  be a chain component with  $r$  the constant value of  $L$  on  $B$ .*

- (a) *If there exists a neighborhood  $U$  of  $B$  such that  $L(x) < r$  for all  $x \in U \setminus B$ , then  $B$  is a terminal chain component.*



- (b) Assume  $L$  is a complete Lyapunov function. If  $B$  is an attractor then it is a terminal chain component and the domain of attraction  $U$  is an open set containing  $B$  such that  $L(x) < r$  for all  $x \in U \setminus B$ .

In particular, if there are only finitely many chain components for  $f$  then by using  $L$ , a complete  $Cf$  Lyapunov function which distinguishes the chain components, we obtain the picture promised at the end of the introduction. The local maxima of  $L$  occur at the terminal chain components which are attractors. The local minima are at the initial chain components which are repellers. The remaining chain components play the role of saddles. As in the gradient case, it can happen that there is an open set of points  $x$  such that  $\omega f(x)$  is contained in one of these saddle chain components. There is a natural topological condition which, when it holds, excludes this pathology.

**Theorem 12** For a homeomorphism  $f$  on  $X$ , assume that  $\mathcal{N}f = Cf$  or, equivalently, that  $\Omega f = \Omega Cf$ .

If  $A$  is a stable closed  $f$  invariant subset of  $X$  then  $A$  is a quasi-attractor.

For a residual set of points  $x$  in  $X$ ,  $\omega f(x)$  is a terminal chain component and  $\alpha f(x)$  is an initial chain component.

*Proof* The result for a stable set  $A$  follows from the characterizations in Theorems 2 and 6

By Theorem 1 the set of  $x$  such that  $\omega f(x) = \Omega f(x)$  is always residual. By assumption this agrees with the set of  $x$  such that  $\omega f(x) = \Omega Cf(x)$ . Corollary 1 implies that for such  $x$  the set  $\omega f(x)$  is a terminal chain component. For  $\alpha f(x)$  apply this result to  $f^{-1}$ .  $\square$

If  $X$  is a compact manifold of dimension at least two, then the condition  $\mathcal{N}f = Cf$  holds for a residual set in the Polish group of homeomorphisms on  $X$  with the uniform topology. If, in addition,  $f$  has only finitely many chain components (and this is not a residual condition on  $f$ ) then it follows that the points which are in the domain of attraction of some terminal chain component and in the domain of repulsion of some initial chain component form a dense, open subset of  $X$ .

### Chaos and Equicontinuity

Any chain transitive system can occur as a chain component, but in the most interesting cases the chain components are topologically transitive subsets. In this section we consider the antithetical phenomena of equicontinuity and chaos in topologically transitive systems.

If  $f$  is a homeomorphism on  $X$  and  $x \in X$  then  $x$  is called a *transitive point* when its orbit  $\mathcal{O}f(x)$  is dense in  $X$ .

As in (32) we denote by  $\text{Trans}_f$  the set of transitive points for  $f$ . The system is minimal exactly when every point is transitive, i. e. when  $\text{Trans}_f = X$ .

To study equicontinuity we introduce a new metric  $d_f$  defined using the original metric  $d$  on  $X$ .

$$d_f(x, y) =_{\text{def}} \sup \{d(f^n(x), f^n(y)) : n = 0, 1, 2, \dots\}. \quad (33)$$

It is easy to check that  $d_f$  is a metric, i. e. it satisfies the conditions of positivity and symmetry as well as the triangle inequality. However, it is usually not topologically equivalent to the original metric  $d$ , generating a topology which is usually strictly finer (= more open sets). We call a point  $x \in X$  an *equicontinuity point* for  $f$  when for every  $\epsilon > 0$  there exists a  $\delta > 0$  so that for all  $y \in X$

$$d(x, y) < \delta \implies d_f(x, y) \leq \epsilon, \quad (34)$$

or, equivalently, if for every  $\epsilon > 0$  there exists a neighborhood  $U$  of  $x$  with  $d_f$  diameter at most  $\epsilon$  (the terms “neighborhood”, “open set”, etc. will always refer to the original topology given by  $d$ ). Here the  $d_f$  diameter of a subset  $A \subseteq X$  is

$$\text{diam}_f(A) =_{\text{def}} \sup \{d_f(x, y) : x, y \in A\}. \quad (35)$$

We denote the, possibly empty, set of equicontinuity points for  $f$  by  $Eq_f$ .

When every point is an equicontinuity point, the system associated with  $f$  is called *equicontinuous*, or we just say that  $f$  is equicontinuous. This coincides with the concept of equicontinuity of the set of functions  $\{f^n : n = 1, 2, \dots\}$ . Recall that any finite set of continuous functions is equicontinuous, but equicontinuity for an infinite set is often, as in this case, a strong condition. Equicontinuity of  $f$  says exactly that the metrics  $d$  and  $d_f$  are topologically equivalent. For compact spaces topologically equivalent metrics are uniformly equivalent. Hence, if  $f$  is equicontinuous, then for every  $\epsilon > 0$  there exists a  $\delta > 0$  such that for all  $x, y \in X$  implication (34) holds.

When  $f$  is equicontinuous then we can replace  $d$  by  $d_f$  since in the equicontinuous case the latter is a metric giving the correct topology. The homeomorphism  $f$  is then an *isometry* of the metric, i. e.  $d_f(x, y) = d_f(f(x), f(y))$  for all  $x, y \in X$ . This equality uses a famous little problem which keeps being rediscovered: If  $f$  is a surjective map on a compact metric space  $X$  with metric  $d$  such that  $d(f(x), f(y)) \leq d(x, y)$  for every  $x, y \in X$  then  $f$  is an isometry on  $X$ , see e.g. Alexopoulos [9] or Akin [3], Proposition 2.4(c).

Conversely, if  $f$  is an isometry of a metric  $d$  on  $X$  (with the correct topology) then  $f$  is clearly equicontinuous.

When  $X$  has a dense set of equicontinuity points we call the system *almost equicontinuous*. If  $f$  is almost equicontinuous but not equicontinuous then the  $\delta$  in (34) will depend upon  $x \in Eq_f$  as well as upon  $\epsilon$ .

On the other hand, we say that the system *has sensitive dependence upon initial conditions*, or simply that  $f$  is *sensitive* when there exists  $\epsilon > 0$  such that  $\text{diam}_f(U) > \epsilon$  for every nonempty open subset  $U$  of  $X$ , or, equivalently, if there exists  $\epsilon > 0$  such that for any  $x \in X$  and  $\delta > 0$ , there exists  $y \in X$  such that  $d(x, y) < \delta$  but for some  $n > 0$   $d(f^n(x), f^n(y)) > \epsilon$ . Here the important issue is that  $\epsilon$  is independent of the choice of  $x$  and  $\delta$ .

Sensitivity is a popular candidate for a definition of chaos in the topological context. Suppose for a sensitive homeomorphism  $f$  you are attempting to estimate analyze the orbit of  $x$ , but may make an – arbitrarily small – positive error for initial point input. Even if you are able to compute the iterates exactly, you still cannot be certain that you always remain  $\epsilon$  close to the orbit you want. If you have chosen a bad point  $y$  as input then at some time  $n$  you will be at  $f^n(y)$  more than distance  $\epsilon$  away from the point  $f^n(x)$  that you want.

For transitive systems, the *Auslander–Yorke Dichotomy Theorem* holds (see [21]):

**Theorem 13** *Let  $f$  be a topologically transitive homeomorphism on a compact metric space  $X$ . Exactly one of the following alternatives is true.*

- (Sensitivity) *The homeomorphism  $f$  is sensitive and there are no equicontinuity points, i. e.  $Eq_f = \emptyset$ .*
- (Almost Equicontinuity) *The set of equicontinuity points coincides with the set of transitive points, i. e.  $Eq_f = Trans_f$ , and so the set of equicontinuity points is residual in  $X$ .*

*Proof* For  $\epsilon > 0$  let  $Eq_{f,\epsilon}$  be the union of all open sets with  $d_f$  diameter at most  $\epsilon$ . So  $x \in Eq_{f,\epsilon}$  if and only if it has a neighborhood with  $d_f$  diameter at most  $\epsilon$ . It is then easy to check that  $x \in Eq_{f,\epsilon}$  if and only if  $f(x) \in Eq_{f,\epsilon}$ . Thus,  $Eq_{f,\epsilon}$  is an  $f$  invariant open set. If  $Eq_{f,\epsilon}$  is nonempty and  $x$  is a transitive point then the orbit eventually enters  $Eq_{f,\epsilon}$ . Since  $f^n(x) \in Eq_{f,\epsilon}$  for some positive  $n$ , it follows that  $x \in Eq_{f,\epsilon}$  because the set is invariant. This shows that

$$Eq_{f,\epsilon} \neq \emptyset \implies Trans_f \subseteq Eq_{f,\epsilon}. \quad (36)$$

If for every  $\epsilon > 0$  we have  $Eq_{f,\epsilon} \neq \emptyset$  then  $Trans_f \subseteq \bigcap_{\epsilon > 0} Eq_{f,\epsilon} = Eq_f$ . We omit the proof that only the transitive points can be equicontinuous in a topologically transitive system.

If, instead, for some  $\epsilon > 0$  the set  $Eq_{f,\epsilon}$  is empty then by definition  $f$  is sensitive.  $\square$

The system is minimal if and only if  $Trans_f = X$  and so a topologically transitive system is equicontinuous if and only if it is both almost equicontinuous and minimal. In particular, we have:

**Corollary 2** *Let  $f$  be a minimal homeomorphism on  $X$ . Either  $f$  is sensitive or  $f$  is equicontinuous.*

Banks et al. [22] showed that a topologically transitive homeomorphism on an infinite space with dense periodic points is always sensitive. On the other hand, there do exist topologically transitive homeomorphisms which are almost equicontinuous but not minimal and hence not equicontinuous. If  $f$  is an almost equicontinuous, topologically transitive homeomorphism then there is a sequence of positive integers  $n_k \rightarrow \infty$  such that the sequence  $\{f^{n_k}\}$  converges uniformly to the identity  $1_X$ , see Glasner and Weiss [37] and Akin, Auslander and Berg [6]. In general, when such a convergent sequence of iterates exists the homeomorphism  $f$  is called *uniformly rigid*.

Other notions of chaos are defined using ideas related to proximality. A pair  $(x, y) \in X \times X$  is called *proximal* for  $f$  if

$$\liminf_{n \rightarrow \infty} d(f^n(x), f^n(y)) = 0, \quad (37)$$

or, equivalently, if  $1_X \cap \omega(f \times f)(x, y) \neq \emptyset$  where  $f \times f$  is the homeomorphism on  $X \times X$  defined by  $(f \times f)(x, y) = (f(x), f(y))$ . If a pair is not proximal then it is called *distal*. The pair is called *asymptotic* if

$$\lim_{n \rightarrow \infty} d(f^n(x), f^n(y)) = 0. \quad (38)$$

We will call  $(x, y)$  a *Li–Yorke pair* if it is proximal but not asymptotic. That is, (37) holds but

$$\limsup_{n \rightarrow \infty} d(f^n(x), f^n(y)) > 0. \quad (39)$$

Li and Yorke [47] called a subset  $A$  of  $X$  a *scrambled set* if every non-diagonal pair in  $A \times A$  is a Li–Yorke pair. Following their definition  $f$  is called *Li–Yorke chaotic* when there is an uncountable scrambled subset.

The homeomorphism  $f$  is called *distal* when every nondiagonal pair in  $X \times X$  is a distal pair. If  $f$  is an isometry then clearly  $f$  is distal and so every equicontinuous homeomorphism is distal. There exist homeomorphisms  $f$  which are minimal and distal but not equicontinuous, as described in e. g. Auslander [17]. Since such an  $f$  is minimal but not equicontinuous, it is sensitive. On the other hand, a distal homeomorphism has no proximal pairs and so is certainly not Li–Yorke chaotic.

There exists an almost equicontinuous, topologically transitive, non-minimal homeomorphism  $f$  such that a fixed point  $e \in X$  is the unique minimal subset of  $X$  and so the pair  $(e, e)$  is the unique subset in  $X \times X$  which is minimal for  $f \times f$ . It follows that  $(e, e) \in \omega(f \times f)(x, y)$  for every pair  $(x, y) \in X \times X$  and so every pair is proximal. On the other hand, because  $f$  is uniformly rigid, every pair  $(x, y)$  is recurrent for  $f \times f$  and it follows that no non-diagonal pair is asymptotic. Thus, the entire space  $X$  is scrambled and  $f$  is Li–Yorke chaotic but not sensitive.

There is a sharpening of topological transitivity which always implies sensitivity as well as most other topological conditions associated with chaos. A homeomorphism  $f$  on  $X$  is called *weak mixing* if the homeomorphism  $f \times f$  on  $X \times X$  is topologically transitive. If  $(x, y)$  is a transitive point for  $f \times f$  then since  $\omega(f \times f)(x, y) = X \times X$  it follows that  $d_f(x, y) = M$ , where  $M$  is the diameter of the entire space  $X$ . Suppose  $f$  is weak mixing and  $U$  is any nonempty open subset of  $X$ . Since  $\text{Trans}_{f \times f}$  is dense in  $X \times X$ , there is a transitive pair  $(x, y)$  in  $U \times U$ . Hence,  $\text{diam}_f(U) = M$ . When  $f$  is weak mixing, there exists an uncountable  $A \subseteq X$  such that every nondiagonal pair in  $A \times A$  is a transitive point for  $f \times f$ , Iwanik [44] and Huang and Ye [42], see also Akin [5]. Since such a set  $A$  is clearly scrambled, it follows that  $f$  is Li–Yorke chaotic.

For any two subsets  $U, V \subseteq X$  we define the *hitting time set* to be the set integers given by:

$$N(U, V) =_{\text{def}} \{n \geq 0: f^n(U) \cap V \neq \emptyset\}. \quad (40)$$

The topological transitivity condition  $\Omega f = X \times X$  says that whenever  $U$  and  $V$  are nonempty open subsets the hitting time set  $N(U, V)$  is infinite. We call  $f$  *mixing* if for every pair of nonempty open subsets  $U, V$  there exists  $k$  such that  $n \in N(U, V)$  for all  $n \geq k$ . As the names suggest, mixing implies weak mixing.

**Example** The most important example of a chaotic homeomorphism is the ubiquitous *shift homeomorphism*. We think of a fixed finite set  $A$  as an *alphabet* and for any positive integer  $k$  the sequences in  $A$  of length  $k$ , i.e. the elements of  $A^k$ , are called *words of length  $k$* . When  $A$  is equipped with the discrete topology it is compact and so by the Tychonoff Product Theorem any product of infinitely many copies of  $A$  is compact when given the product topology. Using the group of integers as our index set, we let  $X = A^{\mathbb{Z}}$ . There is a metric compatible with this topology. For  $x, y \in X$  let

$$d(x, y) =_{\text{def}} \infimum \{2^{-n}: x_i = y_i \text{ for all } i \text{ with } |i| < n\}. \quad (41)$$

The metric  $d$  is an *ultrametric*. That is, it satisfies a strengthening of the triangle inequality. For all  $x, y, z \in X$ :

$$d(x, z) \leq \max(d(x, y), d(y, z)). \quad (42)$$

The ultrametric inequality is equivalent to the condition that for every positive  $\epsilon$  the open relation  $V_\epsilon$  is an equivalence relation.

For any word  $a \in A^k$  and any integer  $j$ , the set  $\{x \in X: x_{i+j} = a_i: i = 1, \dots, k\}$  is a clopen subset of  $X$  called a *cylinder set*, which we will denote  $U_{a,j}$ . For example, with  $k = 2n - 1$  and  $j = -n$  the cylinder sets are precisely the open balls of radius  $\epsilon$  when  $2^{-n} < \epsilon \leq 2^{-n+1}$ . From this we see that  $X$  is a Cantor set with cylinder sets as a countable basis of clopen sets.

On  $X$  the shift homeomorphism  $s$  is defined by the equation

$$s(x)_i = x_{i+1} \text{ for all } i \in \mathbb{Z}. \quad (43)$$

The fixed points  $|s|$  are exactly the constant sequences in  $X$  and the periodic points  $|\mathcal{O}s|$  are the periodic sequences in  $X$ . That is,  $s^t(x) = x$  for some positive integer  $t$  exactly when  $x_{i+t} = x_i$  for all  $i \in \mathbb{Z}$ .

It is easy to see that  $s$  is mixing. For example, if  $a, b \in A^{2n-1}$  and  $j = -n$ , then the hitting time set  $N(U_{a,j}, U_{b,j})$  contains every integer  $t$  larger than  $2n$ . For if  $x_{i-n} = a_i$  and  $x_{i-n+t} = b_i$  for  $i = 1, \dots, 2n - 1$  then  $x \in U_{a,j}$  and  $s^t(x) \in U_{b,j}$ . This illustrates why  $s$  is chaotic in the sense of unpredictable. Given  $x \in X$ , if  $y \in X$  satisfies  $d(x, y) = 2^{-n}$ , then moving right from position 0 the first  $n$  entries of  $y$  are known exactly. They agree with the entries of  $x$ . But after position  $n$ ,  $x$  provides no information about  $y$ . The remaining entries on the right can be chosen arbitrarily.

Since  $s$  is mixing it is certainly topologically transitive, but it is useful to characterize the transitive points of  $s$ . For any point  $x \in X$  and positive integer  $n$  we can scan from the central position  $x_0$ , left and right  $n - 1$  steps and observe a word of length  $2n - 1$ . As we apply  $s$  the central position shifts right. When we have applied  $s^t$  it has shifted  $t$  steps. Scanning left and right we observe a new word. A point is a transitive point when for every  $n$  we can observe every word in  $A^{2n-1}$  in this way by varying  $t$ . To construct a transitive point  $x$  we need only list the – countably many – finite words of every length and lay them out end to end to get the right side of  $x$ . The left side of  $x$  can be arbitrary.

The shift homeomorphism is also expansive. In general, a homeomorphism  $f$  on a compact metric space  $X$  is called *expansive* when the diagonal  $1_X$  is an isolated

invariant set for the homeomorphism  $f \times f$  on  $X \times X$  and so there exists  $\epsilon > 0$  such that  $\bar{V}_\epsilon$  is an isolating neighborhood in the sense of (29). That is, if  $x, y \in X$  and  $d(f^n(x), f^n(y)) \leq \epsilon$  for all  $n \in \mathbb{Z}$  then  $x = y$ . Such a number  $\epsilon > 0$  is called an *expansivity constant*. With respect to the metric given by (41) it is easy to check that  $\frac{1}{2}$  is an expansivity constant for the shift homeomorphism  $s$ .

The shift is interesting in its own right. In addition, the chaotic behavior of other important examples, especially expansive homeomorphisms, are studied by comparing them with the shift.

For a compact metric space  $X$  let  $H(X)$  denote the automorphism group of  $X$ , i.e. the group of homeomorphisms on  $X$ , with the topology of uniform convergence. The metric on  $H(X)$  is defined by the equation, for  $f, g \in H(X)$ :

$$d(f, g) =_{\text{def}} \sup \{d(f(x), g(x)) : x \in X\}. \quad (44)$$

We obtain a topologically equivalent, but complete, metric by using  $\max(d(f, g), d(f^{-1}, g^{-1}))$ . The automorphism group is a Polish (= admits a complete, separable metric) topological group. For  $f \in H(X)$  we define the *translation* homeomorphisms  $\ell_f$  and  $r_f$  on  $H(X)$  by

$$\ell_f(g) =_{\text{def}} f \circ g \quad \text{and} \quad r_f(g) =_{\text{def}} g \circ f. \quad (45)$$

For any  $x \in X$ , the *evaluation map*  $ev_x: H(X) \rightarrow X$  is the continuous map defined by  $ev_x(f) = f(x)$ .

For any  $f \in H(X)$  let  $G_f$  denote the closure in  $H(X)$  of the cyclic subgroup generated by  $f$ . That is,

$$G_f = \overline{\{f^n : n \in \mathbb{Z}\}} \subseteq H(X). \quad (46)$$

Thus,  $G_f$  is a closed, abelian subgroup of  $H(X)$ .

Let  $Iso(X)$  denote the closed subgroup of isometries in  $H(X)$  (In contrast with  $H(X)$  this varies with the choice of metric). It follows from the Arzela–Ascoli Theorem that  $Iso(X)$  is compact in  $H(X)$ . If  $f$  is an isometry on  $X$  then  $G_f$  is a compact, abelian subgroup of  $Iso(X)$ . Furthermore, the translation homeomorphisms  $\ell_f$  and  $r_f$  restrict to isometries on the compact space  $Iso(X)$  and  $G_f$  is a closed invariant subset. In fact, under either  $\ell_f$  or  $r_f$ ,  $G_f$  is just the orbit closure of  $f$  regarded as a point of  $Iso(X)$ .

Recall that if  $f$  is an equicontinuous homeomorphism, then we can replace the original metric by a topologically equivalent one, e.g. replace  $d$  by  $d_f$ , to get one for which  $f$  is an isometry.

If  $f$  is a homeomorphism on  $X$  and  $g$  is a homeomorphism on  $Y$  then we say that a continuous function  $\pi: Y \rightarrow X$  maps  $g$  to  $f$  when  $\pi \circ g = f \circ \pi: Y \rightarrow X$ . If, in addition,  $\pi$  is a homeomorphism then we call  $\pi$  an *isomorphism* from  $g$  to  $f$ .

**Theorem 14** *Let  $f$  be a homeomorphism on  $X$ . Fix  $x \in X$ . The evaluation map  $ev_x: H(X) \rightarrow X$  maps  $\ell_f$  on  $H(X)$  to  $f$  on  $X$ .*

*If  $f$  is an almost equicontinuous, topologically transitive homeomorphism then  $ev_x$  restricts to a homeomorphism from the closed subgroup  $G_f$  of  $H(X)$  onto  $Trans_f$  the residual subset of  $X$  consisting of the transitive points for  $f$ .*

*If  $f$  is a minimal isometry then  $ev_x$  restricts to a homeomorphism from the compact subgroup  $G_f$  of  $H(X)$  onto  $X$  and so is an isomorphism from  $\ell_f$  on  $G_f$  to  $f$  on  $X$ .*

The isometry result is classical, see e.g. Gottschalk and Hedlund [38]. It shows that minimal, equicontinuous homeomorphisms are just translations on compact, monothetic groups, where a topological group is *monothetic* when it has a dense cyclic subgroup. Similarly a topologically transitive, almost equicontinuous homeomorphism which is not minimal is obtained by “compactifying” a translation on a noncompact monothetic group, see Akin and Glasner [7]. In fact, the group cannot even be locally compact and so, for example, is not the discrete group of integers. While examples of such topologically transitive, almost equicontinuous but not equicontinuous systems are known, it is not known whether there are any finite dimensional examples. And it is known that they cannot occur on a zero-dimensional space, i.e. the Cantor set. If finite dimensional examples do not exist then every topologically transitive homeomorphism on a compact manifold is sensitive except for the equicontinuous ones which we now describe.

**Example** We can identify the circle  $S$  with the quotient topological group  $\mathbb{R}/\mathbb{Z}$ . For  $a \in \mathbb{R}$  the translation  $L_a = R_a$  on  $\mathbb{R}$  induces the rotation  $\tau_a$  on the circle  $\mathbb{R}/\mathbb{Z}$ . If  $a \in \mathbb{Z}$  then this is the identity map. If  $a$  is rational then  $\tau_a$  is periodic. But if  $a$  is irrational then  $\tau_a$  is a minimal isometry on  $S$ . More generally, if  $\{1, a_1, \dots, a_n\}$  is linearly independent over the field of rationals  $\mathbb{Q}$  then on the torus  $X = S^n$  the product homeomorphism  $\tau_{a_1} \times \dots \times \tau_{a_n}$  is a minimal isometry. Such systems are sometimes called *quasi-periodic*.

Of course, the translation by 1 on the finite cyclic group  $\mathbb{Z}/k\mathbb{Z}$  of integers modulo  $k$  is just a version of a single periodic orbit, the unique minimal map on a finite space of cardinality  $k$ .

Recall that if  $\{X_1, X_2, \dots\}$  is a sequence of topological spaces and  $\{p_n: X_{n+1} \rightarrow X_n\}$  is a sequence of continuous maps, then the *inverse limit* is the closed subset  $LIM$  of the product space  $\prod_{n=1}^{\infty} X_n$ :

$$LIM = \lim \{X_n, p_n\} =_{\text{def}} \{x: p_n(x_{n+1}) = x_n \text{ for } n = 1, 2, \dots\}. \quad (47)$$



If the spaces are compact and the maps are surjective then the  $n$ th coordinate projection  $\pi_n$  maps the compact space  $LIM$  onto  $X_n$  for every  $n$  (Hint: use Proposition 1). Also, if the spaces are topological groups and the maps are homomorphisms, then  $LIM$  is a closed subgroup of the product topological group. Finally, if  $\{f_n: X_n \rightarrow X_n\}$  is a sequence of homeomorphisms such that  $p_n$  maps  $f_{n+1}$  to  $f_n$  for every  $n$  then the product homeomorphism  $\prod_{n=1}^{\infty} f_n$  restricts to a homeomorphism  $f$  on  $LIM$  and  $\pi_n$  maps  $f$  to  $f_n$  for every  $n$ .

Now let  $\{k_n\}$  be an increasing sequence of positive integers such that  $k_n$  divides  $k_{n+1}$  for every  $n$ . Let  $X_n$  denote the finite cyclic group  $\mathbb{Z}/k_n\mathbb{Z}$  and let  $p_n: X_{n+1} \rightarrow X_n$  be the quotient homomorphism induced by the inclusion map  $k_{n+1}\mathbb{Z} \rightarrow k_n\mathbb{Z}$ . Let  $f_n$  denote the translation by 1 on  $X_n$ . Define  $X$  to be the inverse limit of this system with  $f$  the restriction to  $X$  of the product homeomorphism.  $X$  is a topological group whose underlying space is zero-dimensional and perfect, i. e. the Cantor set. The product homeomorphism is an isometry when we use the metric analogous to the one defined by (41). Furthermore, the restriction  $f$  to  $X$  is a minimal isometry. These systems are usually called *adding machines* or *odometers*. It can be proved that every equicontinuous minimal homeomorphism on a Cantor space is isomorphic to one of these. In general, every equicontinuous minimal homeomorphism is isomorphic to (1) a periodic orbit, (2) an adding machine, (3) a quasi-periodic motion on a torus, or (4) a product with each factor either an irrational rotation on a circle, an adding machine or a periodic orbit. The informal expression *strange attractor* is perhaps best defined as a topologically transitive attractor which is not equicontinuous, i. e. which is not one of these.

The word “chaos” suggests instability and unpredictability. However, for many examples what is most apparent is stability. For the Henon attractor or the Lorenz attractor the word “the” is used because in simulations one begins with virtually any initial point, performs the iterations and, after discarding an initial segment, one observes a particular, repeatable picture. The set as a whole is a predictable feature, stable under perturbation of initial point, an varying continuously with the defining parameters.

However, once the orbit is close enough to the attractor, essentially moving within the attractor itself, the motion is unpredictable, sensitive to arbitrarily small perturbations. What remains is statistical prediction. We cannot exactly predict when the orbit will enter some small subset of the attractor, but we can describe approximately the amount of time it spends in the subset. Such analysis requires an invariant measure and this takes us to the boundary between topological and measurable dynamics.

By a measure  $\mu$  on a compact metric space  $X$  we will mean a Borel probability measure on the space. Such a measure acts, via integration, on the Banach algebra  $C(X)$  of continuous real-valued functions on  $X$ . The set  $\mathcal{P}(X)$  of such measures can thus be regarded as a convex subset of the dual space of  $C(X)$ . Inheriting the weak\* topology on the dual space,  $\mathcal{P}(X)$  becomes a compact, metrizable space. There is a natural inclusion map  $\delta: X \rightarrow \mathcal{P}(X)$  which associates to  $x \in X$  the point mass at  $x$ , denoted  $\delta_x$ .

The *support* of a measure  $\mu$  on  $X$  is the smallest closed set with measure 1. Denoted  $|\mu|$  its complement can be obtained by taking the union out of a countable base for the topology of those members with measure 0. The measure has *full support* (or simply  $\mu$  is *full*) when  $|\mu| = X$ . A measure is full when every nonempty open set has positive measure.

A continuous map  $h: X \rightarrow Y$  induces a map  $h_*: \mathcal{P}(X) \rightarrow \mathcal{P}(Y)$  which associates to  $\mu$  the measure  $h_*\mu$  defined by  $h_*\mu(A) = \mu(h^{-1}(A))$  for every Borel subset  $A$  of  $Y$ . The continuous linear operator  $h_*$  is the dual of  $h^*: C(Y) \rightarrow C(X)$  given by  $u \mapsto u \circ h$ . Furthermore,  $h_*$  is an extension of  $h$ . That is,  $h_*(\delta_x) = \delta_{h(x)}$ . The supports are related by

$$h(|\mu|) = |h_*\mu|. \quad (48)$$

In particular, if  $f$  is a homeomorphism on  $X$  then  $f_*$  is a linear homeomorphism on  $\mathcal{P}(X)$ , extending  $f$  on  $X$ . A measure  $\mu$  such that  $f_*\mu = \mu$  is called an *invariant measure* for  $f$ . Thus,  $|f_*|$ , the set of fixed points for  $f_*$ , is the set of invariant measures for  $f$ . Clearly,  $|f_*|$  is compact, convex subset of  $\mathcal{P}(X)$ . The classical theorem of Krylov and Bogolubov says that this set is nonempty. To prove it, one begins with an arbitrary point  $x \in X$  and considers the sequence of Cesaro averages

$$\sigma_n(f, x) =_{\text{def}} \frac{1}{n+1} \sum_{i=0}^n \delta_{f^i(x)}. \quad (49)$$

It can be shown that every measure in the nonempty set of limit points of this sequence lies in  $|f_*|$ . If the set of limit points consists of a single measure  $\mu$ , i. e. the sequence converges to  $\mu$ , then  $x$  is called a *convergence point* for the invariant measure  $\mu$ . For  $\mu \in |f_*|$  we denote by  $\text{Con}(\mu)$  the – possibly empty – set of convergence points for  $\mu$ .

The extreme points of the compact convex set  $|f_*|$  are called the *ergodic measures* for  $f$ . That is,  $\mu$  is ergodic if it is not in the interior of some line segment connecting a pair of distinct invariant measures. Equivalently,  $\mu$  is ergodic if a Borel subset  $A$  of  $X$  is invariant, i. e.  $f^{-1}(A) = A$ , only when the measure  $\mu(A)$  is either 0 or 1. It follows that if



$u: X \rightarrow \mathbb{R}$  is a Borel measurable, invariant function, then for any ergodic measure  $\mu$  there is a set of measure 1 on which  $u$  is constant. Or more simply,  $u \circ f = u$  implies  $u$  is constant almost everywhere with respect to  $\mu$ .

The central result in measurable dynamics is the Birkhoff Pointwise Ergodic Theorem which says, in this context:

**Theorem 15** *Given a homeomorphism  $f$  on a compact metric space  $X$ , let  $u: X \rightarrow \mathbb{R}$  be a bounded, Borel measurable function. There exists a bounded, invariant, Borel measurable function.  $\hat{u}: X \rightarrow \mathbb{R}$  such that for every  $f$  invariant measure  $\mu$*

$$\int u \, d\mu = \int \hat{u} \, d\mu \quad (50)$$

$$\text{and } \lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{i=0}^n u(f^i(x)) = \hat{u}(x)$$

almost everywhere with respect to  $\mu$ .

In particular, if  $\mu$  is ergodic then

$$\hat{u}(x) = \int u \, d\mu \quad (51)$$

almost everywhere with respect to  $\mu$ .

It is a consequence of the ergodic theorem that

$$\mu \text{ ergodic} \implies \mu(\text{Con}(\mu)) = 1. \quad (52)$$

Thus, with respect to an ergodic measure  $\mu$  for almost every point  $x$ ,

$$\lim_{n \rightarrow \infty} \frac{1}{n+1} \sum_{i=0}^n u(f^i(x)) = \int u \, d\mu, \quad (53)$$

for every  $u \in C(X)$ . The left hand side is the time-average of the function  $u$  along the orbit with initial point  $x$  and it equals the space-average on the right.

For any  $f$  invariant measure  $\mu$  it follows from (48) that the support,  $|\mu|$ , is an  $f$  invariant subset of  $X$ . The Poincaré Recurrence Theorem says that if  $U$  is any open set with  $U \cap |\mu| \neq \emptyset$  then  $U$  is non-wandering. That is, for some positive integer  $n$ ,  $U \cap f^{-n}(U) \neq \emptyset$ . This is clear from the observation that the open sets  $f^{-n}(U)$  all have the same, positive measure and so they cannot all be pairwise disjoint. Applying this to the subsystem obtained by restricting to  $|\mu|$  it follows from Theorem 1 that the set of recurrent points in  $|\mu|$  form a dense  $G_\delta$  subset of  $|\mu|$ .

If  $x \in \text{Con}(\mu)$  then the orbit of  $x$  is dense in  $|\mu|$  and so the subsystem on  $|\mu|$  is topologically transitive whenever  $\text{Con}(\mu)$  is nonempty. By the Birkhoff Ergodic Theorem this applies whenever  $\mu$  is ergodic.

Since the support is a nonempty, closed, invariant subspace it follows that if  $f$  is minimal then every invariant measure has full support. The homeomorphism  $f$  is called *strictly ergodic* when it is minimal and has a unique invariant measure  $\mu$ , which is necessarily ergodic. In that case, it can be shown that every point is a convergence point for  $\mu$ , that is,  $\text{Con}(\mu) = X$ .

We note that in the dynamical systems context the topological notion of residual (that is, a dense  $G_\delta$  subset) is quite different from the measure theoretic idea (a set of full measure). For example,

$$\text{Con} =_{\text{def}} \bigcup_{\mu \in |f_*|} \text{Con}(\mu) \quad (54)$$

is the set of points whose associated Cesaro average sequence is a Cauchy sequence. It follows that  $\text{Con}$  is a Borel set. By (52)  $\mu(\text{Con}) = 1$  for every ergodic measure  $\mu$ . Since every invariant measure is a limit of convex combinations of ergodic measures it follows that  $\text{Con}$  has measure 1 for every invariant measure  $\mu$ .

On the other hand, it can be shown that if  $f$  is the shift homeomorphism on  $X = A^{\mathbb{Z}}$  then for  $x$  in the dense  $G_\delta$  subset of  $X$  the set of limit points of the sequence  $\{\sigma_n(f, x)\}$  is all of  $|f_*|$ , Denker et al. [30] or Akin (Chap. 9 in [1]). Since the shift has many different invariant measures it follows that  $\text{Con}$  is disjoint from this residual subset. In general, if a homeomorphism  $f$  has more than one full, ergodic measure then  $\text{Con}$  is of first category, Akin [1, Theorem 8.11]. In particular, if  $f$  is minimal but not strictly ergodic then  $\text{Con}$  is of first category.

The plethora of invariant measures undercuts somewhat their utility for statistical analysis. Suppose that there are two different ergodic measures  $\mu$  and  $\nu$  with common support, some invariant subset  $A$  of  $X$ . By restricting to  $A$ , we can reduce to the case when  $A = X$  and so  $\mu$  and  $\nu$  are distinct full, ergodic measures. The sets  $\text{Con}(\mu)$  and  $\text{Con}(\nu)$  are disjoint and of measure 1 with respect to  $\mu$  and  $\nu$ , respectively. For an open set  $U$  the average frequency with which an orbit of  $x \in \text{Con}(\mu)$  lies in  $U$  is given by  $\mu(U)$  and similarly for  $\nu$ . These mutually singular measures lead to different statistics.

One solution to this difficulty is to select what seems to be the best invariant measure in some sense, e.g. the measure of maximum entropy (see the article by King) if it should happen to be unique. However, as our introductory discussion illustrates, this somewhat misses the point.

Return to the case of a chaotic attractor  $A$ , a closed invariant subset of  $X$ . It often happens that the state space  $X$  comes equipped with a natural measure  $\lambda$  or at least a Radon–Nikodym equivalence class of measures, all with

the same sets of measure 0. For example, if  $X$  is a manifold then  $\lambda$  is locally Lebesgue measure. The measure  $\lambda$  is usually not  $f$  invariant and it often happens that the set  $A$  of interest has  $\lambda$  measure 0. What we want, an *appropriate* measure  $\mu$  for this situation, would be an invariant measure with support  $A$ , i. e.

$$\mu \in |f_*| \text{ and } |\mu| = A, \quad (55)$$

and an open set  $U$  containing  $A$  such that with respect to  $\lambda$  almost every point of  $U$  is a convergence point for  $\mu$ . That is,

$$\lambda(U \setminus \text{Con}(\mu)) = 0. \quad (56)$$

Notice that for  $x \in \text{Con}(\mu)$ ,  $\omega f(x) = |\mu| = A$  and so by Corollary 1  $A$  is a terminal chain component. That is, for such a measure to exist, the attractor  $A$  must be at least chain transitive. If, in addition,  $\text{Con}(\mu) \cap A \neq \emptyset$  then  $A$  is topologically transitive.

At least when strong hyperbolicity conditions hold this program can be carried out with  $\mu$  the Bowen measure for the invariant set, see Katok–Hasselblatt (Chapter 20 in [45]).

### Minimality and Multiple Recurrence

In this section we provide a sketch of some important topics which were neglected in the above exposition. We first consider the study of minimal systems.

In Theorem 7 we defined a homeomorphism  $f$  on a compact space  $X$  to be *minimal* when every point has a dense orbit. A subset  $A$  of  $X$  is a minimal subset when it is a nonempty, closed, invariant subset such that the restriction  $f|_A$  defines a minimal homeomorphism on  $A$ . The term “minimal” is used because  $f|_A$  is minimal precisely when  $A$  is minimal, with respect to inclusion, in the family of nonempty, closed, invariant subsets. By Zorn’s lemma every such subset contains a minimal subset. Since every system contains minimal systems, the classification of minimal systems provides a foundation upon which to build an understanding of dynamical systems in general.

On the other hand, if you start with the space, homeomorphisms which are minimal – on the whole space – are rather hard to construct. Some spaces have the fixed-point property, i. e. every homeomorphism on the space has a fixed point, and so admit no minimal homeomorphisms. The tori, which are monothetic groups, admit the equicontinuous minimal homeomorphisms described in the previous section. Fathi and Herman [34] constructed a minimal homeomorphism on the 3-sphere. For most other connected, compact manifolds it is not known whether

they admit minimal homeomorphisms or not. It is even difficult to construct topologically transitive homeomorphisms, but for these a beautiful – but nonconstructive – argument due to Oxtoby shows that every such manifold admits topologically transitive homeomorphisms. He uses that Baire Category Theorem to show that if the dimension is at least two then the topologically transitive homeomorphisms are residual in the class of volume preserving homeomorphisms, see Chap. 18 in [53].

Since we understand the equicontinuous minimal systems, we begin by building upon them. This requires a change in our point of view. Up to now we have mostly considered the behavior of a single dynamical system  $(X, f)$  consisting of a homeomorphism  $f$  on a compact metric space  $X$ . Regarding these as our objects of study we turn now to the maps between such systems. A *homomorphism* of dynamical systems, also called an *action map*,  $\pi: (X, f) \rightarrow (Y, g)$  is a continuous map  $\pi: X \rightarrow Y$  such that  $g \circ \pi = \pi \circ f$  and so, inductively,  $g^n \circ \pi = \pi \circ f^n$  for all  $n \in \mathbb{Z}$ . Thus,  $\pi$  maps the orbit of a point  $x \in X$  to the orbit of  $\pi(x) \in Y$ . In general, for the map  $\pi \times \pi: X \times X \rightarrow Y \times Y$  we have that

$$\begin{aligned} \pi \times \pi(\mathcal{A}f) &\subseteq \mathcal{A}g \quad \text{for } \mathcal{A} = \mathcal{O}, \mathcal{R}, \mathcal{N}, \mathcal{G}, \mathcal{C} \\ \pi \times \pi(\Theta f) &\subseteq \Theta g \quad \text{for } \Theta = \omega, \alpha, \Omega, \Omega C. \end{aligned} \quad (57)$$

That is,  $\pi$  maps the various dynamic relations associated with  $f$  to the corresponding relations for  $g$ .

If  $\pi$  is bijective then it is called an *isomorphism* between the two systems and the inverse map  $\pi^{-1}$  is an action map, continuous by compactness.

If  $X$  is a closed invariant subset of  $Y$  with  $f = g|_X$  then the inclusion map  $\pi$  is an action map and then  $(X, f)$  is called the *subsystem* of  $(Y, g)$  determined by the invariant set  $X$ .

On the other hand, if  $\pi$  is surjective then  $(Y, g)$  is called a *factor*, or *quotient system*, of  $(X, f)$  and  $(X, f)$  is called a *lift* of  $(Y, g)$ . A surjective action map is called a *factor map*.

For a factor map  $\pi: (X, f) \rightarrow (Y, g)$  we define an important subset  $R(\pi)$  of  $X \times X$ :

$$\begin{aligned} R(\pi) &=_{\text{def}} \{(x_1, x_2) \in X \times X: \pi(x_1) = \pi(x_2)\} \\ &= (\pi \times \pi)^{-1}(1_Y). \end{aligned} \quad (58)$$

The subset  $R(\pi)$  is an ICER on  $X$ . That is, it is an invariant, closed equivalence relation. In general, if  $R$  is any ICER on  $X$  then on the space of equivalence classes the homeomorphism  $f$  induces a homeomorphism and the natural quotient map is a factor map of dynamical systems. The original factor system  $(Y, g)$  is isomorphic to the quotient system obtained from the ICER  $R(\pi)$ .

Notice that if  $(Y, g)$  is the *trivial system*, meaning that  $Y$  consisting of a single point, then  $R(\pi)$  is the total relation  $X \times X$  on  $X$ .

We use the ICER  $R(\pi)$  to extend various definitions from dynamical systems to action maps between dynamical systems. For example, a factor map  $\pi: (X, f) \rightarrow (Y, g)$  is called *equicontinuous* if for every  $\epsilon > 0$  there exists  $\delta > 0$  such that

$$(x_1, x_2) \in R(\pi) \text{ and } d(x_1, x_2) < \delta \\ \implies d_f(x_1, x_2) < \epsilon. \quad (59)$$

Comparing this with (34) we see that  $(X, f)$  is equicontinuous if and only if the factor map to the trivial system is equicontinuous.

Similarly, recall that  $(x_1, x_2) \in X$  is a *distal pair* for  $(X, f)$  if  $\omega(f \times f)(x_1, x_2)$  is disjoint from the diagonal  $1_X$ , and the system  $(X, f)$  is *distal* when every nondiagonal pair is distal. A factor map  $\pi: (X, f) \rightarrow (Y, g)$  is distal when every nondiagonal pair in  $R(\pi)$  is distal. Again  $(X, f)$  is a distal system if and only if the factor map to the trivial system is distal.

It is easy to check that a distal lift of a distal system is distal. Since an equicontinuous factor map is distal, it follows that an equicontinuous lift of an equicontinuous system is distal. However, it need not be equicontinuous.

*Example* If  $a$  is irrational then the rotation  $\tau_a$  on the circle  $Y = \mathbb{R}/\mathbb{Z}$ , defined by  $x \mapsto a + x$ , is an equicontinuous minimal map. On the torus  $X = \mathbb{R}/\mathbb{Z} \times \mathbb{R}/\mathbb{Z}$  we define  $f$  by

$$f(x, y) =_{\text{def}} (a + x, x + y). \quad (60)$$

It can be shown that  $(X, f)$  is minimal. The projection  $\pi$  to the first coordinate defines an equicontinuous factor map to  $Y, \tau_a$  and so  $(X, f)$  is distal. It is not, however, equicontinuous.

The above projection map is an example of a *group extension*. Let  $G$  be a compact topological group, like  $\mathbb{R}/\mathbb{Z}$ . Given any dynamical system  $(Y, g)$  and any continuous map  $q: Y \rightarrow G$  we let  $X = Y \times G$  and define  $f$  on  $X$  by:

$$f(x, y) =_{\text{def}} (g(x), L_{q(x)}(y)). \quad (61)$$

The homeomorphism commutes with  $1_Y \times R_z$  for any group element  $z$  and from this it easily follows that the projection  $\pi$  to the first coordinate is an equicontinuous factor map. If  $(Y, g)$  is minimal then the restriction of  $\pi$  to any minimal subset of  $X$  defines an equicontinuous fac-

tor map from the associated minimal subsystem. It can be shown that any equicontinuous factor map between minimal systems can be obtained via a factor from such a group extension.

The Furstenberg Structure Theorem says that any distal minimal system can be obtained by a – possibly transfinite – inverse limit construction, beginning with an equicontinuous system and such that each lift is an equicontinuous factor map. For the details of this and for the structure theorem due to Veech for more general minimal systems, we refer the reader to Auslander (Chaps. 7, 14 in [17]) respectively.

The factor maps described above are projections from products. There exist examples which are not product projections even locally. For example, in Auslander (Chap. 1 in [17]) the author uses a construction due to Floyd to build an action map  $\pi$  between minimal systems which is not an isomorphism but which is *almost one-to-one*. That is, for a residual set of points  $x$  in the domain the set of preimages  $\pi^{-1}(\pi(x))$  is a singleton. Such a factor map is the opposite of distal. It is a *proximal* map, meaning that every pair  $(x_1, x_2) \in R(\pi)$  is a proximal pair.

Also, one cannot base all one's constructions upon equicontinuous minimal systems. A dynamical system  $(X, f)$  is called *weak mixing* when the product  $(X \times X, f \times f)$  is topologically transitive. Inclusion (57) implies that if a dynamical system is minimal, topologically transitive or chain transitive then any factor satisfies the corresponding property. It follows that any factor of a weak mixing system is weak mixing. It is clear that only the trivial system is both weak mixing and equicontinuous. Hence for a weak mixing system the trivial system is the only equicontinuous factor. For a minimal system the converse is true: if the trivial system is the only equicontinuous factor then the system is weak mixing. Furthermore, nontrivial weak mixing, minimal systems do exist.

Gottschalk and Hedlund in [38] introduced the idea using various special families of subsets of  $\mathbb{N}$ , the set of nonnegative integers, in order to distinguish different sorts of recurrence. A *family*  $\mathcal{F}$  is a collection of subsets of  $\mathbb{N}$  which is hereditary upwards. That is, if  $A \subseteq B$  and  $A \in \mathcal{F}$  then  $B \in \mathcal{F}$ . The family is called *proper* when it is a proper subset of the entire power set of  $\mathbb{N}$ , i.e. it is neither empty nor the entire power set. The heredity property implies that a family  $\mathcal{F}$  is proper iff  $\mathbb{N} \in \mathcal{F}$  and  $\emptyset \notin \mathcal{F}$ . For a family  $\mathcal{F}$  the dual family, denoted  $\mathcal{F}^*$  (or sometimes  $k\mathcal{F}$ ) is defined by:

$$\mathcal{F}^* =_{\text{def}} \{B \subseteq \mathbb{N} : B \cap A \neq \emptyset \text{ for all } A \in \mathcal{F}\} \\ = \{B \subseteq \mathbb{N} : \mathbb{N} \setminus B \notin \mathcal{F}\}. \quad (62)$$

It is easy to check that  $\mathcal{F}^{**} = \mathcal{F}$  and that  $\mathcal{F}^*$  is proper if and only if  $\mathcal{F}$  is.

A *filter* is a proper family which is closed under pairwise intersection. A family  $\mathcal{F}$  is the dual of a filter when it satisfies what Furstenberg calls the *Ramsey Property*:

$$A \cup B \in \mathcal{F} \implies A \in \mathcal{F} \text{ or } B \in \mathcal{F}. \quad (63)$$

For example, a set is in the dual of the family of infinite sets if and only if it is cofinite, i. e. its complement is finite. The family of cofinite sets is a filter.

A subset  $A \subseteq \mathbb{N}$  is called *thick* when it contains arbitrarily long runs. That is, for every positive integer  $L$  there exists  $i$  such that  $i, i+1, \dots, i+L \in A$ . Dual to the thick sets are the syndetic sets. A subset  $B$  is called *syndetic* or *relatively dense* if there exists a positive integer  $L$  such that every run of length  $L$  meets  $B$ .

In (40) we defined the hitting time set  $N(U, V)$  for  $(X, f)$  when  $U, V \subseteq X$ . When  $U$  is a singleton  $\{x\}$  we omit the braces and so have

$$N(x, V) = \{n \geq 0 : f^n(x) \in V\}. \quad (64)$$

It is clear that  $(X, f)$  is topologically transitive when  $N(U, V)$  is nonempty for every pair of nonempty open sets  $U, V$ . Furstenberg showed that the stronger property that  $(X, f)$  be weak mixing is characterized by the condition that each such  $N(U, V)$  is thick.

Recall that a point  $x \in X$  is recurrent when  $x \in \omega f(x)$ . The point is called a *minimal point* when it is an element of a minimal subset of  $X$  in which case this minimal subset is  $\omega f(x)$ . Clearly, point  $x \in X$  is recurrent when  $N(x, U)$  is nonempty for every neighborhood  $U$  of  $x$ . Gottschalk and Hedlund proved that  $x$  is a minimal point if and only if every such  $N(x, U)$  is relatively dense. For this reason, minimal points are also called *almost periodic* points.

Furstenberg reversed this procedure by using dynamical systems arguments to derive properties about families of sets and more generally to derive results in combinatorial number theory.

If  $f_1, \dots, f_k$  are homeomorphisms on a space  $X$  then  $x \in X$  is a *multiple recurrent point* for  $f_1, \dots, f_k$  if there exists a sequence of positive integers  $n_i \rightarrow \infty$  such that the  $k$  sequences  $\{f_1^{n_i}(x)\}, \dots, \{f_k^{n_i}(x)\}$  all have limit  $x$ . That is, the point  $(x, \dots, x)$  is a recurrent point for the homeomorphism  $f_1 \times \dots \times f_k$  on  $X^k$ . The Furstenberg Multiple Recurrence Theorem says:

**Theorem 16** If  $f_1, \dots, f_k$  are commuting homeomorphisms on a compact metric space  $X$ , i. e.  $f_i \circ f_j = f_j \circ f_i$  for  $i, j = 1, \dots, k$ , then there exists a multiple recurrent point for  $f_1, \dots, f_k$ .

**Corollary 3** If  $f$  is a homeomorphism on a compact metric space  $X$  and  $x \in X$  then for every positive integer  $k$  and every  $\epsilon > 0$  there exist positive integers  $m, n$  such that with  $y = f^m(x)$  the distance between any two of the points  $y, f^n(y), f^{2n}(y), \dots, f^{kn}(y)$  is less than  $\epsilon$ .

*Proof* This follows easily by applying the Multiple Recurrence Theorem to the restrictions of the homeomorphisms  $f, f^2, \dots, f^k$  to the closed invariant subset  $\omega f(x)$ . Obtain a multiple recurrent point  $y' \in \omega f(x)$  and then given  $\epsilon > 0$  choose  $y = f^m(x)$  to approximate  $y'$  sufficiently closely.  $\square$

These results are of great interest in themselves and they have been extended in various directions, see Bergelson–Leibman [24] for example. In addition, Furstenberg used the corollary to obtain a proof of Van der Waerden’s Theorem:

**Theorem 17** If  $B_1, \dots, B_p$  is a partition of  $\mathbb{N}$  then at least one of these sets contains arithmetic progressions of arbitrary length.

*Proof* It suffices to show that for each  $k = 1, 2, \dots$  and some  $a = 1, \dots, p$  the subset  $B_a$  contains an arithmetic progression of length  $k+1$ . For this will then apply to some fixed  $B_a$  for infinitely many  $k$ . Let  $A = \{1, \dots, p\}$  and on  $X = A^{\mathbb{Z}}$  define the shift homeomorphism defined by (43). Using the metric given by (41) and  $\epsilon \leq \frac{1}{2}$  we observe that if  $x, y \in X$  with  $d(x, y) < \epsilon$  then  $x_0 = y_0$ .

Choose  $x \in X$  such that  $x_i = a$  if and only if  $i \in B_a$  for  $i \in \mathbb{N}$ . Apply Corollary 3 to  $x$  and  $\epsilon$ . Choosing positive integers  $m, n$  such that the points  $f^m(x), f^{m+n}(x), f^{m+2n}(x), \dots, f^{m+kn}(x)$  all lie within  $\epsilon$  of each other we have that  $m, m+n, \dots, m+kn$  all lie in  $B_a$  where  $a$  is the common value of the 0 coordinate of these points.  $\square$

One of the great triumphs of modern dynamical systems theory is Furstenberg’s use of the ergodic theory version of these arguments to prove Szemerédi’s Theorem [35, Theorem 3.21].

**Theorem 18** Let  $B$  be a subset of  $\mathbb{N}$  with positive upper Banach density, that is

$$\limsup_{|I| \rightarrow \infty} |B \cap I|/|I| > 0, \quad (65)$$

where  $I$  varies over bounded subintervals of  $\mathbb{N}$  and  $|A|$  denotes the cardinality of  $A \subset \mathbb{N}$ . The set  $B$  contains arithmetic progressions of arbitrary length.

For details and further results, we refer to Furstenberg’s beautiful book [35].

## Future Directions

As far as applied work is concerned, the resolution of the statistical problems, described at the end in the Sect. “Chaos and Equicontinuity”, concerning chaotic attractors would be most useful. In addition, it would be nice to know which compact manifolds admit minimal homeomorphisms.

## Cross References

► Entropy in Ergodic Theory

## Bibliography

The approach to topological dynamics which was described in the initial sections is presented in detail in Akin [1] and [2]. For the deeper, more specialized work in topological dynamics see Auslander [17], Ellis [33] and Akin [4]. A general survey of the field is given in de Vries [65]. Furstenberg [35] is a classic illustration of the applicability of topological dynamics to other fields.

Clark Robinson's text [55] is an excellent introduction to dynamical systems in general. More elementary introductions are Devaney [31] and Alligood et al. [10]. Hirsch et al. [40] provides a nice transition from differential equations to the general theory.

Much of modern differentiable dynamics grows out of the work of Smale and his students, see especially the classic Smale [60], included in the collection [61]. The seminal paper Shub and Smale [58] provides a bridge between this work and the purely topological aspects of attractor theory, see also Shub [57].

The fashionable topic of chaos has generated a large sample of writing whose quality exhibits extremely high variance. For expository surveys I recommend Lorenz [48] and Stewart [62]. An excellent collection of relatively readable, classic papers is Hunt et al. [43].

For applications in biology see Hofbauer and Sigmund [41] and May [49]. Also, don't miss Sigmund's delightful book [59].

- Akin E (1993) The General Topology of Dynamical Systems. Am Math Soc
- Akin E (1996) Dynamical systems: the topological foundations. In: Aulbach B, Colonius F (eds) Six Lectures on Dynamical Systems. World Scientific, Singapore, pp 1–43
- Akin E (1996) On chain continuity. Discret Contin Dyn Syst 2:111–120
- Akin E (1997) Recurrence in Topological Dynamics: Furstenberg Families and Ellis Actions. Plenum Press, New York
- Akin E (2004) Lectures on Cantor and Mycielski sets for dynamical systems. In: Assani I (ed) Chapel Hill Ergodic Theory Workshops. Am Math Soc, Providence, pp 21–80
- Akin E, Auslander J, Berg K (1996) When is a transitive map chaotic? In: Bergelson V, March K, Rosenblatt J (eds) Conference in Ergodic Theory and Probability. Walter de Gruyter, Berlin, pp 25–40
- Akin E, Glasner E (2001) Residual properties and almost equicontinuity. J d'Analyse Math 84:243–286
- Akin E, Hurley M, Kennedy JA (2003) Dynamics of Topologically Generic Homeomorphisms. Memoir, vol 783. Am Math Soc, Providence
- Alexopoulos J (1991) Contraction mappings in a compact metric space. Solution No.6611. Math Assoc Am Mon 98:450
- Alligood KT, Sauer TD, Yorke JA (1996) Chaos: An Introduction to Dynamical Systems. Springer Science Business Media, New York
- Almgren FJ (1966) Plateau's Problem. Benjamin WA, Inc., New York
- Alpern SR, Prasad VS (2001) Typical dynamics of volume preserving homeomorphisms. Cambridge University Press, Cambridge
- Arhangel'skii AV, Pontryagin LS (eds) (1990) General topology I. Springer, Berlin
- Arhangel'skii AV (ed) (1995) General Topology III. Springer, Berlin
- Arhangel'skii AV (ed) (1996) General Topology II. Springer, Berlin
- Auslander J (1964) Generalized recurrence in dynamical systems. In: Contributions to Differential Equations, vol 3. John Wiley, New York, pp 55–74
- Auslander J (1988) Minimal Flows and Their Extensions. North-Holland, Amsterdam
- Auslander J, Bhatia N, Siebert P (1964) Attractors in dynamical systems. Bol Soc Mat Mex 9:55–66
- Auslander J, Siebert P (1963) Prolongations and generalized Liapunov functions. In: International Symposium on Nonlinear Differential Equations and Nonlinear Mechanics. Academic Press, New York, pp 454–462
- Auslander J, Siebert P (1964) Prolongations and stability in dynamical systems. Ann Inst Fourier 14(2):237–268
- Auslander J, Yorke J (1980) Interval maps, factors of maps and chaos. Tohoku Math J 32:177–188
- Banks J, Brooks J, Cairns G, Davis G, Stacey P (1992) On Devaney's Definition of Chaos. Am Math Mon 99:332–333
- Barrow-Green J (1997) Poincaré and the Three Body Problem. American Mathematical Society, Providence
- Bergelson V, Leibman A (1996) Polynomial extensions of Van der Waerden's and Szemerédi's theorems. J Am Math Soc 9:725–753
- Bing RH (1988) The Collected Papers of R. H. Bing. American Mathematical Society, Providence
- Brown M (ed) (1991) Continuum Theory and Dynamical Systems. American Mathematical Society, Providence
- Bhatia N, Szego G (1970) Stability Theory of Dynamical Systems. Springer, Berlin
- Coddington E, Levinson N (1955) Theory of Ordinary Differential Equations. McGraw-Hill, New York
- Conley C (1978) Isolated Invariant Sets and the Morse Index. American Mathematical Society, Providence
- Denker M, Grillenberger C, Sigmund K (1978) Ergodic Theory on Compact Spaces. Lect. Notes in Math, vol 527. Springer, Berlin
- Devaney RL (1989) An Introduction to Chaotic Dynamical Sys-



- tems, 2nd edn. Addison-Wesley Publishing Company, Redwood City
32. Douglas J (1939) Minimal surfaces of higher topological structure. *Ann Math* 40:205–298
  33. Ellis R (1969) *Lectures on Topological Dynamics*. WA Benjamin Inc, New York
  34. Fathi A, Herman M (1977) Existence de difféomorphismes minimaux. *Soc Math de France, Astérisque* 49:37–59
  35. Furstenberg H (1981) *Recurrence in Ergodic Theory and Combinatorial Number Theory*. Princeton Univ. Press, Princeton
  36. Glasner E (2003) *Ergodic Theory Via Joinings*. American Mathematical Society, Providence
  37. Glasner E, Weiss B (1993) Sensitive dependence on initial conditions. *Nonlinearity* 6:1067–1075
  38. Gottschalk W, Hedlund G (1955) *Topological Dynamics*. American Mathematical Society, Providence
  39. Hartman P (1964) *Ordinary Differential Equations*. John Wiley, New York
  40. Hirsch MW, Smale S, Devaney RL (2004) *Differential Equations, Dynamical Systems and an Introduction to Chaos*, 2nd edn. Elsevier Academic Press, Amsterdam
  41. Hofbauer J, Sigmund K (1988) *The Theory of Evolution and Dynamical Systems*. Cambridge Univ. Press, Cambridge
  42. Huang W, Ye X (2002) Devaney's chaos or 2-scattering implies Li–Yorke's chaos. *Topol Appl* 117:259–272
  43. Hunt B, Kennedy JA, Li T-Y, Nusse H (eds) (2004) *The Theory of Chaotic Attractors*. Springer, Berlin
  44. Iwanik A (1989) Independent sets of transitive points. In: *Dynamical Systems and Ergodic Theory*. Banach Cent Publ 23:277–282
  45. Katok A, Hasselblatt B (1995) *Introduction to the Modern Theory of Dynamical Systems*. Cambridge Univ. Press, Cambridge
  46. Kuratowski K (1973) Applications of the Baire-category method to the problem of independent sets. *Fundam Math* 81:65–72
  47. Li TY, Yorke J (1975) Period three implies chaos. *Am Math Mon* 82:985–992
  48. Lorenz E (1993) *The Essence of Chaos*. University of Washington Press, Seattle
  49. May RM (1973) *Stability and Complexity in Model Ecosystems*. Princeton Univ. Press, Princeton
  50. Murdoch J (1991) *Perturbations*. John Wiley and Sons, New York
  51. Mycielski J (1964) Independent sets in topological algebras. *Fundam Math* 55:139–147
  52. Nemytskii V, Stepanov V (1960) *Qualitative Theory of Differential Equations*. Princeton U Press, Princeton
  53. Oxtoby J (1980) *Measure and Category*, 2nd edn. Springer, Berlin
  54. Peterson K (1983) *Ergodic Theory*. Cambridge Univ. Press, Cambridge
  55. Robinson C (1995) *Dynamical Systems: Stability, Symbolic Dynamics and Chaos*. CRC Press, Boca Raton
  56. Rudolph D (1990) *Fundamentals of Measurable Dynamics: Ergodic Theory on Lebesgue Spaces*. Oxford Univ. Press, Oxford
  57. Shub M (1987) *Global Stability of Dynamical Systems*. Springer, New York
  58. Shub M, Smale S (1972) Beyond hyperbolicity. *Ann Math* 96:587–591
  59. Sigmund K (1993) *Games of Life: Explorations in Ecology, Evolution and Behavior*. Oxford Univ. Press, Oxford
  60. Smale S (1967) Differentiable dynamical systems. *Bull Am Math Soc* 73:747–817
  61. Smale S (1980) *The Mathematics of Time*. Springer, Berlin
  62. Stewart I (1989) *Does God Play Dice?* Basil Blackwell, Oxford
  63. Ura T (1953) Sur les courbes définies par les équations différentielles dans l'espace à m dimensions. *Ann Sci Ecole Norm Sup* [3] 70:287–360
  64. Ura T (1959) Sur le courant extérieur à une région invariante. *Funk Ekv* 2:143–200
  65. de Vries J (1993) *Elements of Topological Dynamics*. Kluwer, Dordrecht

## Topological Dynamics of Cellular Automata

PETR KŮRKA<sup>1,2</sup>

<sup>1</sup> Département d'Informatique, Université de Nice Sophia Antipolis, Nice, France

<sup>2</sup> Center for Theoretical Study, Academy of Sciences and Charles University, Prague, Czechia

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Topological Dynamics  
 Symbolic Dynamics  
 Equicontinuity  
 Surjectivity  
 Permutive and Closing Cellular Automata  
 Expansive Cellular Automata  
 Attractors  
 Subshifts and Entropy  
 Examples  
 Future Directions  
 Acknowledgments  
 Bibliography

### Glossary

**Almost equicontinuous CA** has an equicontinuous configuration.

**Attractor** omega-limit of a clopen invariant set.

**Blocking word** interrupts information flow.

**Closing CA** distinct asymptotic configurations have distinct images.

**Column subshift** columns in space-time diagrams.

**Cross section** one-sided inverse map.

**Directional dynamics** dynamics along a direction in the space-time diagram.

**Equicontinuous configuration** nearby configurations remain close.

**Equicontinuous CA** all configurations are equicontinuous.

**Expansive CA** distinct configurations get away.

**Finite time attractor** is attained in finite time from its neighborhood.

**Jointly periodic configuration** is periodic both for the shift and the CA.

**Lyapunov exponents** asymptotic speed of information propagation.

**Maximal attractor** omega-limit of the full space.

**Nilpotent CA** maximal attractor is a singleton.

**Open CA** image of an open set is open.

**Permutive CA** local rule permutes an extremal coordinate.

**Quasi-attractor** a countable intersection of attractors.

**Signal subshift** weakly periodic configurations of a given period.

**Spreading set** clopen invariant set which propagates in both directions.

**Subshift attractor** limit set of a spreading set.

## Definition of the Subject

A **topological dynamical system** is a continuous selfmap  $F: X \rightarrow X$  of a topological space  $X$ . Topological dynamics studies iterations  $F^n: X \rightarrow X$ , or trajectories  $(F^n(x))_{n \geq 0}$ . Basic questions are how trajectories depend on initial conditions, whether they are dense in the state space  $X$ , whether they have limits, or what are their accumulation points. While cellular automata were introduced in the late 1940s by Neumann [36] as regular infinite networks of finite automata, topological dynamics of cellular automata began in 1969 with Hedlund [19] who viewed one-dimensional cellular automata in the context of symbolic dynamics as endomorphisms of the shift dynamical systems. In fact, the term “cellular automaton” never appears in his paper. Hedlund’s main results are the characterizations of surjective and open cellular automata. In the early 1980s Wolfram [41] produced space-time representations of one-dimensional cellular automata and classified them informally into four classes using dynamical concepts like periodicity, stability and chaos. Wolfram’s classification stimulated mathematical research involving all the concepts of topological and measure-theoretical dynamics, and several formal classifications were introduced using dynamical concepts.

There are two well-understood classes of cellular automata with remarkably different stability properties. Equicontinuous cellular automata settle into a fixed or

periodic configuration depending on the initial condition and cannot be perturbed by fluctuations. This is a paradigm of stability. Positively expansive cellular automata, on the other hand, are conjugated (isomorphic) to one-sided full shifts. They have dense orbits, dense periodic configurations, positive topological entropy, and sensitive dependence on the initial conditions. This is a paradigm of chaotic behavior. Between these two extreme classes there are many distinct types of dynamical behavior which are understood much less. Only some specific classes or particular examples have been elucidated and a general theory is still lacking.

## Introduction

Dynamical properties of CA are usually studied in the context of symbolic dynamics. Other possibilities are measurable dynamics (see Pivato ► [Ergodic Theory of Cellular Automata](#)) or non-compact dynamics in Besicovitch or Weyl spaces (see Formenti and Kůrka ► [Dynamics of Cellular Automata in Non-compact Spaces](#)). In symbolic dynamics, the state space is the Cantor space of symbolic sequences. The Cantor space has distinguished topological properties which simplify some concepts of dynamics. This is the case of attractor and topological entropy. Cellular automata can be defined in the context of symbolic dynamics as continuous mappings which commute with the shift map.

Equicontinuous and almost equicontinuous CA can be characterized using the concept of blocking words. While equicontinuous CA are eventually periodic, closely related almost equicontinuous automata, are periodic on a large (residual) subset of the state space. Outside of this subset, however, their behavior can be arbitrarily complex.

A property which strongly constrains the dynamics of cellular automata is surjectivity. Surjective automata preserve the uniform Bernoulli measure, they are bounded-to-one, and their unique subshift attractor is the full space. An important subclass of surjective CA are (left- or right-) closing automata. They have dense sets of periodic configurations. Cellular automata which are both left- and right-closing are open: They map open sets to open sets, are  $n$ -to-one and have cross-sections. A fairly well-understood class is that of positively expansive automata. A positively expansive cellular automaton is conjugated (isomorphic) to a one-sided full shift. Closely related are bijective expansive automata which are conjugated to two-sided subshifts, usually of finite type.

Another important concept elucidating the dynamics of CA is that of an attractor. With respect to attractors, CA classify into two basic classes. In one class there are CA

which have disjoint attractors. They have then countably infinite numbers of attractors and uncountable numbers of quasi-attractors i. e., countable intersections of attractors. In the other class there are CA that have either a minimal attractor or a minimal quasi-attractor which is then contained in any attractor. An important class of attractors are subshift attractors—subsets which are both attractors and subshifts. They have always non-empty intersection, so they form a lattice with maximal element.

Factors of CA that are subshifts are useful because factor maps preserve many dynamical properties while they simplify the dynamics. In a special case of column subshifts, they are formed by sequences of words occurring in a column of a space-time diagram. Factor subshifts are instrumental in evaluating the entropy of CA and in characterizing CA with the shadowing property.

A finer classification of CA is provided by quantitative characteristics. Topological entropy measures the quantity of information available to an observer who can see a finite part of a configuration. Lyapunov exponents measure the speed of information propagation. The minimum preimage number provides a finer classification of surjective cellular automata. Sets of left- and right-expansivity directions provide finer classification for left- and right-closing cellular automata.

## Topological Dynamics

We review basic concepts of topological dynamics as exposed in Kůrka [24]. A **Cantor space** is any metric space which is **compact** (any sequence has a convergent subsequence), **totally disconnected** (distinct points are separated by disjoint **clopen**, i. e., closed and open sets), and **perfect** (no point is isolated). Any two Cantor spaces are homeomorphic. A **symbolic space** is any compact, totally disconnected metric space, i. e., any closed subspace of a Cantor space. A **symbolic dynamical system** (SDS) is a pair  $(X, F)$  where  $X$  is a symbolic space and  $F: X \rightarrow X$  is a continuous map. The  $n$ th **iteration** of  $F$  is denoted by  $F^n$ . If  $F$  is bijective (invertible), the negative iterations are defined by  $F^{-n} = (F^{-1})^n$ . A set  $Y \subseteq X$  is **invariant**, if  $F(Y) \subseteq Y$  and **strongly invariant** if  $F(Y) = Y$ .

A **homomorphism**  $\varphi: (X, F) \rightarrow (Y, G)$  of SDS is a continuous map  $\varphi: X \rightarrow Y$  such that  $\varphi \circ F = G \circ \varphi$ . A **conjugacy** is a bijective homomorphism. The systems  $(X, F)$  and  $(Y, G)$  are **conjugated**, if there exists a conjugacy between them. If  $\varphi$  is surjective, we say that  $(Y, G)$  is a **factor** of  $(X, F)$ . If  $\varphi$  is injective,  $(X, F)$  is a **subsystem** of  $(Y, G)$ . In this case  $\varphi(X) \subseteq Y$  is a closed invariant set. Conversely, if  $Y \subseteq X$  is a closed invariant set, then  $(Y, F)$  is a subsystem of  $(X, F)$ .

We denote by  $d: X \times X \rightarrow [0, \infty)$  the metric and by  $B_\delta(x) = \{y \in X: d(y, x) < \delta\}$  the ball with center  $x$  and radius  $\delta$ . A finite sequence  $(x_i \in X)_{0 \leq i \leq n}$  is a  **$\delta$ -chain** from  $x_0$  to  $x_n$ , if  $d(F(x_i), x_{i+1}) < \delta$  for all  $i < n$ . A point  $x \in X$   **$\varepsilon$ -shadows** a sequence  $(x_i)_{0 \leq i \leq n}$ , if  $d(F^i(x), x_i) < \varepsilon$  for all  $0 \leq i \leq n$ . A SDS  $(X, F)$  has the **shadowing property**, if for every  $\varepsilon > 0$  there exists  $\delta > 0$ , such that every  $\delta$ -chain is  $\varepsilon$ -shadowed by some point.

**Definition 1** Let  $(X, F)$  be a SDS. The **orbit relation**  $\mathcal{O}_F$ , the **recurrence relation**  $\mathcal{R}_F$ , the **non-wandering relation**  $\mathcal{N}_F$ , and the **chain relation**  $\mathcal{C}_F$  are defined by

$$\begin{aligned} (x, y) \in \mathcal{O}_F &\iff \exists n > 0, y = F^n(x) \\ (x, y) \in \mathcal{R}_F &\iff \forall \varepsilon > 0, \exists n > 0, d(y, F^n(x)) < \varepsilon \\ (x, y) \in \mathcal{N}_F &\iff \forall \varepsilon, \delta > 0, \exists n > 0, \exists z \in B_\delta(x), \\ &\quad d(F^n(z), y) < \varepsilon \\ (x, y) \in \mathcal{C}_F &\iff \forall \varepsilon > 0, \exists \varepsilon - \text{chain from } x \text{ to } y. \end{aligned}$$

We have  $\mathcal{O}_F \subseteq \mathcal{R}_F \subseteq \mathcal{N}_F \subseteq \mathcal{C}_F$ . The diagonal of a relation  $S \subseteq X \times X$  is  $|S| := \{x \in X: (x, x) \in S\}$ . We denote by  $S(x) := \{y \in X: (x, y) \in S\}$  the  $S$ -image of a point  $x \in X$ . The orbit of a point  $x \in X$  is  $\mathcal{O}_F(x) := \{F^n(x): n > 0\}$ . It is an invariant set, so its closure  $(\overline{\mathcal{O}_F(x)}, F)$  is a subsystem of  $(X, F)$ . A point  $x \in X$  is **periodic** with period  $n > 0$ , if  $F^n(x) = x$ , i. e., if  $x \in |\mathcal{O}_F|$ . A point  $x \in X$  is **eventually periodic**, if  $F^m(x)$  is periodic for some **preperiod**  $m \geq 0$ . The points in  $|\mathcal{R}_F|$  are called **recurrent**, the points in  $|\mathcal{N}_F|$  are called **non-wandering** and the points in  $|\mathcal{C}_F|$  are called **chain-recurrent**. The sets  $|\mathcal{N}_F|$  and  $|\mathcal{C}_F|$  are closed and invariant, so  $(|\mathcal{N}_F|, F)$  and  $(|\mathcal{C}_F|, F)$  are subsystems of  $(X, F)$ . The set of **transitive** points is  $\mathcal{T}_F := \{x \in X: \mathcal{O}(x) = X\}$ . A system  $(X, F)$  is **minimal**, if  $\mathcal{R}_F = X \times X$ . This happens if each point has a dense orbit, i. e., if  $\mathcal{T}_F = X$ . A system is **transitive**, if  $\mathcal{N}_F = X \times X$ , i. e., if for any non-empty open sets  $U, V \subseteq X$  there exists  $n > 0$  such that  $F^n(U) \cap V \neq \emptyset$ . A system is transitive if it has a transitive point, i. e., if  $\mathcal{T}_F \neq \emptyset$ . In this case the set of transitive points  $\mathcal{T}_F$  is **residual**, i. e., it contains a countable intersection of dense open sets. An infinite system is **chaotic**, if it is transitive and has a dense set of periodic points. A system  $(X, F)$  is **weakly mixing**, if  $(X \times X, F \times F)$  is transitive. It is **strongly transitive**, if  $(X, F^n)$  is transitive for any  $n > 0$ . A system  $(X, F)$  is **mixing**, if for every non-empty open sets  $U, V \subseteq X$ ,  $F^n(U) \cap V \neq \emptyset$  for all sufficiently large  $n$ . A system is **chain-transitive**, if  $\mathcal{C}_F = X \times X$ , and **chain-mixing**, if for any  $x, y \in X$  and any  $\varepsilon > 0$  there exist chains from  $x$  to  $y$  of arbitrary, large enough length. If a system  $(X, F)$  has the shadowing property, then  $\mathcal{N}_F = \mathcal{C}_F$ . It follows that a chain-transitive system with the shadowing property is

transitive, and a chain-mixing system with the shadowing property is mixing.

A **clopen partition** of a symbolic space  $X$  is a finite system of disjoint clopen sets whose union is  $X$ . The **join** of clopen partitions  $\mathcal{U}$  and  $\mathcal{V}$  is  $\mathcal{U} \vee \mathcal{V} = \{U \cap V : U \in \mathcal{U}, V \in \mathcal{V}\}$ . The inverse image of a clopen partition  $\mathcal{U}$  by  $F$  is  $F^{-1}(\mathcal{U}) = \{F^{-1}(U) : U \in \mathcal{U}\}$ . The entropy  $H(X, F, \mathcal{U})$  of a partition and the **entropy**  $h(X, F)$  of a system are defined by

$$\begin{aligned} H(X, F, \mathcal{U}) &= \lim_{n \rightarrow \infty} \frac{\ln |\mathcal{U} \vee F^{-1}(\mathcal{U}) \vee \dots \vee F^{-(n-1)}(\mathcal{U})|}{n}, \\ h(X, F) &= \sup\{H(X, F, \mathcal{U}) : \mathcal{U} \text{ is a clopen partition of } X\}. \end{aligned}$$

## Symbolic Dynamics

An **alphabet** is any finite set with at least two elements. The cardinality of a finite set  $A$  is denoted by  $|A|$ . We frequently use alphabet  $\mathbf{2} = \{0, 1\}$  and more generally  $\mathbf{n} = \{0, \dots, n-1\}$ . A word over  $A$  is any finite sequence  $u = u_0 \dots u_{n-1}$  of elements of  $A$ . The length of  $u$  is denoted by  $|u| := n$  and the word of zero length is denoted by  $\lambda$ . The set of all words of length  $n$  is denoted by  $A^n$ . The set of all non-zero words and the set of all words are

$$A^+ = \bigcup_{n>0} A^n, \quad A^* = \bigcup_{n \geq 0} A^n.$$

We denote by  $\mathbb{Z}$  the set of integers, by  $\mathbb{N}$  the set of non-negative integers, by  $\mathbb{N}^+$  the set of positive integers, by  $\mathbb{Q}$  the set of rational numbers, and by  $\mathbb{R}$  the set of real numbers. The set of one-sided configurations (infinite words) is  $A^{\mathbb{N}}$  and the set of two-sided configurations (biinfinite words) is  $A^{\mathbb{Z}}$ . If  $u$  is a finite or infinite word and  $I = [i, j]$  is an interval of integers on which  $u$  is defined, put  $u_{[i,j]} = u_i \dots u_j$ . Similarly for open or half-open intervals  $u_{[i,j)} = u_i \dots u_{j-1}$ . We say that  $v$  is a **subword** of  $u$  and write  $v \sqsubseteq u$ , if  $v = u_I$  for some interval  $I \subseteq \mathbb{Z}$ . If  $u \in A^n$ , then  $u^\infty \in A^{\mathbb{Z}}$  is the infinite repetition of  $u$  defined by  $(u^\infty)_{kn+i} = u_i$ . Similarly  $x = u^\infty.v^\infty$  is the configuration satisfying  $x_{i+k|u|} = u_i$  for  $k < 0$ ,  $0 \leq i < |u|$  and  $x_{i+k|v|} = v_i$  for  $k \geq 0$ ,  $0 \leq i < |v|$ . On  $A^{\mathbb{N}}$  and  $A^{\mathbb{Z}}$  we have metrics

$$\begin{aligned} d(x, y) &= 2^{-n} \\ \text{where } n &= \min\{i \geq 0 : x_i \neq y_i\}, \quad x, y \in A^{\mathbb{N}} \\ d(x, y) &= 2^{-n} \\ \text{where } n &= \min\{i \geq 0 : x_i \neq y_i \text{ or } x_{-i} \neq y_{-i}\}, \\ &\quad x, y \in A^{\mathbb{Z}}. \end{aligned}$$

Both  $A^{\mathbb{N}}$  and  $A^{\mathbb{Z}}$  are Cantor spaces. In  $A^{\mathbb{N}}$  and  $A^{\mathbb{Z}}$  the **cylinder sets** of a word  $u \in A^n$  are  $[u] := \{x \in A^{\mathbb{N}} : x_{[0,n)} = u\}$ , and  $[u]_k := \{x \in A^{\mathbb{Z}} : x_{[k,k+n)} = u\}$ , where  $k \in \mathbb{Z}$ . Cylinder sets are clopen and every clopen set is a finite union of cylinders. The shift maps  $\sigma : A^{\mathbb{N}} \rightarrow A^{\mathbb{N}}$  and  $\sigma : A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  defined by  $\sigma(x)_i = x_{i+1}$  are continuous. While the two-sided shift is bijective, the one-sided shift is not: every configuration has  $|A|$  preimages. A **one-sided subshift** is any non-empty closed set  $\Sigma \subseteq A^{\mathbb{N}}$  which is shift-invariant, i. e.,  $\sigma(\Sigma) \subseteq \Sigma$ . A **two-sided subshift** is any non-empty closed set  $\Sigma \subseteq A^{\mathbb{Z}}$  which is strongly shift-invariant, i. e.,  $\sigma(\Sigma) = \Sigma$ . Thus, a subshift  $\Sigma$  represents a SDS  $(\Sigma, \sigma)$ . Systems  $(A^{\mathbb{Z}}, \sigma)$  and  $(A^{\mathbb{N}}, \sigma)$  are called **full shifts**. Given a set  $D \subseteq A^*$  of **forbidden words**, the set  $S_D := \{x \in A^{\mathbb{N}} : \forall u \in D, u \not\sqsubseteq x\}$  is a one-sided subshift, provided it is non-empty. Any one-sided subshift has this form. Similarly,  $S_D := \{x \in A^{\mathbb{Z}} : \forall u \in D, u \not\sqsubseteq x\}$  is a two-sided subshift, and any two-sided subshift has this form. A (one- or two-sided) subshift is of **finite type** (SFT), if the set  $D$  of forbidden words is finite. The **language of a subshift**  $\Sigma$  is the set of all subwords of configurations of  $\Sigma$ ,

$$\begin{aligned} \mathcal{L}^n(\Sigma) &= \{u \in A^n : \exists x \in \Sigma, u \sqsubseteq x\}, \\ \mathcal{L}(\Sigma) &= \bigcup_{n \geq 0} \mathcal{L}^n(\Sigma) = \{u \in A^* : \exists x \in \Sigma, u \sqsubseteq x\}. \end{aligned}$$

The entropy of a subshift  $\Sigma$  is  $h(\Sigma, \sigma) = \lim_{n \rightarrow \infty} \ln |\mathcal{L}^n(\Sigma)|/n$ . A word  $w \in \mathcal{L}(\Sigma)$  is **intrinsically synchronizing**, if for any  $u, v \in A^*$  such that  $uw, wv \in \mathcal{L}(\Sigma)$  we have  $uwv \in \mathcal{L}(\Sigma)$ . A subshift is of finite type if all sufficiently long words are intrinsically synchronizing (see Lind and Marcus [29]).

A subshift  $\Sigma$  is **sofic**, if  $\mathcal{L}(\Sigma)$  is a regular language, i. e., if  $\Sigma = \Sigma_{\mathcal{G}}$  is the set of labels of paths of a **labelled graph**  $\mathcal{G} = (V, E, s, t, l)$ , where  $V$  is a finite set of vertices,  $E$  is a finite set of edges,  $s, t : E \rightarrow V$  are the source and target map, and  $l : E \rightarrow A$  is a labelling function. The labelling function extends to a function  $\ell : E^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  defined by  $\ell(x)_i = l(x_i)$ . A graph  $\mathcal{G} = (V, E, s, t, l)$  determines a SFT  $\Sigma_{|\mathcal{G}|} = \{u \in E^{\mathbb{Z}}, \forall i \in \mathbb{Z}, t(u_i) = s(u_{i+1})\}$  and  $\Sigma_{\mathcal{G}} = \{\ell(u) : u \in \Sigma_{|\mathcal{G}|}\}$  so that  $\ell : (\Sigma_{|\mathcal{G}|}, \sigma) \rightarrow (\Sigma_{\mathcal{G}}, \sigma)$  is a factor map. If  $\Sigma = \Sigma_{\mathcal{G}}$ , we say that  $\mathcal{G}$  is a **presentation** of  $\Sigma$ . A graph  $\mathcal{G}$  is **right-resolving**, if different outgoing edges of a vertex are labelled differently, i. e., if  $l(e) \neq l(e')$  whenever  $e \neq e'$  and  $s(e) = s(e')$ . A word  $w$  is **synchronizing** in  $\mathcal{G}$ , if all paths with label  $w$  have the same target, i. e., if  $t(u) = t(u')$  whenever  $\ell(u) = \ell(u') = w$ . If  $w$  is synchronizing in  $\mathcal{G}$ , then  $w$  is intrinsically synchronizing in  $\Sigma_{\mathcal{G}}$ . Any transitive sofic subshift  $\Sigma$  has a unique **minimal right-resolving presentation**  $\mathcal{G}$  which has the

smallest number of vertices. Any word can be extended to a word which is synchronizing in  $\mathcal{G}$  (see Lind and Marcus [29]).

A **deterministic finite automaton (DFA)** over an alphabet  $A$  is a system  $\mathcal{A} = (Q, \delta, q_0, q_1)$ , where  $Q$  is a finite set of states,  $\delta: Q \times A \rightarrow Q$  is a transition function and  $q_0, q_1$  are the initial and rejecting states. The transition function extends to  $\delta: Q \times A^* \rightarrow Q$  by  $\delta(q, \lambda) = q$ ,  $\delta(q, ua) = \delta(\delta(q, u), a)$ . The language accepted by  $\mathcal{A}$  is  $\mathcal{L}(\mathcal{A}) := \{u \in A^*: \delta(q_0, u) \neq q_1\}$  (see e.g., Hopcroft and Ullmann [20]). The DFA of a labelled graph  $\mathcal{G} = (V, E, s, t, l)$  is  $\mathcal{A}(\mathcal{G}) = (\mathcal{P}(V), \delta, V, \emptyset)$ , where  $\mathcal{P}(V)$  is the set of all subsets of  $V$  and  $\delta(q, a) = \{v \in V: \exists u \in q, u \xrightarrow{a} v\}$ . Then  $\mathcal{L}(\mathcal{A}(\mathcal{G})) = \mathcal{L}(\Sigma_{\mathcal{G}})$ . We can reduce the size of  $\mathcal{A}(\mathcal{G})$  by taking only those states that are accessible from the initial state  $V$ .

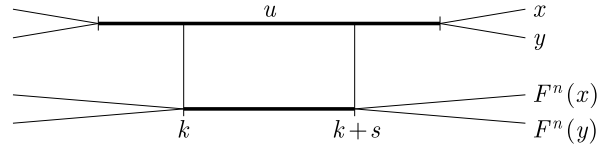
A **periodic structure**  $\mathbf{n} = (n_i)_{i \geq 0}$  is a sequence of integers greater than 1. For a given periodic structure  $\mathbf{n}$ , let  $X_{\mathbf{n}} := \prod_{i \geq 0} \{0, \dots, n_i - 1\}$  be the product space with metric  $d(x, y) = 2^{-n}$  where  $n = \min\{i \geq 0: x_i \neq y_i\}$ . Then  $X_{\mathbf{n}}$  is a Cantor space. The **adding machine** (odometer) of  $\mathbf{n}$  is a SDS  $(X_{\mathbf{n}}, F)$  given by the formula

$$F(x)_i = \begin{cases} (x_i + 1) \bmod n_i & \text{if } \forall j < i, x_j = n_j - 1 \\ x_i & \text{if } \exists j < i, x_j < n_j - 1. \end{cases}$$

Each adding machine is minimal and has zero topological entropy.

**Definition 2** A map  $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  is a **cellular automaton (CA)** if there exist integers  $m \leq a$  (**memory** and **anticipation**) and a **local rule**  $f: A^{a-m+1} \rightarrow A$  such that for any  $x \in A^{\mathbb{Z}}$  and any  $i \in \mathbb{Z}$ ,  $F(x)_i = f(x_{[i-m, i+a]})$ . Call  $r = \max\{|m|, |a|\} \geq 0$  the **radius** of  $F$  and  $d = a - m \geq 0$  its **diameter**.

By a theorem of Hedlund [19], a map  $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  is a cellular automaton if it is continuous and commutes with the shift, i.e.,  $\sigma \circ F = F \circ \sigma$ . This means that  $F: (A^{\mathbb{Z}}, \sigma) \rightarrow (A^{\mathbb{Z}}, \sigma)$  is a homomorphism of the full shift and  $(A^{\mathbb{Z}}, F)$  is a SDS. We can assume that the local rule acts on a symmetric neighborhood of 0, so  $F(x)_i = f(x_{[i-r, i+r]})$ , where  $f: A^{2r+1} \rightarrow A$ . There is a trade-off between the radius and the size of the alphabet. Any CA is conjugated to a CA with radius 1. Any  $\sigma$ -periodic configuration of a CA  $(A^{\mathbb{Z}}, F)$  is  $F$ -eventually periodic. Hence the set of  $F$ -eventually periodic configurations is dense. Thus, a cellular automaton is never minimal, because it has always an  $F$ -periodic configuration. A configuration  $x \in A^{\mathbb{Z}}$  is **weakly periodic**, if there exists  $p \in \mathbb{Z}$  and  $q \in \mathbb{N}^+$  such that  $F^q \sigma^p(x) = x$ . A configuration  $x \in A^{\mathbb{Z}}$  is **jointly peri-**



Topological Dynamics of Cellular Automata, Figure 1

A blocking word

**odic**, if it is both  $F$ -periodic and  $\sigma$ -periodic. A CA  $(A^{\mathbb{Z}}, F)$  is **nilpotent**, if  $F^n(A^{\mathbb{Z}})$  is a singleton for some  $n > 0$ .

### Equicontinuity

A point  $x \in X$  of a SDS  $(X, F)$  is **equicontinuous**, if

$$\forall \varepsilon > 0, \exists \delta > 0, \forall y \in B_{\delta}(x), \forall n \geq 0, \\ d(F^n(y), F^n(x)) < \varepsilon.$$

The set of equicontinuous points is denoted by  $\mathcal{E}_F$ . A system is **equicontinuous**, if  $\mathcal{E}_F = X$ . In this case it is uniformly equicontinuous, i.e.,

$$\forall \varepsilon > 0, \exists \delta > 0, \forall x, y \in X, (d(x, y) < \delta \\ \Rightarrow \forall n \geq 0, d(F^n(x), F^n(y)) < \varepsilon).$$

A system  $(X, F)$  is **almost equicontinuous**, if  $\mathcal{E}_F \neq \emptyset$ . A system is **sensitive**, if

$$\exists \varepsilon > 0, \forall x \in X, \forall \delta > 0, \exists y \in B_{\delta}(x), \exists n \geq 0, \\ d(F^n(y), F^n(x)) \geq \varepsilon.$$

Clearly, a sensitive system has no equicontinuous points. The converse is not true in general but holds for transitive systems (Akin et al. [2]).

**Definition 3** A word  $u \in A^+$  with  $|u| \geq s \geq 0$  is **s-blocking** for a CA  $(A^{\mathbb{Z}}, F)$ , if there exists an **offset**  $k \in [0, |u| - s]$  such that

$$\forall x, y \in [u]_0, \forall n \geq 0, F^n(x)_{[k, k+s]} = F^n(y)_{[k, k+s]}.$$

**Theorem 4 (Kůrka [27])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with radius  $r \geq 0$ . The following conditions are equivalent.

- (1)  $(A^{\mathbb{Z}}, F)$  is not sensitive.
- (2)  $(A^{\mathbb{Z}}, F)$  has an  $r$ -blocking word.
- (3)  $\mathcal{E}_F$  is **residual**, i.e., a countable intersection of dense open sets.
- (4)  $\mathcal{E}_F \neq \emptyset$ .



For a non-empty set  $B \subseteq A^*$  define

$$\begin{aligned}\mathcal{T}_\sigma^n(B) &:= \{x \in A^{\mathbb{Z}} : (\exists j > i > n, x_{[i,j]} \in B) \\ &\quad \text{and } (\exists j < i < -n, x_{[j,i]} \in B)\}, \\ \mathcal{T}_\sigma(B) &:= \bigcap_{n \geq 0} \mathcal{T}_\sigma^n(B).\end{aligned}$$

Each  $\mathcal{T}_\sigma^n(B)$  is open and dense, so the set  $\mathcal{T}_\sigma(B)$  of **B-recurrent** configurations is residual. If  $B$  is the set of  $r$ -blocking words, then  $\mathcal{E}_F = \mathcal{T}_\sigma(B)$ .

**Theorem 5 (Kůrka [27])** *Let  $(A^{\mathbb{Z}}, F)$  be a CA with radius  $r \geq 0$ . The following conditions are equivalent.*

- (1)  $(A^{\mathbb{Z}}, F)$  is equicontinuous, i. e.,  $\mathcal{E}_F = A^{\mathbb{Z}}$ .
- (2) There exists  $k > 0$  such that any  $u \in A^k$  is  $r$ -blocking.
- (3) There exists a preperiod  $q \geq 0$  and a period  $p > 0$ , such that  $F^{q+p} = F^q$ .

In particular every CA with radius  $r = 0$  is equicontinuous. A configuration is equicontinuous for  $F$  if it is equicontinuous for  $F^n$ , i. e.,  $\mathcal{E}_F = \mathcal{E}_{F^n}$ . This fact enables us to consider equicontinuity along rational directions  $\alpha = \frac{p}{q}$ .

**Definition 6** The sets of **equicontinuous directions** and **almost equicontinuous directions** of a CA  $(A^{\mathbb{Z}}, F)$  are defined by

$$\begin{aligned}\mathfrak{E}(F) &= \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} = A^{\mathbb{Z}} \right\}, \\ \mathfrak{A}(F) &= \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, \mathcal{E}_{F^q \sigma^p} \neq \emptyset \right\}.\end{aligned}$$

Clearly,  $\mathfrak{E}(F) \subseteq \mathfrak{A}(F)$ , and both sets  $\mathfrak{E}(F)$  and  $\mathfrak{A}(F)$  are convex (Sablik [37]): If  $\alpha_0 < \alpha_1 < \alpha_2$  and  $\alpha_0, \alpha_2 \in \mathfrak{A}(F)$ , then  $\alpha_1 \in \mathfrak{A}(F)$ . Sablik [37] considers also equicontinuity along irrational directions.

**Proposition 7** *Let  $(A^{\mathbb{Z}}, F)$  be an equicontinuous CA such that there exists  $0 \neq \alpha \in \mathfrak{A}(F)$ . Then  $(A^{\mathbb{Z}}, F)$  is nilpotent.*

*Proof* We can assume  $\alpha < 0$ . There exist  $0 \leq k < m$  and  $w \in A^m$ , such that for all  $x, y \in A^{\mathbb{Z}}$  and for all  $i \in \mathbb{Z}$  we have

$$\begin{aligned}x_{[i, i+m]} &= y_{[i, i+m]} \\ \implies \forall n \geq 0, F^n(x)_{i+k} &= F^n(y)_{i+k}, \\ w &= x_{[i, i+m]} = y_{[i, i+m]} \\ \implies \forall n \geq 0, F^n \sigma^{\lfloor n\alpha \rfloor}(x)_{i+k} &= F^n \sigma^{\lfloor n\alpha \rfloor}(y)_{i+k}.\end{aligned}$$

Take  $n$  such that  $l := \lfloor n\alpha \rfloor + m \leq 0$ . There exists  $a \in A$  such that  $F^n \sigma^{\lfloor n\alpha \rfloor}(z)_k = a$  for every  $z \in [w]_0$ .

Let  $x \in A^{\mathbb{Z}}$  be arbitrary. For a given  $i \in \mathbb{Z}$ , take a configuration  $y \in [x]_{[i, i+m]} \cap [w]_{[i-l, i+m]}$ . Then  $z := \sigma^{i-\lfloor n\alpha \rfloor}(y) = \sigma^{i-l+m}(y) \in [w]_0$  and  $F^n(x)_{i+k} = F^n(y)_{i+k} = F^n \sigma^{\lfloor n\alpha \rfloor}(z)_k = a$ . Thus,  $F^n(x) = a^\infty$  for every  $x \in A^{\mathbb{Z}}$ , and  $F^{n+t}(x) = F^n(F^t(x)) = a^\infty$  for every  $t \geq 0$ , so  $(A^{\mathbb{Z}}, F)$  is nilpotent (see also Sablik [38]).  $\square$

**Theorem 8** *Let  $(A^{\mathbb{Z}}, F)$  be a CA with memory  $m$  and anticipation  $a$ , i. e.,  $F(x)_i = f(x_{[i-m, i+a]})$ . Then exactly one of the following conditions is satisfied.*

- (1)  $\mathfrak{E}(F) = \mathfrak{A}(F) = \mathbb{Q}$ . This happens if  $(A^{\mathbb{Z}}, F)$  is nilpotent.
- (2)  $\mathfrak{E}(F) = \emptyset$  and there exist real numbers  $\alpha_0 < \alpha_1$  such that

$$(\alpha_0, \alpha_1) \subseteq \mathfrak{A}(F) \subseteq [\alpha_0, \alpha_1] \subseteq [-a, -m].$$

- (3) There exists  $-a \leq \alpha \leq -m$  such that  $\mathfrak{A}(F) = \mathfrak{E}(F) = \{\alpha\}$ .
- (4) There exists  $-a \leq \alpha \leq -m$  such that  $\mathfrak{A}(F) = \{\alpha\}$  and  $\mathfrak{E}(F) = \emptyset$ .
- (5)  $\mathfrak{A}(F) = \mathfrak{E}(F) = \emptyset$ .

This follows from Theorems II.2 and II.5 in Sablik [37] and from Proposition 7. The zero CA of Example 1 belongs to class (1). The product CA of Example 4 belongs to class (2). The identity CA of Example 2 belongs to class (3). The Coven CA of Example 18 belongs to class (4). The sum CA of Example 11 belongs to class (5).

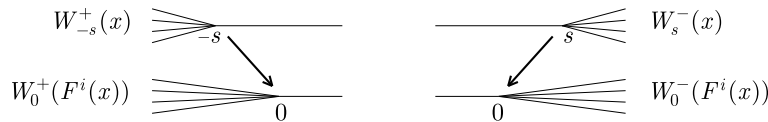
Sensitivity can be expressed quantitatively by **Lyapunov exponents** which measure the speed of information propagation. Let  $(A^{\mathbb{Z}}, F)$  be a CA. The left and right **perturbation sets** of  $x \in A^{\mathbb{Z}}$  are

$$\begin{aligned}W_s^-(x) &= \{y \in A^{\mathbb{Z}} : \forall i \leq s, y_i = x_i\}, \\ W_s^+(x) &= \{y \in A^{\mathbb{Z}} : \forall i \geq s, y_i = x_i\}.\end{aligned}$$

The left and right **perturbation speeds** of  $x \in A^{\mathbb{Z}}$  are

$$\begin{aligned}I_n^-(x) &= \min\{s \geq 0 : \forall i \leq n, F^i(W_s^-(x)) \subseteq W_0^-(F^i(x))\}, \\ I_n^+(x) &= \min\{s \geq 0 : \forall i \leq n, F^i(W_s^+(x)) \subseteq W_0^+(F^i(x))\}.\end{aligned}$$

Thus,  $I_n^-(x)$  is the minimum distance of a perturbation of the left part of  $x$  which cannot reach the zero site by time  $n$ . Both  $I_n^-(x)$  and  $I_n^+(x)$  are non-decreasing. If  $0 < s < t$ , and if  $x_{[s,t]}$  is an  $r$ -blocking word (where  $r$  is the radius), then  $\lim_{n \rightarrow \infty} I_n^-(x) \leq t$ . Similarly, if  $s < t < 0$  and if  $x_{[s,t]}$  is an  $r$ -blocking word, then  $\lim_{n \rightarrow \infty} I_n^+(x) \leq |s|$ . In particular, if  $x \in \mathcal{E}_F$ , then both



**Topological Dynamics of Cellular Automata, Figure 2**  
Perturbation speeds

$I_n^-(x)$  and  $I_n^+(x)$  have finite limit. If  $(A^{\mathbb{Z}}, F)$  is sensitive, then  $\lim_{n \rightarrow \infty} (I_n^-(x) + I_n^+(x)) = \infty$  for every  $x \in A^{\mathbb{Z}}$ .

**Definition 9** The left and right **Lyapunov exponents** of a CA  $(A^{\mathbb{Z}}, F)$  and  $x \in A^{\mathbb{Z}}$  are

$$\lambda_F^-(x) = \liminf_{n \rightarrow \infty} \frac{I_n^-(x)}{n}, \quad \lambda_F^+(x) = \liminf_{n \rightarrow \infty} \frac{I_n^+(x)}{n}.$$

If  $F$  has memory  $m$  and anticipation  $a$ , then  $\lambda_F^-(x) \leq \max\{a, 0\}$  and  $\lambda_F^+(x) \leq \max\{-m, 0\}$  for all  $x \in A^{\mathbb{Z}}$ . If  $x \in \mathcal{E}_F$ , then  $\lambda_F^+(x) = \lambda_F^-(x) = 0$ . If  $F$  is right-permutive (see Sect. “[Permutive and Closing Cellular Automata](#)”) with  $a > 0$ , then  $\lambda_F^-(x) = a$  for every  $x \in A^{\mathbb{Z}}$ . If  $F$  is left-permutive with  $m < 0$ , then  $\lambda_F^+(x) = -m$  for every  $x \in A^{\mathbb{Z}}$ .

**Theorem 10 (Bressaud and Tisseur [9])** For a positively expansive CA (see Sect. “[Expansive Cellular Automata](#)”) there exists a constant  $c > 0$ , such that for all  $x \in A^{\mathbb{Z}}$ ,  $\lambda^-(x) \geq c$  and  $\lambda^+(x) \geq c$ .

**Conjecture 11 (Bressaud and Tisseur [9])** Any sensitive CA has a configuration  $x$  such that  $\lambda^-(x) > 0$  or  $\lambda^+(x) > 0$ .

### Surjectivity

Let  $(A^{\mathbb{Z}}, F)$  be a CA with diameter  $d \geq 0$  and a local rule  $f: A^{d+1} \rightarrow A$ . We extend the local rule to a function  $f: A^* \rightarrow A^*$  by  $f(u)_i = f(u_{[i, i+d]})$  for  $i < |u| - d$ , so  $|f(u)| = \max\{|u| - d, 0\}$ . A **diamond** for  $f$  (Fig. 3 left) consists of words  $u, v \in A^d$  and distinct  $w, z \in A^+$  of the same length, such that  $f(uwv) = f(uzw)$ .

**Theorem 12 (Hedlund [19], Moothathu [33])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with local rule  $f: A^{d+1} \rightarrow A$ . The following conditions are equivalent.

- (1)  $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  is surjective.
- (2) For each  $x \in A^{\mathbb{Z}}$ ,  $F^{-1}(x)$  is finite or countable.
- (3) For each  $x \in A^{\mathbb{Z}}$ ,  $F^{-1}(x)$  is a finite set.
- (4) For each  $x \in A^{\mathbb{Z}}$ ,  $|F^{-1}(x)| \leq |A|^d$ .
- (5)  $f: A^* \rightarrow A^*$  is surjective.
- (6) For each  $u \in A^+$ ,  $|f^{-1}(u)| = |A|^d$ .

(7) For each  $u \in A^+$  with  $|u| \leq d \cdot \log_2 |A| \cdot (2d + |A|^{2d})$ ,  $|f^{-1}(u)| = |A|^d$ .

(8) There exists no diamond for  $f$ .

It follows that any injective CA is surjective and hence bijective. Although (6) asserts equality, the inequality in (4) may be strict. Another equivalent condition states that the uniform Bernoulli measure is invariant for  $F$ . In this form, Theorem 12 has been generalized to CA on mixing SFT (see Theorem 2B.1 in Pivato ▶ [Ergodic Theory of Cellular Automata](#)).

**Theorem 13 (Blanchard and Tisseur [5])**

- (1) Any configuration of a surjective CA is non-wandering, i. e.,  $|\mathcal{N}_F| = A^{\mathbb{Z}}$ .
- (2) Any surjective almost equicontinuous CA has a dense set of  $F$ -periodic configurations.
- (3) If  $(A^{\mathbb{Z}}, F)$  is an equicontinuous and surjective CA, then there exists  $p > 0$  such that  $F^p = \text{Id}$ . In particular,  $F$  is bijective.

**Theorem 14 (Moothathu [32])** Let  $(A^{\mathbb{Z}}, F)$  be a surjective CA.

- (1)  $|\mathcal{R}_F|$  is dense in  $A^{\mathbb{Z}}$ .
- (2)  $F$  is **semi-open**, i. e.,  $F(U)$  has non-empty interior for any open  $U \neq \emptyset$ .
- (3) If  $(A^{\mathbb{Z}}, F)$  is transitive, then it is weakly mixing, and hence totally transitive and sensitive.

**Conjecture 15** Every surjective CA has a dense set of  $F$ -periodic configurations.

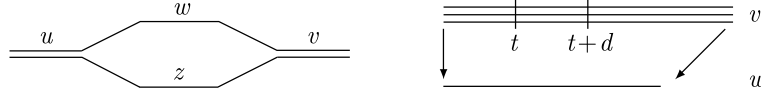
**Proposition 16 (Acerbi et al. [1])** If every mixing CA has a dense set of  $F$ -periodic configurations, then every surjective CA has a dense set of jointly periodic configurations.

**Definition 17** Let  $(A^{\mathbb{Z}}, F)$  be a CA with local rule  $f: A^{d+1} \rightarrow A$ .

- (1) The **minimum preimage number** (Fig. 3 right)  $p(F)$  is defined by

$$p(F, w) = \min_{t \leq |w|} |\{u \in A^d : \exists v \in f^{-1}(w), v_{[t, t+d]} = u\}|,$$

$$p(F) = \min\{p(F, w) : w \in A^+\}.$$



Topological Dynamics of Cellular Automata, Figure 3  
A diamond (left) and a magic word (right)

(2) A word  $w \in A^+$  is **magic**, if  $p(F, w) = \mathbf{p}(F)$ .

Recall that  $\mathcal{T}_\sigma(w)$  is the (residual) set of configurations which contain an infinite number of occurrences of  $w$  both in  $x_{(-\infty, 0)}$  and in  $x_{(0, \infty)}$ . Configurations  $x, y \in A^\mathbb{Z}$  are **d-separated**, if  $x_{[i, i+d]} \neq y_{[i, i+d]}$  for all  $i \in \mathbb{Z}$ .

**Theorem 18 (Hedlund [19], Kitchens [23])** Let  $(A^\mathbb{Z}, F)$  be a surjective CA with diameter  $d$  and minimum preimage number  $\mathbf{p}(F)$ .

- (1) If  $w \in A^+$  is a magic word, then any  $z \in \mathcal{T}_\sigma(w)$  has exactly  $\mathbf{p}(F)$  preimages. These preimages are pairwise  $d$ -separated.
- (2) Any configuration  $z \in A^\mathbb{Z}$  has at least  $\mathbf{p}(F)$  pairwise  $d$ -separated preimages.
- (3) If every  $y \in A^\mathbb{Z}$  has exactly  $\mathbf{p}(F)$  preimages, then all long enough words are magic.

**Theorem 19** Let  $(A^\mathbb{Z}, F)$  be a CA and  $\Sigma \subseteq A^\mathbb{Z}$  a sofic subshift. Then both  $F(\Sigma)$  and  $F^{-1}(\Sigma)$  are sofic subshifts. In particular, the first image subshift  $F(A^\mathbb{Z})$  is sofic.

See e. g., Formenti and Kůrka [16] for a proof. The **first image graph** of a local rule  $f: A^{d+1} \rightarrow A$  is  $\mathcal{G}(f) = (A^d, A^{d+1}, s, t, f)$ , where  $s(u) = u_{[0, d]}$  and  $t(u) = u_{[1, d]}$ . Then  $F(A^\mathbb{Z}) = \Sigma_{\mathcal{G}(f)}$ . It is algorithmically decidable whether a given CA is surjective. One decision procedure is based on the Moothathu result in Theorem 12(7). Another procedure is based on the construction of the DFA  $\mathcal{A}(\mathcal{G}(f))$  (see Sect. “Symbolic Dynamics”). A CA with local rule  $f: A^{d+1} \rightarrow A$  is surjective if the rejecting state  $\emptyset$  cannot be reached from the initial state  $A^d$  in  $\mathcal{A}(\mathcal{G}(f))$ . See Morita ► [Reversible Cellular Automata](#) for further information on bijective CA.

## Permutive and Closing Cellular Automata

**Definition 20** Let  $(A^\mathbb{Z}, F)$  be a CA, and let  $f: A^{d+1} \rightarrow A$  be the local rule for  $F$  with smallest diameter.

- (1)  $F$  is **left-permutive** if  $\forall u \in A^d, \forall b \in A, \exists! a \in A, f(au) = b$ .
- (2)  $F$  is **right-permutive** if  $\forall u \in A^d, \forall b \in A, \exists! a \in A, f(ua) = b$ .

(3)  $F$  is **permutive** if it is either left-permutive or right-permutive.

(4)  $F$  is **bipermutive** if it is both left- and right-permutive.

Permutive CA can be seen in Examples 8, 10, 11, 18.

**Definition 21** Let  $(A^\mathbb{Z}, F)$  be a CA.

- (1) Configurations  $x, y \in A^\mathbb{Z}$  are **left-asymptotic**, if  $\exists n, x_{(-\infty, n)} = y_{(-\infty, n)}$ .
- (2) Configurations  $x, y \in A^\mathbb{Z}$  are **right-asymptotic**, if  $\exists n, x_{(n, \infty)} = y_{(n, \infty)}$ .
- (3)  $(A^\mathbb{Z}, F)$  is **right-closing** if  $F(x) \neq F(y)$  for any left-asymptotic  $x \neq y \in A^\mathbb{Z}$ .
- (4)  $(A^\mathbb{Z}, F)$  is **left-closing** if  $F(x) \neq F(y)$  for any right-asymptotic  $x \neq y \in A^\mathbb{Z}$ .
- (5) A CA is **closing** if it is either left- or right-closing.

**Proposition 22**

- (1) Any right-permutive CA is right-closing.
- (2) Any right-closing CA is surjective.
- (3) A CA  $(A^\mathbb{Z}, F)$  is right-closing if there exists  $m > 0$  such that for any  $x, y \in A^\mathbb{Z}$ ,  $x_{[-m, 0]} = y_{[-m, 0]}$  and  $F(x)_{[-m, m]} = F(y)_{[-m, m]} \implies x_0 = y_0$  (see Fig. 4 left).

See e. g., Kůrka [24] for a proof. The proposition holds with obvious modification for left-permutive and left-closing CA. The multiplication CA from Example 14 is both left- and right-closing but neither left-permutive nor right-permutive. The CA from Example 15 is surjective but not closing.

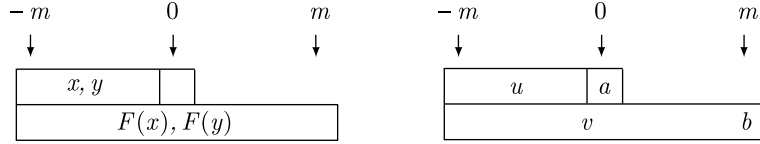
**Proposition 23** Let  $(A^\mathbb{Z}, F)$  be a right-closing CA. For all sufficiently large  $m > 0$ , if  $u \in A^m, v \in A^{2m}$ , and if  $F([u]_{-m}) \cap [v]_{-m} \neq \emptyset$ , then (Fig. 4 right)

$$\forall b \in A, \exists! a \in A, F([ua]_{-m}) \cap [vb]_{-m} \neq \emptyset.$$

See e. g., Kůrka [24] for a proof.

**Theorem 24 (Boyle and Kitchens [6])** Any closing CA  $(A^\mathbb{Z}, F)$  has a dense set of jointly periodic configurations.

**Theorem 25 (Coven, Pivato and Yassawi [14])** Let  $F$  be a left-permutive CA with memory 0.



**Topological Dynamics of Cellular Automata, Figure 4**  
Closingness

- (1) If  $\mathcal{O}(x)$  is infinite and  $x_{[0,\infty)}$  is fixed, i.e., if  $F(x)_{[0,\infty)} = x_{[0,\infty)}$ , then  $(\mathcal{O}(x), F)$  is conjugated to an adding machine.
- (2) If  $F$  is not bijective, then the set of configurations such that  $(\mathcal{O}(x), F)$  is conjugated to an adding machine is dense.

A SDS  $(X, F)$  is **open**, if  $F(U)$  is open for any open  $U \subseteq X$ . A **cross section** of a SDS  $(X, F)$  is any continuous map  $G: X \rightarrow X$  such that  $F \circ G = \text{Id}$ . If  $F$  has a cross section, it is surjective. In particular, any bijective SDS has a cross section.

**Theorem 26 (Hedlund [19])** Let  $(A^{\mathbb{Z}}, F)$  be a CA. The following conditions are equivalent.

- (1)  $(A^{\mathbb{Z}}, F)$  is open.
- (2)  $(A^{\mathbb{Z}}, F)$  is both left- and right-closing.
- (3) For any  $x \in A^{\mathbb{Z}}$ ,  $|F^{-1}(x)| = \mathbf{p}(F)$
- (4) There exist cross sections  $G_1, \dots, G_{\mathbf{p}(F)}: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$ , such that for any  $x \in A^{\mathbb{Z}}$ ,  $F^{-1}(x) = \{G_1(x), \dots, G_{\mathbf{p}(F)}(x)\}$  and  $G_i(x) \neq G_j(x)$  for  $i \neq j$ .

In general, the cross sections  $G_i$  are not CA as they need not commute with the shift. Only when  $\mathbf{p}(F) = 1$ , i.e., when  $F$  is bijective, the inverse map  $F^{-1}$  is a CA. Any CA which is open and almost equicontinuous is bijective (Kůrka [24]).

### Expansive Cellular Automata

**Definition 27** Let  $(A^{\mathbb{Z}}, F)$  be a CA.

- (1)  $F$  is **left-expansive**, if there exists  $\varepsilon > 0$  such that if  $x_{(-\infty, 0]} \neq y_{(-\infty, 0]}$ , then  $d(F^n(x), F^n(y)) \geq \varepsilon$  for some  $n \geq 0$ .
- (2)  $F$  is **right-expansive**, if there exists  $\varepsilon > 0$  such that if  $x_{[0, \infty)} \neq y_{[0, \infty)}$ , then  $d(F^n(x), F^n(y)) \geq \varepsilon$  for some  $n \geq 0$ .
- (3)  $F$  is **positively expansive**, if it is both left- and right-expansive, i.e., if there exists  $\varepsilon > 0$  such that for all  $x \neq y \in A^{\mathbb{Z}}$ ,  $d(F^n(x), F^n(y)) \geq \varepsilon$  for some  $n > 0$ .

Any left-expansive or right-expansive CA is sensitive and (by Theorem 12) surjective, because it cannot contain a di-amond. A bijective CA is **expansive**, if

$$\exists \varepsilon > 0, \forall x \neq y \in A^{\mathbb{Z}}, \exists n \in \mathbb{Z}, d(F^n(x), F^n(y)) \geq \varepsilon.$$

**Proposition 28** Let  $(A^{\mathbb{Z}}, F)$  be a CA with memory  $m$  and anticipation  $a$ .

- (1) If  $m < 0$  and if  $F$  is left-permutive, then  $F$  is left-expansive.
- (2) If  $a > 0$  and if  $F$  is right-permutive, then  $F$  is right-expansive.
- (3) If  $m < 0 < a$  and if  $F$  is bipermutive, then  $F$  is positively expansive.

See e. g., Kůrka [24] for a proof.

**Theorem 29 (Nasu [34,35])**

- (1) Any positively expansive CA is conjugated to a one-sided full shift.
- (2) A bijective expansive CA with memory 0 is conjugated to a two-sided SFT.

**Conjecture 30** Every bijective expansive CA is conjugated to a two-sided SFT.

**Definition 31** Let  $(A^{\mathbb{Z}}, F)$  be a CA. The left- and right-expansivity direction sets are defined by

$$\begin{aligned} \mathcal{X}^-(F) &= \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, F^q \sigma^p \text{ is left-expansive} \right\}, \\ \mathcal{X}^+(F) &= \left\{ \frac{p}{q} : p \in \mathbb{Z}, q \in \mathbb{N}^+, F^q \sigma^p \text{ is right-expansive} \right\}, \\ \mathcal{X}(F) &= \mathcal{X}^-(F) \cap \mathcal{X}^+(F). \end{aligned}$$

All these sets are convex and open. Moreover,  $\mathcal{X}^-(F) \cap \mathcal{X}(F) = \mathcal{X}^+(F) \cap \mathcal{X}(F) = \emptyset$  (Sablik [37]).

**Theorem 32 (Sablik [37])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with memory  $m$  and anticipation  $a$ .

- (1) If  $F$  is left-permutive, then  $\mathcal{X}^-(F) = (-\infty, -m)$ .
- (2) If  $F$  is right-permutive, then  $\mathcal{X}^+(F) = (-a, \infty)$ .
- (3) If  $\mathcal{X}^-(F) \neq \emptyset$  then there exists  $\alpha \in \mathbb{R}$  such that  $\mathcal{X}^-(F) = (-\infty, \alpha) \subseteq (-\infty, -m)$ .
- (4) If  $\mathcal{X}^+(F) \neq \emptyset$  then there exists  $\alpha \in \mathbb{R}$  such that  $\mathcal{X}^+(F) = (\alpha, \infty) \subseteq (-a, \infty)$ .
- (5) If  $\mathcal{X}(F) \neq \emptyset$  then there exists  $\alpha_0, \alpha_1 \in \mathbb{R}$  such that  $\mathcal{X}(F) = (\alpha_0, \alpha_1) \subseteq (-a, -m)$ .

**Theorem 33** Let  $(A^{\mathbb{Z}}, F)$  be a cellular automaton.

- (1)  $F$  is left-closing if  $\mathcal{X}^-(F) \neq \emptyset$ .
- (2)  $F$  is right-closing if  $\mathcal{X}^+(F) \neq \emptyset$ .
- (3) If  $\mathcal{A}(F)$  is an interval, then  $F$  is not surjective and  $\mathcal{X}^-(F) = \mathcal{X}^+(F) = \emptyset$ .

*Proof* (1) The proof is the same as the following proof of (2).

(2 $\Leftarrow$ ) If  $F$  is not right-closing and  $\varepsilon = 2^{-n}$ , then there exist distinct left-asymptotic configurations such that  $x_{(-\infty, n]} = y_{(-\infty, n]}$  and  $F(x) \neq F(y)$ . It follows that  $d(F^i(x), F^i(y)) < \varepsilon$  for all  $i \geq 0$ , so  $F$  is not right-expansive. The same argument works for any  $F^q \sigma^p$ , so  $\mathcal{X}^+(F) = \emptyset$ .

(2 $\Rightarrow$ ) Let  $F$  be right-closing, and let  $m > 0$  be the constant from Proposition 23. Assume that

$$\begin{aligned} F^n(x)_{[-m+(m+1)n, m+(m+1)n]} \\ = F^n(y)_{[-m+(m+1)n, m+(m+1)n]} \end{aligned}$$

for all  $n \geq 0$ . By Proposition 23,

$$\begin{aligned} F^{n-1}(x)_{m+(m+1)(n-1)+1} \\ = F^{n-1}(y)_{m+(m+1)(n-1)+1} . \end{aligned}$$

By induction we get  $x_{[-m, m+n]} = y_{[-m, m+n]}$ . This holds for every  $n > 0$ , so  $x_{[0, \infty)} = y_{[0, \infty)}$ . Thus,  $F \sigma^{m+1}$  is right-expansive, and therefore  $\mathcal{X}^+(F) \neq \emptyset$ .

(3) If there are blocking words for two different directions, then the CA has a diamond and therefore is not surjective by Theorem 12.  $\square$

**Corollary 34** Let  $(A^{\mathbb{Z}}, F)$  be an equicontinuous CA. There are three possibilities.

- (1) If  $F$  is surjective, then  $\mathcal{A}(F) = \{0\}$ ,  $\mathcal{X}^-(F) = (-\infty, 0)$ ,  $\mathcal{X}^+(F) = (0, \infty)$ .
- (2) If  $F$  is neither surjective nor nilpotent, then  $\mathcal{A}(F) = \{0\}$ ,  $\mathcal{X}^-(F) = \mathcal{X}^+(F) = \emptyset$ .

- (3) If  $F$  is nilpotent, then  $\mathcal{A}(F) = \mathcal{E}(F) = \mathcal{R}$ ,  $\mathcal{X}^-(F) = \mathcal{X}^+(F) = \emptyset$ .

The proof follows from Proposition 7 and Theorem 13 (see also Sablik [38]). The identity CA is in class (1). The product CA of Example 3 is in class (2). The zero CA of Example 1 is in class (3).

## Attractors

Let  $(X, F)$  be a SDS. The **limit set** of a clopen invariant set  $V \subseteq X$  is  $\Omega_F(V) := \bigcap_{n \geq 0} F^n(V)$ . A set  $Y \subseteq X$  is an **attractor**, if there exists a non-empty clopen invariant set  $V$  such that  $Y = \Omega_F(V)$ . We say that  $Y$  is a **finite time attractor**, if  $Y = \Omega_F(V) = F^n(V)$  for some  $n > 0$  (and a clopen invariant set  $V$ ). There exists always the largest attractor  $\Omega_F := \Omega_F(X)$ . Finite time maximal attractors are also called **stable limit sets** in the literature. The number of attractors is at most countable. The union of two attractors is an attractor. If the intersection of two attractors is non-empty, it contains an attractor. The **basin** of an attractor  $Y \subseteq X$  is the set  $\mathcal{B}(Y) = \{x \in X; \lim_{n \rightarrow \infty} d(F^n(x), Y) = 0\}$ . An attractor  $Y \subseteq X$  is a **minimal attractor**, if no proper subset of  $Y$  is an attractor. An attractor is a minimal attractor if it is chain-transitive. A periodic point  $x \in X$  is **attracting** if its orbit  $\mathcal{O}(x)$  is an attractor. Any attracting periodic point is equicontinuous. A **quasi-attractor** is a non-empty set which is an intersection of a countable number of attractors.

**Theorem 35 (Hurley [22])**

- (1) If a CA has two disjoint attractors, then any attractor contains two disjoint attractors and an uncountably infinite number of quasi-attractors.
- (2) If a CA has a minimal attractor, then it is a subshift, it is contained in any other attractor, and its basin of attraction is a dense open set.
- (3) If  $x \in A^{\mathbb{Z}}$  is an attracting  $F$ -periodic configuration, then  $\sigma(x) = x$  and  $F(x) = x$ .

**Corollary 36** For any CA, exactly one of the following statements holds.

- (1) There exist two disjoint attractors and a continuum of quasi-attractors.
- (2) There exists a unique quasi-attractor. It is a subshift and it is contained in any attractor.
- (3) There exists a unique minimal attractor contained in any other attractor.

Both equicontinuity and surjectivity yield strong constraints on attractors.



**Theorem 37 (Kůrka [24])**

- (1) A surjective CA has either a unique attractor or a pair of disjoint attractors.
- (2) An equicontinuous CA has either two disjoint attractors or a unique attractor which is an attracting fixed configuration.
- (3) If a CA has an attracting fixed configuration which is a unique attractor, then it is equicontinuous.

We consider now subshift attractors of CA, i.e., those attractors which are subshifts. Let  $(A^{\mathbb{Z}}, F)$  be a CA. A clopen  $F$ -invariant set  $U \subseteq A^{\mathbb{Z}}$  is **spreading**, if there exists  $k > 0$  such that  $F^k(U) \subseteq \sigma^{-1}(U) \cap \sigma(U)$ . If  $U$  is a clopen invariant set, then  $\Omega_F(U)$  is a subshift iff  $U$  is spreading (Kůrka [25]). Recall that a language is **recursively enumerable**, if it is a domain (or a range) of a recursive function (see e.g., Hopcroft and Ullmann [20]).

**Theorem 38 (Formenti and Kůrka [17])** Let  $\Sigma \subseteq A^{\mathbb{Z}}$  be a subshift attractor of a CA  $(A^{\mathbb{Z}}, F)$ .

- (1)  $A^* \setminus \mathcal{L}(\Sigma)$  is a recursively enumerable language.
- (2)  $\Sigma$  contains a jointly periodic configuration.
- (3)  $(\Sigma, \sigma)$  is chain-mixing.

**Theorem 39 (Formenti and Kůrka [17])**

- (1) The only subshift attractor of a surjective CA is the full space.
- (2) A subshift of finite type is an attractor of a CA if it is mixing.
- (3) Given a CA  $(A^{\mathbb{Z}}, F)$ , the intersection of all subshift attractors of all  $F^q \sigma^p$ , where  $q \in \mathbb{N}^+$  and  $p \in \mathbb{Z}$ , is a non-empty  $F$ -invariant subshift called the **small quasi-attractor**  $\mathcal{Q}_F$ .  $(\mathcal{Q}_F, \sigma)$  is chain-mixing and  $F: \mathcal{Q}_F \rightarrow \mathcal{Q}_F$  is surjective.

The system of all subshift attractors of a given CA forms a lattice with join  $\Sigma_0 \cup \Sigma_1$  and meet  $\Sigma_0 \wedge \Sigma_1 := \Omega_F(\Sigma_0 \cap \Sigma_1)$ . There exist CA with infinite number of subshift attractors (Kůrka [26]).

**Proposition 40 (Di Lena [28])** The basin of a subshift attractor is a dense open set.

By a theorem of Hurd [21], if  $\Omega_F$  is SFT, then it is stable, i.e.,  $\Omega_F = F^n(A^{\mathbb{Z}})$  for some  $n > 0$ . We generalize this theorem to subshift attractors.

**Theorem 41** Let  $U$  be a spreading set for a CA  $(A^{\mathbb{Z}}, F)$ .

- (1) There exists a spreading set  $W \subseteq U$  such that  $\Omega_F(W) = \Omega_F(U)$  and  $\widetilde{\Omega}_\sigma(W) := \bigcap_{i \in \mathbb{Z}} \sigma^i(W)$  is a mixing subshift of finite type.

- (2) If  $\Omega_F(W)$  is a SFT, then  $\Omega_F(W) = F^n(\widetilde{\Omega}_\sigma(W))$  for some  $n \geq 0$ .

*Proof*

- (1) See Formenti and Kůrka [17].
- (2) Let  $D$  be a finite set of forbidden words for  $\Omega_F(W)$ . For each  $u \in D$  there exists  $n_u > 0$  such that  $u \notin \mathcal{L}(F^{n_u}(\widetilde{\Omega}_\sigma))$ . Take  $n := \max\{n_u : u \in D\}$ .  $\square$

By Theorem 5, every equicontinuous CA has a finite time maximal attractor.

**Definition 42** Let  $\Sigma \subseteq A^{\mathbb{Z}}$  be a mixing sofic subshift, let  $\mathcal{G} = (V, E, s, t, l)$  be its minimal right-resolving presentation with factor map  $\ell: (\Sigma_{|\mathcal{G}|}, \sigma) \rightarrow (\Sigma, \sigma)$ .

- (1) A homogenous configuration  $a^\infty \in \Sigma$  is **receptive**, if there exist intrinsically synchronizing words  $u, v \in \mathcal{L}(\Sigma)$  and  $n \in \mathbb{N}$  such that  $ua^m v \in \mathcal{L}(\Sigma)$  for all  $m > n$ .
- (2)  $\Sigma$  is **almost of finite type (AFT)**, if  $\ell: \Sigma_{|\mathcal{G}|} \rightarrow \Sigma$  is one-to-one on a dense open set of  $\Sigma_{|\mathcal{G}|}$ .
- (3)  $\Sigma$  is **near-Markov**, if  $\{x \in \Sigma : |\ell^{-1}(x)| > 1\}$  is a finite set of  $\sigma$ -periodic configurations.

- (3) is equivalent to the condition that  $\ell$  is left-closing, i.e., that  $\ell(u) \neq \ell(v)$  for distinct right-asymptotic paths  $u, v \in E^{\mathbb{Z}}$ . Each near-Markov subshift is AFT.

**Theorem 43 (Maass [30])** Let  $\Sigma \subseteq A^{\mathbb{Z}}$  be a mixing sofic subshift with a receptive configuration  $a^\infty \in \Sigma$ .

- (1) If  $\Sigma$  is either SFT or AFT, then there exists a CA  $(A^{\mathbb{Z}}, F)$  such that  $\Sigma = \Omega_F = F(A^{\mathbb{Z}})$ .
- (2) A near-Markov subshift cannot be an infinite time maximal attractor of a CA.

On the other hand, a near-Markov subshift can be a finite time maximal attractor (see Example 21). The language  $\mathcal{L}(\Omega_F)$  can have arbitrary complexity (see Culik et al. [15]). A CA with non-sofic mixing maximal attractor has been constructed in Formenti and Kůrka [17].

**Definition 44** Let  $f: A^{d+1} \rightarrow A$  be a local rule of a cellular automaton. We say that a subshift  $\Sigma \subseteq A^{\mathbb{Z}}$  has **decreasing preimages**, if there exists  $m > 0$  such that for each  $u \in A^* \setminus \mathcal{L}(\Sigma)$ , each  $v \in f^{-m}(u)$  contains as a subword a word  $w \in A^* \setminus \mathcal{L}(\Sigma)$  such that  $|w| < |u|$ .

**Proposition 45 (Formenti and Kůrka [16])** If  $(A^{\mathbb{Z}}, F)$  is a CA and  $\Sigma \subseteq A^{\mathbb{Z}}$  has decreasing preimages, then  $\Omega_F \subseteq \Sigma$ .

For more information about attractor-like objects in CA see Sect. “Invariance of Maxentropy Measures” of Pivato  
 ► [Ergodic Theory of Cellular Automata](#).

### Subshifts and Entropy

**Definition 46** Let  $F: A^{\mathbb{Z}} \rightarrow A^{\mathbb{Z}}$  be a cellular automaton.

- (1) Denote by  $\mathcal{S}_{(p,q)}(F) := \{x \in A^{\mathbb{Z}} : F^q \sigma^p(x) = x\}$  the set of all weakly periodic configurations of  $F$  with period  $(p, q) \in \mathbb{Z} \times \mathbb{N}^+$ .
- (2) A **signal subshift** is any non-empty  $\mathcal{S}_{(p,q)}(F)$ .
- (3) The **speed subshift** of  $F$  with speed  $\alpha = \frac{p}{q} \in \mathbb{Q}$  is
 
$$\mathcal{S}_{\alpha}(F) = \bigcup_{n>0} \overline{\mathcal{S}_{(np,nq)}(F)}.$$

Note that both  $\mathcal{S}_{(p,q)}(F)$  and  $\mathcal{S}_{\alpha}(F)$  are closed and  $\sigma$ -invariant. However,  $\mathcal{S}_{(p,q)}(F)$  can be empty, so it need not be a subshift.

**Theorem 47** Let  $(A^{\mathbb{Z}}, F)$  be a cellular automaton with memory  $m$  and anticipation  $a$ , so  $F(x)_i = f(x_{[i+m, i+a]})$ .

- (1) If  $\mathcal{S}_{(p,q)}(F)$  is non-empty, then it is a subshift of finite type.
- (2) If  $\mathcal{S}_{(p,q)}(F)$  is infinite, then  $-a \leq p/q \leq -m$ .
- (3) If  $p_0/q_0 < p_1/q_1$ , then  $\mathcal{S}_{(p_0,q_0)}(F) \cap \mathcal{S}_{(p_1,q_1)}(F) \subseteq \{x \in A^{\mathbb{Z}} : \sigma^p(x) = x\}$ , where  $p = q(p_1/q_1 - p_0/q_0)$  and  $q = \text{lcm}(q_0, q_1)$  (the least common multiple).
- (4)  $\mathcal{S}_{(p,q)}(F) \subseteq \mathcal{S}_{\frac{p}{q}}(F) \subseteq \Omega_F$  and  $\mathcal{S}_{\frac{p}{q}}(F) \neq \emptyset$ .
- (5) If  $\mathcal{X}(F) \neq \emptyset$  or if  $(A^{\mathbb{Z}}, F)$  is nilpotent, then  $F$  has no infinite signal subshifts.

*Proof* (1), (2) and (3) have been proved in Formenti and Kůrka [16].

(4) Since  $F$  is bijective on each signal subshift, we get  $\mathcal{S}_{(p,q)}(F) \subseteq \Omega_F$ , and therefore  $\mathcal{S}_{p/q}(F) \subseteq \Omega_F$ . Since every  $F^q \sigma^p$  has a periodic point, we get  $\mathcal{S}_{p/q}(F) \neq \emptyset$ .

(5) It has been proved in Kůrka [25], that a positively expansive CA has no signal subshifts. This property is preserved when we compose  $F$  with a power of the shift map. If  $(A^{\mathbb{Z}}, F)$  is nilpotent, then each  $\mathcal{S}_{(p,q)}(F)$  contains at most one element.  $\square$

The Identity CA has a unique infinite signal subshift  $\mathcal{S}_{(0,1)}(\text{Id}) = A^{\mathbb{Z}}$ . The CA of Example 17 has an infinite number of infinite signal subshifts of the same speed. A CA with infinitely many infinite signal subshifts with infinitely many speeds has been constructed in Kůrka [25]. In some cases, the maximal attractor can be constructed from signal subshifts (see Theorem 49).

**Definition 48** Given an integer  $c \geq 0$ , the  **$c$ -join**  $\Sigma_0 \dot{\vee} \Sigma_1$  of subshifts  $\Sigma_0, \Sigma_1 \subseteq A^{\mathbb{Z}}$  consists of all configurations  $x \in A^{\mathbb{Z}}$  such that either  $x \in \Sigma_0 \cup \Sigma_1$ , or there exist integers  $b, a$  such that  $b - a \geq c$ ,  $x_{(-\infty, b)} \in \mathcal{L}(\Sigma_0)$ , and  $x_{[a, \infty)} \in \mathcal{L}(\Sigma_1)$ .

The operation of join is associative, and the  $c$ -join of sofic subshifts is sofic.

**Theorem 49 (Formenti, Kůrka [16])** Let  $(A^{\mathbb{Z}}, F)$  be a CA and let  $\mathcal{S}_{(p_1,q_1)}(F), \dots, \mathcal{S}_{(p_n,q_n)}(F)$  be signal subshifts with decreasing speeds, i. e.,  $p_i/q_i > p_j/q_j$  for  $i < j$ . Set  $q := \text{lcm}\{q_1, \dots, q_n\}$  (the least common multiple). There exists  $c \geq 0$  such that for  $\Sigma := \mathcal{S}_{(p_1,q_1)}(F) \dot{\vee} \dots \dot{\vee} \mathcal{S}_{(p_n,q_n)}(F)$  we have  $\Sigma \subseteq F^q(\Sigma)$  and therefore  $\Sigma \subseteq \Omega_F$ . If moreover  $F^{nq}(\Sigma)$  has decreasing preimages for some  $n \geq 0$ , then  $F^{nq}(\Sigma) = \Omega_F$ .

**Definition 50** Let  $(A^{\mathbb{Z}}, F)$  be a CA.

- (1) The  **$k$ -column homomorphism**  $\varphi_k: (A^{\mathbb{Z}}, F) \rightarrow ((A^k)^{\mathbb{N}}, \sigma)$  is defined by  $\varphi_k(x)_i = F^i(x)_{[0,k]}$ .
- (2) The  **$k$ th column subshift** is  $\Sigma_k(F) = \varphi_k(A^{\mathbb{Z}}) \subseteq (A^k)^{\mathbb{N}}$ .
- (3) If  $\psi: (A^{\mathbb{Z}}, F) \rightarrow (\Sigma, \sigma)$  is a factor map, where  $\Sigma$  is a one-sided subshift, we say that  $\Sigma$  is a **factor subshift** of  $(A^{\mathbb{Z}}, F)$ .

Thus, each  $(\Sigma_k(F), \sigma)$  is a factor of  $(A^{\mathbb{Z}}, F)$  and each factor subshift is a factor of some  $\Sigma_k(F)$ . Any positively expansive CA  $(A^{\mathbb{Z}}, F)$  with radius  $r > 0$  is conjugated to  $(\Sigma_{2r+1}(F), \sigma)$ . This is an SFT which by Theorem 29 is conjugated to a full shift.

**Proposition 51 (Shereshevski and Afraimovich [39])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with negative memory and positive anticipation  $m < 0 < a$ . Then  $(A^{\mathbb{Z}}, F)$  is bipermutive if it is positively expansive and  $\Sigma_{a-m+1}(F) = (A^{a-m+1})^{\mathbb{N}}$  is the full shift.

**Theorem 52 (Blanchard and Maass [4], Di Lena [28])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with radius  $r$  and memory  $m$ .

- (1) If  $m \geq 0$  and  $\Sigma_r(F)$  is sofic, then any factor subshift of  $(A^{\mathbb{Z}}, F)$  is sofic.
- (2) If  $\Sigma_{2r+1}(F)$  is sofic, then any factor subshift of  $(A^{\mathbb{Z}}, F)$  is sofic.

If  $(x_i)_{i \geq 0}$  is a  $2^{-m}$ -chain in a CA  $(A^{\mathbb{Z}}, F)$ , then for all  $i$ ,  $F(x_i)_{[-m,m]} = (x_{i+1})_{[-m,m]}$ , so  $u_i = (x_i)_{[-m,m]}$  satisfy  $F([u_i]_{-m}) \cap [u_{i+1}]_{-m} \neq \emptyset$ . Conversely, if a sequence  $(u_i \in A^{2m+1})_{i \geq 0}$  satisfies this property and  $x_i \in [u_i]_{-m}$ , then  $(x_i)_{i \geq 0}$  is a  $2^{-m}$ -chain.

**Theorem 53 (Kůrka [27])** Let  $(A^{\mathbb{Z}}, F)$  be a CA.

- (1) If  $\Sigma_k(F)$  is an SFT for any  $k > 0$ , then  $(A^{\mathbb{Z}}, F)$  has the shadowing property.
- (2) If  $(A^{\mathbb{Z}}, F)$  has the shadowing property, then any factor subshift is sofic.

Any factor subshift of the Coven CA from Example 18 is sofic, but the CA does not have the shadowing property (Blanchard and Maass [4]). A CA with shadowing property whose factor subshift is not SFT has been constructed in Kůrka [24].

**Proposition 54** Let  $(A^{\mathbb{Z}}, F)$  be a CA.

- (1)  $\mathbf{h}(A^{\mathbb{Z}}, F) = \lim_{k \rightarrow \infty} \mathbf{h}(\Sigma_k(F), \sigma)$ .
- (2) If  $F$  has radius  $r$ , then

$$\begin{aligned} \mathbf{h}(A^{\mathbb{Z}}, F) &\leq 2 \cdot \mathbf{h}(\Sigma_r(F), \sigma) \\ &\leq 2r \cdot \mathbf{h}(\Sigma_1(F), \sigma) \leq 2r \cdot \ln |A|. \end{aligned}$$

- (3) If  $0 \leq m \leq a$ , then  $\mathbf{h}(A^{\mathbb{Z}}, F) = \mathbf{h}(\Sigma_a(F), \sigma)$ .

See e. g., Kůrka [24] for a proof.

**Conjecture 55** If  $(A^{\mathbb{Z}}, F)$  is a CA with radius  $r$ , then  $\mathbf{h}(A^{\mathbb{Z}}, F) = \mathbf{h}(\Sigma_{2r+1}(F), \sigma)$ .

**Conjecture 56 (Moothathu [32])** Any transitive CA has positive topological entropy.

**Definition 57** The **directional entropy** of a CA  $(A^{\mathbb{Z}}, F)$  along a rational direction  $\alpha = p/q$  is  $\mathbf{h}_\alpha(A^{\mathbb{Z}}, F) := \mathbf{h}(A^{\mathbb{Z}}, F^q \sigma^p)/q$ .

The definition is based on the equality  $\mathbf{h}(X, F^n) = n \cdot \mathbf{h}(X, F)$  which holds for every SDS. Directional entropies along irrational directions have been introduced in Milnor [31].

**Proposition 58 (Courbage and Kamiński [11], Sablik [37])** Let  $(A^{\mathbb{Z}}, F)$  be a CA with memory  $m$  and anticipation  $a$ .

- (1) If  $\alpha \in \mathcal{E}(F)$ , then  $\mathbf{h}_\alpha(F) = 0$ .
- (2) If  $\alpha \in \mathcal{X}^-(F) \cup \mathcal{X}^+(F)$ , then  $\mathbf{h}_\alpha(F) > 0$ .
- (3)  $\mathbf{h}_\alpha(F) \leq (\max(a + \alpha, 0) - \min(m + \alpha, 0)) \cdot \ln |A|$ .
- (4) If  $F$  is bipermutive, then  $\mathbf{h}_\alpha(F) = (\max\{a + \alpha, 0\} - \min\{m + \alpha, 0\}) \cdot \ln |A|$ .
- (5) If  $F$  is left-permutive, and  $\alpha < -a$ , then  $\mathbf{h}_\alpha(F) = |m + \alpha| \cdot \ln |A|$ .
- (6) If  $F$  is right-permutive, and  $\alpha > -m$ , then  $\mathbf{h}_\alpha(F) = (a + \alpha) \cdot \ln |A|$ .

The directional entropy is not necessarily continuous (see Smillie [40]).

**Theorem 59 (Boyle and Lind [8])** The function  $\alpha \mapsto \mathbf{h}_\alpha(A^{\mathbb{Z}}, F)$  is convex and continuous on  $\mathcal{X}^-(F) \cup \mathcal{X}^+(F)$ .

## Examples

Cellular automata with binary alphabet  $\mathbf{2} = \{0, 1\}$  and radius  $r = 1$  are called **elementary** (Wolfram [42]). Their local rules are coded by numbers between 0 and 255 by

$$\begin{aligned} f(000) + 2 \cdot f(001) + 4 \cdot f(010) + \dots \\ + 32 \cdot f(101) + 64 \cdot f(110) + 128 \cdot f(111). \end{aligned}$$

**Example 1 (The zero rule ECA0)**  $F(x) = 0^\infty$ .

The zero CA is an equicontinuous nilpotent CA. Its equicontinuity directions are  $\mathcal{E}(F) = \mathcal{U}(F) = (-\infty, \infty)$ .

**Example 2 (The identity rule ECA204)**  $\text{Id}(x) = x$ .

The identity is an equicontinuous surjective CA which is not transitive. Every clopen set is an attractor and every configuration is a quasi-attractor. The equicontinuity and expansivity directions are  $\mathcal{U}(\text{Id}) = \mathcal{E}(\text{Id}) = \{0\}$ ,  $\mathcal{X}^-(\text{Id}) = (-\infty, 0)$ ,  $\mathcal{X}^+(\text{Id}) = (0, \infty)$ . The directional entropy is  $\mathbf{h}_\alpha(\text{Id}) = |\alpha|$ .

**Example 3 (An equicontinuous rule ECA12)**  $F(x)_i = (1 - x_{i-1})x_i$ .

$$\begin{array}{cccccc} 000 : 0, & 001 : 0, & 010 : 1, & 011 : 1, & 100 : 0, \\ & & & & & 101 : 0, & 110 : 0, & 111 : 0. \end{array}$$

The ECA12 is equicontinuous: the preperiod and period are  $m = p = 1$ . The automaton has finite time maximal attractor  $\Omega_F = F(A^{\mathbb{Z}}) = \Sigma_{\{11\}} = \mathcal{S}_{(0,1)}(F)$  which is called the **golden mean subshift**.

**Example 4 (A product rule ECA128)**  $F(x)_i = x_{i-1}x_ix_{i+1}$ .

The ECA128 is almost equicontinuous and 0 is a 1-blocking word. The first column subshift is  $\Sigma_1(F) = \mathcal{S}_{\{01\}}$ . Each column subshift  $\Sigma_k(F)$  is an SFT with zero entropy, so  $F$  has the shadowing property and zero entropy. The  $n$ th image

$$F^n(2^{\mathbb{Z}}) = \{x \in 2^{\mathbb{Z}} : \forall m \in [1, 2n], 10^m 1 \not\sqsubseteq x\}$$

is a SFT. The first image graph can be seen in Fig. 10 left. The maximal attractor  $\Omega_F = \mathcal{S}_{\{10^n 1 : n > 0\}}$  is a sofic subshift and has decreasing preimages. The only other attractor is the minimal attractor  $\{0^\infty\} = \Omega_F([0]_0)$ , which is also the minimal quasi-attractor. The equicontinuity directions are



Topological Dynamics of Cellular Automata, Figure 5  
ECA12



Topological Dynamics of Cellular Automata, Figure 6  
Signal subshifts of ECA128

$\mathfrak{C}(F) = \emptyset$  and  $\mathfrak{A}(F) = [-1, 1]$ . For Lyapunov exponents we have  $\lambda_F^-(0^\infty) = \lambda_F^+(0^\infty) = 0$  and  $\lambda_F^-(1^\infty) = \lambda_F^+(1^\infty) = 1$ . The only infinite signal subshifts are non-transitive subshifts  $\mathcal{S}_{(1,1)}(F) = S_{\{10\}}$  and  $\mathcal{S}_{(-1,1)}(F) = S_{\{01\}}$ . The maximal attractor can be constructed using the join construction  $\Omega_F = F(\mathcal{S}_{(1,1)}(F) \checkmark \mathcal{S}_{(-1,1)}(F))$  (Fig. 6).

*Example 5 (A product rule ECA136)*  $F(x)_i = x_i x_{i+1}$ .

The ECA136 is almost equicontinuous since 0 is a 1-blocking word. As in Example 4, we have  $\Omega_F = S_{\{10^k 1; k > 0\}}$ . For any  $m \in \mathbb{Z}$ ,  $[0]_m$  is a clopen invariant set, which is spreading to the left but not to the right. Thus,  $Y_m = \Omega_F([0]_m) = \{x \in \Omega_F : \forall i \leq m, x_i = 0\}$  is an attractor but not a subshift. We have  $Y_{m+1} \subset Y_m$  and  $\bigcap_{m \geq 0} Y_m = \{0^\infty\}$  is the unique minimal quasi-attractor. Since  $F^2 \sigma^{-1}(x)_i = x_{i-1} x_i x_{i+1}$  is the ECA128 which has a minimal subshift attractor  $\{0^\infty\}$ ,  $F$  has the small quasi-attractor  $\mathcal{Q}_F = \{0^\infty\}$ . The almost equicontinuity directions are  $\mathfrak{A}(F) = [-1, 0]$ .

*Example 6 (A unique attractor)*  $(2^\mathbb{Z}, F)$  where  $F(x)_i = x_{i+1} x_{i+2}$ .

The system is sensitive and has a unique attractor  $\Omega_F = S_{\{10^k 1; k > 0\}}$  which is not  $F$ -transitive. If  $x \in [10]_0 \cap \Omega_F(2^\mathbb{Z})$ , then  $x_{[0, \infty)} = 10^\infty$ , so for any  $n > 0$ ,  $F^n(x) \notin$

$[11]_0$ . However,  $(A^\mathbb{Z}, F)$  is chain-transitive, so it does not have the shadowing property. The small quasi-attractor is  $\mathcal{Q}_F = \{0^\infty\}$ . The topological entropy is zero. The factor subshift  $\Sigma_1(F) = \{x \in 2^\mathbb{N} : \forall n \geq 0, (x_{[n, n+1]} = 10 \implies x_{[n, 2n+1]} = 10^{n+1})\}$  is not sofic (Gilman [18]).

*Example 7 (The majority rule ECA232)*  $F(x)_i = \lfloor (x_{i-1} + x_i + x_{i+1})/2 \rfloor$ .

000 : 0,    001 : 0,    010 : 0,    011 : 1,    100 : 0,  
101 : 1,    110 : 1,    111 : 1.

The majority rule has 2-blocking words 00 and 11, so it is almost equicontinuous. More generally, let  $E = \{u \in 2^* : |u| \geq 2, u_0 = u_1, u_{|u|-2} = u_{|u|-1}, 010 \not\sqsubseteq u, 101 \not\sqsubseteq u\}$ . Then for any  $u \in E$  and for any  $i \in \mathbb{Z}$ ,  $[u]_i$  is a clopen invariant set, so its limit set  $\Omega_F([u]_i)$  is an attractor. These attractors are not subshifts. There exists a subshift attractor given by the spreading set  $U := 2^\mathbb{Z} \setminus ([010]_0 \cup [101]_0)$ . We have  $\Omega_F(U) = \mathcal{S}_{(0,1)}(F) = S_{\{010, 101\}}$ . There are two more infinite signal subshifts  $\mathcal{S}_{(-1,1)}(F) = S_{\{001, 110\}}$  and  $\mathcal{S}_{(1,1)}(F) = S_{\{011, 100\}}$ . The maximal attractor is  $\Omega_F = \mathcal{S}_{(1,0)}(F) \cup (\mathcal{S}_{(1,1)}(F) \checkmark \mathcal{S}_{(-1,1)}(F)) = S_{\{010^k 1, 10^k 10, 01^k 01, 101^k 0; k > 1\}}$ . All column subshifts are SFT, for example  $\Sigma_1(F) = S_{\{001, 110\}}$  and the entropy is zero. The equicontinuity directions are  $\mathfrak{C}(F) = \emptyset$ ,  $\mathfrak{A}(F) = \{0\}$ .



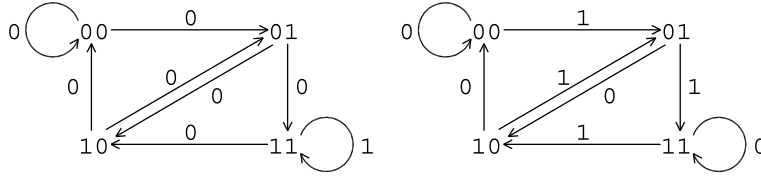
Topological Dynamics of Cellular Automata, Figure 7  
ECA136



Topological Dynamics of Cellular Automata, Figure 8  
A unique attractor  $F(x)_i = x_{i+1}x_{i+2}$



Topological Dynamics of Cellular Automata, Figure 9  
The majority rule ECA232



Topological Dynamics of Cellular Automata, Figure 10  
First image subshift of ECA128 (left) and ECA106 (right)

*Example 8 (A right-permutive rule ECA106)*  $F(x)_i = (x_{i-1}x_i + x_{i+1}) \bmod 2$ .

$$\begin{array}{ccccccc} 000 : 0, & 001 : 1, & 010 : 0, & 011 : 1, & 100 : 0, \\ & & 101 : 1, & 110 : 1, & 111 : 0. \end{array}$$

The ECA106 is transitive (see Kůrka [24]). The first image graph is in Fig. 10 right. The minimum preimage number is  $\mathbf{p}_F = 1$  and the word  $u = 0101$  is magic. Its preimages are  $f^{-1}(0101) = \{010101, 100101, 000101, 111001\}$  and for every  $v \in f^{-1}(u)$  we have  $v_{[4,5]} = 01$ . This can be seen in Fig. 11 bottom left, where all paths in the first image graph with label 0101 are displayed. Accordingly,  $(01)^\infty$  has a unique preimage  $F^{-1}((01)^\infty) = \{(10)^\infty\}$ . On the other hand  $0^\infty$  has two preimages  $F^{-1}(0^\infty) = \{0^\infty, 1^\infty\}$  (Fig. 11 bottom right) and  $1^\infty$  has three preimages  $F^{-1}(1^\infty) = \{(011)^\infty, (110)^\infty, (101)^\infty\}$ . We have  $\mathcal{X}^-(F) = \emptyset$  and  $\mathcal{X}^+(F) = (-1, \infty)$  and there are no equicontinuity directions. For every  $x$  we have  $\lambda_F^-(x) = 1$ . On the other hand the right Lyapunov exponents are not constant. For example,  $\lambda_F^+(0^\infty) = 0$  while  $\lambda_F^+((01)^\infty) = 1$ . The only infinite signal subshift is the golden mean subshift  $\mathcal{S}_{(-1,1)}(F) = \mathcal{S}_{\{11\}}$ .

*Example 9 (The shift rule ECA170)*  $\sigma(x)_i = x_{i+1}$ .

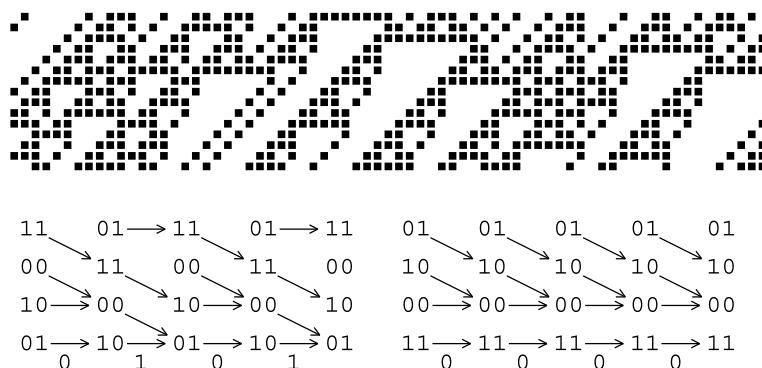
The shift rule is bijective, expansive and transitive. It has a dense set of periodic configurations, so it is chaotic. Its only signal subshift is  $\mathcal{S}_{(-1,1)}(\sigma) = 2^\mathbb{Z}$ . The equicontinuity and expansivity directions are  $\mathcal{E}(\sigma) = \mathcal{A}(\sigma) = \{-1\}$ ,  $\mathcal{X}^-(\sigma) = (-\infty, -1)$ ,  $\mathcal{X}^+(\sigma) = (-1, \infty)$ . For any  $x \in 2^\mathbb{Z}$  we have  $\lambda_\sigma^-(x) = 0$ ,  $\lambda_\sigma^+(x) = 1$ .

*Example 10 (A bipermutive rule ECA102)*  $F(x)_i = (x_i + x_{i+1}) \bmod 2$ .

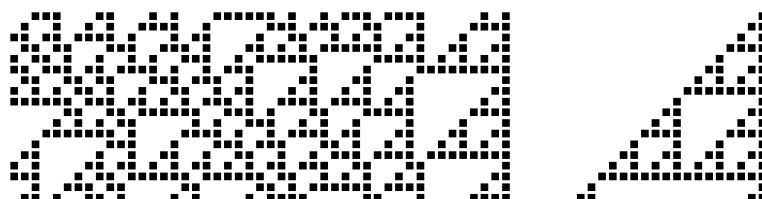
$$\begin{array}{ccccccc} 000 : 0, & 001 : 1, & 010 : 1, & 011 : 0, & 100 : 0, \\ & & 101 : 1, & 110 : 1, & 111 : 0. \end{array}$$

The ECA102 is bipermutive with memory 0, so it is open but not positively expansive. The expansivity directions are  $\mathcal{X}^-(F) = (-\infty, 0)$ ,  $\mathcal{X}^+(F) = (-1, \infty)$ ,  $\mathcal{X}(F) = (-1, 0)$ . If  $x \in Y := W_0^+(0^\infty.10^\infty)$ , i.e., if  $x_0 = 1$  and  $x_i = 0$  for all  $i > 0$ , then  $(\mathcal{O}(x), F) = (Y, F)$  is conjugated to the adding machine with periodic structure  $\mathbf{n} = (2, 2, 2, \dots)$ . If  $-2^n < i \leq -2^{n-1}$ , then  $(F^m(x)_i)_{m \geq 0}$  is periodic with period  $2^n$ . There are no signal subshifts. The minimum preimage number is  $\mathbf{p}_F = 2$  and the two cross sections  $G_0, G_1$  are uniquely determined by





## Topological Dynamics of Cellular Automata, Figure 11



Topological Dynamics of Cellular Automata, Figure 12  
ECA102

the conditions  $G_0(x)_0 = 0$ ,  $G_1(x)_0 = 1$ . The entropy is  $\mathbf{h}(A^{\mathbb{Z}}, F) = \ln 2$ .

*Example 11 (The sum rule ECA90)*  $F(x)_i = (x_{i-1} + x_{i+1}) \bmod 2$ .

$$000 : 0, \quad 001 : 1, \quad 010 : 0, \quad 011 : 1, \quad 100 : 1, \\ 101 : 0, \quad 110 : 1, \quad 111 : 0.$$

The sum rule is bipermutive with negative memory and positive anticipation. Thus it is open, positively expansive and mixing. It is conjugated to the full shift on four symbols  $\Sigma_2(F) = \{00, 01, 10, 11\}^{\mathbb{N}}$ . It has four cross-sections  $G_0, G_1, G_2, G_3$  which are uniquely determined by the conditions  $G_0(x)_{[0,1]} = 00$ ,  $G_1(x)_{[0,1]} = 01$ ,  $G_2(x)_{[0,1]} = 10$ , and  $G_3(x)_{[0,1]} = 11$ . For every  $x \in 2^{\mathbb{Z}}$  we have  $\lambda_F^-(x) = \lambda_F^+(x) = 1$ . The system has no almost equicontinuous directions and  $\mathcal{X}^-(F) = (-\infty, 1)$ ,  $\mathcal{X}^+(F) = (-1, \infty)$ . The directional entropy is continuous and piecewise linear (Fig. 13).

*Example 12 (The traffic rule ECA184)*  $F(x)_i = 1$  if  $x_{[i-1,i]} = 10$  or  $x_{[i,i+1]} = 11$ .

$$\begin{array}{ccccccccc} 000:0 & 001:0 & 010:0 & 011:1 & 100:1 & & & & \\ & & & & & 101:1 & 110:0 & 111:0. & \end{array}$$

The ECA184 has three infinite signal subshifts

$$\begin{aligned} \mathbb{S}_{(1,1)}(F) &= S_{\{11\}} \cup \{1^\infty\}, & \mathbb{S}_{(0,1)}(F) &= S_{\{10\}}, \\ \mathbb{S}_{(-1,1)}(F) &= S_{\{00\}} \cup \{0^\infty\} \end{aligned}$$

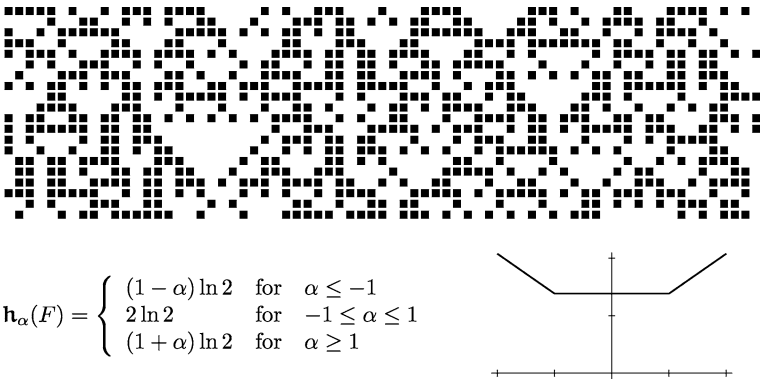
and a unique  $F$ -transitive attractor  $\Omega_F = S_{(1,1)}(F) \hat{\vee} S_{(-1,1)}(F) = S_{\{1(10)^n 0; n > 0\}}$  which is sofic. The system has neither almost equicontinuous nor expansive directions. The directional entropy is continuous, but neither piecewise linear nor convex (Smillie [40]).

*Example 13 (ECA62)*  $F(x)_i = x_{i-1}(x_i + 1) + x_i x_{i+1}$ .

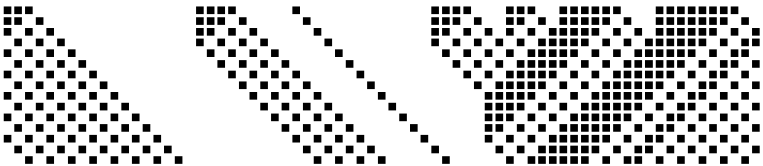
$$000:0, \quad 001:1, \quad 010:1, \quad 011:1, \quad 100:1, \\ 101:1, \quad 110:0, \quad 111:0.$$

There exists a spreading set  $U = \mathbb{A}^{\mathbb{Z}} \setminus ([0^6]_2 \cup [1^7]_1) \cup \bigcup_{v \in f^{-1}(17)} [v]_0$  and  $\Omega_F(U)$  is a subshift attractor which contains  $\sigma$ -transitive infinite signal subshifts  $\mathcal{S}_{(1,2)}(F)$  and  $\mathcal{S}_{(0,3)}(F)$  as well as their join. It follows  $\mathcal{Q}_F = \Omega_F(U) = F^2(\mathcal{S}_{(1,2)}(F) \overset{\vee}{\cup} \mathcal{S}_{(0,3)}(F))$ . The only other infinite signal subshifts are  $\mathcal{S}_{(4,4)}(F)$  and  $\mathcal{S}_{(-1,1)}(F)$  and

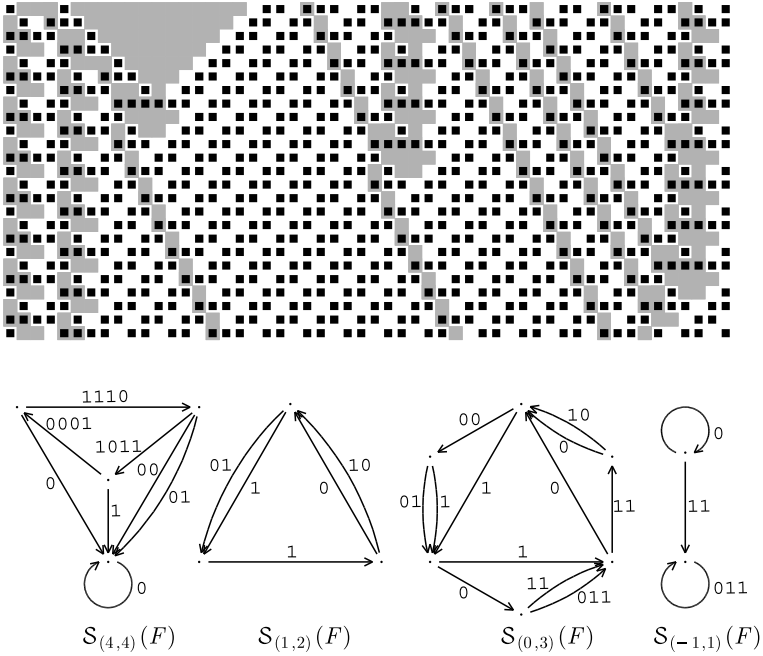
$$\Omega_F = F^2(\mathcal{S}_{(4,4)}(F) \mathbin{\dot{\vee}} \mathcal{S}_{(1,2)}(F) \mathbin{\dot{\vee}} \mathcal{S}_{(0,3)}(F) \mathbin{\dot{\vee}} \mathcal{S}_{(-1,1)}(F)).$$



Topological Dynamics of Cellular Automata, Figure 13  
Directional entropy of ECA90



Topological Dynamics of Cellular Automata, Figure 14  
The traffic rule ECA184



Topological Dynamics of Cellular Automata, Figure 15  
ECA 62 and its signal subshifts

000010000300001230000  
 000020001200003120000  
 000100003000012300000  
 000200012000031200000

**Topological Dynamics of Cellular Automata, Figure 16**  
**The multiplication CA**

In the space-time diagram in Fig. 15, the words 00, 111 and 010, which do not occur in the intersection  $\mathcal{S}_{(1,2)}(F) \cap \mathcal{S}_{(0,3)}(F) = \{(110)^\infty, (101)^\infty, (011)^\infty\}$  are displayed in grey (Kůrka [25]).

*Example 14 (A multiplication rule)*  $(4^{\mathbb{Z}}, F)$ , where

$$\begin{aligned} F(x)_i &= \left(2x_i + \left\lfloor \frac{x_{i+1}}{2} \right\rfloor\right) \bmod 4 \\ &= 2x_i + \left\lfloor \frac{x_{i+1}}{2} \right\rfloor - 4 \left\lfloor \frac{x_i}{2} \right\rfloor. \end{aligned}$$

00	01	02	03	10	11	12	13	20	21	22	23	30	31	32	33
0	0	1	1	2	2	3	3	0	0	1	1	2	2	3	3

We have

$$\begin{aligned} F^2(x)_i &= \left(4x_i + 2 \cdot \left\lfloor \frac{x_{i+1}}{2} \right\rfloor + (x_{i+1}) \bmod 4\right) \bmod 2 \\ &= x_{i+1} = \sigma(x)_i. \end{aligned}$$

Thus, the CA is a “square root” of the shift map. It is bijective and expansive, and its entropy is  $\ln 2$ . The system expresses multiplication by two in base four. If  $x \in A^{\mathbb{Z}}$  is left-asymptotic with  $0^\infty$ , then  $\varphi(x) = \sum_{i=-\infty}^{\infty} x_i 4^{-i}$  is finite and  $\varphi(F(x)) = 2\varphi(x)$ .

*Example 15 (A surjective rule)*  $(4^{\mathbb{Z}}, F)$ ,  $m = 0$ ,  $a = 1$ , and the local rule is

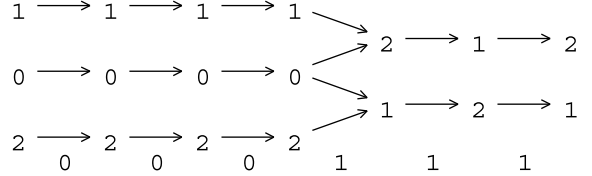
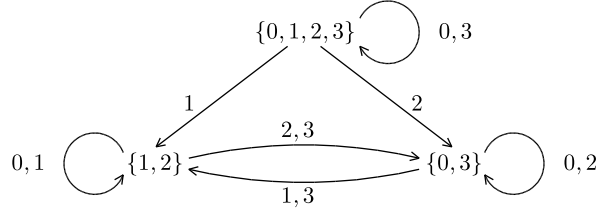
00	11	22	33	01	02	12	21	03	10	13	30	20	23	31	32
0	0	0	0	1	1	1	1	2	2	2	2	3	3	3	3

The system is surjective but not closing. The first image automaton is in Fig. 17 top. We see that  $F(4^{\mathbb{Z}})$  is the full shift, so  $F$  is surjective. The configuration  $0^\infty.1^\infty$  has left-asymptotic preimages  $0^\infty.(21)^\infty$  and  $0^\infty.(12)^\infty$ , so  $F$  is not right-closing. This configuration has also right-asymptotic preimages  $0^\infty.(12)^\infty$  and  $2^\infty.(12)^\infty$ , so  $F$  is not left-closing (Fig. 17 bottom). Therefore,  $\mathcal{X}^-(F) = \mathcal{X}^+(F) = \emptyset$ .

*Example 16*  $(2^{\mathbb{Z}} \times 2^{\mathbb{Z}}, \text{Id} \times \sigma)$ , i. e.,  $F(x, y)_i = (x_i, y_{i+1})$ .

The system is bijective and sensitive but not transitive.  $\mathcal{A}(F) = \mathcal{E}(F) = \emptyset$ ,  $\mathcal{X}^-(F) = (-\infty, 1)$ ,  $\mathcal{X}^+(F) = (0, \infty)$ . There are infinite signal subshifts

$$\mathcal{S}_{(0,n)} = 2^{\mathbb{Z}} \times |\mathcal{O}_\sigma^n|, \quad \mathcal{S}_{(-n,n)} = |\mathcal{O}_\sigma^n| \times 2^{\mathbb{Z}}$$



**Topological Dynamics of Cellular Automata, Figure 17**  
**Asymptotic configurations**

where  $|\mathcal{O}_\sigma^n| = \{x \in 2^{\mathbb{Z}} : \sigma^n(x) = x\}$ , so the speed subshifts are  $\mathcal{S}_0(F) = \mathcal{S}_{-1}(F) = A^{\mathbb{Z}}$ .

*Example 17 (A bijective CA)*  $(A^{\mathbb{Z}}, F)$ , where  $A = \{000, 001, 010, 011, 100\}$ , and

$$\begin{aligned} F(x, y, z)_i &= (x_i, (1 + x_i)y_{i+1} + x_{i-1}z_i, \\ &\quad (1 + x_i)z_{i-1} + x_{i+1}y_i) \bmod 2. \end{aligned}$$

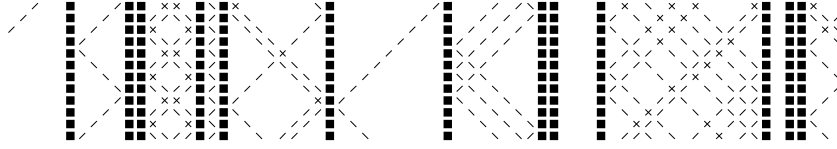
The dynamics is conveniently described as movement of three types of particles,  $1 = 001$ ,  $2 = 010$  and  $4 = 100$ . Letter  $0 = 000$  corresponds to an empty cell and  $3 = 011$  corresponds to a cell occupied by both  $1 = 001$  and  $2 = 010$ . The particle  $4 = 100$  is a wall which neither moves nor changes. Particle 1 goes to the left and when it hits a wall 4, it changes to 2. Particle 2 goes to the right and when it hits a wall, it changes to 1. Clearly 4 is a 1-blocking word, so the system is almost equicontinuous. It is bijective and its inverse is

$$\begin{aligned} F^{-1}(x, y, z)_i &= (x_i, (1 + x_i)y_{i-1} + x_{i+1}z_i, \\ &\quad (1 + x_i)z_{i+1} + x_{i-1}y_i) \bmod 2. \end{aligned}$$

The first column subshift is  $\Sigma_1(F) = \{0, 1, 2, 3\}^{\mathbb{N}} \cup \{4^\infty\}$ . We have infinite signal subshifts  $\mathcal{S}_{(-1,1)}(F) = \{0, 1\}^{\mathbb{Z}}$ ,  $\mathcal{S}_{(1,1)}(F) = \{0, 2\}^{\mathbb{Z}}$ ,  $\mathcal{S}_{(0,1)}(F) = \{0, 4\}^{\mathbb{Z}}$ . For  $q > 0$  we get

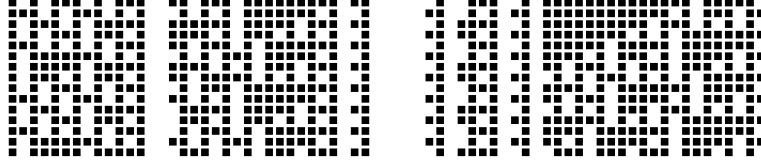
$$\begin{aligned} \mathcal{S}_{(0,q)}(F) &= \{x \in A^{\mathbb{Z}} : \forall u \in 4^*, (4u4 \sqsubseteq x \implies \\ &\quad (\exists m, 2m|u| = q) \text{ or } u \in \{0\}^*\} \end{aligned}$$

so the speed subshift is  $\mathcal{S}_0(F) = A^{\mathbb{Z}}$ . The equicontinuity directions are  $\mathcal{A}(F) = \{0\}$ ,  $\mathcal{E}(F) = \emptyset$ . The expansivity directions are  $\mathcal{X}^-(F) = (-\infty, -1)$ ,  $\mathcal{X}^+(F) = (1, \infty)$ .



Topological Dynamics of Cellular Automata, Figure 18

A bijective almost equicontinuous CA



Topological Dynamics of Cellular Automata, Figure 19

The Coven CA

*Example 18 (The Coven CA, Coven and Hedlund [13], Coven [12])*  $(2^{\mathbb{Z}}, F)$  where  $F(x)_i = x_i + x_{i+1}(x_{i+2} + 1) \bmod 2$ .

The CA is left-permutive with zero memory. It is not right-closing, since it does not have a constant number of preimages:  $F^{-1}(0^\infty) = \{0^\infty\}$ ,  $F^{-1}(1^\infty) = \{(01)^\infty, (10)^\infty\}$ . It is almost equicontinuous with 2-blocking word 000 with offset 0. It is not transitive but it is chain-transitive and its unique attractor is the full space (Blanchard and Maass [4]). While  $\Sigma_1(F) = 2^{\mathbb{Z}}$ , the two-column factor subshift

$$\Sigma_2(F) = \{10, 11\}^{\mathbb{N}} \cup \{11, 01\}^{\mathbb{N}} \cup \{01, 00\}^{\mathbb{N}}$$

is sofic but not SFT and the entropy is  $h(A^{\mathbb{Z}}, F) = \ln 2$ . For any  $a, b \in 2$  we have  $f(1a1b) = 1c$  where  $c = a + b + 1$  (here  $f$  is the local rule and the addition is modulo 2). Define a CA  $(2^{\mathbb{Z}}, G)$  by  $G(x)_i = (x_i + x_{i+1} + 1) \bmod 2$  and a map  $\varphi: 2^{\mathbb{Z}} \rightarrow 2^{\mathbb{Z}}$  by  $\varphi(x)_{2i} = 1$ ,  $\varphi(x)_{2i+1} = x_i$ . Then  $\varphi: (2^{\mathbb{Z}}, G) \rightarrow (2^{\mathbb{Z}}, F)$  is an injective morphism and  $(2^{\mathbb{Z}}, G)$  is a transitive subsystem of  $(2^{\mathbb{Z}}, F)$ . If  $x_i = 0$  for all  $i > 0$  and  $x_{2i} = 1$  for all  $i \leq 0$ , then  $(\mathcal{O}(x), F)$  is conjugated to the adding machine with periodic structure  $\mathbf{n} = (2, 2, 2, \dots)$ , and  $I_n^+((10)^\infty) = 2$ . We have  $\mathfrak{E}(F) = \emptyset$ ,  $\mathfrak{A}(F) = \{0\}$ ,  $\mathfrak{X}^-(F) = (-\infty, 0)$  and  $\mathfrak{X}^+(F) = \emptyset$ . We have  $\mathcal{S}_0(F) = 2^{\mathbb{Z}}$  and there exists an increasing sequence of non-transitive infinite signal subshifts  $\mathcal{S}_{(0,2^n)}(F)$ :

$$\begin{aligned} \mathcal{S}_{(0,1)}(F) &= \mathcal{S}_{\{10\}} \subset \mathcal{S}_{(0,2)}(F) \\ &= \mathcal{S}_{\{1010, 1110\}} \subset \mathcal{S}_{(0,4)}(F) \subset \dots \end{aligned}$$

*Example 19 (Gliders and Walls, Milnor [31])* The alphabet is  $A = \{0, 1, 2, 3\}$ ,  $F(x)_i = f(x_{[i-1, i+1]})$ , where the lo-

cal rule is

$$\begin{aligned} x3x : 3, \quad 12x : 3, \quad 1x2 : 3, \quad x12 : 3, \quad 1xx : 1, \\ x1x : 0, \quad xx2 : 2, \quad x2x : 0. \end{aligned}$$

Directional entropy has discontinuity at  $\alpha = 0$  (see Fig. 20).

*Example 20 (Golden mean subshift attractor: ECA132)*

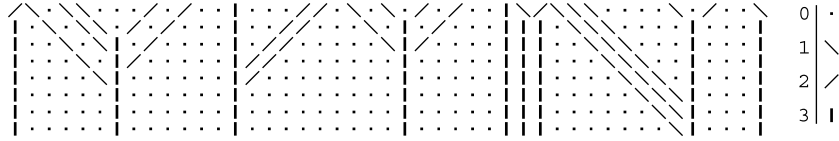
$$\begin{aligned} 000 : 0, \quad 001 : 0, \quad 010 : 1, \quad 011 : 0, \quad 100 : 0, \\ 101 : 0, \quad 110 : 0, \quad 111 : 1. \end{aligned}$$

While the golden mean subshift  $\mathcal{S}_{\{11\}}$  is the finite time maximal attractor of ECA12 (see Example 3), it is also an infinite time subshift attractor of ECA132. The clopen set  $U := [00]_0 \cup [01]_0 \cup [10]_0$  is spreading and  $\mathcal{Q}_F = \mathcal{S}_{\{11\}} = \mathcal{S}_{(0,1)}$ . There exist infinite signal subshifts  $\mathcal{S}_{(1,1)}(F) = \mathcal{S}_{\{10\}}$ ,  $\mathcal{S}_{(-1,1)}(F) = \mathcal{S}_{\{01\}}$  and the maximal attractor is their join  $\mathcal{Q}_F = \mathcal{S}_{(1,1)}(F) \hat{\vee} \mathcal{S}_{(0,1)}(F) \hat{\vee} \mathcal{S}_{(-1,1)}(F) = \mathcal{S}_{\{11\}} \cup (\mathcal{S}_{(1,1)}(F) \hat{\vee} \mathcal{S}_{(-1,1)}(F))$ .

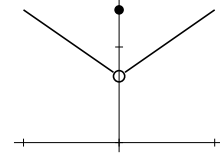
*Example 21 (A finite time sofic maximal attractor)*  $(2^{\mathbb{Z}}, F)$ , where  $m = -1$ ,  $a = 2$  and the local rule is

$$\begin{aligned} 0000 : 0, \quad 0001 : 0, \quad 0010 : 0, \quad 0011 : 1, \\ 0100 : 1, \quad 0101 : 0, \quad 0110 : 1, \quad 0111 : 1, \\ 1000 : 1, \quad 1001 : 1, \quad 1010 : 0, \quad 1011 : 1, \\ 1100 : 1, \quad 1101 : 0, \quad 1110 : 0, \quad 1111 : 0. \end{aligned}$$

The system has finite time maximal attractor  $\mathcal{Q}_F = F(2^{\mathbb{Z}}) = \mathcal{S}_{\{01^{2n+1}0 : n \geq 0\}}$  (Fig. 22 left). This is the **even**



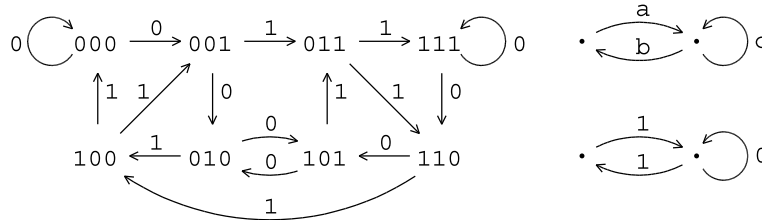
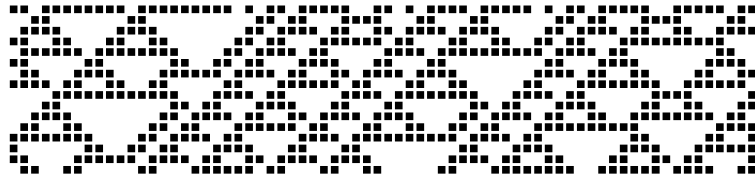
$$h_{\alpha}(A^{\mathbb{Z}}, F) = \begin{cases} (1 + |\alpha|) \ln 2 & \text{for } \alpha \neq 0 \\ 2 \ln 2 & \text{for } \alpha = 0 \end{cases}$$



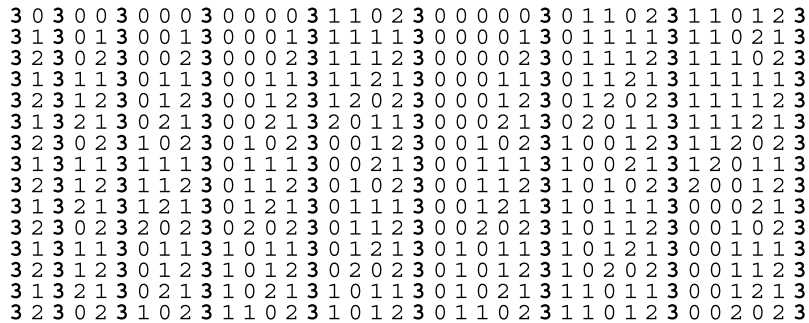
Topological Dynamics of Cellular Automata, Figure 20  
Directional entropy of Gliders and walls



Topological Dynamics of Cellular Automata, Figure 21  
ECA132



Topological Dynamics of Cellular Automata, Figure 22  
The first image graph and the even subshift



Topological Dynamics of Cellular Automata, Figure 23  
Logarithmic perturbation speeds



3	0	3	0	0	3	0	0	0	3	0	0	0	0	3	1	1	0	2	3	0	0	0	0	3	0	0	0	0	3	0	3		
3	1	3	0	1	3	0	0	1	3	0	0	0	1	3	1	1	1	1	3	0	0	0	0	1	3	0	0	0	1	3	1	3	
3	2	3	0	2	3	0	0	2	3	0	0	0	2	3	1	1	2	3	0	0	0	0	2	3	0	0	0	0	2	3	2	3	
3	1	3	1	1	3	0	1	1	3	0	0	1	1	3	1	1	2	1	3	0	0	0	1	1	3	0	0	0	1	0	3	1	3
3	2	3	1	2	3	0	1	2	3	0	0	1	2	3	1	2	0	2	3	0	0	0	1	2	3	0	0	0	1	1	3	2	3
3	1	3	2	1	3	0	2	1	3	0	0	2	1	3	2	0	1	1	3	0	0	0	2	1	3	0	0	0	1	1	3	1	3
3	1	3	0	2	3	1	0	2	3	0	1	0	1	3	0	0	1	2	3	0	0	1	0	2	3	0	0	0	1	2	3	2	3
3	2	3	1	1	3	1	1	1	3	0	1	0	2	3	0	0	2	1	3	0	0	1	1	1	3	0	0	0	2	0	3	1	3
3	1	3	1	2	3	1	1	2	3	0	1	1	1	3	0	1	0	2	3	0	0	1	1	2	3	0	0	1	0	1	3	2	3
3	2	3	2	1	3	1	2	1	3	0	1	1	2	3	0	1	1	1	3	0	0	1	2	1	3	0	0	1	0	1	3	1	3
3	0	3	0	2	3	2	0	2	3	0	1	2	1	3	0	1	1	2	3	0	0	2	0	2	3	0	0	1	0	2	3	2	3
3	1	3	1	0	3	0	1	1	3	0	2	0	2	3	0	1	2	1	3	0	1	0	1	1	3	0	0	1	1	0	3	1	3
3	2	3	1	1	3	0	1	2	3	1	0	1	1	3	0	2	0	2	3	0	1	0	1	2	3	0	0	1	1	1	3	2	3
3	1	3	1	2	3	0	2	1	3	1	0	1	2	3	1	0	1	1	3	0	1	0	2	1	3	0	0	1	1	1	3	1	3
3	2	3	2	1	3	1	0	2	3	1	0	2	1	3	1	0	1	2	3	0	1	1	0	2	3	0	0	1	1	2	3	2	3
3	0	3	0	2	3	1	1	1	3	1	1	0	2	3	1	0	2	1	3	0	1	1	1	1	3	0	0	1	2	0	3	1	3

Topological Dynamics of Cellular Automata, Figure 24

Sensitive system with logarithmic perturbation speeds

**subshift** whose minimal right-resolving presentation is in Fig. 22 right. We have  $E = \{a, b, c\}$ ,  $l(a) = l(b) = 1$ ,  $l(c) = 0$ . A word is synchronizing in  $\mathcal{G}$  (and intrinsically synchronizing) if it contains 0. The factor map  $\ell$  is right-resolving and also left-resolving. Thus  $\ell$  is left-closing and the even subshift is AFT. We have  $\ell^{-1}(1^\infty) = \{(ab)^\infty, (ba)^\infty\}$ , and  $|\ell^{-1}(x)| = 1$  for each  $x \neq 1^\infty$ . Thus, the even subshift is near-Markov, and it cannot be an infinite time maximal attractor.

*Example 22 (Logarithmic perturbation speeds)*  $(4^\mathbb{Z}, F)$  where  $m = 0$ ,  $a = 1$ , and the local rule is

$$\begin{array}{llllll} 00 : 0, & 01 : 0, & 02 : 1, & 03 : 1, & 10 : 1, & 11 : 1, \\ 12 : 2, & 13 : 2, & & & & \\ 20 : 0, & 21 : 0, & 22 : 1, & 23 : 1, & 30 : 3, & 31 : 3, \\ 32 : 3, & 33 : 3. \end{array}$$

The letter 3 is a 1-blocking word. Assume  $x_i = 3$  for  $i > 0$  and  $x_i \neq 3$  for  $i \leq 0$ . If  $\varphi(x) = \sum_{i=1}^{\infty} x_{-i} \cdot 2^i$  is finite, then  $\varphi(F(x)) = \varphi(x) + 1$ . Thus,  $(\mathcal{O}(x), F)$  is conjugated to the adding machine with periodic structure  $\mathbf{n} = (2, 2, 2, \dots)$ , although the system is not left-permutive. If  $x = 0^\infty.3^\infty$ , then for any  $i < 0$ ,  $(F^n(x))_{n \geq 0}$  has preperiod  $-i$  and period  $2^{-i}$ . For the zero configuration we have

$$\begin{aligned} n < 2^s + s &\implies F^n(W_s^-(0^\infty)) \subseteq W_0^-(0^\infty) \\ 2^{s-1} + s - 1 \leq n < 2^s + s &\implies I_n^-(0^\infty) = s \end{aligned}$$

and therefore  $\log_2 n - 1 < I_n^-(0^\infty) < \log_2 n + 1$ . More generally, for any  $x \in \{0, 1, 2\}^\mathbb{Z}$  we have  $\lim_{n \rightarrow \infty} I_n^-(x) / \log_2 n = 1$ .

*Example 23 (A sensitive CA with logarithmic perturbation speeds)*  $(4^\mathbb{Z}, F)$  where  $m = 0$ ,  $a = 2$  and the local rule is

$$\begin{array}{llll} 33x : 0, & 032 : 0, & 132 : 1, & 232 : 0, \quad 02x : 1, \\ 03x : 1, & & & \\ 12x : 2, & 13x : 2, & 20x : 0, & 21x : 0, \quad 22x : 1, \\ 23x : 1. \end{array}$$

A similar system is constructed in Bressaud and Tisseur [9]. If  $i < j$  are consecutive sites with  $x_i = x_j = 3$ , then  $F^n(x)_{(i,j)}$  acts as a counter machine whose binary value is increased by one unless  $x_{j+1} = 2$ :

$$B_{ij}(x) = \sum_{k=1}^{j-i-1} x_{i+k} \cdot 2^{j-i-1-k}$$

$$B_{ij}(F(x)) = B_{ij}(x) + 1 - \xi_2(x_{j+1}) - 2^{j-i-1} \cdot \xi_2(x_{i+1})$$

Here  $\xi_2(2) = 1$  and  $\xi_2(x) = 0$  for  $x \neq 2$ . If  $x \in \{0, 1, 2\}^\mathbb{Z}$ , then  $\lim_{n \rightarrow \infty} I_n^-(x) / \log_2(n) = 1$ . For periodic configurations which contain 3, Lyapunov exponents are positive. We have  $\lambda^-((30^n)^\infty) \approx 2^{-n}$ .

### Future Directions

There are two long-standing problems in topological dynamics of cellular automata. The first is concerned with expansivity. A positively expansive CA is conjugated to a one-sided full shift (Theorem 29). Thus, the dynamics of this class is completely understood. An analogous assertion claims that bijective expansive CA are conjugated to two-sided full shifts or at least to two-sided subshifts of finite type (Conjecture 30). Some partial results have been obtained in Nasu [35].

Another open problem is concerned with chaotic systems. A dynamical system is called chaotic, if it is topologically transitive, sensitive, and has a dense set of periodic

points. Banks et al. [3] proved that topological transitivity and density of periodic points imply sensitivity, provided the state space is infinite. In the case of cellular automata, transitivity alone implies sensitivity (Codenotti and Margara [10] or a stronger result in Moothathu [32]). By Conjecture 15, every transitive CA (or even every surjective CA) has a dense set of periodic points. Partial results have been obtained by Boyle and Kitchens [6], Blanchard and Tisseur [5], and Acerbi et al. [1]. See Boyle and Lee [7] for further discussion.

Interesting open problems are concerned with topological entropy. For CA with non-negative memory, the entropy of a CA can be obtained as the entropy of the column subshift whose width is the radius (Proposition 54). For the case of negative memory and positive anticipation, an analogous assertion would say that the entropy of the CA is the entropy of the column subshift of width  $2r + 1$  (Conjecture 55). Some partial results have been obtained in Di Lena [28]. Another aspect of information flow in CA is provided by Lyapunov exponents which have many connections with both topological and measure-theoretical entropies (see Bressaud and Tisseur [5] or Pivato ► [Ergodic Theory of Cellular Automata](#)). Conjecture 11 states that each sensitive CA has a configuration with a positive Lyapunov exponent.

## Acknowledgments

I thank Marcus Pivato and Mathieu Sablik for careful reading of the paper and many valuable suggestions. The research was partially supported by the Research Program CTS MSM 0021620845.

## Bibliography

### Primary Literature

- Acerbi L, Dennunzio A, Formenti E (2007) Shifting and lifting of a cellular automaton. In: Computational logic in the real world. Lecture Notes in Computer Sciences, vol 4497. Springer, Berlin, pp 1–10
- Akin E, Auslander J, Berg K (1996) When is a transitive map chaotic? In: Bergelson, March, Rosenblatt (eds) Conference in Ergodic Theory and Probability. de Gruyter, Berlin, pp 25–40
- Banks J, Brook J, Cairns G, Davis G, Stacey P (1992) On Devaney's definition of chaos. *Am Math Monthly* 99:332–334
- Blanchard F, Maass A (1996) Dynamical behaviour of Coven's aperiodic cellular automata. *Theor Comput Sci* 291:302
- Blanchard F, Tisseur P (2000) Some properties of cellular automata with equicontinuous points. *Ann Inst Henri Poincaré* 36:569–582
- Boyle M, Kitchens B (1999) Periodic points for onto cellular automata. *Indag Math* 10(4):483–493
- Boyle M, Lee B (2007) Jointly periodic points in cellular automata: computer explorations and conjectures. (manuscript)
- Boyle M, Lind D (1997) Expansive subdynamics. *Trans Am Math Soc* 349(1):55–102
- Bressaud X, Tisseur P (2007) On a zero speed cellular automaton. *Nonlinearity* 20:1–19
- Codenotti B, Margara L (1996) Transitive cellular automata are sensitive. *Am Math Mon* 103:58–62
- Courbage M, Kamiński B (2006) On the directional entropy of  $\mathbb{Z}^2$ -actions generated by cellular automata. *J Stat Phys* 124(6):1499–1509
- Coven EM (1980) Topological entropy of block maps. *Proc Am Math Soc* 78:590–594
- Coven EM, Hedlund GA (1979) Periods of some non-linear shift registers. *J Comb Theor A* 27:186–197
- Coven EM, Pivato M, Yassawi R (2007) Prevalence of odometers in cellular automata. *Proc Am Math Soc* 815:821
- Culik K, Hurd L, Yu S (1990) Computation theoretic aspects of cellular automata. *Phys D* 45:357–378
- Formenti E, Kůrka P (2007) A search algorithm for the maximal attractor of a cellular automaton. In: Thomas W, Weil P (eds) STACS Lecture Notes in Computer Science, vol 4393. Springer, Berlin, pp 356–366
- Formenti E, Kůrka P (2007) Subshift attractors of cellular automata. *Nonlinearity* 20:105–117
- Gilman RH (1987) Classes of cellular automata. *Ergod Theor Dynam Syst* 7:105–118
- Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. *Math Syst Theory* 3:320–375
- Hopcroft JE, Ullmann JD (1979) Introduction to Automata Theory, Languages and Computation. Addison-Wesley, London
- Hurd LP (1990) Recursive cellular automata invariant sets. *Complex Syst* 4:119–129
- Hurley M (1990) Attractors in cellular automata. *Ergod Theor Dynam Syst* 10:131–140
- Kitchens BP (1998) Symbolic Dynamics. Springer, Berlin
- Kůrka P (2003) Topological and symbolic dynamics, Cours spécialisés, vol 11. Société Mathématique de France, Paris
- Kůrka P (2005) On the measure attractor of a cellular automaton. *Discret Contin Dyn Syst* 2005:524–535; supplement
- Kůrka P (2007) Cellular automata with infinite number of subshift attractors. *Complex Syst* 17:219–230
- Kůrka P (1997) Languages, equicontinuity and attractors in cellular automata. *Ergod Theor Dynam Syst* 17:417–433
- Lena PD (2007) Decidable and computational properties of cellular automata, Ph D thesis. Università de Bologna e Padova, Bologna
- Lind D, Marcus B (1995) An Introduction to Symbolic Dynamics and Coding. Cambridge University Press, Cambridge
- Maass A (1995) On the sofic limit set of cellular automata. *Ergod Theor Dyn Syst* 15:663–684
- Milnor J (1988) On the entropy geometry of cellular automata. *Complex Syst* 2(3):357–385
- Moothathu TKS (2005) Homogeneity of surjective cellular automata. *Discret Contin Dyn Syst* 13(1):195–202
- Moothathu TKS (2006) Studies in topological dynamics with emphasis on cellular automata, Ph D thesis. University of Hyderabad, Hyderabad
- Nasu M (1995) Textile Systems for Endomorphisms and Automorphisms of the Shift. *Mem Am Math Soc* 546
- Nasu M (2006) Textile systems and one-sided resolving automorphisms and endomorphisms of the shift. American Mathematical Society, Providence

36. von Neumann J (1951) The general and logical theory of automata. In: Jeffers LA (ed) *Cerebral Mechanics of Behaviour*. Wiley, New York
37. Sablik M (2006) Étude de l'action conjointe d'un automate cellulaire et du décalage: une approche topologique et ergodique, Ph D thesis. Université de la Méditerranée, Marseille
38. Sablik M (2007) Directional dynamics of cellular automata: a sensitivity to initial conditions approach. *Theor Comput Sci* 400(1–3):1–18
39. Shereshevski MA, Afraimovich VS (1992) Bipermutative cellular automata are topologically conjugated to the one-sided Bernoulli shift. *Random Comput Dynam* 1(1):91–98
40. Smillie J (1988) Properties of the directional entropy function for cellular automata. In: *Dynamical systems*, vol 1342 of *Lecture Notes in Mathematics*. Springer, Berlin, pp 689–705
41. Wolfram S (1984) Computation theory of cellular automata. *Comm Math Phys* 96:15–57
42. Wolfram S (1986) *Theory and Applications of Cellular Automata*. World Scientific, Singapore

### Books and Reviews

- Burks AW (1970) *Essays on Cellular automata*. University of Illinois Press, Chicago
- Delorme M, Mazoyer J (1998) *Cellular automata: A parallel Model*. Kluwer, Amsterdam
- Demongeot J, Goles E, Tchuente M (1985) *Dynamical systems and cellular automata*. Academic Press, New York
- Farmer JD, Toffoli T, Wolfram S (1984) *Cellular automata*. North-Holland, Amsterdam
- Garzon M (1995) *Models of Massive Parallelism: Analysis of Cellular automata and Neural Networks*. Springer, Berlin
- Goles E, Martinez S (1994) *Cellular automata, Dynamical Systems and Neural Networks*. Kluwer, Amsterdam
- Goles E, Martinez S (1996) *Dynamics of Complex Interacting Systems*. Kluwer, Amsterdam
- Gutowitz H (1991) *Cellular Automata: Theory and Experiment*. MIT Press/Bradford Books, Cambridge Mass. ISBN 0-262-57086-6
- Kitchens BP (1998) *Symbolic Dynamics*. Springer, Berlin
- Kůrka P (2003) *Topological and symbolic dynamics*. Cours spécialisés, vol 11. Société Mathématique de France, Paris
- Lind D, Marcus B (1995) *An Introduction to Symbolic Dynamics and Coding*. Cambridge University Press, Cambridge
- Macucci M (2006) *Quantum Cellular automata: Theory, Experimentation and Prospects*. World Scientific, London
- Manneville P, Boccara N, Vichniac G, Bidaux R (1989) *Cellular automata and the modeling of complex physical systems*. Springer, Berlin
- Moore C (2003) *New Constructions in Cellular automata*. Oxford University Press, Oxford
- Toffoli T, Margolus N (1987) *Cellular Automata Machines: A New Environment for Modeling*. Mass MIT Press, Cambridge
- von Neumann J (1951) The general and logical theory of automata. In: Jeffers LA (ed) *Cerebral Mechanics of Behaviour*. Wiley, New York
- Wolfram S (1986) *Theory and applications of cellular automata*. World Scientific, Singapore
- Wolfram S (2002) *A new kind of science*. Media Inc, Champaign
- Wuensche A, Lesser M (1992) The Global Dynamics of Cellular Automata. In: *Santa Fe Institute Studies in the Sciences of Complexity*, vol 1. Addison-Wesley, London

## Topological Magnetohydrodynamics and Astrophysics

MITCHELL A. BERGER<sup>1,2</sup>

<sup>1</sup> Mathematics, University of Exeter, Devon, UK

<sup>2</sup> Mullard Space Science Laboratory, UCL, London, UK

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Topology of Curves](#)

[Magnetohydrodynamics](#)

[Magnetic Helicity](#)

[Applications to Astrophysical Fields](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Corona** The atmosphere of the sun or a star. The solar corona is generally much hotter (at temperatures up to 2 million K) than the photosphere (solar surface), due to magnetic heating.

**Crossing number** The projection of a three-dimensional curve onto a plane will exhibit a certain number of crossings (where the projected curve passes over itself). This is the *crossing number* of the curve; in general the crossing number depends on the orientation of the plane as well as the original three-dimensional curve. The *average crossing number* averages over all possible planar projections. This removes the dependence on orientation, but there is still a dependence on the geometry of the curve. Given arbitrary distortions of a curve, without letting the curve pass through itself, there will be a minimum number of crossings, the *minimum crossing number*. For closed curves (*knots*), and for collections of curves (*links*) the minimum crossing number provides a measure of topological complexity. Crossing numbers can also be defined for vector fields in terms of the crossings seen between all pairs of field lines.

**Force-free field** A force-free magnetic field does not impart any magnetic forces. In the absence of other forces, such as pressure or gravity, a magnetized fluid in equilibrium will possess a force-free field.

**Helicity integral** The *helicity integral*  $H(\mathbf{V}, \mathbf{W})$  of two vector fields  $\mathbf{V}$  and  $\mathbf{W}$  measures the net linking of the field lines of  $\mathbf{V}$  with those of  $\mathbf{W}$ . Common exam-

ples in a fluid with magnetic field  $\mathbf{B}$  and vorticity  $\boldsymbol{\omega}$  include the *kinetic helicity*  $H(\boldsymbol{\omega}, \boldsymbol{\omega})$ , which measures the self linking of the vorticity  $\boldsymbol{\omega}$ ; the *magnetic helicity*  $H(\mathbf{B}, \mathbf{B})$ , which measures the self-linking of magnetic field  $\mathbf{B}$  with itself; and the *cross-helicity*  $H(\mathbf{B}, \boldsymbol{\omega})$ , the linking between the magnetic field with the vortex field.

**Linking number** The *linking number*  $L$  of two oriented closed curves measures how much they intertwine about each other. If the two curves are projected onto a plane,  $L$  equals one half the number of (signed) crossings. The linking number is invariant to arbitrary deformations of the curves, as long as the two curves do not pass through each other.

**Magnetohydrodynamics (MHD)** Magnetohydrodynamics studies the evolution of a fluid carrying electrical currents and subjected to magnetic forces. Typical examples are liquid metals and ionized plasmas. In the latter case, MHD neglects small-scale effects, such as those arising from fluctuations about charge neutrality or from details of particle trajectories. *Ideal MHD* assumes a perfectly conducting plasma. The magnetic lines of force in ideal MHD are convected by the fluid motions without slipping through the fluid, breaking, or passing through each other.

**Photosphere** The surface of the sun. Near the photosphere there is a steep change in pressure, density, and optical depth. Also the temperature reaches a minimum (on average 5 800 K).

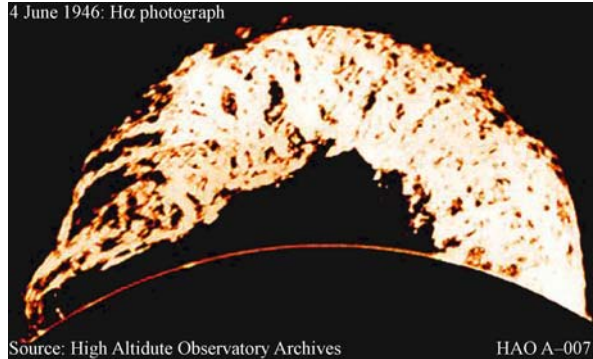
**Reconnection** Reconnection occurs when field lines change their connectivity. In a simple reconnection event two bundles of field lines meet in a small region. The field lines are cut in this region, allowing the two pieces of a field line from the first bundle to connect to two corresponding pieces from the second bundle. Reconnection events in nature may involve several simple events.

**Twist number** The *twist number*  $\mathcal{T}_w$  applies to ribbons and tubes. For ribbons, it measures the extent to which one side of the ribbon rotates about the other. For magnetic flux tubes, it measures how much field lines within the tube rotate about the axis of the tube.

**Winding number** Consider two curves extending between parallel planes. The *winding number*  $w$  measures the angle (divided by  $2\pi$ ) through which the two curves rotate about each other.

**Writhe** The *writhe number*  $\mathcal{W}_r$  is a property of a single curve. It measures three-dimensional geometrical structure, and can change if the curve is distorted. For a ribbon, the writhe of the axis of the ribbon plus the twist of an edge of the ribbon about the axis equals

4 June 1946:  $H\alpha$  photograph



Source: High Altitude Observatory Archives

HAO A-007

**Topological Magnetohydrodynamics and Astrophysics, Figure 1** The “Grand Daddy” erupting prominence of June 1946. The hot plasma in the prominence traces out coiled magnetic field lines. The plasma is photographed in the  $H\alpha$  line, sensitive to plasma of temperatures about 20 000 K. The structure extends some 200 000 km above the solar surface (the arc at the bottom)

the linking number between the edge and the axis (Călugăreanu theorem).

## Definition of the Subject

Many important processes in astrophysics involve magnetic fields, from solar flares and eruptions (Fig. 1) to the formation of galactic jets. Magnetic fields are often highly structured by their *field lines*; a particularly striking example of this can be seen in the fibrous appearance of clouds in the solar atmosphere (Fig. 10). These linear structures can be twisted, kinked, and interlinked. Several geometrical quantities exist which measure the amounts of such structural features. When some of these geometrical quantities are left unchanged by some set of deformations of the field, they are called *topological invariants*. Astrophysicists have paid particular attention to the magnetic helicity integral, because it is approximately conserved in highly conducting plasmas. In our own solar system, magnetic helicity is transported from the interior of the sun, through the solar surface and atmosphere and into interplanetary space. Measurements of this transport assist in the understanding of basic physical processes in these regions.

## Introduction

Recent advances in numerical simulations, analytic theory, and astrophysical observations have allowed us to reveal nature in its full three-dimensional glory (plus time as well). Geometrical and topological measures help us understand the complicated three-dimensional structures that we find. Some quantities, such as linking and helicity integrals, can be positive or negative. The sign changes for



a mirror or parity transformation, so these quantities measure handedness as well as topological aspects of a structure. Other quantities, like minimum energy and crossing number, are positive definite, and measure topological complexity without regard to handedness.

Section “[The Topology of Curves](#)” will briefly review the kinds of topological invariants relevant to magnetic field studies. Emphasis will be placed on the Gauss linking integral, which provides the template for the magnetic helicity invariant. The next section provides a brief introduction to magnetohydrodynamics, especially concerning the structure of magnetic equilibria. Section “[Magnetic Helicity](#)” covers magnetic helicity in some detail. Astrophysical applications of helicity and other topological concepts will be found in Sect. “[Applications to Astrophysical Fields](#)”. The review ends with a brief discussion of future directions.

## The Topology of Curves

### Types of Topological Invariants

Figure 2 shows three configurations. In the first, two curves link each other. The Gauss linking integral (described below) detects this linking. For example it evaluates to  $-3$  for the curves shown but will always give zero for unlinked curves. The Gauss integral is an example of an *isotopic* invariant: it does not change no matter how much we distort the curves (as long as the curves are not allowed to pass through each other).

The middle configuration demonstrates a more subtle topological structure. The figure shows the *Borromean rings*, where any two rings are unlinked but the configuration as a whole cannot be taken apart. The Gauss integral is useless here. Isotopic invariants such as ‘higher order linking numbers’ do exist which detect this topology [10,13,36,42,47,53,64].

The configuration on the right shows two curves extending between parallel planes. Unlike the previous two examples, these are not closed curves: they have end points on boundaries. Nevertheless, we can still find isotopic invariants and complexity measures. For the two curves shown, the *winding number*  $w$  gives the net angle (divided by  $2\pi$ ) through which the two curves rotate about each other. The winding number is an isotopic invariant, given the restriction of fixed boundaries (i. e. the endpoints of the curves are fixed on the boundary planes, and the rest of the curves are not allowed to pass through the planes).

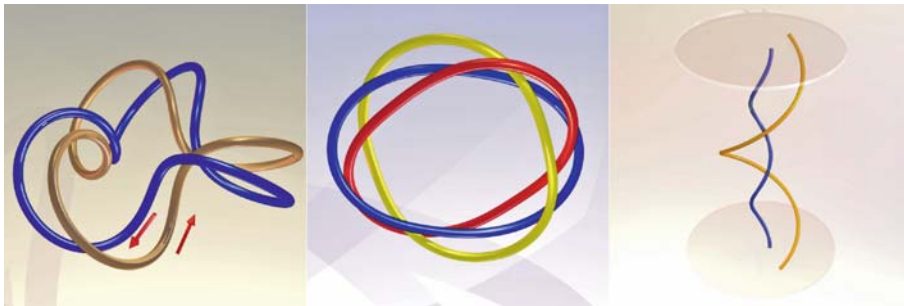
We can also look for positive definite numbers which can be minimized over all possible geometrical configurations. In particular, we can count the number of crossings of the curves as seen in a plane projection. If we distort the curves by introducing extra twist and kinks the number of crossings may go up. For the Borromean rings, for example, over the set of all configurations (reachable by arbitrary distortions without curves passing through each other) the minimum number of crossings is six. The *minimum crossing number* is not an isotopic invariant; instead it gives a measure of complexity.

### The Gauss Linking Number

Here we give more detail on perhaps the simplest but most powerful isotopic invariant, the Gauss linking number. For more detail and proofs of the properties of the linking number describe below, see [1,15,25,80]. Early in the 19th century Gauss discovered an integral formula that measures how many times two closed curves link each other [28]. Let  $\mathbf{x}(\sigma)$  be a point on curve 1 and  $\mathbf{y}(\tau)$  a point on curve 2, where  $\sigma$  and  $\tau$  parametrize curves 1 and 2. Also let  $\mathbf{r} = \mathbf{y} - \mathbf{x}$  be the relative position vector.

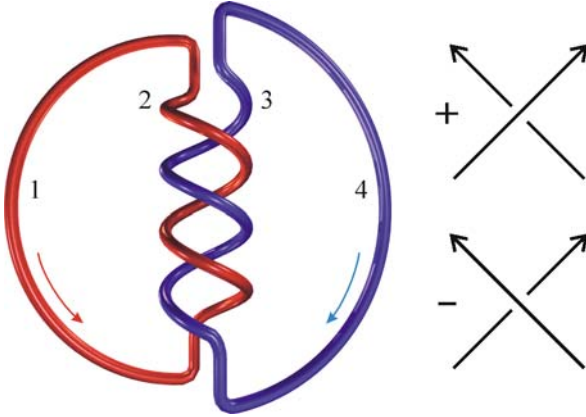
**Definition 1** The Gauss linking integral  $L_{12}$  is given by

$$L_{12} = \frac{1}{4\pi} \oint_1 \oint_2 \frac{d\mathbf{x}}{d\sigma} \cdot \frac{\mathbf{r}}{r^3} \times \frac{d\mathbf{y}}{d\tau} d\tau d\sigma. \quad (1)$$



Topological Magnetohydrodynamics and Astrophysics, Figure 2  
Left: two linked curves. Middle: Borromean rings. Right: two braided curves





**Topological Magnetohydrodynamics and Astrophysics, Figure 3**  
Two linked curves with four negative crossings. The linking number is  $L = -2$

**Theorem 1** The Gauss linking integral gives an integer for any two closed curves. This integer measures how many times (in a right-handed sense) curve 1 crosses a surface bounded by curve 2.

The Gauss linking integral is invariant to deformations of the two curves, as long as the two curves do not pass through each other. This follows from the fact that the integral counts points of intersection of one curve with a surface bounded by the other. Outside those points of intersection, the curves are free to move without changing the sum. Even if they do move, the sum will not change as long as the curves do not pass through each other.

A direct evaluation of the Gauss linking integral would be difficult (although, as it always gives an integer a numerical evaluation does not need to be very precise). Fortunately one can evaluate the integral simply by projecting the curves onto a plane and counting the number of crossings (see Fig. 3).

**Theorem 2** Let  $N_+$  and  $N_-$  be the numbers of positive and negative crossings of a link as seen in a plane projection. Then

$$L_{12} = (N_+ - N_-)/2. \quad (2)$$

This theorem shows that linking number can be calculated directly from a picture of the two curves; often we do not need to deal directly with the formidable looking linking integral. Similarly, helicity integrals for vector fields (Sect. “Magnetic Helicity”) often can be calculated directly from the morphology of the field, without ever evaluating the integral itself.

### Twist and Writhe

The self-helicity of a single flux tube can be decomposed into two contributions called *twist*  $\mathcal{T}_w$  and *writhe*  $\mathcal{W}_r$ . First we define twist and writhe for closed curves, then generalize to magnetic flux tubes.

Recall that the Gauss linking integral consists of a double line integral, integrated along two separate curves. Suppose we try doing both line integrals along the same curve?

**Definition 2** The *writhe* of a closed curve  $\gamma$  is given by

$$\mathcal{W}_r = \frac{1}{4\pi} \oint_{\gamma} \oint_{\gamma} \frac{d\mathbf{x}}{d\sigma} \cdot \frac{\mathbf{r}}{r^3} \times \frac{d\mathbf{y}}{d\tau} d\tau d\sigma. \quad (3)$$

Unlike the linking integral  $L$ , however, the writhe  $\mathcal{W}_r$  is not a topological invariant – it changes if we distort the curve.

Suppose we consider a *ribbon*, which can be described mathematically by specifying two almost parallel curves: the axis curve running along the center of the ribbon, and a secondary curve (one of the edges of the ribbon). A DNA molecule provides an important application of the theory of ribbons [1,85]. The two curves may twist about each other; in addition the axis may itself be coiled.

We can define the twist  $\mathcal{T}_w$  as the net amount that the secondary curve twists about the direction of the axis:

**Definition 3** Let  $\mathbf{T}(s)$  be the tangent vector to the axis curve at arclength  $s$ , and let  $\hat{\mathbf{v}}$  be a unit vector at  $s$  pointing to the secondary curve. Then the *twist number* is

$$\mathcal{T}_w = \frac{1}{2\pi} \oint \mathbf{T}(s) \cdot \hat{\mathbf{v}}(s) \times \frac{d\hat{\mathbf{v}}(s)}{ds} ds. \quad (4)$$

If the ribbon closes upon itself, we may calculate the linking number of the two curves. In 1961 Călugăreanu [21] proved a remarkable theorem:

**Theorem 3**

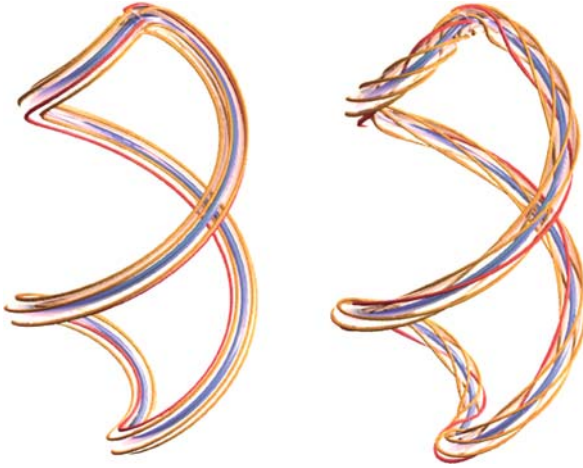
$$L = \mathcal{T}_w + \mathcal{W}_r. \quad (5)$$

We can also define twist, writhe, and linking for a tube filled with curves (such as magnetic field lines). For every field line within the tube, we can define a twist number with respect to the central axis (see Fig. 4).

### Magnetohydrodynamics

#### Field Lines

We briefly review magnetohydrodynamics (MHD) [81, 83,84,87], with an emphasis on the geometrical structure



**Topological Magnetohydrodynamics and Astrophysics, Figure 4**  
Two tubes with the same axis geometry. The central axes have a writhe of  $W_r = -0.72$ . The left tube has zero internal twist; i. e.  $\mathcal{T}_w = 0$ . On the right  $\mathcal{T}_w = 5$

of the *magnetic field lines*. A magnetic field line is a curve which follows the direction of the magnetic field vector  $\mathbf{B}$ : if  $s$  denotes arclength along the curve and  $\mathbf{x}(s)$  gives the position vector on the curve at  $s$ , then

$$\frac{d\mathbf{x}}{ds} = \frac{\mathbf{B}}{|\mathbf{B}|}. \quad (6)$$

Of course, field lines are infinitely thin, and there are infinitely many of them. Often people prefer to consider bundles of field lines, thin enough to be more or less coherent over their length, called flux elements or *flux tubes*. In many cases a field can be approximated as a finite set of flux tubes [80]. Often topological quantities like magnetic helicity can be calculated directly from an examination of the field line or flux tube structure, without ever directly seeing the equations for  $\mathbf{B}$ .

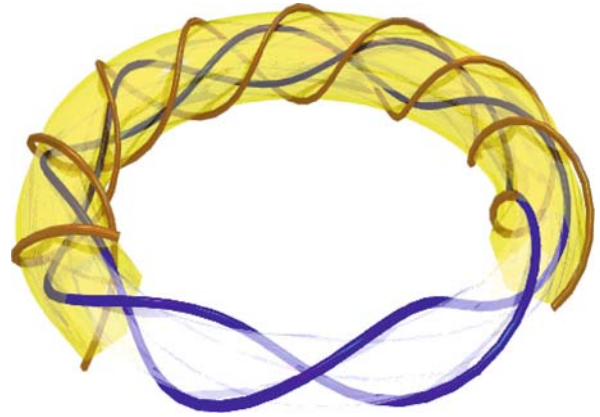
### The Maxwell Equations

Let us start with the Maxwell equations (e. g. [86]). The internal equations involve the electric and magnetic fields  $\mathbf{B}$  and  $\mathbf{E}$  only; matter terms involving charges and currents do not appear.

$$\nabla \cdot \mathbf{B} = 0, \quad (7)$$

$$\nabla \times \mathbf{E} + \partial_t \mathbf{B} = 0. \quad (8)$$

Equation (7) says that there are no magnetic monopoles. This lack of source charges has a profound topological consequence: lines of magnetic flux have no endpoints. Magnetic field have four options: they can either



**Topological Magnetohydrodynamics and Astrophysics, Figure 5**  
A magnetic field where the field lines form nested toroidal surfaces. Suppose a field line closes upon itself after encircling the torus  $n$  times the short way around and  $m$  times the long way around. Then it has the (rational) periodicity ratio  $T = n/m$ . Here  $T$  gives the number of times the field line twists about the toroidal axis in one circuit the long way around

form closed curves, extend to infinity, wander ergodically within a volume, or wrap around a toroidal surface. The last option is illustrated in Fig. 5. Generally, a magnetic field may fill a torus with a family of nested surfaces, with smoothly varying twist  $T$ . When  $T$  is irrational, field lines never exactly close upon themselves (although they will come arbitrarily close).

The Maxwell source equations in vacuum with charge density  $\rho$  and electrical current  $\mathbf{J}$  (in SI units with permittivity  $\epsilon_0$  and permeability  $\mu_0$ ) are

$$\nabla \cdot \epsilon_0 \mathbf{E} = \rho, \quad (9)$$

$$\nabla \times \frac{1}{\mu_0} \mathbf{B} - \epsilon_0 \partial_t \mathbf{E} = \mathbf{J}. \quad (10)$$

Equation (9) implies that electric field lines start and stop at electric charges. For non-relativistic applications,  $\partial_t \mathbf{E}$  is small, and Eq. (10) gives

$$\nabla \times \frac{1}{\mu_0} \mathbf{B} \approx \mathbf{J} \quad (11)$$

i. e. magnetic field lines circle electric currents.

### Magnetized Fluids

In magnetohydrodynamics (MHD), we consider a neutral fluid which can carry electrical currents. Physically, this may be liquid metal or a neutral plasma. For basic references emphasizing applications in Tokomaks and other

fusion devices, see [81]. For solar physics applications, see [82] and [78].

Electrical currents tend to dissipate in physical media. In MHD we assume some functional relation between the electric field  $\mathbf{E}$  and the electrical current  $\mathbf{J}$ . The simplest relation is the linear Ohm's law

$$\mathbf{E} = \eta \mathbf{J} \quad (12)$$

where  $\eta$  is the resistivity. Some MHD studies employ a *generalized Ohm's law* with more terms on the right-hand side [79]; for example, the Hall effect can be important for collisionless plasmas in the Earth's magnetosphere [24].

Consider a small region of fluid moving at speed  $\mathbf{V}$  with respect to some fixed or laboratory coordinate frame. Because of the assumption of charge neutrality, the electrostatic electric field vanishes in the rest frame of this fluid element. Thus  $\mathbf{E} = \eta \mathbf{J}$  in this frame. However, if we Lorentz transform to the fixed frame, the electric field becomes

$$\mathbf{E} = \mathbf{B} \times \mathbf{V} + \eta \mathbf{J}. \quad (13)$$

If we assume uniform resistivity  $\eta$ , by the Maxwell equation Eq. (8) we obtain the *magnetic induction equation*

$$\partial_t \mathbf{B} = \nabla \times (\mathbf{V} \times \mathbf{B}) + \eta \nabla^2 \mathbf{B}. \quad (14)$$

The last (dissipation) term shows that a concentration of magnetic field will diffuse into the fluid. Often in astrophysical plasmas this term is small, except near intense electrical currents. If we set this term to zero, then we have the situation of *ideal MHD*. In ideal MHD, magnetic field lines are convected by the fluid like material strings. Field lines cannot be cut or cross through each other. This preserves the topology of the field lines; for example if two closed field lines are linked, then they will stay linked no matter how much the velocity  $\mathbf{V}$  moves them around. Thus we have a magnetic analogy to the Helmholtz circulation theorem: consider a closed curve  $C$  (which may or may not be a field line) moving with the flow. Then the net magnetic flux encircled by  $C$  is a constant of the motion.

Ideal MHD allows us to follow the evolution of field lines in time; in a relativistic description field lines trace out surfaces in space-time [35,37]. When  $\eta \neq 0$  then strictly speaking magnetic field lines cannot be unambiguously followed in time; however some understanding of field line evolution is still possible [70]. In regions of intense currents  $\eta \nabla^2 \mathbf{B}$  may be large even if  $\eta$  is small. In this case field lines can break and reconnect with each other, thus changing the field topology. This phenomenon of *reconnection* is an active area of research (e. g. [41,70,83]).

In addition to the magnetic induction equation, we need to describe the evolution of the velocity field. The Navier–Stokes equation in the presence of magnetic forces gives

$$\rho (\partial_t \mathbf{V} + \mathbf{V} \cdot \nabla \mathbf{V}) = -\nabla p + \mathbf{J} \times \mathbf{B} + \nu \nabla^2 \mathbf{V} \quad (15)$$

where  $p$  and  $\rho$  are the pressure and density of the flow, and  $\nu$  is the kinematic viscosity.

### Magnetostatic Equilibria

Of particular interest are equilibrium solutions where  $\mathbf{V} = \partial_t \mathbf{V} = 0$ . In this case

$$\mathbf{J} \times \mathbf{B} = \nabla p. \quad (16)$$

As  $\mu_0 \mathbf{J} = \nabla \times \mathbf{B}$ , this is a nonlinear equation for  $\mathbf{B}$ . The theory of magnetic equilibria is of particular importance in studies of the solar atmosphere, where magnetic structures can last on timescales up to months (in contrast, dynamic timescales such as the wave travel time across a coronal loop may only be seconds) [3,45].

First we consider pressure-balanced equilibria. Note that solutions with non-uniform pressure satisfy  $\mathbf{B} \cdot \nabla p = \mathbf{J} \cdot \nabla p = 0$ , so field lines and current lines lie within level surfaces of the pressure. A *magnetic surface* is a surface  $S$  which no field lines cross, i. e.  $\mathbf{B} \cdot \hat{n} = 0$ , where  $\hat{n}$  is the unit normal. For fields completely contained within a finite magnetic surface, the pressure surfaces must lie on nested torii [23,49]. (Note that a sphere can be filled with nested torii, as in a Hill's vortex; for the outermost torus the central hole degenerates into a line segment extending from the top of the sphere to the bottom.)

Pressure gradients can be neglected in diffuse plasmas (for example in the solar corona typical magnetic forces are three orders of magnitude higher than pressure forces). In this case equilibria take the form of force-free fields [44,50,71]. If pressure  $p = 0$  then the electrical current is parallel to the field, so that

$$\nabla \times \mathbf{B} = \lambda(\mathbf{x}) \mathbf{B} \quad (17)$$

for some scalar function  $\lambda(\mathbf{x})$  of position  $\mathbf{x}$ . Taking the divergence of this equation gives  $\mathbf{B} \cdot \nabla \lambda = 0$ . Hence field lines (and current lines too) lie within level surfaces, this time of the function  $\lambda$  rather than  $p$ . Again inside a finite magnetic surface field lines lie on nested torii.

The one exception occurs when  $\lambda$  is uniform in space. In this case field lines have the possibility of ergodically filling a volume. The topology of such fields is of great interest [31,80]. These fields are eigenfunctions of the curl operator,  $\nabla \times \mathbf{B} = \lambda \mathbf{B}$ . They are sometimes called *linear*

*force-free fields* because this is a linear equation, and hence much easier to solve than in the case of variable  $\lambda$ .

Force-free solutions minimize the magnetic energy subject to topological constraints: given the set of all magnetic configurations with the same field line topology, local minima in the magnetic energy will be free of magnetic forces. It is usually difficult to determine whether a particular solution is a local or global (true) minimum of the energy. The energetics of magnetic equilibria have application in knot theory. The minimum energy of a magnetic field in the form of a knotted flux tube provides, like minimum crossing number (Sect. “Types of Topological Invariants”) a measure of topological complexity [49,50].

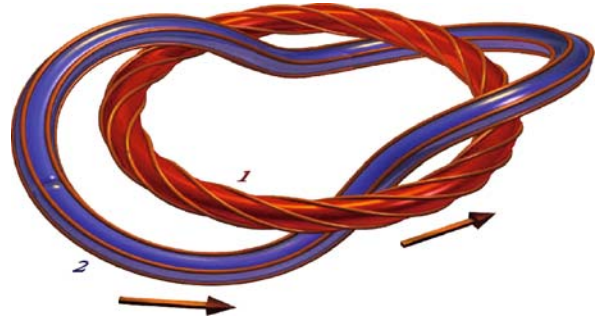
Linear force-free fields minimize the energy for all fields with the same magnetic helicity (see Sect. “Magnetic Helicity” below), and the same boundary conditions [8,27,71]. Note that fields with different field line topologies can have the same magnetic helicity. Thus reconnection is generally needed to enable a field to relax to the linear force-free state [34,67].

## Magnetic Helicity

### Overview

Helicity integrals measure the net interlinking of the field lines of a vector field. Physicists, astronomers, and mathematicians calculate the helicity of magnetic fields in a variety of situations, from plasma experiments to descriptions of galactic magnetic fields. Key properties of helicity include its relations to field structure and its conservation in highly conducting plasmas. Magnetic energy remains the most popular integral quantity employed in analysis of magnetic fields. However, magnetic helicity has some advantages compared to energy. First, energy can convert between several physical forms – ideal (non-resistive) forces exchange energy between the magnetic field and the fluid flow, while non-ideal effects create thermal energy or accelerate non-thermal cosmic ray particles. In contrast, magnetic helicity only lives in the magnetic field; it cannot convert into other quantities. Secondly, magnetic energy can dissipate rapidly during reconnection, whereas helicity is approximately conserved, with rigorous estimates on what ‘approximately’ means. Helicity conservation will be described in detail in Sects. “Evolution of Helicity” and “Magnetic Helicity and Reconnection”. Third, the input of magnetic helicity through the boundaries of a volume can be simpler to observe than energy input. This applies particularly to flow of helicity from the interior of the sun into the solar atmosphere.

First we describe how magnetic helicity measures some simple structural features of the field. Figure 6 shows two



**Topological Magnetohydrodynamics and Astrophysics, Figure 6**

Two linked flux tubes. Tube 1 has axial flux  $\Phi_1$ , while tube 2 has flux  $\Phi_2$ . The tubes link each other twice in a right handed sense, so  $H_{12} = H_{21} = 2\Phi_1\Phi_2$ , for a total mutual helicity of  $4\Phi_1\Phi_2$ . Tube 1 has twist  $T_1 = 4$  so its self helicity is  $4\Phi_1^2$ ; tube 2 has zero net self helicity. Thus the total helicity of the configuration is  $H = 4\Phi_1\Phi_2 + 4\Phi_1^2$

linked flux tubes. As in calculations of magnetic induction, we can express the magnetic helicity as a sum of a mutual helicity between the two tubes, and the self helicities of each tube. The mutual helicity expresses the linking of the tubes (Sect. “The Gauss Linking Number”), while the self helicities measure the twist of field lines within the tubes (Sect. “Twist and Writhe”). The self helicity of a tube actually involves not only simple twisting of field lines, but also the writhe of the tube, which measures knotting and kinking of the tube axis [1,15,52].

Many applications concern a field residing in some subregion of space. In general, the field lines or flux tubes will cross the boundary of the region. Figure 10 shows hot plasma loops in the solar corona. The plasma traces out the shape of the magnetic field. (Transport of plasma across field lines is much slower than transport along a line. Also, thermal conductivity is high along the field direction. Thus density and temperature can vary from field line to field line, but are roughly uniform along each line.) Each loop has ends (called footpoints) in the surface of the sun (the photosphere). The helicity of fields which cross a boundary such as the photosphere will be discussed in Sect. “Helicity in Open Volumes”. Again we can describe the helicity as a sum of mutual and self helicities, but linking number gives way to a more general measure of entanglement.

### Helicity Integrals

We will consider the mutual linking of two vector fields  $\mathbf{V}$  and  $\mathbf{W}$  which reside in a volume  $\mathcal{D}$ . Any  $\mathbf{V}$  or  $\mathbf{W}$  lines that cross the boundary  $\partial\mathcal{D}$  of the volume form loops outside  $\mathcal{D}$ . If we are only given information about the field structure inside  $\mathcal{D}$ , we will not know the shape of these



loops. The loops outside  $\partial\mathcal{D}$  may be highly linked! The Gauss integral, if restricted to the parts of the field lines within  $\mathcal{D}$ , will not detect this external linking. In fact, it will not give a topological invariant. Section “[Helicity in Open Volumes](#)” gives a general definition for helicity integrals which eliminates this problem. For now we simply assume all field lines close within  $\mathcal{D}$ .

The helicity integral  $H(\mathbf{V}, \mathbf{W})$  measures how much the flux of  $\mathbf{V}$  links the flux of  $\mathbf{W}$  inside  $\mathcal{D}$ . The form of the integral mirrors that of the Gauss integral Eq. (1):

**Definition 4** (Closed field regions) Consider two closed divergence-free vector fields  $\mathbf{V}$  and  $\mathbf{W}$  inside a domain  $\mathcal{D}$  with boundary  $\partial\mathcal{D}$ . We assume that the field lines of  $\mathbf{V}$  and  $\mathbf{W}$  close within  $\mathcal{D}$ :

$$\mathbf{V} \cdot \hat{\mathbf{n}}|_{\partial\mathcal{D}} = \mathbf{W} \cdot \hat{\mathbf{n}}|_{\partial\mathcal{D}} = 0. \quad (18)$$

Then the *helicity integral*  $H(\mathbf{V}, \mathbf{W})$  is given by

$$H(\mathbf{V}, \mathbf{W}) \equiv \frac{1}{4\pi} \int_{\mathcal{D}} \int_{\mathcal{D}} \mathbf{V}(\mathbf{x}) \cdot \frac{\mathbf{r}}{r^3} \times \mathbf{W}(\mathbf{y}) d^3x d^3y. \quad (19)$$

We have taken Eq. (1), replaced the field line tangent vectors  $d\mathbf{x}/d\sigma$  and  $d\mathbf{y}/d\tau$  with  $\mathbf{V}$  and  $\mathbf{W}$ , and then replaced the line integrals with volume integrals. In effect, the volume integrals sum over all pairs of field lines of  $\mathbf{V}$  and  $\mathbf{W}$  (weighted by flux) [48,80]. The volume element  $d^3x$  can be regarded as the product of a differential line element  $d\sigma$  along a  $\mathbf{V}$  field line and an area element  $d^2x$  perpendicular to that line. The area element combines with the  $\mathbf{V}$  field strength to give the flux weighting.

We will be principally interested in the magnetic helicity  $H(\mathbf{B}, \mathbf{B})$ , which measures the self linking of magnetic flux. But several other quadratic quantities appearing in physics and fluid mechanics can be interpreted as helicity integrals. The kinetic helicity  $H(\boldsymbol{\omega}, \boldsymbol{\omega})$  measures the self linking of vorticity, while  $H(\boldsymbol{\omega}, \mathbf{B})$ , the cross helicity, measures the mutual linking between vorticity and magnetic flux.

The definition for  $H(\mathbf{V}, \mathbf{W})$  above is closest in form to the Gauss linking integral, but in practice would be difficult to calculate directly. Fortunately, we can simplify the integral through the use of vector potentials.

## Vector Potentials

**Definition 5** For a divergence free field  $\mathbf{V}$  inside a closed field region  $\mathcal{D}$ , we define an inverse curl operation by

$$\text{curl}^{-1}\mathbf{W}(\mathbf{x}) \equiv -\frac{1}{4\pi} \int_{\mathcal{D}} \frac{\mathbf{r}}{r^3} \times \mathbf{W}(\mathbf{y}) d^3y. \quad (20)$$

A simple calculation shows  $\nabla \times \text{curl}^{-1}\mathbf{W} = \mathbf{W}$ . The operator inverting the curl within  $\mathcal{D}$  is called the modified Biot-Savart operator [85] (the unmodified operator being integrated over all  $\mathbb{R}^3$ ). Then

$$H(\mathbf{V}, \mathbf{W}) = \int_{\mathcal{D}} \mathbf{V} \cdot \text{curl}^{-1}\mathbf{W} d^3x = \int_{\mathcal{D}} \text{curl}^{-1}\mathbf{V} \cdot \mathbf{W} d^3y. \quad (21)$$

For example, when integrated over all space, the magnetic energy  $E = \frac{1}{2\mu_0} \int B^2 d^3x$  can be written as a helicity integral, as soon as we realize that  $\mathbf{B}$  is a vector potential for the electric current  $\mathbf{J}$ ,  $\nabla \times \mathbf{B} = \mu_0\mathbf{J}$ . Thus we can write

$$E = \frac{1}{2} H(\mathbf{B}, \mathbf{J}) \quad (22)$$

i. e. the linking of magnetic flux and electric current gives twice the magnetic energy.

## Gauge-Invariance of Closed Helicity

Equation (20) gives a unique expression for the inversion of the curl operator. However, for any divergence free vector field  $\mathbf{W}$ , there exist an infinite number of vector fields  $\mathbf{A}_W$  whose curl gives  $\nabla \times \mathbf{A}_W = \mathbf{W}$ . Suppose we let

$$\mathbf{A}_W = \text{curl}^{-1}\mathbf{W} + \nabla\psi \quad (23)$$

for some gauge function  $\psi$ . Transforming a vector field by adding a zero-curl field is called a *gauge transformation*. For closed fields the helicity can readily be shown to be invariant to gauge transformations.

## Helicity in Open Volumes

So far we have considered helicity integrals inside closed volumes, where the field lines never cross the boundary. However, this requirement is far too restrictive for many physical and mathematical applications. Solar magnetic fields, for example, cross a natural boundary at the photosphere. Scientists studying the interior of the sun will see this as an outer boundary, while those who study the solar atmosphere regard the photosphere as an inner boundary. We will need a form of helicity integral which works just for the interior, or just for the exterior. Furthermore, the sum of interior and exterior helicities should sensibly relate to the helicity integral of all space. Helicity should be allowed to flow across boundaries as well, just like energy. Laboratory physicists studying confined fusion energy devices also deal with magnetic fields not completely enclosed inside the plasma. In fact, magnetic helicity can be injected into a plasma (to improve its stability) via the field lines that cross the boundary [62]. Mathematicians



also interest themselves in topological objects which do not close upon themselves (unlike knots and links). Tangles and braids, for example, involve curves with fixed end-points on a boundary [18].

Here we will consider the magnetic helicity  $H(\mathbf{B}) = H(\mathbf{B}, \mathbf{B})$  inside an arbitrary volume  $\mathcal{D}$ . We give a definition of helicity which retains topological meaning and is gauge invariant. First we divide space, and our magnetic field, into pieces inside and outside our domain of interest  $\mathcal{D}$ .

**Definition 6** Let space be divided into domains  $\mathcal{D}$  and  $\mathcal{D}'$ , containing magnetic fields  $\mathbf{B}$  and  $\mathbf{B}'$ . At the boundary  $\partial\mathcal{D}$ ,  $\mathbf{B} \cdot \hat{n} = \mathbf{B}' \cdot \hat{n}$ . The magnetic field defined in all space is

$$\{\mathbf{B}, \mathbf{B}'\}(\mathbf{x}) = \begin{cases} \mathbf{B}, & \mathbf{x} \in \mathcal{D}; \\ \mathbf{B}', & \mathbf{x} \in \mathcal{D}'. \end{cases} \quad (24)$$

Unfortunately,  $H(\{\mathbf{B}, \mathbf{B}'\})$  includes information about all the helical structure in the exterior region  $\mathcal{D}'$ . We need to subtract this extra information. Simply integrating Eq. (19) or Eq. (21) over  $\mathcal{D}$  as before will not do, as  $\mathcal{D}$  is no longer a closed field region. The integral will no longer be gauge invariant or topologically meaningful. Instead, we measure the helicity relative to a minimal base state. This procedure is similar to measuring voltage relative to ground, or potential gravitational energy relative to sea level. Thus we will look for some simple vector field  $\mathbf{P}$  inside  $\mathcal{D}$  for which we can calculate the reference helicity  $H(\{\mathbf{P}, \mathbf{B}'\})$ . Once we subtract this reference helicity, the dependence on the external field will vanish [14].

The boundary information  $\mathbf{B} \cdot \hat{n}$  tells us the distribution of flux crossing the boundary  $\partial\mathcal{D}$ . It also determines a unique vector field, the vacuum (or potential) field  $\mathbf{P}$ :

**Definition 7** The vacuum (potential) field  $\mathbf{P}$  in  $\mathcal{D}$  satisfies

$$\mathbf{P} \cdot \hat{n}|_S = \mathbf{B} \cdot \hat{n}; \quad (25)$$

$$\nabla \times \mathbf{P}(\mathbf{x}) = 0, \quad \mathbf{x} \in \mathcal{D}. \quad (26)$$

If  $\mathcal{D}$  is multiply connected, the net flux of  $\mathbf{P}$  through any closed curve on  $\partial\mathcal{D}$  should also be the same for  $\mathbf{B}$  and  $\mathbf{P}$ .

A simple variational calculation will show that the vacuum field is the minimum energy state consistent with the boundary data  $\mathbf{B} \cdot \hat{n}$ . Magnetic field lines spiral about current lines. As the vacuum field has zero current, it has minimum helical structure. It also requires a minimum of information for its specification. These qualities make it the ideal choice for the reference field.

**Definition 8** The magnetic helicity inside an arbitrary volume  $\mathcal{D}$  is given by

$$H_{\mathcal{D}} = H(\{\mathbf{B}, \mathbf{B}'\}) - H(\{\mathbf{P}, \mathbf{B}'\}). \quad (27)$$

**Theorem 4** [14,29]

1.  $H_{\mathcal{D}}$  is gauge invariant;
2.  $H_{\mathcal{D}}$  can be expressed as an integral over  $\mathcal{D}$  alone:

$$H_{\mathcal{D}} = \int_{\mathcal{V}} (\mathbf{A} + \mathbf{A}_{\mathbf{P}}) \cdot (\mathbf{B} - \mathbf{P}) d^3x \quad (28)$$

where  $\nabla \times \mathbf{A}_{\mathbf{P}} = \mathbf{P}$ .

3.  $H_{\mathcal{D}}$  is independent of the field  $\mathbf{B}'$  outside of  $\mathcal{D}$ .

### Self and Mutual Helicity

The total helicity of a collection of magnetic structures can be decomposed into sums of self helicities and mutual helicities. The self helicity of a single structure measures properties of field lines within the structure like twist and kinking. The mutual helicity between two structures measures their linking.

As an example, consider two solar coronal loops 1 and 2 with fluxes  $\Phi_1, \Phi_2$ . We will assume that the loops are uniformly twisted through  $T_1$  and  $T_2$  turns (e.g. a field line rotates about the axis of loop 1  $2\pi T_1$  times). The two loops will then have self-helicities due to their twist

$$H_{11} = T_1 \Phi_1^2; \quad H_{22} = T_2 \Phi_2^2. \quad (29)$$

However, two loops will in general also share a mutual helicity if they cross over each other or are misaligned. The total helicity of the two loops is the sum of the self and mutual helicities

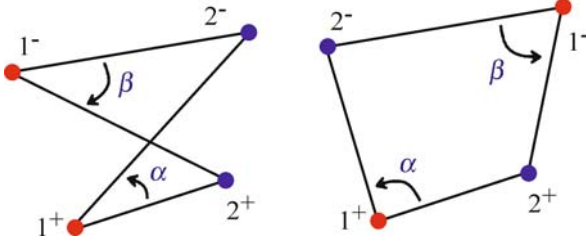
$$H = H_{11} + H_{22} + 2H_{12}. \quad (30)$$

For the moment, let the corona be the upper half space  $\{z > 0\}$ . The photospheric ( $z = 0$ ) ends of the loops 1 and 2 are assumed to be small and located at the points  $1^+, 1^-, 2^+$ , and  $2^-$ . Here  $B_z > 0$  at  $1^+$  and  $2^+$ . If the loops cross as seen from above, we assume that loop 1 is the upper loop. Consider the quadrilateral  $1_+ 2_+ 1_- 2_-$ . Let  $\alpha$  and  $\beta$  be the angles of at vertices  $1_+$  and  $1_-$ , respectively (as defined in Fig. 7). Then the helicity integral gives [7]

$$H_{12} = H_{21} = \frac{\Phi_1 \Phi_2}{2\pi} (\alpha + \beta). \quad (31)$$

Next consider  $N$  loops. In this case there are  $N$  self helicities and  $N(N-1)$  mutual helicities:

$$H = \sum_{i=1}^N H_{ii} + \sum_{i=1}^N \sum_{\substack{j=1 \\ j \neq i}}^N H_{ij}. \quad (32)$$



Topological Magnetohydrodynamics and Astrophysics, Figure 7  
Definition of the angles in Eq. (31)

Thus the total helicity  $H$  equals the sum of the entries in a matrix  $H_{ij}$ . If  $N$  is large then there will be many more mutual helicity terms. In this case ignoring the self helicities (if they are difficult to observe) may only give a small error.

### Helicity of Toroidal and Poloidal Fields

In Cartesian or spherical geometries it is often useful to decompose a magnetic field into toroidal and poloidal components. Let

$$\mathcal{L} \equiv \begin{cases} -\hat{z} \times \nabla, & \text{Cartesian geometry;} \\ -\mathbf{r} \times \nabla, & \text{Spherical geometry.} \end{cases} \quad (33)$$

Then we can write

$$\mathbf{B} = \mathcal{L}T + \nabla \times \mathcal{L}P, \quad (34)$$

where  $T$  is the toroidal function and  $P$  is the poloidal function.

**Theorem 5 [8]** Consider a magnetic field in a region  $\mathcal{D}$ . Assume that  $\mathcal{D}$  is either 1) all space, 2) a halfspace bounded by a plane, 3) a layer bounded by two planes, 4) the interior or exterior of a sphere, or 5) a spherical shell bounded by two concentric spheres. Then inside  $\mathcal{D}$

1. A purely poloidal field ( $T = 0$ ) has  $H = 0$ .
2. A purely toroidal field ( $P = 0$ ) has  $H = 0$ .
3. In general,

$$H = 2 \int_{\mathcal{D}} \mathcal{L}T \cdot \mathcal{L}P d^3x. \quad (35)$$

### Evolution of Helicity

Suppose the plasma inside  $\mathcal{D}$  has magnetic diffusivity  $\eta$  and a velocity field  $\mathbf{V}$ . Then one can derive a Poynting theorem for helicity [7]. First we will need a vector potential for  $\mathbf{P}$  which simplifies the expressions (as everything

is gauge-invariant we are allowed to choose a convenient gauge to do our calculations). Here  $\hat{\mathbf{n}}$  points out of  $\mathcal{D}$  at the boundary  $\partial\mathcal{D}$ .

**Definition 9** Let  $\mathbf{A}_P$  be the unique vector potential satisfying

$$\nabla \times \mathbf{A}_P = \mathbf{P}, \quad (36)$$

$$\nabla \cdot \mathbf{A}_P = 0, \quad (37)$$

$$\hat{\mathbf{n}} \cdot \nabla \times \mathbf{A}_P = B_n \quad \text{at } \partial\mathcal{D}, \quad (38)$$

$$\mathbf{A}_P \cdot \hat{\mathbf{n}} = 0 \quad \text{at } \partial\mathcal{D}. \quad (39)$$

**Theorem 6 [6,7,27]**

$$\begin{aligned} \frac{dH}{dt} = & -2 \int_{\mathcal{V}} \eta \mathbf{J} \cdot \mathbf{B} d^3x \\ & + 2 \oint_{\partial\mathcal{D}} ((\mathbf{A}_P \cdot \mathbf{V})\mathbf{B} - (\mathbf{A}_P \cdot \mathbf{B})\mathbf{V}) \cdot \hat{\mathbf{n}} d^2x. \end{aligned} \quad (40)$$

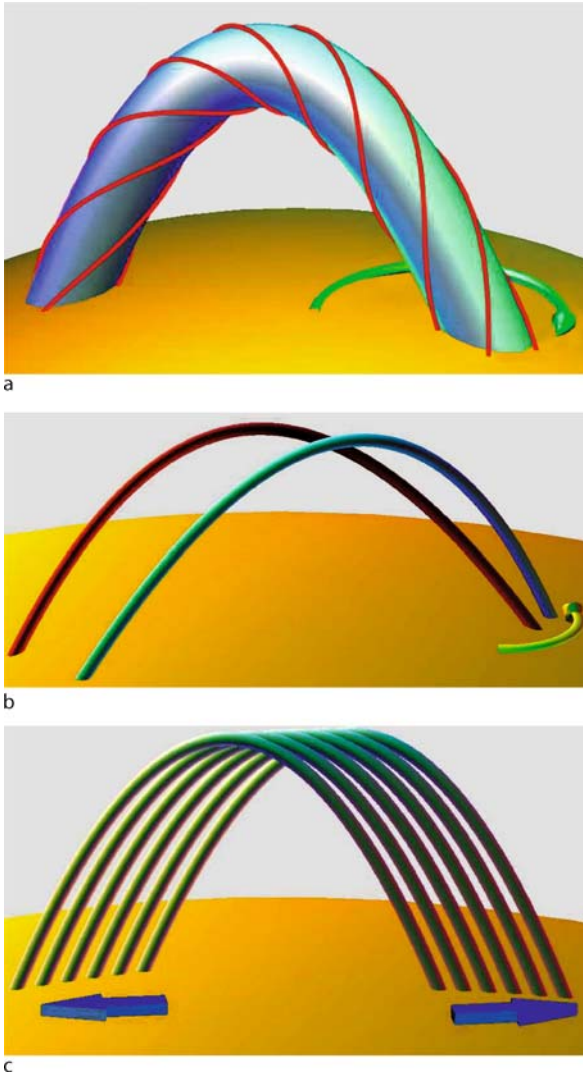
The first term gives dissipation of magnetic helicity, while the second gives flow of helicity across the boundary (see Fig. 8).

### Magnetic Helicity and Reconnection

The helicity dissipation term in Theorem 6 is generally very small. We can readily compare it to magnetic energy dissipation [7]: The magnetic energy is  $E = 1/(2\mu_0) \int B^2 d^3x$ , while magnetic energy dissipates at the rate  $dE/dt = \int \eta J^2 d^3x$ . Thus a Schwartz inequality gives

$$\left( \frac{dH}{dt} \right)^2 \leq 8\mu_0 \eta E \frac{dE}{dt}. \quad (41)$$

For astrophysical conditions this is a very sharp bound. For example, when integrated over the lifetime of a solar flare, it gives a change in helicity less than one part in  $10^5$  compared to the change in magnetic energy. While this is a rigorous inequality, in practice it may overestimate helicity loss. The inequality assumes the helicity dissipation occurs over the entire integration volume. If the dissipation is in a small subvolume (as frequently occurs in reconnection events), then the helicity loss may be substantially less than the upper bound. The relative conservation of helicity compared to energy is of importance in both fusion plasmas [67] and solar coronal fields [34]. Almost exact helicity conservation during reconnection implies that



**Topological Magnetohydrodynamics and Astrophysics, Figure 8**  
Three ways in which helicity can flow through a boundary. A single flux tube can become twisted due to rotational motions at the boundary. Two flux tubes can braid if their end points move around each other. An arcade can acquire helicity due to shearing motions at the boundary

mutual helicity can be converted into self helicity, or vice-versa [7]. For example, flux transfer events at the Earth's magnetopause transfer the mutual helicity of interplanetary and terrestrial magnetic fields into internal twist of magnetic flux ropes [72].

The concept of writhe can be extended to curves with endpoints [15,30,65]. For example, tubes of magnetic flux in the atmosphere of the sun can become highly twisted, leading to a kinking of the flux tube geometry (see Fig. 9).

In this situation the twist number of magnetic lines of force about the central axis of the tube decreases, with a corresponding increase in the writhe of the axis [15,33].

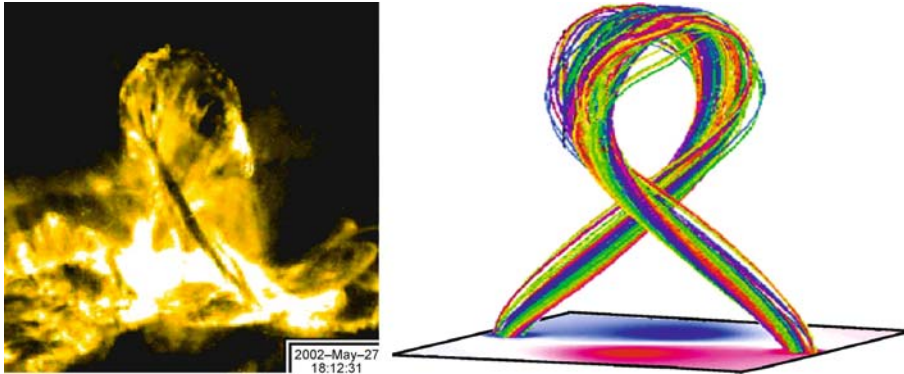
### Applications to Astrophysical Fields

Magnetic fields are ubiquitous in the universe. Stars and the interstellar medium consist predominantly of highly conducting ionized plasma. On all but the smallest scales the fluid MHD approximation works well (exceptions include the fine structure of reconnection events, the acceleration of cosmic ray particles, and the effects of neutral atoms on the plasma). Magnetic fields generate a wide range of activity [73,78,79]: the atmospheres of the sun and stars are subject to violent magnetic storms known as *flares*; large parts of the solar atmosphere can simply lift off into space as *coronal mass ejections*; magnetic fields structure stellar winds, accretion disks and galactic jets.

The solar magnetic field is thought to be generated (and regenerated every 11 years in the solar cycle) by a *dynamo* converting the kinetic energy of convection into magnetic energy [76,87]. Magnetic helicity  $H(\mathbf{B}, \mathbf{B})$  and electrical current helicity  $H(\mathbf{J}, \mathbf{J})$  moderate the growth of the magnetic field in a dynamo [19,39]. Dynamo models in which the magnetic field keeps its structure but grows in magnitude must have zero magnetic helicity; otherwise magnetic helicity would not be conserved [32,51].

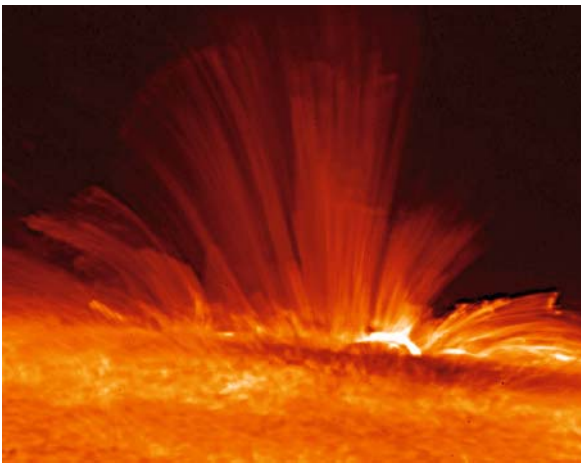
After new magnetic flux is generated in the interior of the sun, it can reach the surface (*photosphere*) via the effects of convection and buoyancy. Measurements of helicity flow across the photosphere (as illustrated in Fig. 8) can give us clues about the generation process. Present day measurements involve direct measurement of the helicity Poynting flux [22,54,59], matching best-fit linear force-free fields with observations [4], and calculating the effects of solar differential rotation on the solar magnetic field [16,26]. Structures in the corona tend to have negative magnetic helicity in the Northern hemisphere and positive helicity in the southern hemisphere [60]. See [77] for a collection of helicity studies relevant to the sun.

Magnetic field lines rising from the photosphere can either form loops in the corona (see Fig. 10) and reenter the photosphere, or wander off into the interplanetary medium. The loops provide a mathematical mapping of the surface to itself. This connectivity is invariant to ideal motions in the atmosphere, but slowly changes due to motions of field line endpoints in the photosphere. Reconnection events such as flares can rapidly change the connectivity, of course. Discontinuities in the connectivity are called *separatrices*. Two separatrix surfaces can interest in a *separator* line. Separators in turn can inter-



Topological Magnetohydrodynamics and Astrophysics, Figure 9

An erupting magnetic structure in the solar corona. On the left is an image from the TRACE mission (195 Å line). The right picture is from a numerical simulation of this event by Török and Kliem 2005 [69]



Topological Magnetohydrodynamics and Astrophysics, Figure 10

Loops of hot plasma in the solar atmosphere (*corona*), as seen by the Hinode mission solar optical telescope. As the thermal conductivity is much greater parallel to the magnetic field than across the field, individual field lines tend to have an almost uniform temperature. As a consequence, the plasma loops seen at particular temperatures trace the magnetic field lines. Dark areas also have magnetic fields, but with the wrong plasma temperature for the telescope's wavelength band. Image credit: Hinode JAXA/NASA

sect in a *magnetic neutral point*. The complex of separatrices, separators, and neutral points is called the *skeleton* of the magnetic field. Recent years have seen extensive research into the structure of the magnetic skeleton in solar fields, and its relevance to solar activity such as flares and mass ejections [20,43,68] (a more extensive review appears in [75]). Sometimes the connectivity of the field lines can have sharp gradients without actual discontinuities.

The regions of sharp gradients may still have enhanced activity [5,9,61].

The solar corona is heated to over 1 million degrees, even though the surface is only 5000 K. The likely mechanism for this heating involves many small tiny reconnection events called microflares or nanoflares. Parker [56, 57,73,74] has developed a theory of *topological dissipation* to account for this. Photospheric motions continually increase the complexity of the coronal field topology by moving the endpoints of the field lines (as in Fig. 8). This leads to an increase in minimum crossing numbers and minimum magnetic energy [11,12]. The corona can only relieve these stresses by reconnecting the field lines in flare events.

As flares only change the connectivity, but do not destroy magnetic helicity, the helicity of a coronal region can build up over time. It is possible that coronal mass ejections provide the mechanism of allowing the sun to shed excess helicity into outer space [40,45,63].

### Future Directions

Astrophysicists have developed a great body of knowledge and experience concerning the topology of potential fields (e. g. [17,20,75]). However, the topology of force-free fields [3,8,38] needs further exploration. Global field line descriptions in terms of the magnetic skeleton, topological invariants like helicity, and complexity measures like minimum crossing number have evolved separately; their interrelationships require further study.

Magnetic equilibrium studies have concentrated mostly on smooth solutions to the equilibrium equations. However, smooth solutions may not always exist or be available to an evolving astrophysical plasma [2,46,55,56,



57,58,74]. In this case current sheets will form, leading to activity like flares and coronal mass ejections. The mathematical properties of equilibrium solutions with current sheets could be a direction of fruitful future work.

Another future direction concerns detailed analyzes of the flow of helicity throughout the heliosphere, from the interior of the sun through the photosphere, into the corona, and out to magnetic clouds in the interplanetary medium, using the helicity Poynting Theorem 6. Helicity flow may also be of particular importance in dynamo studies [66].

## Bibliography

### Primary Literature

- Aldinger J, Klapper I, Tabor M (1995) Formulae for the calculation and estimation of writhe. *J Knot Theory Ram* 4:343–372
- Aly JJ (2005) A uniqueness result for a simple force-free magnetic field submitted to a topological constraint. *Astronom Astrophys* 429:15–18
- Antiochos SK (1987) The topology of force-free magnetic fields and its implications for coronal activity. *Astrophys J* 312: 886–894
- Aulanier G, Srivastava N, Martin SF (2000) Model Prediction for an Observed Filament. *Astrophys J* 543:447–456
- Aulanier G, Parlat E, Demoulin P (2005) Current sheet formation in quasiseparatrix layers and hyperbolic flux tubes. *Astronom Astrophys* 444:961–976
- Barnes DC (1988) Mechanical injection of magnetic helicity. *Phys Fluids* 31:2214–2220
- Berger MA (1984) Rigorous new limits on magnetic helicity dissipation in the solar corona. *Geophys Astrophys Fluid Dyn* 30:79–104
- Berger MA (1985) Structure and stability of constant  $\alpha$  force-free fields. *Astrophys J* 59:433–444
- Berger MA (1989) Three-dimensional reconnection from a global viewpoint. In: Guyenne TD, Hunt JJ (eds) *Reconnection in Space Plasma*. ESA Paris SP-285 II:83–86
- Berger MA (1990) Third order link integrals. *J Phys A: Math Gen* 23:2787–2793
- Berger MA (1993) Energy-crossing number relations for braided magnetic fields. *Phys Rev Lett* 70:705–708
- Berger MA (1994) Coronal Heating by Dissipation of Magnetic Structure. *Space Sci Rev* 68:3–14
- Berger MA (2001) Topological invariants in braid theory. *Lett Math Phys* 55:181–192
- Berger MA, Field GB (1984) The topological properties of magnetic helicity. *J Fluid Mech* 147:133–148
- Berger MA, Prior P (2006) The writhe of open and closed curves. *J Phys A: Math Gen* 39:8321–8348
- Berger MA, Ruzmaikin A (2000) Magnetic helicity production by differential rotation. *J Geophys Res* 105:10481–10490
- Beveridge C, Longcope DW (2005) On Three-Dimensional Magnetic Skeleton Elements due to Discrete Flux Sources. *Solar Phys* 227:193–206
- Birman JS (1991) Recent Developments In Braid And Link Theory *Math Intell* 13:52–60
- Blackman EG, Field GB (2000) Constraints on the Magnitude of  $\alpha$  in Dynamo Theory. *Astrophys J* 534:984–988
- Brown DS, Priest ER (2001) The topological behaviour of 3D null points in the Sun's corona. *Astronom Astrophys* 367: 339–346
- Călugăreanu (1961) On isotopy classes of three-dimensional knots and their invariants. *Czechoslovak Math J* T11:588–598
- Chae J (2001) Observational Determination of the Rate of Magnetic Helicity Transport through the Solar Surface via the Horizontal Motion of Field Line Footpoints. *Astrophys J* 560: L95–L98
- Chui AYY, Moffatt HK (1995) The energy and helicity of knotted magnetic flux tubes. *Phil Trans R Soc London A* 451:609–618
- Deng XH, Matsumoto H (2001) Rapid magnetic reconnection in the Earth's magnetosphere mediated by whistler waves. *Nature* 410:557–560
- Dennis MR, Hannay JH (2005) Geometry of Calugareanu's theorem. *Proc Roy Soc A* 2062:3245–3254
- DeVore CR (2000) Magnetic helicity generation by solar differential rotation. *Astrophys J* 539:944–953
- Dixon A, Berger MA, Browning P, Priest ER (1989) A generalization of the Woltjer minimum energy principle. *Astronom Astrophys* 225:156–166
- Eppele M (1998) Orbits of asteroids, a braid, and the first link invariant. *Math Intell* 20:45–52
- Finn J, Antonsen TM (1985) Magnetic helicity: what is it, and what is it good for? *Comments Plasma Phys Contr Fusion* 9:111–120
- Fuller FB (1978) Decomposition of the linking of a ribbon: a problem from molecular biology. *Proc Natl Acad Sci USA* 75:3557–3561
- Ghrist R, Komendarczyk R (2006) Overtwisted energy-minimizing curl eigenfields. *Nonlinearity* 19:41–52
- Gilbert AD (2002) Magnetic Helicity in Fast Dynamos. *Geophys Astrophys Fluid Dyn* 96:135–151
- Green LM, Kliem B, Török T, van Driel-Gesztelyi L, Attrill GDR (2007) Transient Coronal Sigmoids and Rotating Erupting Flux Ropes. *Sol Phys* 246:365–391
- Heyvaerts J, Priest ER (1984) Coronal heating by reconnection in DC current systems – a theory based on Taylor hypothesis. *Astronom Astrophys* 137:63–78
- Hornig G (1997) The covariant transport of electromagnetic fields and its relation to magnetohydrodynamics. *Phys Plasmas* 4:646–654
- Hornig G, Mayer C (2002) Towards a third-order topological invariant for magnetic fields. *J Phys A: Math Gen* 35:3945–3959
- Hornig G, Schindler K (1996) Magnetic topology and the problem of its invariant definition. *Phys Plasmas* 3:781–791
- Hudson TS, Wheatland MS (1999) Topological Differences Between Force-Free Field Models. *Sol Phys* 186:301–310
- Ji HT (1999) Turbulent dynamos and magnetic helicity. *Phys Rev Lett* 83:3198–3201
- Kumar A, Rust DM (1996) Interplanetary magnetic clouds, helicity conservation, and current-core flux ropes. *J Geophys Res* 101:15667–15684
- Lau YT, Finn JM (1990) Three-dimensional kinematic reconnection in the presence of field nulls and closed field lines. *Astrophys J* 350:672–691
- Laurence P, Stredulinsky E (2000) Asymptotic Massey products, induced currents and Borromean torus links. *J Math Phys* 41:3170–3191



43. Longcope DW, Klapper I (2002) A general theory of connectivity and current sheets in coronal magnetic fields. *Astrophys J* 579:468–481
44. Low BC (1990) Equilibrium and dynamics of coronal magnetic fields. *Ann Rev Astron Astrophys* 28:491–524
45. Low BC (1994) Magnetohydrodynamic processes in the solar corona – flares, coronal mass ejections, and magnetic helicity. *Phys Plasmas* 1:1684–1690
46. Low BC, Wolfson R (1988) Spontaneous formation of electric current sheets and the origin of solar flares. *Astrophys J* 324:574–581
47. Massey WS (1998) Higher order linking numbers. *J Knot Theory Ram* 7:393–414
48. Moffatt HK (1969) The degree of knottedness of tangled vortex lines. *J Fluid Mech* 35:117–129
49. Moffatt HK (1985) Magnetostatic equilibria and analogous Euler flows of arbitrarily complex topology, Part 1: Fundamentals. *J Fluid Mech* 159:359–378
50. Moffatt HK (1990) The energy spectrum of knots and links. *Nature* 347:367–369
51. Moffatt HK, Proctor MRE (1985) Topological constraints associated with fast dynamo action. *J Fluid Mech* 154:493–507
52. Moffatt HK, Ricca RL (1992) Helicity and the Călugăreanu invariant. *Proc Royal Society London A* 439:411–429
53. Monastyrski MI, Sasarov PV (1987) Topological invariants in magnetohydrodynamics. *Sov Phys JETP* 66:683–688
54. Moon YJ, Chae J, Choe GS et al. (2002) Flare activity and magnetic helicity injection by photospheric horizontal motions. *Astrophys J* 574:1066–1073
55. Ng CS, Bhattacharjee A (1998) Nonequilibrium and current sheet formation in line-tied magnetic fields. *Phys Plasmas* 5:4028–4040
56. Parker EN (1972) Topological dissipation and the small-scale fields in turbulent gases. *Astrophys J* 174:499–510
57. Parker EN (1983) Magnetic neutral sheets in evolving fields, II – Formation of the solar corona. *Astrophys J* 264:642–647
58. Parker EN (2004) Tangential discontinuities in untidy magnetic topologies. *Phys Plasmas* 11:2328–2332
59. Parlat E, Nindos A, Démoulin P, Berger MA (2006) What is the spatial distribution of magnetic helicity injected in a solar active region? *Astronom Astrophys* 52:623–630
60. Pevtsov AA, Canfield RC, Metcalf TR (1995) Latitudinal variation of helicity of photospheric magnetic fields. *Astrophys J* 440:L109–L112
61. Priest ER, Démoulin P (1995) Three-dimensional magnetic reconnection without null points, 1: Basic theory of magnetic flipping. *J Geophys Res* 100(A9):23443–23464
62. Redd AJ, Jarboe TR, Hamp WT et al (2007) Overview of the Helicity Injected Torus (HIT) program. *J Fusion Energy* 26: 163–168
63. Rust DM (1994) Spawning and shedding helical magnetic fields in the solar atmosphere. *Geophys Res Lett* 21:241–244
64. Ruzmaikin A, Akhmetiev P (1994) Topological invariants of magnetic fields, and the effects of reconnection. *Phys Plasmas* 1:331–336
65. Starostin EL (2005) On the writhing number of a non-closed curve. In: Calvo J, Millett K, Rawdon E, Stasiak A (eds) *Physical and numerical models in knot theory including applications to the life sciences. Series on Knots and Everything*. World Scientific Publishing, Singapore, pp 525–545
66. Subramanian K, Brandenburg A (2006) Magnetic Helicity Density And Its Flux In Weakly Inhomogeneous Turbulence. *Astrophys J* 648:L71–L74
67. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. *Phys Rev L* 33:1139–1141
68. Titov VS, Hornig G, Démoulin P (2002) Theory of magnetic connectivity in the solar corona. *J Geophys Res Space Phys* 107:SSH3-1
69. Török T, Kliem B (2005) Confined and ejective eruptions of kink-unstable flux ropes. *Astrophys J* 630:L97–L100
70. Wilmot-Smith AL, Priest ER, Hornig G (2005) Magnetic diffusion and the motion of field lines. *Geophys Astrophys Fluid Dyn* 99:177–197
71. Woltjer L (1958) A theorem on force-free magnetic fields. *Proc Natl Acad Sci USA* 44:489–491
72. Wright A, Berger MA (1989) The effect of reconnection upon the linkage and interior structure of magnetic flux tubes. *J Geophys Res* 94:1295–1302
73. Parker EN (1979) *Cosmical Magnetic Fields: Their Origin and Their Activity*. International Series of Monographs on Physics, Clarendon Press, Oxford University Press, Oxford
74. Parker EN (1994) *Spontaneous Current Sheets in Magnetic Fields: With Applications to Stellar X-Rays*. International Series on Astronomy and Astrophysics, vol 1. Oxford University Press, New York
75. Longcope DW (2005) Topological Methods for the Analysis of Solar Magnetic Fields. *Living Reviews of Solar Physics* 2:7 Max-Planck-Gesellschaft, Katlenburg-Lindau
76. Brandenburg A, Subramanian K (2005) Astrophysical magnetic fields and nonlinear dynamo theory. *Phys Rep* 417:1–205
77. Brown MR, Canfield RC, Pevtsov AA (eds) (1999) *Magnetic Helicity in Space and Laboratory Plasmas*. Geophysical Monograph 111, American Geophysical Union, Washington
78. Aschwanden M (2006) *Physics of the Solar Corona*. Springer, New York
79. Choudhuri AR (1998) *The Physics of Fluids and Plasmas: An Introduction for Astrophysicists*. Cambridge University Press, Cambridge
80. Arnold VI, Khesin BA (1998) *Topological Methods in Hydrodynamics*. Springer, New York
81. Bateman G (1978) *MHD Instabilities*. MIT Press, Cambridge
82. Priest ER (1984) *Solar Magnetohydrodynamics*. Springer, Berlin
83. Priest ER, Forbes TG (1999) *Magnetic Reconnection: MHD Theory and Applications*. Cambridge University Press, Cambridge
84. Biskamp D (1997) *Nonlinear Magnetohydrodynamics*. Cambridge Monographs on Plasma Physics 1, Cambridge
85. Cantarella J, DeTurck D, Gluck H (2001) The Biot-Savart operator for application to knot theory, fluid dynamics, and plasma physics. *J Math Phys* 42:876–905
86. Jackson JD (1998) *Classical Electrodynamics*. Wiley, New York
87. Moffatt HK (1978) *Magnetic Field Generation in Electrically Conducting Fluids*. Cambridge Univ Press, Cambridge

## Books and Reviews

- Kulsrud RM (2004) *Plasma Physics for Astrophysics*. Princeton University Press, Princeton
- Moffatt HK (1983) Transport effects associated with turbulence, with particular attention to the influence of helicity. *Rep Prog Phys* 46:621–664

- Moffatt HK, Tsinober A (1992) Helicity in laminar and turbulent flow. *Ann Rev Fluid Mech* 24:281–312
- Parker EN (2007) *Conversations on Electric and Magnetic Fields in the Cosmos*. Princeton University Press, Princeton

## Traffic Breakdown, Probabilistic Theory of

BORIS S. KERNER<sup>1</sup>, SERGEY L. KLENOV<sup>2</sup>

<sup>1</sup> Group Research GR/ETI, HPC: G021, Daimler AG, Sindelfingen, Germany

<sup>2</sup> Moscow Institute of Physics and Technology, Department of Physics, Dolgoprudny, Russia

### Article Outline

Glossary

Definition of the Subject

Introduction

Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow

Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model

Probabilistic Description of Traffic Breakdown Based on Master Equation

Capacity of Free Flow at Bottlenecks

Conclusions

Future Directions

Acknowledgments

Bibliography

### Glossary

**Traffic breakdown** Traffic breakdown is the onset of congested traffic in an initial free traffic flow. Traffic breakdown occurs mostly at effectual road bottlenecks like on- and off-ramps, roadworks, road gradients, reduction of road lanes, etc. Traffic breakdown results in the emergence of the synchronized flow phase of congested traffic, i. e., traffic breakdown is a phase transition from the free flow traffic phase to synchronized flow traffic phase ( $F \rightarrow S$  transition for short). Thus the terms *traffic breakdown* and an  $F \rightarrow S$  transition are synonyms related to the same phenomenon of the onset of congestion in free flow.

**Effectual bottleneck** An effectual bottleneck is a freeway bottleneck at which an  $F \rightarrow S$  transition (traffic breakdown) occurs during many days and years of observations. Because only effectual bottlenecks are considered in the article, the term *bottleneck* is used below for an effectual bottleneck.

**Nucleation and probabilistic features of traffic breakdown** Traffic breakdown is a *local first-order*  $F \rightarrow S$  transition, which exhibits a nucleation feature:

- (i) At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.
- (ii) Onset and dissolution of congestion are accompanied by a *hysteresis* effect.
- (iii) Traffic breakdown exhibits the following *probabilistic* features: (a) at the same flow rates in some of the realizations (days) traffic breakdown occurs but in other realizations it does not occur; (b) empirical probability of the breakdown is a growing function of the flow rate downstream of the bottleneck;
- (iv) In a realization, the flow rate at which traffic breakdown is observed can be considerably smaller than flow rates at which no traffic breakdown has occurred before.

**Nucleation model for traffic breakdown** In three-phase traffic theory, it is suggested that a bottleneck introduces a speed and density disturbance that is spatially limited: at the same flow rates, the average speed is lower and vehicle density is greater within the disturbance. The disturbance is on average motionless and it is localized in a neighborhood of the bottleneck. The total number of vehicles within the disturbance can be considered a local vehicle cluster (cluster for short). Traffic breakdown occurs when a critical cluster appears whose subsequent growth leads to synchronized flow emergence.

### Definition of the Subject

Traffic breakdown is the onset of congested traffic in an initial free traffic flow. Thus traffic breakdown restricts free flow conditions in vehicular traffic. In observations, traffic breakdown exhibits the probabilistic nature. For this reason, the development of the probabilistic theory of traffic breakdown is of a great importance for all kind of traffic management and control in transportation networks and for other traffic engineering applications.

### Introduction

The onset of congestion in an initial free flow is accompanied by a sharp decrease in average vehicle speed in the free flow to a considerably lower speed in congested traffic. This speed breakdown occurs mostly at freeway bottlenecks and is called the breakdown phenomenon or traffic breakdown (see, e. g., [5,10,11,31,35]). The traffic breakdown occurs mostly at effectual road bottlenecks like

on- and off-ramps, roadworks, road gradients, reduction of road lanes, etc. An effectual bottleneck is a bottleneck where traffic breakdown most frequently occurs on many different days.

It has been found that the breakdown phenomenon has a probabilistic nature [5,35]. The probabilistic nature of traffic breakdown means the following: at a given flow rate traffic breakdown can occur but it should not necessarily occur. This means that on one day the onset of congestion occurs, however, on another day at the same flow rate and at the same traffic conditions the onset of congestion is not observed.

There are a huge number of publications devoted to theoretical investigations of traffic breakdown (see references in [11,17,30,31]). However, the puzzle of empirical features of traffic breakdown has been understood only recently [14,16]. Consequently, as explained in ► [Traffic Congestion, Modeling Approaches to](#) in this Encyclopedia, earlier traffic flow theories and models reviewed in [2,3,4,8,9,12,26,30,31,32,34,36,37,38] cannot explain and predict the set of the fundamental features of traffic breakdown.

For these reasons, Kerner introduced an alternative traffic flow theory called three-phase traffic theory, which overcomes the disadvantages of other traffic flow models and theories in the explanation of spatiotemporal empirical features of traffic breakdown [17]. For this reason, the three-phase traffic theory is the basis for the probabilistic theory of traffic breakdown, which explains empirical spatiotemporal probabilistic features of traffic breakdown.

In three-phase traffic theory, besides the free flow traffic phase there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus there are three traffic phases in this theory: 1. Free flow. 2. Synchronized flow. 3. Wide moving jam. The synchronized flow and wide moving jam phases in congested traffic are defined through spatiotemporal empirical criteria. The definition of the wide moving jam phase [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. The definition of the synchronized flow phase [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

It turns out that traffic breakdown is a local phase transition from the free flow traffic phase to synchronized flow traffic phase ( $F \rightarrow S$  transition) [17]. Thus the terms traffic breakdown, breakdown phenomenon, speed breakdown

and an  $F \rightarrow S$  transition are synonyms that are related to the same phenomenon of the onset of congestion in free flow.

Reviews of empirical features of traffic breakdown and resulting traffic congested patterns as well as modeling approaches to traffic congestion have already been made in this Encyclopedia (► [Traffic Congestion, Modeling Approaches to](#), ► [Traffic Congestion, Spatiotemporal Features of](#)). In this article, we discuss approaches to probabilistic description of traffic breakdown.

The article is organized as follows. In Sect. “[Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow](#)”, a qualitative discussion of a nucleation model for traffic breakdown, which can explain fundamental empirical features of traffic breakdown, is made. Cellular automata probabilistic description of traffic breakdown is reviewed in Sect. “[Probabilistic Description of Traffic Breakdown with Cellular Automata \(CA\) Traffic Flow Model](#)”. In Sect. “[Probabilistic Description of Traffic Breakdown Based on Master Equation](#)” based on the nucleation model of Sect. “[Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow](#)”, we consider a probabilistic description of traffic breakdown with master equation. A link between traffic breakdown, freeway capacity, and the diagram of congested patterns at bottlenecks is considered in Sect. “[Capacity of Free Flow at Bottlenecks](#)”.

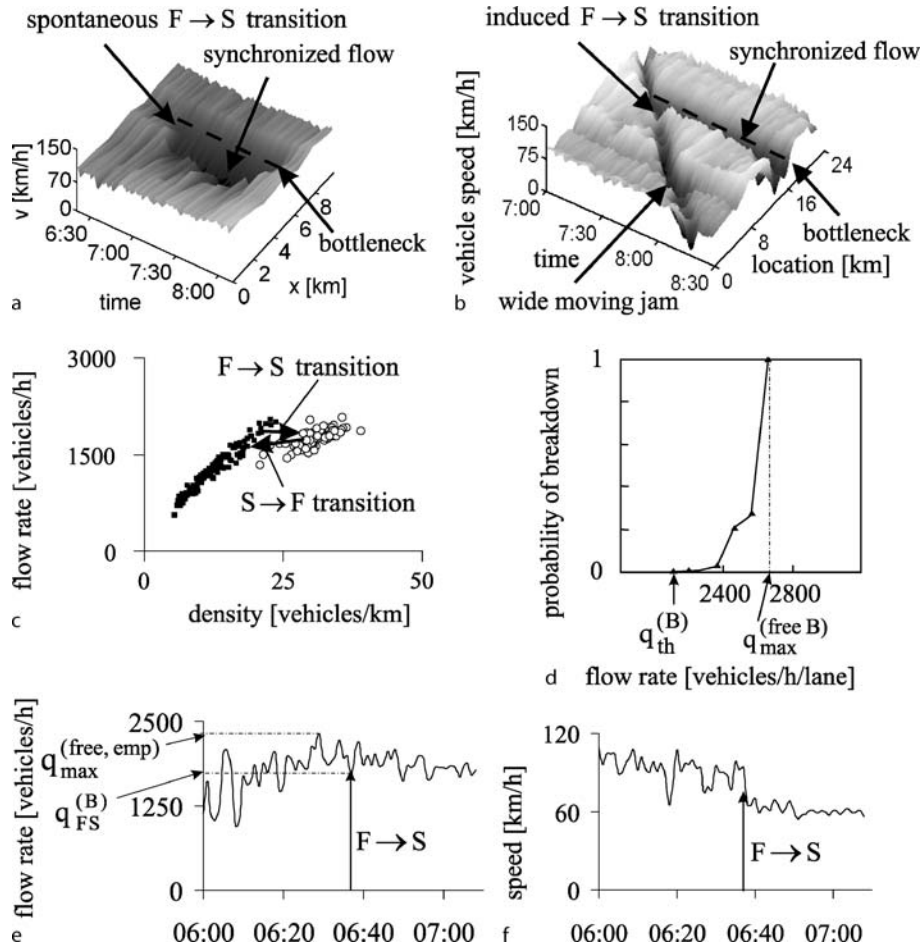
## Traffic Breakdown – First-Order Phase Transition from Free Flow to Synchronized Flow

### Fundamental Empirical Features of Traffic Breakdown

Empirical features of traffic breakdown ( $F \rightarrow S$  transition) and their explanation have already been considered in ► [Traffic Congestion, Modeling Approaches to](#) in this Encyclopedia. For this reason, we make here only a brief discussion of the fundamental empirical features of this phenomenon, which are as follows (Fig. 1):

- A. Traffic breakdown is a local phase transition from the free flow phase to the synchronized flow phase ( $F \rightarrow S$  transition).
- B. At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.
- C. Onset and dissolution of congestion are accompanied by a *hysteresis* effect.
- D. Traffic breakdown exhibits the *probabilistic* nature.

The fundamental feature **A** has already been explained above. Feature **B**. In empirical observations, the spontaneous traffic breakdown (spontaneous  $F \rightarrow S$  transition) occurs within an initial free flow at a bottleneck due to an *internal* speed disturbance in the free flow in a neighbor-



**Traffic Breakdown, Probabilistic Theory of, Figure 1**

Fundamental empirical features of traffic breakdown ( $F \rightarrow S$  transition): a, b Spontaneous (a) and induced (b) traffic breakdowns; vehicle speed in space and time; taken from [17]. c Hysteresis effect associated with a (arrows labeled  $F \rightarrow S$  and  $S \rightarrow F$  show traffic breakdown and  $S \rightarrow F$  transition at bottleneck, respectively). d Probability of traffic breakdown as function of flow rate downstream of bottleneck; 10 min average data; taken from [35]. e, f Flow rate (e) and speed (f) at bottleneck as time-function before and after breakdown (arrows labeled  $F \rightarrow S$  show the time instant of the breakdown). Taken from [17]

hood of the bottleneck (Fig. 1a). In contrast, the induced traffic breakdown (induced  $F \rightarrow S$  transition) is caused by a short-time *external* speed disturbance in traffic flow in a neighborhood of the bottleneck. This external disturbance is usually related to the propagation of a moving spatiotemporal congested pattern that initially occurs at a *different* freeway location than that of the induced  $F \rightarrow S$  transition. The induced  $F \rightarrow S$  transition can occur, after the congested pattern has reached the bottleneck. In particular, the breakdown can be induced by a wide moving jam propagating upstream through a bottleneck (Fig. 1b). Induced  $F \rightarrow S$  transition at a bottleneck can also occur when a region of synchronized flow first occurs downstream of this bottleneck, and the region later reaches the

bottleneck due to the upstream propagation of synchronized flow (see an empirical example in ► [Traffic Congestion, Modeling Approaches to](#)).

Feature C. The  $F \rightarrow S$  transition, which leads to congested pattern emergence, and the  $S \rightarrow F$  transition, which leads to the dissolution of this congested pattern, are accompanied by a well-known *hysteresis effect* and hysteresis loop in the flow-density plane (see references in [10,11]): a congested pattern emerges usually at a greater flow rate downstream of the bottleneck than this flow rate is at which the congested pattern dissolves (Fig. 1a,c).

Feature D. The probability of traffic breakdown at a bottleneck is an increasing function of flow rate (Fig. 1d)



[35]. At given traffic parameters (weather, etc.), the maximum flow rate downstream of an on-ramp bottleneck (denoted  $q_{\max}^{(\text{free,emp})}$ ), which was measured on a specific day before congestion occurred, can be greater than the flow rate (denoted  $q_{\text{FS}}^{(\text{B})}$ ) at which traffic breakdown occurs (Fig. 1e,f).

### Model of Traffic Breakdown at Highway Bottleneck

In three-phase traffic theory [17], the fundamental empirical features of traffic breakdown are explained as follows. The possibility of induced and spontaneous traffic breakdowns at the bottleneck (Fig. 1a,b) can be explained by the *nucleation* nature of breakdown phenomenon. This means that there should be some critical speed in free flow required for traffic breakdown. The breakdown occurs, if due to a speed disturbance in free flow in the neighborhood of a bottleneck the speed decreases below the critical one. Otherwise, if free flow speed within the disturbance is greater than the critical one, the breakdown does not occur.

The critical speed within a speed disturbance required for the breakdown should depend on the flow rate downstream of the bottleneck. The smaller the flow rate, the lower the critical speed required for the breakdown in free flow. In contrast, when the flow rate increases, the critical speed at the bottleneck required for the breakdown increases. Obviously, the lower the critical speed, the smaller the probability for the breakdown. This explains the increasing character of the probability of traffic breakdown on the flow rate (Fig. 1d).

In accordance with other systems of natural science, the disturbance in free flow at the bottleneck within which the critical speed is reached can be called a critical speed disturbance or a *nucleus* for traffic breakdown. Only if a disturbance appears randomly at the bottleneck within which the speed is equal or lower than the speed within the critical disturbance, traffic breakdown occurs. In other words, the probability for traffic breakdown is the probability of random critical speed disturbance appearance at the bottleneck. These nucleation features of traffic breakdown are general ones for first-order phase transitions observed in a diverse variety of physical, chemical, biological and other complex spatiotemporal systems.

Thus, the empirical features A-D of traffic breakdown mean that the breakdown is a *first-order*  $F \rightarrow S$  transition. The flow rate in free flow downstream of the bottleneck should be a control parameter for this phase transition: at the same other traffic characteristics, the greater the flow rate, the greater the probability for the  $F \rightarrow S$  transition. For this reason, there is the maximum flow rate

(denoted  $q_{\max}^{(\text{free B})}$  and that is shown in Fig. 1d) downstream of the bottleneck at which breakdown probability denoted by  $P_{\text{FS}}^{(\text{B})}$  reaches one, i. e.,  $P_{\text{FS}}^{(\text{B})} = 1$ . There is also a threshold flow rate (denoted  $q_{\text{th}}^{(\text{B})}$ ). At the flow rate in free flow downstream of the bottleneck that is smaller than  $q_{\text{th}}^{(\text{B})}$ , breakdown probability  $P_{\text{FS}}^{(\text{B})} = 0$ . Both  $q_{\max}^{(\text{free B})}$  and  $q_{\text{th}}^{(\text{B})}$  are mean values, which found in many different realizations (days) at which traffic breakdowns have occurred.

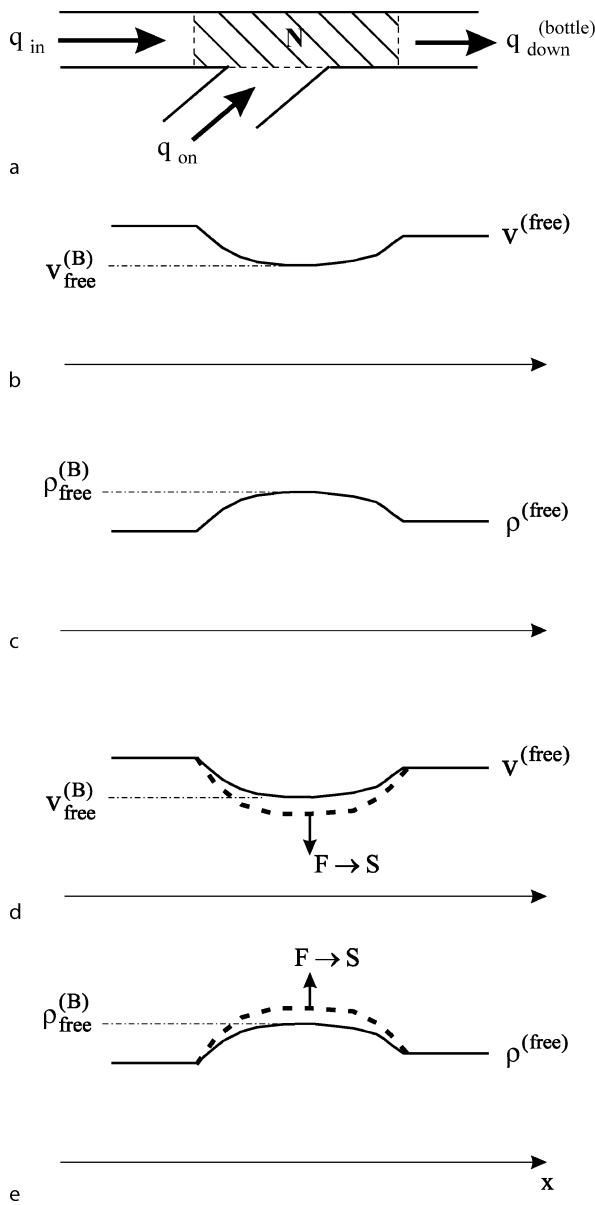
**Model of Local Vehicle Cluster at Bottleneck** In a nucleation model of traffic breakdown based on three-phase traffic theory [15,17,20,21,22], we assume that the breakdown is associated with the existence of an *initial non-homogeneity* of free flow at the bottleneck. Due to this non-homogeneity, the speed is lower and density is greater in a neighborhood of the bottleneck than outside the bottleneck. This local speed and density disturbance at the bottleneck can be considered a local *vehicle cluster* in free flow at the bottleneck. Traffic breakdown probability within the cluster should be considerably greater than outside the cluster. This should explain why traffic breakdown is observed mostly at bottlenecks.

The non-homogeneity in free flow at the bottleneck, i. e., the vehicle cluster at the bottleneck exists even in a hypothetical case in which there were *no random fluctuations* in traffic flow. In this hypothetical case, this permanent non-homogeneity can be considered a *deterministic* vehicle cluster (a *deterministic* local disturbance) localized at the bottleneck.

Let us firstly discuss an on-ramp bottleneck. In this case, the non-homogeneity of an initial free flow at the bottleneck is caused by two flows, which merge at the bottleneck: (i) An on-ramp inflow with the rate  $q_{\text{on}}$ . (ii) A flow on the main road upstream of the bottleneck with the rate  $q_{\text{in}}$ . This flow merging occurs permanent and on the same freeway location (within an on-ramp merging region). For this reason, the non-homogeneity of free flow at the bottleneck is on average motionless and permanent (Fig. 2).

To explain features of this non-homogeneity, note that at a given high enough flow rate  $q_{\text{in}}$  in free flow on the main road upstream of the bottleneck, vehicles that merge from the on-ramp onto the main road force the vehicles on the main road to decelerate in the vicinity of an on-ramp merging region. This leads to a local decrease in speed and consequently to a local increase in density, i. e. to vehicle cluster occurrence and existence in free flow in the vicinity of the bottleneck. One of the characteristics of the cluster is the cluster size, i. e., the *total* number of vehicles (denoted by  $N$ ) within the cluster location (Fig. 2a).





**Traffic Breakdown, Probabilistic Theory of, Figure 2**

Qualitative explanations of vehicle cluster in free flow at on-ramp bottleneck: **a** Qualitative schema of on-ramp bottleneck with a vehicle cluster (dashed region) localized at the bottleneck. **b, c** Qualitative spatial speed (**b**) and density (**c**) distributions within deterministic vehicle cluster (deterministic local disturbance). **d, e** Dashed curves show qualitative spatial speed (**d**) and density (**e**) distributions within a vehicle cluster (local disturbance) at a fixed time instant for the case when fluctuations in free flow lead to an increase in the cluster size  $N$  in comparison with the size of the deterministic cluster  $N^{(determ)}$ . Arrows labeled  $F \rightarrow S$  symbolize traffic breakdown that occurs when a critical cluster appears

It must be noted that here and below, the flow rates  $q_{in}$ , and  $q_{on}$ , as well as all other flow rates used in this article below, are total flow rates across associated road locations, which are divided by the number of lanes on the main road downstream of the bottleneck. As the flow rates, the cluster size  $N$  is also considered per a highway lane: the total vehicle number of vehicles within the cluster is divided by the same number of lanes on the main road downstream of the bottleneck as that used for flow rate unit definition.

As mentioned above, in the hypothetical free flow at a bottleneck in which there were no random fluctuations, a deterministic cluster occurs at the bottleneck. We denote the size of this deterministic cluster by  $N = N^{(determ)}$ . The speed  $v_{free}^{(B)}$  and density  $\rho_{free}^{(B)}$  within this deterministic cluster (Fig. 2b,c) correspond to the conditions

$$v_{free}^{(B)} < v^{(free)}, \quad \rho_{free}^{(B)} > \rho^{(free)}, \quad (1)$$

where  $v^{(free)}$  and  $\rho^{(free)}$  are the average vehicle speed and density in homogeneous free flow on the main road downstream of the cluster (Fig. 2). Because the deterministic cluster is permanent and motionless, for the on-ramp bottleneck at given  $q_{in}$  and  $q_{on}$ , the flow rate

$$q_{sum} = q_{in} + q_{on} \quad (2)$$

does not depend on the spatial co-ordinate in the *hypothetical* state of the free flow. For this reason, the speed and density on the main road satisfy the following condition:

$$q_{sum} = v^{(free)} \rho^{(free)} = v_{free}^{(B)} \rho_{free}^{(B)}. \quad (3)$$

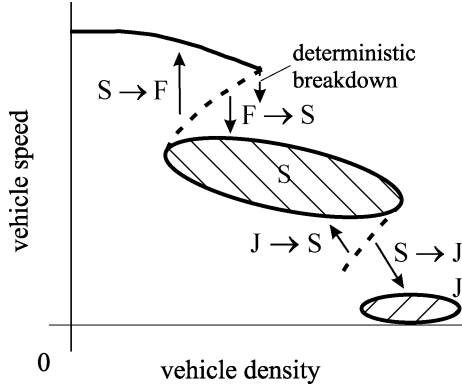
The flow rate  $q_{sum}$  downstream of the bottleneck in the state of free flow under consideration depends on the bottleneck type. In the cases of off-ramp and merge bottlenecks (a merge bottleneck is a bottleneck caused by a decrease in the number of highway lanes in the flow direction), we get

$$q_{sum} = q_{in}. \quad (4)$$

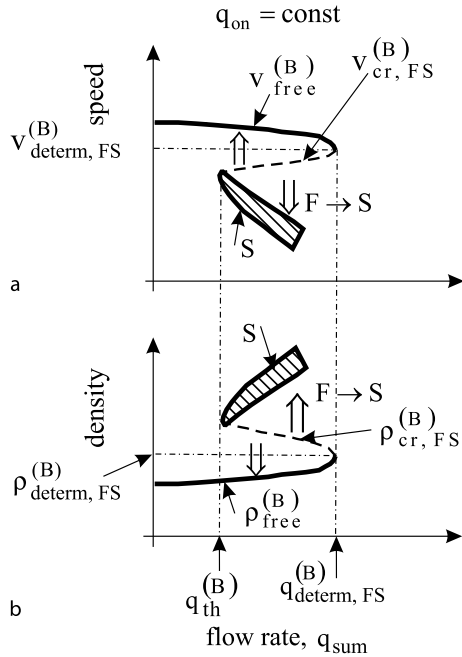
Under condition (4), the formula (1) and (3) are also valid.

### Explanation of Traffic Breakdown Within Vehicle Cluster

To explain traffic breakdown within the vehicle cluster, we use the hypothesis of three-phase traffic theory about a double Z-speed-density characteristic (Fig. 3) [17], [Traffic Congestion, Modeling Approaches to](#). In this theory, the first Z-speed-density characteristic between states  $F$  and  $S$  explains traffic breakdown (arrow  $F \rightarrow S$  in Fig. 3). In accordance with the double Z-characteristic (Fig. 3), at a given value  $q_{on}$  there should be non-



**Traffic Breakdown, Probabilistic Theory of, Figure 3**  
Qualitative double Z speed-density characteristic for phase transitions of three-phase traffic theory. Taken from [17], ► [Traffic Congestion, Modeling Approaches to](#)



**Traffic Breakdown, Probabilistic Theory of, Figure 4**  
Qualitative shapes of speed (a) and density (b) dependences on the flow rate  $q_{\text{sum}}$  associated with the location of vehicle cluster

monotonous dependence of speed on the flow rate  $q_{\text{sum}}$  at the location of vehicle cluster (Fig. 4a); this speed-flow characteristic leads to a S-shaped density dependence on this flow rate (Fig. 4b).

When  $q_{\text{on}}$  is a given value and the flow rate  $q_{\text{sum}}$  increases due to an increase in  $q_{\text{in}}$ , the vehicle speed in free flow within the deterministic cluster  $v_{\text{free}}^{(B)}$  decreases and in accordance with (3) the associated density  $\rho_{\text{free}}^{(B)}$  increases.

However, this increase is limited by some critical density  $\rho_{\text{free}}^{(B)} = \rho_{\text{determ,FS}}^{(B)}$  (Fig. 4b) within the deterministic cluster associated with a critical flow rate

$$q_{\text{sum}} = q_{\text{determ,FS}}^{(B)} \quad (5)$$

After this critical deterministic cluster is reached, the further increase in  $q_{\text{sum}}$  should lead to *deterministic traffic breakdown* at the bottleneck causing spontaneous synchronized flow emergence at the bottleneck. The critical deterministic cluster can be considered *deterministic nucleus* for traffic breakdown. After the critical deterministic cluster is reached, deterministic traffic breakdown occurs at the bottleneck even if there were no random fluctuations in free flow at the bottleneck.

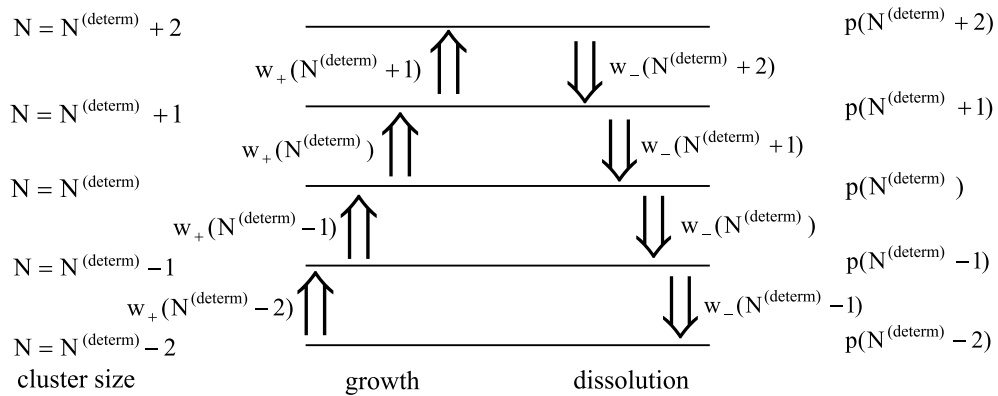
Due to fluctuations in real free flow, spontaneous traffic breakdown ( $F \rightarrow S$  transition) can occur within the flow rate range

$$q_{\text{th}}^{(B)} \leq q_{\text{sum}} < q_{\text{determ,FS}}^{(B)} \quad (6)$$

i.e., before the deterministic nucleus for traffic breakdown is reached (arrows labeled  $F \rightarrow S$  in Fig. 4). In (6), the characteristic flow rate  $q_{\text{th}}^{(B)}$  is the threshold flow rate for traffic breakdown, which is considerably less than  $q_{\text{determ,FS}}^{(B)}$ : as mentioned above, at  $q_{\text{sum}} < q_{\text{th}}^{(B)}$ , the breakdown probability is equal to zero.

To explain the physics of this traffic breakdown, we note that due to fluctuations in real free flow at the bottleneck, the cluster size  $N$  can exceed the deterministic cluster size  $N^{(\text{determ})}$ . In this case, the vehicle density increases and speed decreases within the cluster in free flow at the bottleneck (Fig. 2d,e) in comparison with the density and speed within the deterministic cluster (Fig. 2b,c). In other words, due to fluctuations in real free flow, the speed and density within the vehicle cluster at the bottleneck and therefore the cluster size  $N$  can differ considerably from, respectively, the values  $v_{\text{free}}^{(B)}$ ,  $\rho_{\text{free}}^{(B)}$ , and  $N^{(\text{determ})}$  for the deterministic cluster. In general, the vehicle cluster localized at the bottleneck with the size  $N$  can be considered consisting of two components: (a) the deterministic cluster with the size  $N^{(\text{determ})}$  and (b) a random cluster component with the size  $N - N^{(\text{determ})}$ .

There are a critical speed denoted by  $v_{\text{cr,FS}}^{(B)}$  and the associated critical density  $\rho_{\text{cr,FS}}^{(B)}$ , which depend on the flow rate (Fig. 4). When within the cluster the speed is equal to or lower than  $v_{\text{cr,FS}}^{(B)}$ , respectively the density is equal to or greater than  $\rho_{\text{cr,FS}}^{(B)}$ , then traffic breakdown occurs spontaneously. Thus when the cluster size  $N$  exceeds some critical cluster size associated with a critical density  $\rho_{\text{cr,FS}}^{(B)}$  within the cluster (Fig. 4b), then traffic breakdown occurs.



**Traffic Breakdown, Probabilistic Theory of, Figure 5**

Schematic illustration of vehicle cluster transformation due to fluctuations in an initially non-homogeneous free flow at bottleneck. Taken from [20,21,22]

The cluster with the critical size and therefore critical density within the cluster can be considered a critical vehicle cluster or a nucleus for traffic breakdown at the bottleneck. Otherwise, if the cluster size  $N$  is smaller than the critical one, the cluster decays towards the initial deterministic cluster with the cluster size  $N^{(\text{determ})}$ . Thus fluctuations in the cluster size in the neighborhood of  $N = N^{(\text{determ})}$  are responsible for traffic breakdown under condition (6) (Fig. 5; a more detailed explanation of this figure appears in Subsect. “Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory”).

### Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model

The Kerner–Klenov–Wolf (KKW) cellular automata (CA) model is the first CA traffic flow model developed in the framework of three-phase traffic theory, which can show and predict empirical features of traffic breakdown and resulting congested patterns [23]. Simulations of this model show that in accordance with the nucleation model of traffic breakdown considered above there is a vehicle cluster, i.e., a local density (and speed) disturbance in free flow, which is localized at the bottleneck. In the KKW-model, as in the nucleation model of Subsect. “Model of Traffic Breakdown at Highway Bottleneck”, traffic breakdown in free flow at the bottleneck occurs mostly at the location of the vehicle cluster.

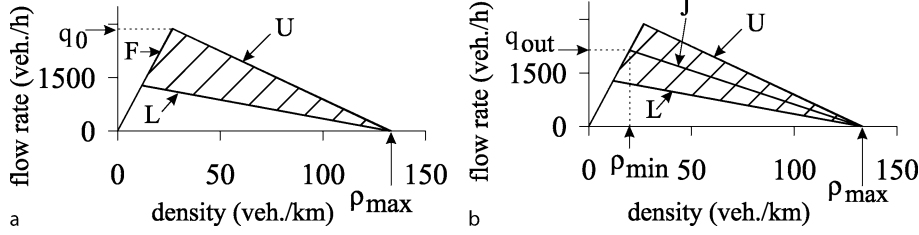
Moreover, in the KKW-model traffic breakdown is also a first-order  $F \rightarrow S$  transition: (i) in a finite flow rate range, there can be spontaneous and induced breakdowns at the same bottleneck, (ii) there is a hysteresis effect associated with traffic breakdown ( $F \rightarrow S$  transition) and a re-

turn  $S \rightarrow F$  transition (iii) there is a random time delay of traffic breakdown, and (iv) the breakdown probability is an increasing function of the flow rate  $q_{\text{sum}}$  (see below). For these reasons, we briefly discuss this traffic flow model as well as a probabilistic theory of traffic breakdown associated with the model.

### KKW Three-Phase Traffic Flow CA Model

Basic driver behavioral assumptions of the KKW CA model are similar to those of the our stochastic three-phase traffic flow model [18,19] discussed in ► [Traffic Congestion, Modeling Approaches to:](#)

- (i) *Fundamental hypothesis of three-phase traffic theory.* In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap (space gap is the net distance between vehicles) to the preceding vehicle. This means that hypothetical steady model states of synchronized flow cover a 2D-region in the flow-density plane (Fig. 6a). The boundaries of this 2D region  $F$ ,  $L$ , and  $U$  are respectively associated with free flow, a synchronization space gap, and a safe space gap determined through a safe speed. The 2D region of the steady states is associated with a driver behavioral assumption that in synchronized flow a driver is able to recognize whether the space gap is increasing or decreasing regardless of the speed difference to the preceding vehicle.
- (ii) *Line J and 2D region of steady states.* In the model, the line  $J$ , which represents the motion of the downstream jam front in the flow-density plane (the line  $J$



**Traffic Breakdown, Probabilistic Theory of, Figure 6**  
Steady states of the KKW CA model and the line  $J$ . Taken from [23]

is explained in ► [Traffic Congestion, Modeling Approaches to](#)), is between the boundaries  $L$  and  $U$  (Fig. 6b), i.e., the line  $J$  divides the 2D region of steady states of synchronized flow onto two classes: the states on and above the line  $J$  and the states below the line  $J$ , which are metastable and stable states with respect to wide moving jam formation, respectively. Thus at a given steady speed, a driver behavioral assumption is that the space gap in synchronized flow associated with the line  $J$ , i.e., between the jams is greater than the safe one and it is smaller than the synchronization gap.

- (iii) *Speed adaptation effect in synchronized flow.* The speed adaptation effect takes place, when the vehicle cannot pass the preceding vehicle, within the space gap range:

$$g_{s,n} \leq g_n \leq G_n, \quad (7)$$

where  $g_n = x_{\ell,n} - x_n - d$  is the space gap,  $x_n$  is the vehicle co-ordinate, the lower index  $\ell$  marks functions and values related to the preceding vehicle;  $G_n$  is a synchronization gap;  $g_{s,n}$  is a safe gap determined from the equation  $v_n = v_{s,n}$ , in which  $v_n$  is the vehicle speed,  $v_{s,n} = g_n/\tau$  is a safe speed taken as in the Nagel-Schreckenberg (NaSch) CA model [33]; all vehicles have the same length  $d$ , which includes the minimum space gap between vehicles within a wide moving jam; index  $n$  corresponds to the discrete time  $t = n\tau$ ,  $n = 0, 1, 2, \dots$ ; and  $\tau$  is time step. Under condition (7), the vehicle tends to adjust its speed to the preceding vehicle without caring, what the precise space gap is, as long as it is safe. For example, at a given time-independent speed of the preceding vehicle  $v_{\ell,n} = v_\ell = \text{const}$ , this speed adaptation leads to car following with  $v_n = v = v_\ell$  at a time-independent space gap  $g_n = g$ . There is a multitude of these gaps associated with the same speed  $v = v_\ell$ . These gaps lie between the synchronization gap and safe

gap, i.e., there is no desired (or optimal) space gap in synchronized flow.

- (iv) *Over-acceleration effect.* In synchronized flow of a lower density, a driver searches for the opportunity to accelerate and to pass. A competition between the speed adaptation (item (iii)) and over-acceleration effects simulates traffic breakdown ( $F \rightarrow S$  transition). The over-acceleration is simulated as a collective effect, which occurs on average in traffic flow, through the use of lane changing to a faster lane as well as through the use of random vehicle acceleration. In KKW CA model, both random acceleration and random deceleration are simulated by adding of the same random term  $a\tau\eta_n$  to vehicle speed, where

$$\eta_n = \begin{cases} -1 & \text{if } r < p_b, \\ 1 & \text{if } p_b \leq r < p_b + p_a, \\ 0 & \text{otherwise,} \end{cases} \quad (8)$$

$p_a$  is probability of random acceleration,  $p_b$  is probability of random deceleration,  $p_a + p_b \leq 1$ ,  $r = \text{rand}(0, 1)$  is a random number uniformly distributed between 0 and 1,  $a$  is the vehicle acceleration. The random vehicle acceleration is described by the second line in (8).

- (v) *Pinch effect in synchronized flow.* Moving in synchronized flow, a driver comes on average closer to the preceding vehicle over time that should explain the pinch effect. This effect is simulated through the increase in random vehicle acceleration at low vehicle speeds. For this purpose, the probability  $p_a$  is chosen to be a decreasing speed function

$$p_a(v_n) = \begin{cases} p_{a1} & \text{if } v_n < v_p \\ p_{a2} & \text{if } v_n \geq v_p, \end{cases} \quad (9)$$

where  $p_{a1}$ ,  $p_{a2}$ , and  $v_p$  are constants;  $p_{a1} > p_{a2}$ .

- (vi) *Over-deceleration effect.* In the model, a competition between a well-known over-deceleration associated with driver reaction time and the speed adaptation

effect determines moving jam emergence in synchronized flow. As in the NaSch CA model [33], the over-deceleration is also simulated as a collective effect through the use of random fluctuations in vehicle deceleration. The random vehicle deceleration is described by the first line in formula (8).

- (vii) *Driver time delay in acceleration.* In the model, this well-known effect should describe driver delay in acceleration at the downstream front of synchronized flow or wide moving jam after the preceding vehicle has begun to accelerate. A driver time delay in acceleration is simulated as a collective effect through the use of a random vehicle deceleration (8). Indeed, a vehicle that has to accelerate at the current time step can remain its speed due to applying of random vehicle deceleration. In the KKW CA model, as in a version of the NaSch CA model made in [1], the slow-and-start rules are used to describe a longer driver time delay in acceleration when the vehicles begins to accelerate from a standstill:

$$p_b(v_n) = \begin{cases} p_0 & \text{if } v_n = 0 \\ p & \text{if } v_n > 0, \end{cases} \quad (10)$$

where  $p, p_0 > p$  are constants [1]; consequently, the mean time in vehicle acceleration is

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_0}. \quad (11)$$

Thus the basis of the KKW three-phase CA model are driver behavioral assumptions made in three-phase traffic theory (items (i)-(v)). In addition, over-deceleration (item vi) and driver time delay in acceleration (item vii) introduced in the NaSch-model in the framework of the fundamental diagram approach [1,33] (see review ► [Traffic Congestion, Modeling Approaches to](#) in this Encyclopedia) have also been incorporated. Finally, the KKW model reads as follows:

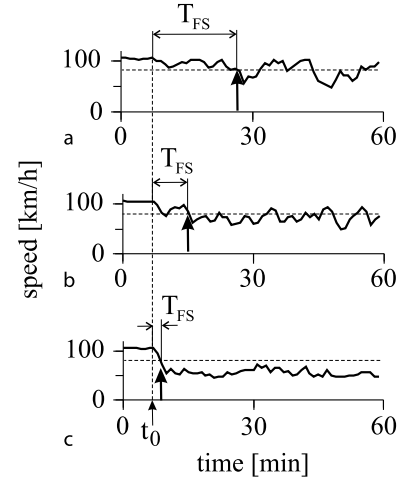
$$v_{n+1} = \max(0, \min(\tilde{v}_{n+1} + a\tau\eta_n, v_n + a\tau, v_{\text{free}}, v_{s,n})), \quad (12)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (13)$$

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})), \quad (14)$$

$$v_{c,n} = \begin{cases} v_n + a\tau & \text{at } g_n > G_n, \\ v_n + a\tau \text{sign}(v_{\ell,n} - v_n) & \text{at } g_n \leq G_n, \end{cases} \quad (15)$$

where  $\tilde{v}_n$  is the speed at time step  $n$  without fluctuating part,  $v_{\text{free}}$  is the maximum speed in free flow. The synchro-



**Traffic Breakdown, Probabilistic Theory of, Figure 7**

Time delays  $T_{FS}$  of traffic breakdown ( $F \rightarrow S$ -transition) at on-ramp bottleneck at the same  $q_{in}$  for different flow rates to the on-ramp  $q_{on}$ : a 70, b 90, c 120 vehicles/h. Up-arrows mark the time  $t_0 + T_{FS}$  at which traffic breakdown has occurred. At time  $t = t_0$  on-ramp inflow is switched on. Taken from [23].

nization gap  $G_n$  is taken as speed function:

$$G(v_n) = kv_n\tau, \quad (16)$$

where  $k$  is constant.

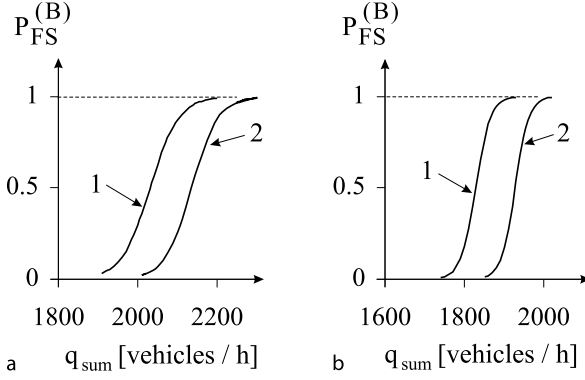
### Time Delay and Probability of Traffic Breakdown

Simulations of the KKW CA model [23] show that at a given flow rate in free flow on the main road upstream of the bottleneck  $q_{in}$ , free flow is metastable with respect to a local first-order  $F \rightarrow S$  transition, which occurs spontaneously at the bottleneck, when the flow rate  $q_{sum}$  in free flow downstream of the bottleneck is satisfied condition (6).

Traffic breakdown is characterized by a random time delay  $T_{FS}^{(B)}$  (Fig. 7): in different realizations, at the same  $q_{on}$  and  $q_{in}$  usually different  $T_{FS}^{(B)}$  are found. The smaller  $q_{on}$  and/or  $q_{in}$  are, the longer the mean time delay  $T_{FS}^{(B,mean)}$  of traffic breakdown is.

Theoretical probability for an  $F \rightarrow S$  transition at the bottleneck (Fig. 8) shows the same features as empirical probability for  $F \rightarrow S$  transition [35] (see ► [Traffic Congestion, Modeling Approaches to](#)). To study the  $F \rightarrow S$  transition probability (Fig. 8), a large number of realizations (runs) have been calculated for given flow rates  $q_{sum} = q_{on} + q_{in}$  and  $q_{on}$  during a given time interval  $T_0$ . At the beginning of each run there was free flow at the on-ramp. For each run it was checked, whether the  $F \rightarrow S$





**Traffic Breakdown, Probabilistic Theory of, Figure 8**

Probability  $P_{FS}^{(B)}$  that an  $F \rightarrow S$  transition occurs at the bottleneck within  $T_0 = 30$  min (curve 1) or already within  $T_0 = 15$  min (curve 2), after the on-ramp inflow was switched on, versus the traffic demand upstream of the on-ramp,  $q_{sum} = q_{in} + q_{on}$ . Results are shown for two different flow rates to the on-ramp,  $q_{on} = 60$  vehicles/h in (a) and  $q_{on} = 200$  vehicles/h in (b). Taken from [23]

transition at the on-ramp occurred during the time interval  $T_0$  or not. The result of these simulations is the number of realizations  $n_p$  where the  $F \rightarrow S$  transition at the on-ramp had occurred in comparison with the number of all realizations  $N_p$ . Then the probability for an  $F \rightarrow S$  transition is  $P_{FS}^{(B)} = n_p/N_p$ .

The flow rate  $q_{sum}$  was changed and the procedure with all realizations was repeated at the same flow rate to the on-ramp  $q_{on}$ . The flow rate  $q_{sum}$  at which the  $F \rightarrow S$  transition at the on-ramp occurred in all realization is therefore related to the probability of the  $F \rightarrow S$  transition  $P_{FS}^{(B)} = 1$ . We found that lower flow rates  $q_{sum}$  correspond to  $P_{FS}^{(B)} < 1$ . As expected, we found an increase of the probability  $P_{FS}^{(B)}$  as a function of the flow rate  $q_{sum}$  at a given flow rate to the on-ramp  $q_{on}$ ; the related function found numerically can be approximated by the formula [23]

$$P_{FS}^{(B)} = \frac{1}{1 + \exp[\alpha(q_p - q_{sum})]}, \quad (17)$$

where  $\alpha$  and  $q_p$  are functions of  $q_{on}$  and  $T_0$ .

### Probabilistic Description of Traffic Breakdown Based on Master Equation

The nucleation model for traffic breakdown of Subsect. “[Model of Traffic Breakdown at Highway Bottleneck](#)”, which is confirmed by numerical simulations of the KKW CA model discussed above, can be used for an analytical study of the nucleation rate, breakdown probability, and mean time delay of traffic breakdown at a bottleneck

based on the well-known master equation approach for probabilistic description of first-order phase transitions in complex systems of diverse nature [7].

### Master Equation and One-Step Processes

Traffic flow variables (speed, density, occupancy, flow rate) within the vehicle cluster localized at a bottleneck (Subsect. “[Model of Traffic Breakdown at Highway Bottleneck](#)”) are time dependent random variables. A time dependent random variable  $X(t)$  can be characterized by a set of joint probabilities [7]

$$p(x_1, t_1; x_2, t_2; \dots; x_n, t_n), \quad n = 1, 2, \dots, \quad (18)$$

where  $x_1, x_2, \dots, x_n$  are random values of variable  $X$  at time instants  $t_1 \leq t_2 \leq \dots \leq t_n$ . The joint probabilities  $p(x_1, t_1; \dots; x_n, t_n)$  and  $p(x_1, t_1; \dots; x_{n-1}, t_{n-1})$  are connected each other by a conditional probability

$$p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}), \quad n = 2, 3, \dots, \quad (19)$$

as follows [7]

$$\begin{aligned} p(x_1, t_1; \dots; x_n, t_n) \\ = p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}) \\ \cdot p(x_1, t_1; \dots; x_{n-1}, t_{n-1}). \end{aligned} \quad (20)$$

The basic assumption of *Marcovian* stochastic process is that the conditional probability  $p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1})$  depends on the value of random variable  $X$  at the last previous time moment  $t_{n-1}$  only [7]

$$p(x_n, t_n | x_1, t_1; \dots; x_{n-1}, t_{n-1}) = p(x_n, t_n | x_{n-1}, t_{n-1}). \quad (21)$$

In this case, according to (20) all joint probabilities  $p(x_1, t_1; \dots; x_n, t_n)$  (18) at  $n > 1$  can be expressed in terms of the conditional probability (21) and the probability  $p(x_1, t_1)$  as

$$\begin{aligned} p(x_1, t_1; \dots; x_n, t_n) \\ = p(x_1, t_1) \prod_{k=2}^n p(x_k, t_k | x_{k-1}, t_{k-1}). \end{aligned} \quad (22)$$

In the nucleation model of Subsect. “[Model of Traffic Breakdown at Highway Bottleneck](#)” is assumed that there is a vehicle cluster localized at a bottleneck and this cluster exists in free flow even if there were no fluctuations in the flow. In the model, the cluster dynamics describes traffic breakdown ( $F \rightarrow S$  transition) at the bottleneck. For

an analytical study of this nucleation model, we will further make the following additional assumptions: (i) The cluster dynamics is associated with a Markovian stochastic process; (ii) the random variable  $X$  of this process is associated with the total number of vehicles in the cluster  $N$  only (see Subsect. “Outflow Rate from Cluster” for more detail). Then  $X$  is a discrete variable. In this case, the Chapman–Kolmogorov equation for the conditional probability  $p(x, t|x', t')$  reads [7]:

$$p(x, t|x', t') = \sum_{x''} p(x, t|x'', t'') p(x'', t'|x', t'), \quad (23)$$

where  $t \geq t'' \geq t'$ , the summation is taken over all possible values of  $x''$ . In turn, the conditional probability  $p(x, t|x', t')$  determines the time evolution of the probability  $p(x, t)$  as

$$p(x, t) = \sum_{x'} p(x, t|x', t') p(x', t'). \quad (24)$$

The differential equation for the conditional probability  $p(x, t|x', t')$  that follows from (23) can be written as [7]

$$\frac{\partial p(x, t|x', t')}{\partial t} = \sum_{y \neq x} \left[ W(x, y, t) p(y, t|x', t') - W(y, x, t) p(x, t|x', t') \right], \quad (25)$$

where

$$W(x, y, t) = \lim_{\Delta t \rightarrow 0} p(x, t + \Delta t|y, t)/\Delta t \quad \text{at } x \neq y \quad (26)$$

is the transition rate, i.e., the probability per unit time for the transition from  $y$  to  $x \neq y$  at time  $t$ . The Eq. (25) is usually called *master equation*. Multiplying (25) by  $p(x', t')$  and using (24), one can write (25) as the equation for probability  $p(x, t)$  [7]

$$\frac{\partial p(x, t)}{\partial t} = \sum_{y \neq x} [W(x, y, t) p(y, t) - W(y, x, t) p(x, t)]. \quad (27)$$

In the model of traffic breakdown (Subsect. “[Model of Traffic Breakdown at Highway Bottleneck](#)” and Subsect. “Outflow Rate from Cluster”), we will make once more assumption that at each time instant the total number of vehicles within the cluster  $N$  can be changed by one vehicle only. This assumption is associated with a *one-step process*, also called *birth-and-death process*, when transitions between neighboring states  $X = x$  and  $X = x \pm 1$

can occur only [7]. The transition rate  $W(x, y, t)$  for one-step processes is [7]

$$W(x, y, t) = w_+(y) \delta_{x, y+1} + w_-(y) \delta_{x, y-1}, \quad (28)$$

where  $w_+(y)$  is the rate of transition from  $y$  to  $x = y + 1$ ,  $w_-(y)$  is the rate of transition from  $y$  to  $x = y - 1$ ;  $\delta_{x, y}$  is Kroneker symbol.

Substituting (28) into the master Eq. (27), one finds the master equation for one-step processes [7]

$$\begin{aligned} \frac{\partial p(x, t)}{\partial t} = & w_+(x-1) p(x-1, t) \\ & + w_-(x+1) p(x+1, t) \\ & - [w_+(x) + w_-(x)] p(x, t). \end{aligned} \quad (29)$$

For the master Eq. (29), the mean first passage time  $T$  for transition from an initial point  $x = a$  to a finite point  $x = b$  is given by formula [7]

$$T = \sum_{x=a}^b \left[ (w_+(x) p_s(x))^{-1} \sum_{y=0}^x p_s(y) \right], \quad (30)$$

where  $p_s(x)$  is a steady solution of (29):

$$p_s(x) = p_s(0) \prod_{y=1}^x \frac{w_+(y-1)}{w_-(y)} \quad \text{at } x > 0, \quad (31)$$

the reflecting boundary condition

$$w_-(0) = 0 \quad (32)$$

is assumed to be satisfied at the point  $x = 0$ .

If the points  $a$  and  $b$  are separated by a potential barrier and therefore  $p_s^{-1}(x)$  has a strong maximum at a point  $x = c$  between  $a$  and  $b$ , the formula (30) is reduced to [7]:

$$T = \frac{1}{w_+(c)} \sum_{x=0}^c p_s(x) \sum_{x=a}^b p_s^{-1}(x). \quad (33)$$

We will use this formula for the evaluation of the mean time delay for traffic breakdown  $T_{\text{FS}}^{(\text{B}, \text{mean})}$  in Subsect. “[Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks](#)”.

### Critical Discussion of Earlier Probabilistic Traffic Breakdown Models

It must be noted that models for vehicle cluster nucleation in traffic flow based on the master Eq. (29), which should describe traffic breakdown, have firstly been in-

troduced by Mahnke et al. and Kühne et al. [24,25,28,29] (see review [30]). However, in [21,22] we have explained in detail that and why the models of Ref. [24,25,28,29,30] do not explain the fundamental empirical features of traffic breakdown, i.e., they cannot describe the probabilistic traffic breakdown in real traffic flow. In particular, the nucleation models of Ref. [24,28,29,30] describe the nucleation of a wide moving jam in free flow ( $F \rightarrow J$  transition). This is in contrast with the fundamental empirical feature A of traffic breakdown (Subsect. “Fundamental Empirical Features of Traffic Breakdown”).

Briefly, the most crucial differences of the nucleation models of Ref. [24,25,28,29,30] with the model of traffic breakdown of Subsect. “Model of Traffic Breakdown at Highway Bottleneck” [15,17,20,21,22] are as follows.

The basic assumption of the nucleation models of Ref. [24,25,28,29,30] is that there are *no* vehicle clusters on a road without fluctuations. This basic assumption is made both for a homogeneous road [24,28,29,30] and for a road with a bottleneck [25,30]. It is suggested that for vehicle cluster emergence firstly a random pre-cluster should occur from fluctuations in free flow. This pre-cluster should forego subsequent cluster evolution towards a critical cluster (nucleus) whose growth should lead to traffic breakdown. In other words, it is suggested that without random pre-cluster the number of vehicles within the cluster, i.e., the cluster size is exactly equal to *zero* and, consequently, without random fluctuations no traffic breakdown is possible.

In contrast with the nucleation models of Ref. [24,25,28,29,30], in the nucleation model of Subsect. “Model of Traffic Breakdown at Highway Bottleneck” [20,21,22] the cluster size  $N$  is equal to the *total* number of vehicles within the cluster. Therefore, even if there were no random fluctuations in free flow, rather than the cluster size is equal to zero, the cluster size  $N = N^{(determ)}$ , i.e., this is equal to the size of the deterministic cluster (Fig. 5) and, consequently, the deterministic traffic breakdown is possible [20,21,22]. The nucleation model of Subsect. “Model of Traffic Breakdown at Highway Bottleneck” results from the model assumption that an initial free flow at a bottleneck is non-homogeneous, i.e., a vehicle cluster exists in free flow even if random fluctuations in traffic flow were not taken into account.

In the model of Subsect. “Model of Traffic Breakdown at Highway Bottleneck” [20,21,22], due to fluctuations in real free flow, the cluster size  $N$  changes in a neighborhood of  $N = N^{(determ)}$  (Fig. 5). These fluctuations in the cluster size lead to traffic breakdown before the condition for the deterministic breakdown is satisfied. Below we will see that this nucleation model explains all fundamental empirical

features of traffic breakdown (Subsect. “Fundamental Empirical Features of Traffic Breakdown”).

### Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory

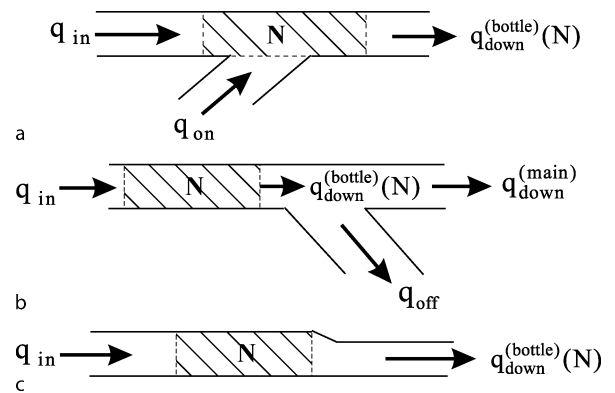
**Outflow Rate from Cluster** An important characteristic of the nucleation model of Subsect. “Model of Traffic Breakdown at Highway Bottleneck” is the outflow rate from the vehicle cluster  $q_{down}^{(bottle)}$  (Figs. 2a and 9). In contrast with the on-ramp and merge bottlenecks for which there is only one outflow from the cluster, for the off-ramp bottleneck, there are two different outflows from the cluster, which are related to the flow rate on the main road downstream of the merging region of the off-ramp denoted by  $q_{down}^{(main)}$  and to the flow rate to the off-ramp road denoted by  $q_{off}$ . This means that for the off-ramp bottleneck (Fig. 9b)

$$q_{down}^{(bottle)} = q_{down}^{(main)} + q_{off}. \quad (34)$$

In the nucleation model for traffic breakdown discussed in Subsect. “Model of Traffic Breakdown at Highway Bottleneck”, we make further an assumption that the outflow rate from the cluster depends on the total number of vehicles within the cluster, i.e., on the cluster size  $N$  (Fig. 9) [20,21,22]:

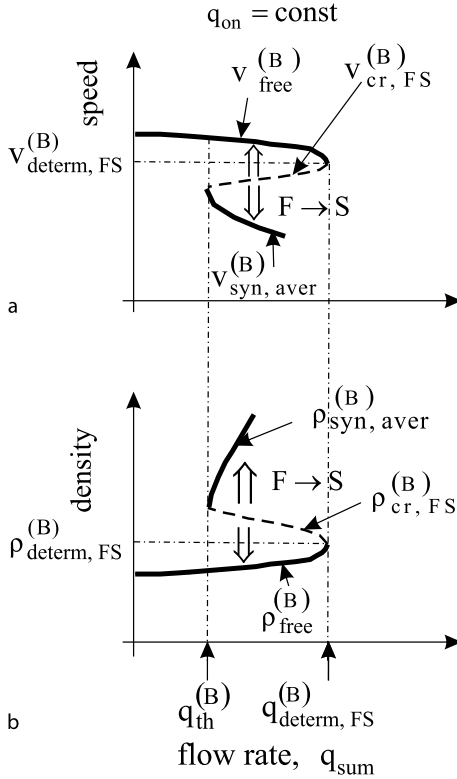
$$q_{down}^{(bottle)} = q_{down}^{(bottle)}(N). \quad (35)$$

To determine  $q_{down}^{(bottle)}(N)$ , we can use the flow rate dependence of the density (Fig. 4b). However, in the nucleation model for traffic breakdown [20,21,22], which we discuss below, we make an approximation in which we average the infinite synchronized flow states for each density



**Traffic Breakdown, Probabilistic Theory of, Figure 9**

Qualitative explanation of vehicle cluster (dashed regions) at bottlenecks: **a** On-ramp bottleneck. **b** Off-ramp bottleneck. **c** Merge bottleneck



**Traffic Breakdown, Probabilistic Theory of, Figure 10**

Simplified qualitative shapes of speed (a) and density (b) dependences on the flow rate  $q_{\text{sum}}$  associated with location of vehicle cluster. Taken from [20,21,22]

to one average speed. Thus in this approximation, rather than Fig. 4, we use the related simplified non-monotonous dependences shown in Fig. 10.

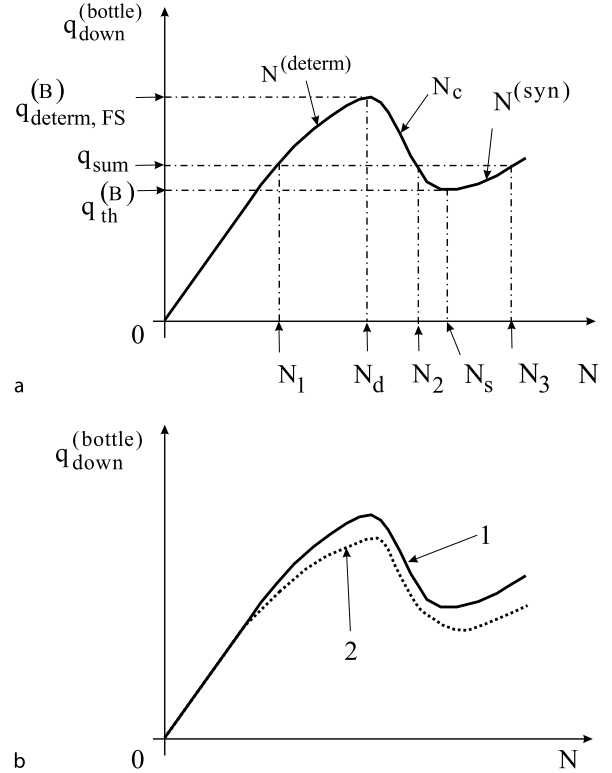
In the model, it is further assumed that the shape of the characteristic  $q_{\text{down}}^{(\text{bottle})}(N)$  (Fig. 11) follows from the S-shaped density-flow characteristic shown in Fig. 10b. In this nucleation model, the characteristic  $q_{\text{down}}^{(\text{bottle})}(N)$  has at least two different branches  $q_{\text{down}}^{(\text{bottle})}(N)$  labeled  $N^{(\text{determ})}$  and  $N_c$  in Fig. 11a. These branches are related to the vehicle number ranges, respectively, given by the conditions

$$0 \leq N \leq N_d \quad (36)$$

and

$$N_d < N \leq N_s. \quad (37)$$

The branches  $N^{(\text{determ})}$  and  $N_c$  in Fig. 11a are associated with the density branches  $\rho_{\text{free}}^{(\text{B})}$  and  $\rho_{\text{cr,FS}}^{(\text{B})}$  of the S-shaped density-flow characteristic in Fig. 10b, respectively. The branch  $N^{(\text{determ})}$  is associated with the case in which at



**Traffic Breakdown, Probabilistic Theory of, Figure 11**

Qualitative dependencies of the outflow rate  $q_{\text{down}}^{(\text{bottle})}$  on the total vehicle number  $N$  within the cluster localized at the bottleneck (a), and possible dependencies of the  $N$ -shaped function  $q_{\text{down}}^{(\text{bottle})}(N)$  on  $q_{\text{on}}$  for two different values  $q_{\text{on}}$  (b); curve 1 for  $q_{\text{on}} = q_{\text{on}}^{(1)}$ , curve 2 for  $q_{\text{on}} = q_{\text{on}}^{(2)} > q_{\text{on}}^{(1)}$

a high enough flow rate  $q_{\text{sum}}$  and the on-ramp flow rate  $q_{\text{on}} > 0$  the deterministic cluster exists at the bottleneck under free flow conditions. The branch  $N_c$  is associated with the case in which the critical cluster whose growth leads to an  $F \rightarrow S$  transition occurs at the bottleneck.

In addition, from the S-shaped density-flow characteristic (Fig. 10b) can be seen that at

$$N > N_s, \quad (38)$$

there can be a third branch  $N^{(\text{syn})}$  on the characteristic  $q_{\text{down}}^{(\text{bottle})}(N)$  (Fig. 11a) associated with the branch  $\rho_{\text{syn,aver}}^{(\text{B})}$  for averaged synchronized flow states in Fig. 10b. In this case,  $q_{\text{down}}^{(\text{bottle})}(N)$  (35) is a  $N$ -shaped flow-vehicle-number characteristic. It should be noted that the branch for synchronized flow  $v_{\text{syn,aver}}^{(\text{B})}$  in Fig. 10b and the associated branch  $N^{(\text{syn})}$  on the characteristic  $q_{\text{down}}^{(\text{bottle})}(N)$  (Fig. 11a) follow from the microscopic three-phase traffic theory, rather than from the nucleation model. This branch, which

can be drawn for the case only in which a localized SP (LSP) occurs as a result of the breakdown, is shown with the aim of a qualitative illustration of a possible traffic flow state after synchronized flow nucleation. This branch has no influence on the nucleation rate and probability of an  $F \rightarrow S$  transition discussed below. In other words, the nucleation effect leading to traffic breakdown and its characteristics are fully independent of possible congested patterns resulting from this  $F \rightarrow S$  transition. These congested patterns, which cannot be described by the nucleation model, are discussed in [17], ► [Traffic Congestion, Spatiotemporal Features of](#).

At the critical point  $N = N_d$  at which the branches  $N^{(\text{determ})}$  and  $N_c$  merges, the function  $q_{\text{down}}^{(\text{bottle})}(N)$  has its maximum point. At the threshold point  $N = N_s$  at which the branches  $N_c$  and  $N^{(\text{syn})}$  merges, the function  $q_{\text{down}}^{(\text{bottle})}(N)$  has its minimum point.

In the case of the on-ramp bottleneck, quantitative characteristics of the  $N$ -shaped function  $q_{\text{down}}^{(\text{bottle})}(N)$  (e.g., values  $N_d$  and  $N_s$ ) can depend on the flow rate to the on-ramp  $q_{\text{on}}$ . This is because the on-ramp inflow and the flow upstream of the bottleneck can make a considerable different influence on the cluster size and the outflow rate from the cluster  $q_{\text{down}}^{(\text{bottle})}$ . In particular, it can turn out that at the same  $N$  the greater  $q_{\text{on}}$ , the more difficult for vehicles to escape from the cluster, i.e., the less  $q_{\text{down}}^{(\text{bottle})}$  is. This is confirmed by microscopic simulations [23] and reflected in Fig. 11b in which it is assumed that the greater  $q_{\text{on}}$ , the less  $q_{\text{down}}^{(\text{bottle})}$  and the greater  $N_d$  and  $N_s$  are. Thus, in a case of the on-ramp bottleneck instead of (35) we should use

$$q_{\text{down}}^{(\text{bottle})} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}), \quad (39)$$

where  $q_{\text{on}}$  is a parameter.

In accordance with the nucleation model (Fig. 11a), critical cluster occurrence describes an  $F \rightarrow S$  transition at the bottleneck. There are two reasons for this statement: (i) A vehicle cluster is on average motionless, i.e., fixed at the bottleneck. This is also related to the definition of synchronized flow whose downstream front is usually fixed at the bottleneck, whereas the downstream front of a wide moving jam propagates through any bottleneck while maintaining the mean downstream jam velocity [17]. (ii) The shape of the chosen function (35) in the nucleation model associated with this motionless cluster (Fig. 11a) follows from a Z-shaped speed-flow characteristic for traffic breakdown (Fig. 10b) found in a microscopic traffic flow theory. In this theory has been shown that if these two requirements are satisfied, then rather than an  $F \rightarrow J$  transition, an  $F \rightarrow S$  transition occurs at the bottleneck, as found in empirical observations.

**Steady States** Steady states of the vehicle number  $N$  within the cluster at a given  $q_{\text{sum}}$  are associated with solutions of the equation

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}}), \quad (40)$$

for the on-ramp bottleneck and

$$q_{\text{sum}} = q_{\text{down}}^{(\text{bottle})}(N), \quad (41)$$

for the off-ramp and merge bottlenecks. As can be seen from Fig. 11a, at given flow rates  $q_{\text{on}}$  and  $q_{\text{in}}$  that satisfy condition (6) there can be at least two steady states:  $N = N_1$  associated with the deterministic cluster and  $N = N_2$  associated with the critical cluster. These steady states are the roots of Eq. (40) (or Eq. (41)), i.e., they are associated with the intersection points of the horizontal line  $q = q_{\text{sum}}$  with the branches  $N^{(\text{determ})}$  and  $N_c$  of the characteristic  $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$  (Fig. 11a), respectively. In addition, as mentioned above if an LSP occurs as a result of an  $F \rightarrow S$  transition, then there is a third root of Eq. (40) (or Eq. (41)),  $N = N_3$ , associated with the intersection point of the horizontal line  $q = q_{\text{sum}}$  with the branch  $N^{(\text{syn})}$  of the characteristic  $q_{\text{down}}^{(\text{bottle})}(N, q_{\text{on}})$ .

If the flow rate  $q_{\text{sum}}$  increases, then the critical vehicle number difference within the cluster

$$\Delta N_c = N_2 - N_1 \quad (42)$$

decreases. This critical vehicle number difference is associated with the vehicle number difference within the critical cluster and within the deterministic cluster at the bottleneck. The growth of the critical cluster leads to traffic breakdown at the bottleneck.

At the critical flow rate (5), we get  $\Delta N_c = 0$ : the steady states  $N_1$  and  $N_2$  merge into one point with the critical vehicle number  $N = N_d$  at which

$$q_{\text{determ,FS}}^{(B)} = q_{\text{down}}^{(\text{bottle})}(N_d, q_{\text{on}}). \quad (43)$$

At  $q_{\text{sum}} \geq q_{\text{determ,FS}}^{(B)}$  the deterministic traffic breakdown should occur even if there were no random increase in the vehicle number within the deterministic cluster at the bottleneck.

If the flow rate  $q_{\text{sum}}$  decreases gradually, then the threshold flow rate

$$q_{\text{sum}} = q_{\text{th}}^{(B)} \quad (44)$$

is reached at which the steady states  $N_2$  and  $N_3$  merge into one threshold steady state  $N = N_s$  at which

$$q_{\text{th}}^{(B)} = q_{\text{down}}^{(\text{bottle})}(N_s, q_{\text{on}}). \quad (45)$$



Based on the nucleation model discussed above and on the master Eq. (29), we consider the dynamics of the vehicle cluster at a bottleneck (dashed regions in Fig. 9) [20,21,22]. The probability  $p(N, t)$  that  $N$  vehicles occur within the cluster at time  $t$  is given by the master Eq. (29) in which the variable  $x$  is the total vehicle number within the cluster  $N$ :

$$\begin{aligned} \frac{\partial p(N, t)}{\partial t} = & w_+(N-1)p(N-1, t) \\ & + w_-(N+1)p(N+1, t) \\ & - [w_+(N) + w_-(N)]p(N, t), \quad \text{at } N > 0. \end{aligned} \quad (46)$$

The master Eq. (46) describes probability  $p$  for the total vehicle number within the cluster, i. e., the cluster size  $N$ , which randomly changes due to fluctuations in a neighborhood of the size  $N^{(\text{determ})}$  of the deterministic cluster (Fig. 5) [20,21,22]. The size of the deterministic cluster  $N^{(\text{determ})}$  does not depend on fluctuations. Random fluctuations cause either a decrease or an increase in the cluster size  $N$  in comparison with  $N^{(\text{determ})}$ . The fluctuations, which decrease the cluster size  $N$ , are associated with an increase in speed within the cluster. Therefore, these fluctuations can prevent the breakdown. In contrast, fluctuations, which increase the cluster size  $N$ , i. e., decrease the speed within the cluster, can cause traffic breakdown before the deterministic nucleus for traffic breakdown associated with condition (5) is reached.

In the master Eq. (46),  $w_+$  and  $w_-$  are the attachment rate onto and detachment rate from the cluster, respectively.  $w_+$  does not depend on  $N$ , i. e., regardless of  $N$

$$w_+ = q_{\text{sum}}. \quad (47)$$

In contrast,  $w_-$  is given by the formula

$$w_- = q_{\text{down}}^{(\text{bottle})}, \quad (48)$$

i. e.,  $w_-$  depends on  $N$  in accordance with (35) for off-ramp and merge bottlenecks, or else in accordance with (39) for the case of an on-ramp bottleneck; in the latter case,  $w_-$  depends on  $N$  and  $q_{\text{on}}$ . When  $N = 0$ , the following boundary condition should be satisfied in (46)

$$w_-(0) = 0, \quad (49)$$

i. e., no vehicles can leave the cluster. In this case,

$$\frac{\partial p(0, t)}{\partial t} = w_-(1)p(1, t) - w_+(0)p(0, t), \quad \text{at } N = 0. \quad (50)$$

### Nucleation Rate and Mean Time Delay of Traffic Breakdown at Bottlenecks

As follows from the analysis of the model (46)-(50) and formula (33) made in [20,21,22], within the flow rate range (6) the mean time delay of an  $F \rightarrow S$  transition at the bottleneck is

$$T_{\text{FS}}^{(\text{B,mean})} = 2\pi\tau_b \exp\{\Delta\Phi\}, \quad (51)$$

where a potential barrier for the  $F \rightarrow S$  transition

$$\Delta\Phi = \Phi(N_2) - \Phi(N_1), \quad (52)$$

the potential  $\Phi(N)$

$$\Phi(N) = \begin{cases} \sum_{n=1}^N \ln \frac{w_-(n)}{w_+} & \text{at } N > 0, \\ 0 & \text{at } N = 0, \end{cases} \quad (53)$$

$$\tau_b = \left( w'_-(N_1) | w'_-(N_2) | \right)^{-\frac{1}{2}}, \quad (54)$$

$w'_-(N) = dw_-/dN$ . Respectively, the nucleation rate for traffic breakdown is

$$G_{\text{FS}}^{(\text{B})} = \frac{1}{T_{\text{FS}}^{(\text{B,mean})}} = \frac{1}{2\pi\tau_b} \exp\{-\Delta\Phi\}. \quad (55)$$

It can be seen from (51) that the mean time delay for traffic breakdown decreases exponentially with increase in potential barrier  $\Delta\Phi$  (52). If in Fig. 12 the total flow rate increases from  $q_{\text{sum}} = q_{\text{sum}}^{(1)}$  to  $q_{\text{sum}} = q_{\text{sum}}^{(2)}$ , which is close to the critical flow rate (5) for deterministic traffic breakdown, then the potential barrier  $\Delta\Phi$  (52) decreases from  $\Delta\Phi_1$  to  $\Delta\Phi_2$ .

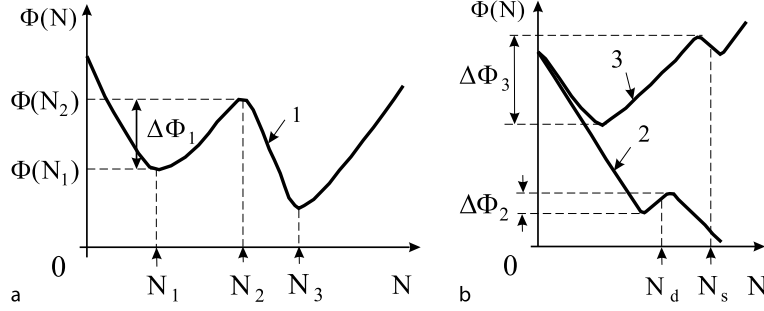
In contrast, if the total flow rate decreases from  $q_{\text{sum}} = q_{\text{sum}}^{(1)}$  to  $q_{\text{sum}} = q_{\text{sum}}^{(3)}$ , which is close to the threshold flow rate  $q_{\text{th}}^{(\text{B})}$  (44) for random traffic breakdown, then the potential barrier  $\Delta\Phi$  (52) increases from  $\Delta\Phi_1$  to  $\Delta\Phi_3$  (Fig. 12). At the threshold point  $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$  (44), the potential barrier  $\Delta\Phi(N)$  reaches the maximum value

$$\Delta\Phi = \Phi(N_s) - \Phi(N_{\text{th}}), \quad (56)$$

where  $N_{\text{th}} = N_1$  at  $q_{\text{sum}} = q_{\text{th}}^{(\text{B})}$ . As a result, the mean time delay  $T_{\text{FS}}^{(\text{B,mean})}$  (51) strongly increases as  $q_{\text{sum}}$  approaches the threshold point  $q_{\text{th}}^{(\text{B})}$ . Under the condition

$$q_{\text{sum}} < q_{\text{th}}^{(\text{B})} \quad (57)$$

no traffic breakdown at the bottleneck regardless of a random increase in the vehicle number within the cluster is possible.



**Traffic Breakdown, Probabilistic Theory of, Figure 12**

Qualitative shape of the potential  $\Phi(N)$  (53) for different flow rates  $q_{\text{sum}}$ : curves 1, 2, and 3 are related to the corresponding flow rates  $q_{\text{sum}}^{(1)}$ ,  $q_{\text{sum}}^{(2)}$ , and  $q_{\text{sum}}^{(3)}$  satisfying the condition  $q_{\text{sum}}^{(3)} < q_{\text{sum}}^{(1)} < q_{\text{sum}}^{(2)}$ . Characteristic values of the vehicle number within the cluster  $N = N_i$ ,  $i = 1, 2, 3$ ,  $N = N_s$  and  $N = N_d$  are discussed in Subsect. “Probabilistic Traffic Breakdown Model Based on Three-Phase Traffic Theory” (Fig. 11). Taken from [20,21,22]

If in the vicinity of the critical vehicle number  $N_d$  the function  $w_-(N)$  can be approximated by a parabolic function of  $N$ , then the following approximate formula can be derived from (51)-(54):

$$T_{\text{FS}}^{(\text{B}, \text{mean})} = \frac{\sqrt{2\pi} N_d}{q_{\text{determ}, \text{FS}}^{(\text{B})} (\xi_d \Delta_c)^{1/2}} \left( \frac{1 + \Delta_c^{1/2}}{1 - \Delta_c^{1/2}} \right)^{2\sqrt{2/\xi_d} N_d} \times \exp \left( -\frac{4\sqrt{2} N_d \Delta_c^{1/2}}{\sqrt{\xi_d}} \right), \quad (58)$$

where

$$\xi_d = -(N^2 d^2 \ln w_- / dN^2) \big|_{N=N_d} \quad (59)$$

is a dimensionless value of the order of 1,

$$\Delta_c = \frac{q_{\text{determ}, \text{FS}}^{(\text{B})} - q_{\text{sum}}}{q_{\text{determ}, \text{FS}}^{(\text{B})}}, \quad (60)$$

i. e.,  $\Delta_c$  is the relative difference between the critical flow rate  $q_{\text{determ}, \text{FS}}^{(\text{B})}$  for the deterministic  $F \rightarrow S$  transition and the total flow rate  $q_{\text{sum}}$  (2). If in (58)  $\Delta_c \ll 1$ , then we get

$$T_{\text{FS}}^{(\text{B}, \text{mean})} = \frac{\sqrt{2\pi} N_d}{q_{\text{determ}, \text{FS}}^{(\text{B})} (\xi_d \Delta_c)^{1/2}} \exp \left( \frac{8N_d \Delta_c^{3/2}}{3\sqrt{2}\xi_d} \right). \quad (61)$$

Respectively, the nucleation rate  $G_{\text{FS}}^{(\text{B})} = 1/T_{\text{FS}}^{(\text{B}, \text{mean})}$  for traffic breakdown at the bottleneck associated with (61) is

$$G_{\text{FS}}^{(\text{B})} = \frac{q_{\text{determ}, \text{FS}}^{(\text{B})} (\xi_d \Delta_c)^{1/2}}{\sqrt{2\pi} N_d} \exp \left( -\frac{8N_d \Delta_c^{3/2}}{3\sqrt{2}\xi_d} \right). \quad (62)$$

Note that  $q_{\text{determ}, \text{FS}}^{(\text{B})}$ ,  $N_d$  and  $\xi_d$  for the on-ramp bottleneck depend  $q_{\text{on}}$ ; in this case,  $T_{\text{FS}}^{(\text{B}, \text{mean})}$  (61) and  $G_{\text{FS}}^{(\text{B})}$  (62) are functions of  $q_{\text{sum}}$  and  $q_{\text{on}}$ .

As usual for each first-order phase transition observed in many other systems in natural science [7], the nucleation rate for traffic breakdown (62) is an exponential function of  $\Delta_c$  (60). For traffic flow, in accordance with (62) and (60) the exponential growth of the nucleation rate with  $\Delta_c$  (60) is very sensible to the critical value for the deterministic breakdown phenomenon  $q_{\text{determ}, \text{FS}}^{(\text{B})}$ . This emphasizes the important impact of the deterministic cluster, which occurs at the on-ramp bottleneck at  $q_{\text{on}} > 0$ , on the nucleation rate for traffic breakdown (62) at a given flow rate  $q_{\text{sum}}$ .

### Probability of Breakdown

The mean time delay  $T_{\text{FS}}^{(\text{B}, \text{mean})}$  of traffic breakdown at the bottleneck enables us to find probability  $P_{\text{FS}}^{(\text{B})}$  that traffic breakdown occurs at the bottleneck during a time interval for observing traffic flow  $T_{\text{ob}}$ . Let us consider probability

$$P_C^{(\text{B})} = 1 - P_{\text{FS}}^{(\text{B})} \quad (63)$$

that during a chosen time interval  $T_{\text{ob}}$  there is *no*  $F \rightarrow S$  transition, i. e., free flow remains at the bottleneck. In accordance with a stochastic theory of dynamic systems with first-order phase transitions [7], a dependence of the probability  $P_C^{(\text{B})}$  on  $T_{\text{ob}}$  governs by the equation

$$\frac{dP_C^{(\text{B})}}{dT_{\text{ob}}} = -\frac{P_C^{(\text{B})}}{T_{\text{FS}}^{(\text{B}, \text{mean})}}, \quad (64)$$

where  $T_{\text{FS}}^{(\text{B}, \text{mean})}$  is constant at given values of  $q_{\text{sum}}$  and  $q_{\text{on}}$ . This equation derived with Kramer's method [7] is valid under condition that  $T_{\text{FS}}^{(\text{B}, \text{mean})}$  is long enough in comparison with a relaxation time towards a metastable steady state with the deterministic cluster of the size

$N^{(\text{determ})} = N_1$ . This relaxation time is of the order of  $\tau_b$  (54), i. e., Eq. (64) is valid, if

$$\tau_b \ll T_{\text{FS}}^{(\text{B,mean})}. \quad (65)$$

Note that as follows from (51), condition (65) can be satisfied only if the potential barrier  $\Delta\Phi$  in (51) is relatively high, i. e.,  $\Delta\Phi \gg 1$  (52).

In accordance with [7], the solution of (64) is

$$P_C^{(\text{B})}(T_{\text{ob}}) = P_0 \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right), \quad (66)$$

where  $P_0$  is constant. If the time interval  $T_{\text{ob}}$  is short enough, i. e.,  $T_{\text{ob}} \ll T_{\text{FS}}^{(\text{B,mean})}$ , but the condition

$$\tau_b \ll T_{\text{ob}} \quad (67)$$

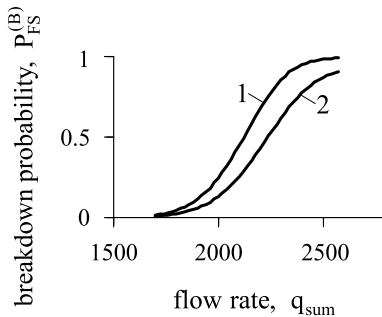
is also satisfied, then probability  $P_C^{(\text{B})}$  should be close to one, therefore,  $P_0 \approx 1$  in (66). Then under conditions (65), (67) probability  $P_C^{(\text{B})}$  is given by the approximate formula

$$P_C^{(\text{B})}(T_{\text{ob}}) = \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right). \quad (68)$$

In this case, in accordance with (63) probability  $P_{\text{FS}}^{(\text{B})}$  of traffic breakdown at the bottleneck during the time interval  $T_{\text{ob}}$  is

$$P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}) = 1 - \exp\left(-T_{\text{ob}}/T_{\text{FS}}^{(\text{B,mean})}\right). \quad (69)$$

Since  $T_{\text{FS}}^{(\text{B,mean})}$  in (69) is a function of  $q_{\text{sum}}$  and  $q_{\text{on}}$ , breakdown probability  $P_{\text{FS}}^{(\text{B})} = P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$ . Probability of traffic breakdown  $P_{\text{FS}}^{(\text{B})}$  for two different time intervals  $T_{\text{ob}}$  as functions of the flow rate  $q_{\text{sum}}$  calculated through formula (58), (60), and (69) is shown in Fig. 13. These dependences of breakdown probability on the flow rate are qualitatively the same as those found with the KKW-model (see Fig. 8).



**Traffic Breakdown, Probabilistic Theory of, Figure 13**

Probability of traffic breakdown  $P_{\text{FS}}^{(\text{B})}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}})$  for different time intervals for observing traffic flow  $T_{\text{ob}} = 30$  (curve 1) and 15 min (curve 2).  $q_{\text{determ,FS}}^{(\text{B})} = 2800$  vehicles/h,  $N_d = 16$ ,  $\xi_d = 2$

## Capacity of Free Flow at Bottlenecks

### Definition of Free Flow Capacity

Freeway capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck. During a given time interval  $T_{\text{ob}}$ , traffic breakdown can occur with probability  $P_{\text{FS}}^{(\text{B})}$ . In other words, an attribute of this *probabilistic free-flow capacity at the bottleneck* is the probability  $P_C^{(\text{B})}$  (63) that free flow remains at the bottleneck during the time interval  $T_{\text{ob}}$  [17]. Freeway capacity is reached when

$$P_C^{(\text{B})} < 1. \quad (70)$$

Thus we define freeway capacity in free flow at a bottleneck as follows [17]:

- Freeway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability  $P_C^{(\text{B})} < 1$  (70) during a given time interval  $T_{\text{ob}}$  for observing traffic flow. This means that at this flow rate with the probability  $P_{\text{FS}}^{(\text{B})} > 0$  an  $F \rightarrow S$  transition (traffic breakdown) occurs at the bottleneck during the time interval  $T_{\text{ob}}$ .

Because there can be an infinite number of flow rates downstream of the bottleneck for which the condition (70) is satisfied, there can also be an infinite number of freeway capacities in free flow at the bottleneck. Each of the freeway capacities has two attributes:

- (1) The probability  $P_C^{(\text{B})}$  (70) that free flow remains at the bottleneck during a given time interval  $T_{\text{ob}}$  for observing traffic flow.
- (2) The time interval  $T_{\text{ob}}$ .

When for free flow at the bottleneck rather than the condition (70) the condition

$$P_C^{(\text{B})} = 1 \quad (71)$$

is satisfied, then the flow rate downstream of the bottleneck in this free flow is lower than any of the freeway capacities.

It must be noted that flow rates are usually complicated time-functions in real traffic. For this reason, in empirical observations measured data is usually averaged for a time interval  $T_{\text{av}}$ . Thus empirical definition of freeway capacity at a bottleneck is as follows (for more detail, see Sect. 10.3.1 of the book [17] and ► [Traffic Congestion, Modeling Approaches to](#)):

- Freeway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability  $P_C^{(B)} < 1$  (70) during a given averaging time interval  $T_{av}$ .

### Infinite Free Flow Capacities and Diagram of Congested Patterns

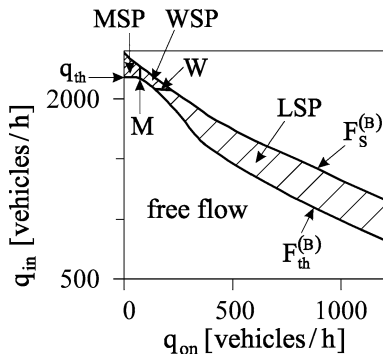
A diagram of congested patterns determines regions of the emergence and existence of various congested patterns occurring at a bottleneck as functions of flow rates upstream of the bottleneck. In particular, a part of the diagram for synchronized flow patterns (SP), which can occur at an on-ramp bottleneck due to traffic breakdown, in the flow-flow plane whose co-ordinates are  $q_{in}$  and  $q_{on}$  is presented in Fig. 14 found in a microscopic three-phase traffic theory. Let us show that boundaries on this diagram associated with emergence and dissolution of the SPs determine infinite free flow capacities at the bottleneck [17].

At the boundary  $F_S^{(B)}$  of the diagram an  $F \rightarrow S$  transition occurs during the time  $T_{ob}$  with breakdown probability

$$P_{FS}^{(B)} = 1. \quad (72)$$

This means that the boundary  $F_S^{(B)}$  determines an infinite number of *maximum* freeway capacities of free flow at the bottleneck  $q_{max}^{(free B)}$  for different flow rates  $q_{in}$  and  $q_{on}$  at the given  $T_{ob}$ .

There is also a region in the diagram of congested patterns at a bottleneck below and left of the boundary  $F_S^{(B)}$



**Traffic Breakdown, Probabilistic Theory of, Figure 14**

Freeway capacity and metastable region (hatched region) with respect to SP emergence and existence in the diagram of congested patterns at the on-ramp bottleneck found in a microscopic three-phase traffic theory. MSP is a moving SP, WSP is a widening SP, LSP is a localized SP (see explanations of SPs in ► [Traffic Congestion, Spatiotemporal Features of](#)). Taken from [17]

within which an  $F \rightarrow S$  transition occurs during the time  $T_{ob}$  with probability  $0 < P_{FS}^{(B)} < 1$ . The more distant a point within this region in the diagram from the boundary  $F_S^{(B)}$ , the smaller the probability  $P_{FS}^{(B)}$  of traffic breakdown.

For this reason, there is a *threshold* boundary  $F_{th}^{(B)}$  in the diagram (Fig. 14), which restricts the above region of the diagram within which  $0 < P_{FS}^{(B)} < 1$ . Below and left of the threshold boundary  $F_{th}^{(B)}$  in the diagram, breakdown probability

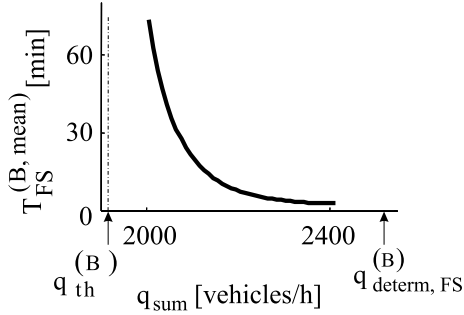
$$P_{FS}^{(B)} = 0, \quad (73)$$

i. e., no traffic breakdown is possible during the time  $T_{ob}$ . There are an infinite amount of *threshold* flow rates downstream of an on-ramp bottleneck  $q_{th}^{(B)} = q_{on} + q_{in}$  for traffic breakdown at the bottleneck associated with different points  $(q_{on}, q_{in})$  in the diagram of congested patterns at the threshold boundary  $F_{th}^{(B)}$ . As above mentioned, at the flow rate  $q_{sum} < q_{th}^{(B)}$  (57) no traffic breakdown is possible, i. e., the flow rate is smaller than any of the freeway capacities in free flow at the bottleneck. This means that the threshold boundary  $F_{th}^{(B)}$  determines the infinite number of *minimum* freeway capacities in free flow at the bottleneck  $q_{th}^{(B)}$  for different  $q_{in}$  and  $q_{on}$  at the given  $T_{ob}$ .

Different points  $(q_{on}, q_{in})$  for an on-ramp bottleneck between the boundaries  $F_S^{(B)}$  and  $F_{th}^{(B)}$  in the diagram of congested patterns are related to the infinite amount of the flow rates in free flow downstream of the bottleneck associated with the infinite number of freeway capacities in free flow at the bottleneck. This is because for each of these flow rates the probability  $P_C^{(B)}$  (63) satisfies the condition

$$0 < P_C^{(B)} < 1. \quad (74)$$

Thus free flow, which is observed during the time interval  $T_{ob}$ , is in a metastable state with respect to traffic breakdown on and between the boundaries  $F_S^{(B)}$  and  $F_{th}^{(B)}$  in the diagram of congested patterns. This metastability of free flow means that either spontaneous or induced traffic breakdown is possible. In the case of a spontaneous breakdown, a critical cluster (nucleus) for the breakdown can occur spontaneously at the bottleneck leading to traffic breakdown. In the case of an induced breakdown, the role of the cluster plays a congested pattern that has initially occurred at another location as the bottleneck location; due to pattern propagation, the induced traffic breakdown occurs after the pattern reaches the bottleneck (see a more detailed explanation of spontaneous and induced traffic breakdowns in ► [Traffic Congestion, Modeling Approaches to](#)).



**Traffic Breakdown, Probabilistic Theory of, Figure 15**  
Mean time delay for traffic breakdown  $T_{FS}^{(B, \text{mean})}$  (51) as functions of the total flow rate  $q_{\text{sum}}$  for  $q_{\text{on}} = 300$  vehicles/h

### Probability of Breakdown and Diagram of Congested Patterns

A link between the mean time delay  $T_{FS}^{(B, \text{mean})}$  in traffic breakdown, traffic breakdown probability  $P_{FS}^{(B)}$ , and the diagram of congestion patterns (Fig. 14) discussed above is confirmed by numerical simulations presented in Figs. 15 and 16 [20,21,22]. For the numerical simulations, we used an example of the function  $w_{-}(N)$  (48)

$$w_{-}(N) = N \left[ \frac{a(q_{\text{on}})}{1 + (N/N_0(q_{\text{on}}))^4} + b(q_{\text{on}}) \right], \quad (75)$$

which exhibits qualitatively the same shape as that shown in Fig. 11. The analytical function (75) allows us to perform a numerical analysis of the mean time delay (51) and the associated nucleation rate (55) for the breakdown phenomenon ( $F \rightarrow S$  transition). For the analysis of (51) and (55), only branches  $N^{(\text{determ})}$  and  $N_c$  (Fig. 11) of the function (75) associated with the deterministic and critical clusters within which  $N \leq N_s$  are relevant. This is because

the maximum possible value of  $N = N_2$  in the potential barrier  $\Delta\Phi$  (52) that determines the nucleation rate (55) is equal to  $N_s$ . The nucleation model cannot be applied for  $q_{\text{on}} = 0$ ; therefore, in Fig. 16 points in the vicinity of  $q_{\text{on}} = 0$  show only the tendency of the boundaries and the flow rates for the limiting case of small values  $q_{\text{on}}$  in which, however, the on-ramp inflow rate  $q_{\text{on}} > 0$ , specifically, it is assumed that the deterministic cluster still exists at the bottleneck.

A numerical study shows that the flow rate dependence of the mean time delay  $T_{FS}^{(B, \text{mean})}$  (51) (Fig. 15) exhibits qualitative features found in a microscopic three-phase traffic theory [19,23].

In the diagram of congested patterns at the bottleneck (Fig. 16a), there are boundaries  $F_{S, \zeta}^{(B)}$  (curves 1 and 2) each of them is found from the condition that the breakdown occurs during the time  $T_{\text{ob}}$  with the probability  $P_{FS}^{(B)}$  that is equal to a given value  $\zeta$ :

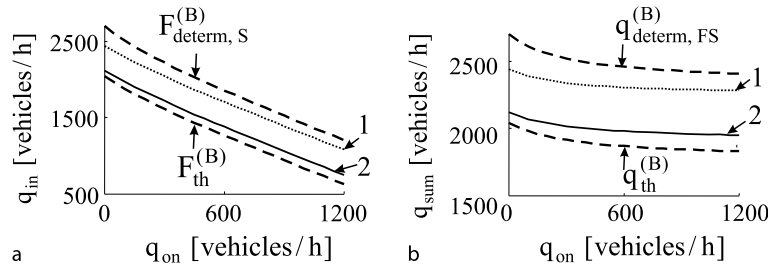
$$P_{FS}^{(B)}(T_{\text{ob}}, q_{\text{sum}}, q_{\text{on}}) = \zeta, \quad \zeta = \text{const}. \quad (76)$$

At the boundary  $F_{S, \zeta}^{(B)}$ , the flow rate

$$q_{\text{sum}} = q_G^{(B)}(q_{\text{on}}, T_{\text{ob}}) \quad (77)$$

depends on  $q_{\text{on}}$  and  $T_{\text{ob}}$ . The boundary  $F_{S, \zeta}^{(B)}$  at  $\zeta \approx 1$  (curve 1 in Fig. 16a) is qualitatively similar with the boundary  $F_S^{(B)}$  found in numerical simulation of a three-phase microscopic traffic flow model shown in Fig. 14. In the diagram, there is also the threshold boundary  $F_{\text{th}}^{(B)}$  (curve  $F_{\text{th}}^{(B)}$  in Fig. 16a) at which (44) is satisfied. The boundary  $F_{\text{determ}, S}^{(B)}$  determines the on-ramp inflow dependence of the critical flow rate for the deterministic breakdown.

Boundaries for traffic breakdown can also be calculated for another congested pattern diagram whose co-



**Traffic Breakdown, Probabilistic Theory of, Figure 16**

Boundaries of constant values of breakdown probability  $P_{FS}^{(B)}$  in diagrams of congested patterns in the flow-flow planes  $q_{\text{in}}(q_{\text{on}})$  (a) and  $q_{\text{sum}}(q_{\text{on}})$  (b).  $F_{\text{determ}, S}^{(B)}$  and  $q_{\text{determ}, FS}^{(B)}$  (5) are the critical boundary for deterministic traffic breakdown and the critical flow rate for deterministic traffic breakdown, respectively.  $F_{\text{th}}^{(B)}$  and  $q_{\text{th}}^{(B)}$  (44) are the threshold boundary and the threshold flow rate for traffic breakdown, respectively. Curves 1, 2 are related to different chosen values  $P_{FS}^{(B)}$  in (76):  $\zeta = 0.986$  (curves 1) and  $\zeta = 0.22$  (curves 2) at  $T_{\text{ob}} = 15$  min



ordinates are  $q_{\text{sum}}$  and  $q_{\text{on}}$  (Fig. 16b). Curves 1 and 2 in Fig. 16b show on-ramp inflow dependences of the critical flow rate for traffic breakdown associated with two different values  $\zeta$  of breakdown probability (76). In this diagram, the boundaries  $q_{\text{determ,FS}}^{(B)}$  and  $q_{\text{th}}^{(B)}$  determine the on-ramp inflow dependences of the critical flow rate for the deterministic breakdown and the threshold flow rate for traffic breakdown, respectively.

## Conclusions

Probabilistic theory of traffic breakdown [15,17,20,21,22] discussed in this review article explains fundamental empirical features of traffic breakdown A-D (Subsect. “Fundamental Empirical Features of Traffic Breakdown”). The nucleation theory of traffic breakdown allows us to conclude that in accordance with measured traffic data traffic breakdown is a local phase transition from free flow to synchronized flow ( $F \rightarrow S$  transition). The  $F \rightarrow S$  transition exhibits features of first-order phase transitions observed in diverse complex systems of natural science: (i) an  $F \rightarrow S$  transition can be either spontaneous or induced; (ii) emergence and dissolution of synchronized flow is accompanied by a hysteresis effect; (iii) breakdown probability is an increasing function of control parameters (flow rates).

## Future Directions

Whereas empirical *macroscopic* features of traffic breakdown have been found (Subsect. “Fundamental Empirical Features of Traffic Breakdown”), there are almost no empirical investigations of *microscopic* empirical characteristics of the breakdown. In particular, behavior of speed fluctuations in real free flow at bottlenecks, appearance and growth of critical fluctuations (critical vehicle clusters) leading to the breakdown have not been understood. In general, empirical and theoretical analyzes of various sources of traffic flow fluctuations and their behavior near traffic breakdown should be a very important direction in the field of traffic flow dynamics.

Although some *spatiotemporal* features of nuclei for traffic breakdown and synchronized flow are incorporated into the KKW CA traffic flow model (Sect. “Probabilistic Description of Traffic Breakdown with Cellular Automata (CA) Traffic Flow Model”), this is not made for the analytical study of traffic breakdown made in Sect. “Probabilistic Description of Traffic Breakdown Based on Master Equation”. In the latter case, to use the well-known probabilistic theory of first-order phase transitions based on a master equation and one step processes [7], possible var-

ious spatial shapes of nuclei have been ignored and suggested that the cluster dynamics can be described by the dynamic behavior of the vehicle number within the cluster only. Therefore, we believe that there is a great potential for further analytical probabilistic theories of traffic breakdown in which spatiotemporal character of vehicle clusters at bottlenecks is taken into account.

## Acknowledgments

We would like to thank Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald-Hiller and Olivia Brickley for help and useful suggestions.

## Bibliography

1. Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
2. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
3. Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
4. Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier, New York
5. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Transp Res Rec* 1484:80–89
6. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and granular flow' 01*. Springer, Heidelberg
7. Gardiner CW (1994) *Handbook of stochastic methods for physics, chemistry, and the natural sciences*. Springer, Berlin
8. Gartner NH, Messer CJ, Rathi A (eds) (1997) Special report 165: Revised monograph on traffic flow theory. Trans Res Board, Washington DC
9. Haight FA (1963) *Mathematical theories of traffic flow*. Academic Press, New York
10. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Trans Res Rec* 1320:91–98
11. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
12. Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
13. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and granular flow' 99*. Springer, Heidelberg
14. Kerner BS (1998) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
15. Kerner BS (2000) Theory of breakdown phenomenon at highway bottlenecks. *Trans Res Rec* 1710:136–144
16. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
17. Kerner BS (2004) *The physics of traffic*. Springer, Berlin, New York
18. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35:L31–L43

19. Kerner BS, Klenov SL (2003) Microscopic theory of spatial-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
20. Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase traffic theory. [arXiv:cond-mat/0502281](https://arxiv.org/abs/cond-mat/0502281)
21. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic. *Physica A* 364:473–492
22. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. *Transp Res Rec* 1965:70–78
23. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35:9971–10013
24. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2002) Probabilistic description of traffic breakdown. *Phys Rev E* 65:066125
25. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2004) Probabilistic description of traffic breakdown caused by on-ramp. [arXiv:cond-mat/0405163](https://arxiv.org/abs/cond-mat/0405163)
26. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
27. Lorenz M, Eleftheriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. *Trans Res Cir E-C018*:84–95
28. Mahnke R, Kaupužs J (1999) Stochastic theory of freeway traffic. *Phys Rev E* 59:117–125
29. Mahnke R, Pieret N (1997) Stochastic master-equation approach to aggregation in freeway traffic. *Phys Rev E* 56:2666–2671
30. Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
31. May AD (1990) Traffic flow fundamentals. Prentice-Hall, New Jersey
32. Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
33. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys France I* 2:2221–2229
34. Nagel K, Wagner P, Woesler R (2003) Still flowing: Approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–710
35. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Trans Res Rec* 1634:64–69
36. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
37. Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
38. Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451

## Traffic Congestion, Modeling Approaches to

BORIS S. KERNER  
GR/ETI, HPC: G021, Daimler AG,  
Sindelfingen, Germany

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Free and Congested Traffic  
 Three Traffic Phases  
 Characteristic Parameters of Wide Moving Jam  
   Propagation and Line *J*  
 Traffic Breakdown and Highway Capacity  
 Moving Jam Emergence in Synchronized Flow  
   (*S* → *J* Transition)  
 Empirical Double *Z*-Characteristic  
   for Phase Transitions in Traffic Flow  
 Hypotheses of Three-Phase Traffic Theory  
   as the Result of Traffic Phase Definitions  
 Critical Discussion of Fundamental Diagram Modeling  
   Approach to Traffic Congestion  
 Three-Phase Traffic Flow Models  
 Link Between Three-Phase Traffic Theory  
   and Fundamental Diagram Approach  
 Conclusions  
 Future Directions  
 Cross References  
 Acknowledgments  
 Bibliography

### Glossary

**Free flow** Free flow is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

**Traffic breakdown** In empirical observations, when density in free flow increases and becomes great enough, the phenomenon of the onset of congestion is observed in this free flow: The average speed decreases sharply to a lower speed in congested traffic. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown.

**Congested traffic** Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow.

**Bottleneck** Traffic breakdown occurs mostly at highway bottlenecks. Just as defects and impurities are important for phase transitions in complex spatially distributed systems of various nature, so are freeway bottlenecks in freeway traffic. The bottleneck can be a result of roadworks, on- and off-ramps, a decrease in the number of freeway lanes, road curves and road gradients, etc.

**Moving jams** A moving jam is a localized structure of great vehicle density spatially limited by two jam fronts. Within the downstream jam front vehicles accelerate escaping from the jam; within the upstream jam front, vehicles slow down approaching the jam. The jam as a whole structure propagates upstream in traffic flow. Within the jam (between the jam fronts) vehicle density is great and speed is very low (sometimes as low as zero). A sequence of moving jams is often called “stop-and-go” traffic.

**Traffic flow model** A traffic flow model is devoted to the explanation and simulation of traffic flow phenomena, which are observed in measured data of real traffic flow, and to the prediction of new traffic flow phenomena that could be found in real traffic flow. First of all, a traffic flow model should explain and predict empirical (measured) features of traffic breakdown.

**Fundamental diagram of traffic flow** The fundamental diagram is a relationship between the flow rate and density in vehicle traffic. Because the flow rate is the product of the average speed and density, the fundamental diagram is associated with a relationship between these traffic variables. In accordance with an obvious result of traffic measurements, on average the speed decreases when the density increases. Thus in the flow-density plane, the fundamental diagram should pass through the origin (when the density is zero so is the flow rate) and should have at least one maximum. The fundamental diagram gives also a connection between the average space gap (net distance) between vehicles and the average speed.

**Steady states of traffic flow** Steady states of traffic flow are hypothetical states of homogeneous (in time and space) traffic flow of identical vehicles (and identical drivers) in which all vehicles move with the same time-independent speed and have the same space gaps.

**Fundamental diagram approach to traffic flow theory and modeling** The fundamental hypothesis of the fundamental diagram approach to traffic flow theory and modeling suggests that steady states of both free flow and congested traffic lie on a one-dimension curve(s) (i. e., on a theoretical fundamental diagram of traffic flow) in the flow-density plane. At each given time-independent speed of the preceding vehicle, the theoretical fundamental diagram determines a single desired space gap at which a vehicle moves with this time-independent speed while following the preceding vehicle. This model vehicle behavior is related to a steady state of traffic flow associated with a hypothetical noiseless and acceleration less (deceleration less) model limit.

**Three-phase traffic theory** In the author’s three-phase traffic theory, besides the free flow phase there are two phases in congested traffic, the synchronized flow and wide moving jam phases. The synchronized flow and wide moving jam phases in congested traffic are defined through empirical spatiotemporal criteria (definitions) [S] and [J]. In contrast with the fundamental diagram approach, the fundamental hypothesis of three-phase traffic theory suggests that in steady states of *synchronized flow*, at each given time-independent speed of the preceding vehicle, there is an infinite number of space gaps at which the vehicle can move with this speed, while following the preceding vehicle. Thus hypothetical steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane, i. e., there is no desired space gap in hypothetical steady states of synchronized flow in this theory.

**Wide moving jam traffic phase** In three-phase traffic theory, the following definition of the wide moving jam phase [J] in congested traffic based on measured traffic data is made. A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks. This is the characteristic feature of the wide moving jam phase.

**Synchronized flow traffic phase** In three-phase traffic theory, the following definition of the synchronized flow phase [S] in congested traffic based on measured traffic data is made. The downstream front of synchronized flow does not exhibit the above mentioned characteristic feature of wide moving jams; specifically, the latter front is often fixed at a bottleneck.

**Microscopic criterion for traffic phases** Within wide moving jams, there are regions in which traffic flow is interrupted: the inflow into the jam has no influence on the jam outflow. A sufficient condition for flow interruption determines the microscopic criterion for the wide moving jam phase. Based on this criterion both the synchronized flow and wide moving jam phases can be identified in congested traffic from single vehicle data measured even at one freeway location. This is because if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase.

**F → S transition** In all known observations, traffic breakdown is a phase transition from the free flow phase to synchronized flow phase (F → S transition). Thus the

terms *traffic breakdown* and an  $F \rightarrow S$  transition are synonyms related to the same phenomenon of the onset of congestion in free flow.

**Highway capacity of free flow at bottleneck** Highway (freeway) capacity of free flow at a bottleneck (called also bottleneck capacity) is limited by traffic breakdown, i.e.,  $F \rightarrow S$  transition at the bottleneck, which occurs with probability (denoted by  $P_{FS}^{(B)}$ ) during a given averaging time interval (denoted by  $T_{av}$ ) for traffic variables. Highway capacity in free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck during the time interval  $T_{av}$  with probability  $P_C^{(B)} = 1 - P_{FS}^{(B)}$ , which is less than one; thus highway capacity has two attributes  $P_C^{(B)}$  and  $T_{av}$ . At each given  $T_{av}$ , there are infinite highway capacities in a limited range associated with different probabilities  $P_C^{(B)}$ .

**Spontaneous traffic breakdown (spontaneous  $F \rightarrow S$  transition) at bottleneck** If before traffic breakdown free flow has been at a bottleneck as well as upstream and downstream in a neighborhood of the bottleneck, the breakdown is caused by occurrence and subsequent growth of speed disturbances (fluctuations) within the free flow at the bottleneck. Such traffic breakdown is called a *spontaneous* traffic breakdown at the bottleneck. A speed disturbance begins to grow, i.e., the speed within the disturbance begins to decrease over time with the subsequent breakdown at the bottleneck, if within the disturbance the initial speed is equal to or lower than a critical speed for an  $F \rightarrow S$  transition. Such a critical speed disturbance can be considered a nucleus for the breakdown at the bottleneck. There can be various sources for speed disturbances whose occurrence and subsequent growth lead to spontaneous traffic breakdown: unexpected vehicle braking and/or lane changing in free flow, fluctuations in flow rates upstream of the bottleneck, vehicle merging from other roads in the bottleneck neighborhood (e.g., at on-ramp bottlenecks), etc.

**Induced traffic breakdown (induced  $F \rightarrow S$  transition) at Bottleneck** In contrast with spontaneous traffic breakdown, which occurs when before the breakdown free flow is in a neighborhood of the bottleneck, an induced traffic breakdown at the bottleneck is caused by the propagation of a moving spatiotemporal *congested* traffic pattern, which has initially occurred at a *different* road location (e.g., at another bottleneck). When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The congested pattern whose propaga-

tion causes the breakdown at the bottleneck can be considered an external speed disturbance at the bottleneck.

**Spontaneous wide moving jam emergence in synchronized flow** Within synchronized flow, there can be speed disturbances (fluctuations) whose growth leads to wide moving jam emergence. Such wide moving jam emergence is called *spontaneous* wide moving jam emergence in synchronized flow (spontaneous  $S \rightarrow J$  transition). A speed disturbance begins to grow, if within the disturbance the initial speed is equal to or lower than a critical speed for an  $S \rightarrow J$  transition. Such a critical speed disturbance can be considered a nucleus for the  $S \rightarrow J$  transition. The growing speed disturbance in synchronized flow propagates upstream; for this reason, in contrast with spontaneous traffic breakdown at a bottleneck, the  $S \rightarrow J$  transition occurs usually upstream of the road location at which the critical speed disturbance has initially appeared. There can be various sources for speed disturbances whose occurrence and subsequent growth lead to spontaneous  $S \rightarrow J$  transition: unexpected vehicle braking and/or lane changing within synchronized flow, fluctuations in upstream flow rates, vehicle merging from other roads, etc.

### Definition of the Subject

Understanding of vehicular traffic congestion is the key for effective traffic management, traffic control, organization, and other engineering applications, which should improve traffic safety and result in high-quality mobility. In empirical observations, traffic congestion occurs as the result of a so-called traffic breakdown: the vehicle speed decreases sharply and vehicle density increases in an initially free traffic flow. The subsequent development of congested traffic exhibits a very complex spatiotemporal behavior. To explain traffic breakdown and resulting traffic congested patterns a huge number of traffic theories and models have been developed. It is clear that if a traffic flow model cannot show empirical features of traffic breakdown, the model cannot also show and predict many other traffic phenomena observed in real congested traffic. Thus an assessment of modeling approaches to show and predict traffic breakdown in vehicle traffic is the starting point for a further consideration of the reliability of these modeling approaches to show and predict traffic flow phenomena observed in real congested traffic. Furthermore, the development of modeling approaches to traffic congestion, which are consistent with empirical results, is of great importance for development of traffic control meth-



ods as well as management strategies, which should either prevent traffic breakdown or lead to dissolution of existing congested patterns.

Modeling approaches for the explanation of traffic breakdown include traffic flow models based on a diverse variety of methods of complexity and systems science. This is explained by an extremely complex non-linear character of real vehicular traffic. Moreover, there are links between spatiotemporal phenomena associated with vehicular traffic congestion and traffic phenomena observed in many other complex systems like pedestrian traffic, which is responsible for the complexity of crowd and evacuation dynamics, and air traffic. For this reason, the understanding of vehicular traffic congestion is also very important for the further progress in understanding of many other many-particle complex systems.

## Introduction

Earlier traffic flow models, which claim to explain and predict the onset of traffic congestion, are reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. All these models are associated with the fundamental diagram approach to traffic flow modeling.

In 1955 Lighthill and Whitham [86] wrote in their classic work (see p. 319 in [86]): “...The fundamental hypothesis of the theory is that at any point of the road the flow (vehicles per hour) is a function of the concentration (vehicles per mile)...” (Fig. 1a). This hypothesis is also the fundamental one for all traffic flow theories and models reviewed in [7,16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. The fundamental diagram for a traffic flow model means that in all these traffic flow models it is suggested that for a given time-independent speed of the preceding vehicle there is a single desired (or optimal) space gap for the following vehicle that the vehicle chooses in a hypothetical acceleration less (deceleration less) and noiseless model limit case.

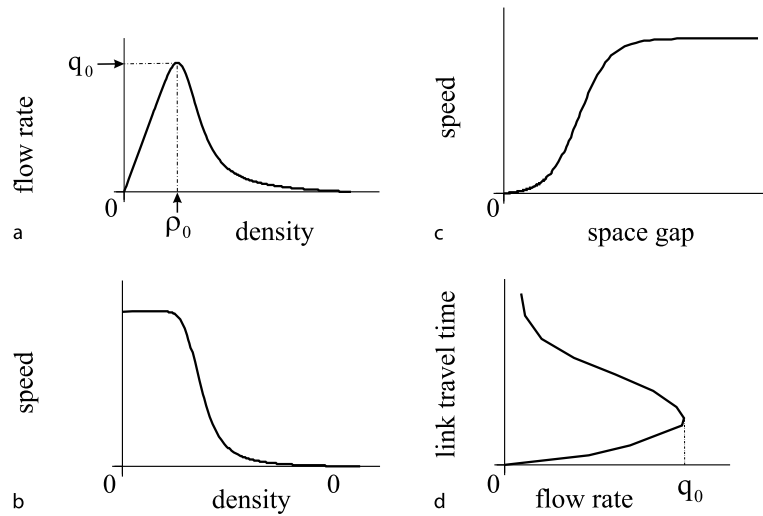
The models within the framework of the fundamental diagram can be classified into two main classes. The first model class refers to the classic LWR (Lighthill–Whitham–Richards) model introduced in 1955–1956 [86,109] (see also references in [16,21,33,37,40,84,93,126]). Examples for this model class are the cell-transmission models of Daganzo [20]. The basic idea of the LWR-model is that maximum flow rate (denoted  $q_0$  in Fig. 1a), associated with the maximum point at the fundamental diagram, determines capacity of free flow at a bottleneck, i. e., the bottleneck capacity. Thus if the flow rate upstream of the bottleneck exceeds the bottleneck capacity, then traffic breakdown should occur.

The second model class refers to the classic GM (General Motors) model of Herman, Gazis, Rothery, Montroll and Potts introduced in 1959–1961 [34,43]. The basic idea of the GM-model approach is as follows. Beginning at a critical density, there is instability of steady model states at the fundamental diagram. This model instability that should explain traffic breakdown is associated with a finite value of driver reaction time. The instability can qualitatively be explained, if we suggest that in a homogeneous traffic flow a vehicle decelerates unexpectedly, which causes a local speed decrease (speed disturbance) in the flow. Due to the driver reaction time, the following vehicle can decelerate stronger than it would be necessary to avoid collision. This effect is called the over-deceleration effect (or human over-reaction). As a result of over-deceleration, the speed of this vehicle can become lower than the speed of the preceding vehicle. The same over-deceleration effect can occur for many other following vehicles that can lead to the growth of the initial speed disturbance, i. e., to traffic flow instability. Examples for the GM model class are the optimal velocity (OV) models by Newell, by Whitham, and by Bando et al., the Payne macroscopic model and its variants, the Wiedemann psychophysical traffic flow model and its variants, the Gipps-model, the NaSch (Nagel–Schreckenberg) cellular automata (CA) model and its variants, the Krauß-model, the IDM-model (Intelligent Driver Model) of Treiber et al. as well as a huge number of other traffic flow models (see below and references in [16,18,33,37,40,84,92,93,95,98,102,126,127,128]).

Naturally, traffic flow models should explain empirical spatiotemporal features of traffic congestion. However, it is only recently that spatiotemporal features of congested traffic patterns have been adequately understood [52]. Consequently, these earlier traffic flow theories and models cannot explain and predict many of these empirical spatiotemporal traffic pattern features. When these traffic flow models are used for simulations of freeway control and management strategies, the related simulations cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated control strategy.

To understand this critical conclusion, let us consider the empirical basis of a theoretical fundamental diagram. In all observations of traffic flow the greater the vehicle density, the lower the average vehicle speed (Fig. 1b). This should prove the theoretical fundamental diagram as the basis for traffic flow theories and models. However, this obvious and correct empirical result is associated with the *averaging* of measured traffic data. In particular, if traffic data is associated with measurements of the speed and flow rate at a freeway location(s) during some long enough av-





**Traffic Congestion, Modeling Approaches to, Figure 1**

Qualitative example of fundamental diagram: **a** Flow-density relationship (fundamental diagram). **b,c** The speed-density (**b**), speed-space-gap (**c**) and link-travel-time-flow (**d**) relationships associated with **a**

eraging time interval, then to find a point at the fundamental diagram related to a chosen density, all measured data for different flow rates and speeds associated with densities within a small density range around the chosen density should be averaged to one value of the flow rate (Fig. 1a) and associated average speed (Fig. 1b). By this data averaging almost all information about the *spatiotemporal* structure of traffic patterns propagating through the freeway location(s) at which these measurements have been made is *lost*. The remaining information that determines a fundamental diagram shape is associated with averaging characteristics of traffic patterns, which are most frequently observed during the averaging time interval at the related freeway location(s).

Because of the failure of earlier traffic flow theories and models in explanation of traffic congestion, in 1996–1999 the author introduced a three-phase traffic theory [52], which explains empirical features of traffic breakdown and resulting congested patterns.

In this article, we will try to explain this dramatic development of traffic flow theories to traffic congestion. The article is organized as follows. In Sects. “Free and Congested Traffic”–“Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow”, we consider empirical features of traffic breakdown and moving jam emergence in real freeway traffic; here we give also a brief explanation of these empirical traffic flow features based on three-phase traffic theory. In Sect. “Hypotheses of Three-Phase Traffic Theory as The Result of Traffic Phase Definitions”, we show that hypotheses of three-phase traffic

theory, which explain the empirical features of traffic congestion, are the result of the traffic phase definitions made in this theory. In Sect. “Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion”, based on a consideration of traffic flow theories and models in the context of the fundamental diagram approach we explain that and why these models cannot show features of real traffic breakdown. We discuss also achievements of the fundamental diagram approach in explanation of empirical wide moving jam propagation and characteristic parameters of wide moving jams. In Sect. “Three-Phase Traffic Flow Models”, we briefly discuss a three-phase traffic flow model, which has been used to derive simulation results explaining empirical features of traffic congestion in Sects. “Free and Congested Traffic”–“Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow”. A link between three-phase traffic theory and the fundamental diagram approach is discussed in Sect. “Link Between Three-Phase Traffic Theory and Fundamental Diagram Approach”.

### Free and Congested Traffic

At small enough vehicle density, interactions between vehicles in free flow are negligible. Therefore, vehicles have an opportunity move with their desired maximum speeds (if this speed is not restricted by road conditions or traffic regulations). When the density increases, vehicle interaction cannot be neglected any more. As a result of vehicle interaction in free flow, the average speed decreases with

increase in density (measured data left of the dashed line  $FC$  in Fig. 2a).

This common vehicle behavior in free flow leads to the *fundamental diagram for free flow*, i. e., a certain curve with a positive slope between the flow rate and density associated with averaging of measured data shown left of the dashed line  $FC$  in Fig. 2a to one average flow rate for each density (curve  $F$  in Fig. 2b) (e. g., [16,18,21,40,42,44,89,93,101,108,126,127]).

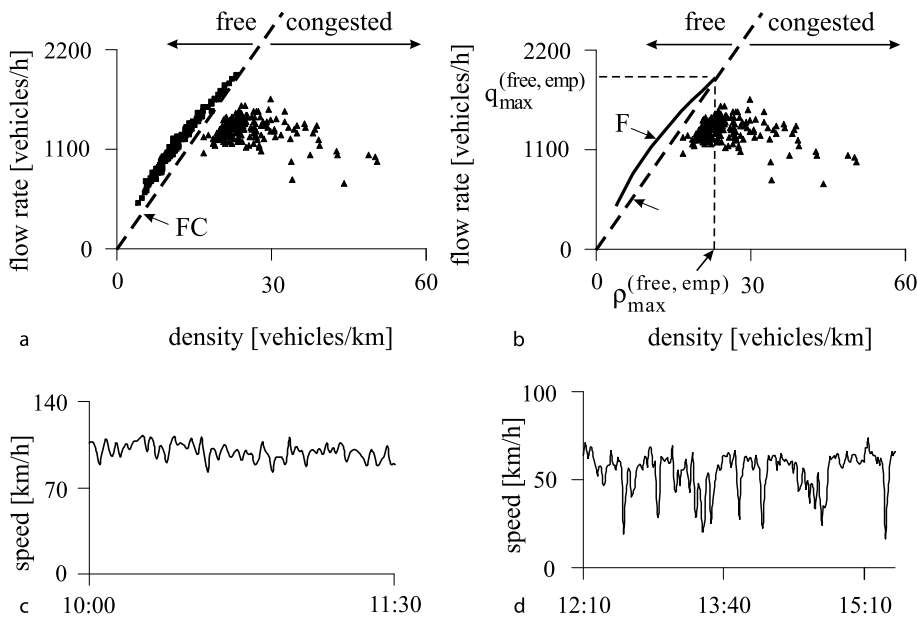
The fundamental diagram for free flow is cut off at some empirical limit (maximum) point of free flow (denoted  $(\rho_{\max}^{(\text{free,emp})}, q_{\max}^{(\text{free,emp})})$  in Fig. 2b) (e. g., [39,93]). At this limit point of free flow, the average vehicle speed has a minimum possible value for free flow. Thus empirical points of free flow as well as the associated fundamental diagram lie to the left of the dashed line  $FC$  in the flow-density plane whose slope equals this minimum possible speed in free flow  $v_{\min}^{(\text{free,emp})}$ .

When the vehicle density is great enough, traffic is usually in a congested state (e. g., [16,18,21,40,42,44,89,93,101,108,126,127]). We define traffic phenomena in terms of their empirical features found in measured data rather than their genesis. Often, congested traffic is defined as a traffic state, which merely results from a freeway bottleneck as the inevitable consequence of an upstream flow

that exceeds the bottleneck capacity. In this definition it seems obvious to suggest that congested flow is not self-organized, but rigidly controlled by external constraints. However, it has been shown that a variety self-organized processes are responsible for features of real traffic congestion [52]. For these reasons, we use the following *definition* of congested traffic.

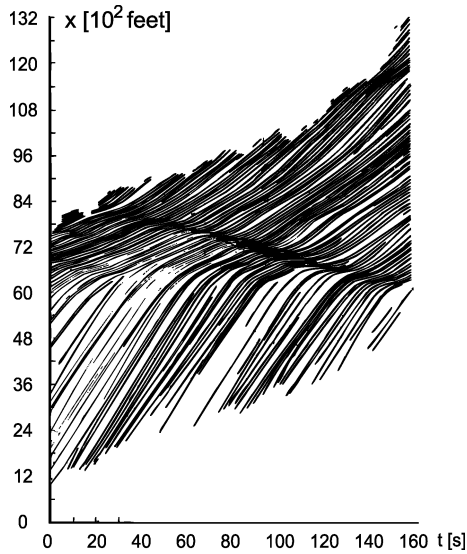
Congested traffic states will be defined as complementary to states of free flow (e. g., [76]). It has already been noted that empirical points related to free flow can be approximately presented by a curve with a positive slope in the flow-density plane (Fig. 2b). Congested traffic will be defined, therefore, as a state of traffic where the average vehicle speed is *lower* than the minimum possible speed in free flow, which is related to the limit point  $(\rho_{\max}^{(\text{free,emp})}, q_{\max}^{(\text{free,emp})})$  in free flow (Fig. 2b). Thus empirical points of congested traffic lie to the right of the dashed line  $FC$  in the flow-density plane whose slope equals the minimum possible speed in free flow.

In this definition of congested traffic, nothing is said about the origin of the congestion. It is related to empirical facts that a congested state can occur spontaneously also away from bottlenecks and that many diverse self-organizing effects are responsible for traffic congestion [52], some of them will be considered in the article.



**Traffic Congestion, Modeling Approaches to, Figure 2**

Free flow and congested traffic (e. g., [76]). **a** Data for free flow (points left of the dashed line  $FC$ ) and for congested traffic (points right of the dashed line  $FC$ ) measured at a road location. **b** The fundamental diagram for free flow (curve  $F$ ) and the same measured data for congested traffic as those in **a**. **c,d** Vehicle speed in free flow (**c**) and congested traffic (**d**), related to points left and right of the line  $FC$  in **a**, respectively. 1-min average data



**Traffic Congestion, Modeling Approaches to, Figure 3**  
An empirical moving jam: traffic dynamics derived from aerial photography. Taken from Treiterer [121]

In congested traffic, the “stop-and-go” phenomenon is observed, i. e., a sequence of different moving traffic jams can appear [26,27,28,29,75,76,121,122,123]. Moving jams have been studied empirically by many authors, in particular, in classic empirical works by Edie et al. [26,27,28,29], Treiterer et al. [121,122,123] (Fig. 3) and Koshi et al. [75,76].

### Three Traffic Phases

A traffic phase is a traffic state considered in space and time that possesses specific *empirical spatiotemporal* features. These features are specific (unique) only to this traffic phase. Note that a traffic state is characterized by a certain set of statistical properties of traffic variables. Examples of traffic variables are the flow rate  $q$  (vehicles/h), vehicle speed  $v$  (km/h), vehicle space gap (vehicle space gap is also called as net distance or space headway)  $g$  (m), time headway (s) that is also called time space or time gap between vehicles, and vehicle density  $\rho$  (vehicles/km).

Based on investigations of congested spatiotemporal patterns measured on different freeways over many days and years (see references in [52]), the three-phase traffic theory introduced by the author suggests that besides the free flow traffic phase there are two other traffic phases in congested traffic: synchronized flow and wide moving jam.

Traffic flow consists of free flow and congested traffic. Congested traffic consists of the synchronized flow phase

and the wide moving jam phase. Thus, there are three traffic phases in three-phase traffic theory:

- Free flow
- Synchronized flow
- Wide moving jam

### Empirical Macroscopic Criteria for Defining Phases in Congested Traffic

The synchronized flow and wide moving jam phases in congested traffic are defined through the use of the following *empirical (objective) criteria for the traffic phases in congested traffic*. These empirical criteria are related to some characteristic spatiotemporal empirical features of the traffic phases [52]:

- [J] The wide moving jam traffic phase is defined as follows. A wide moving jam is a moving jam that maintains the mean velocity of the downstream *front* of the jam,  $v_g$ , as the jam propagates. Vehicles accelerate within the downstream jam front from low speed states (sometimes as low as zero) inside the jam to higher speeds downstream of the jam. On average a wide moving jam maintains the mean velocity of the downstream jam front, even as it propagates through other (possibly very complex) traffic states or freeway bottlenecks. This is a characteristic feature of the wide moving jam traffic phase.
- [S] The synchronized flow traffic phase is defined as follows. In contrast to the wide moving jam traffic phase, the downstream front of the synchronized flow traffic phase does *not* maintain the mean velocity of the downstream front. In particular, the downstream front of synchronized flow is often *fixed* at a freeway bottleneck. In other words, the synchronized flow traffic phase does not show aforementioned characteristic feature of the wide moving jam traffic phase.

The downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow upstream of the front to higher speeds in free flow downstream of the front.

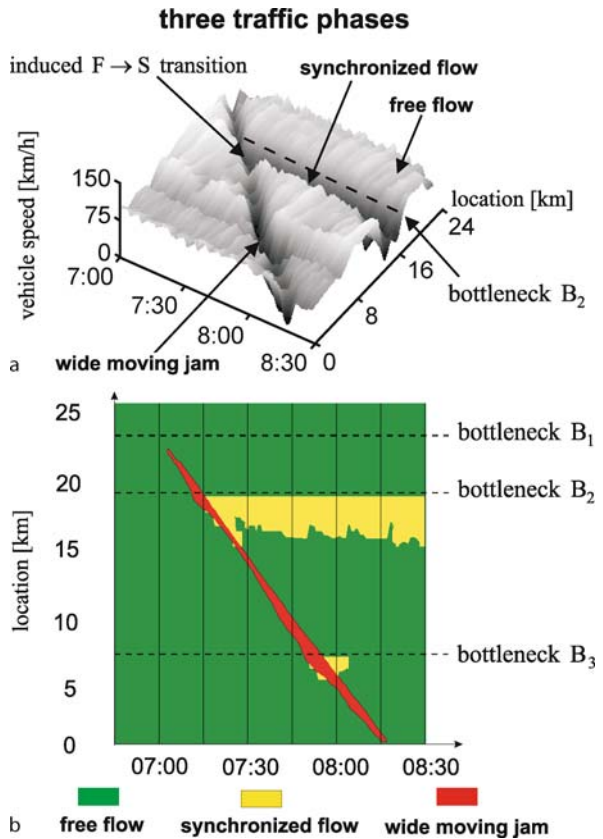
Thus the definitions of the traffic phases in congested traffic are made via the spatiotemporal empirical criteria [J] and [S]. The definitions [J] and [S] are associated with dynamic behavior of the downstream front of these phases, while a wide moving jam or synchronized flow propagates through other traffic states or bottlenecks.

The definitions [J] and [S] of the traffic phases in congested traffic mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is associated with the synchronized flow phase. This is because congested traffic can be either within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. In particular, it must be stressed that a moving jam can be associated either with a wide moving jam or with synchronized flow. In accordance with the phase definitions [J] and [S], this depends on the dynamic behavior of the downstream jam front, while the jam propagates through other traffic states or bottlenecks.

One example of application of the objective criteria (definitions) for the traffic phases in congested traffic [J] and [S] is shown in Fig. 4. In Fig. 4, a moving jam propagating through the location of bottleneck  $B_2$  induces the synchronized flow phase at this bottleneck (this induced phase transition is labeled “induced  $F \rightarrow S$  transition” in Fig. 4a). Indeed, in accordance with the definition [S], the downstream front of synchronized flow is fixed at the bottleneck  $B_2$ . In contrast, the moving jam propagates through this bottleneck with the mean velocity of the downstream jam front remaining unchanged. Thus in accordance with the definition [J], this moving jam is associated with the wide moving jam phase. Synchronized flow is self-sustaining for a very long time (more than an hour) upstream of bottleneck  $B_2$ . The possibility of the induced  $F \rightarrow S$  transition means that an  $F \rightarrow S$  transition exhibits nucleation character that is a characteristic feature of first-order phase transitions.

Another empirical example is shown in Fig. 5. It can be seen that a sequence of two moving jams propagates through different states of traffic flow and through a bottleneck while maintaining the downstream jam front velocity. Therefore, these moving jams belong to the wide moving jam phase. In contrast, there is a congested traffic in which speed is much lower than in free flow (compare vehicle speeds in Fig. 5c, d). The downstream front of this congested traffic flow, where vehicles accelerate to free flow, is fixed at the bottleneck (dashed line in Fig. 5a). Therefore, this congested traffic belongs to the synchronized flow phase.

We can see that whereas in wide moving jams both the speed and flow rate are very low (sometimes as low as zero), in synchronized flow the flow rate is high (compare the flow rates within the wide moving jams and within synchronized flow in Fig. 5d). The vehicle speed in syn-



**Traffic Congestion, Modeling Approaches to, Figure 4**

Measured three traffic phases. **a** Vehicle speed in space and time. **b** A graph of **a** with the free flow phase (green), the synchronized flow phase (yellow), and the wide moving jam phase (red).  $B_1$ ,  $B_2$ , and  $B_3$  are locations of bottlenecks. Taken from [52]

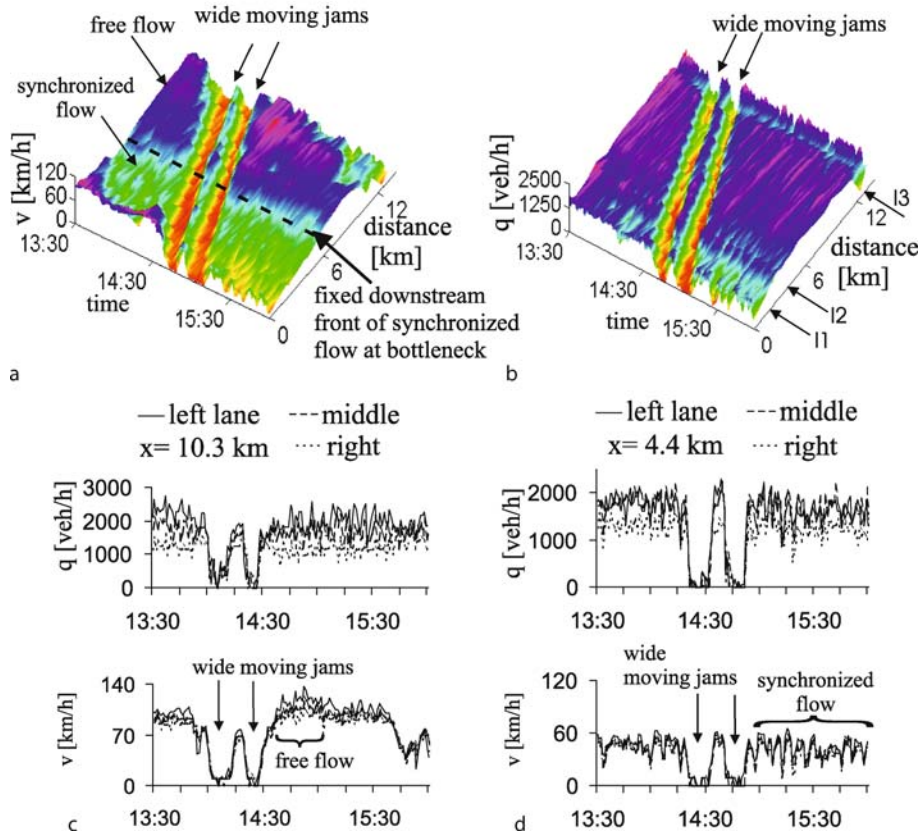
chronized flow is considerably lower than in free flow. However, the flow rates in the free flow and synchronized flow phases can be close to one another (Fig. 5c, d).

### Microscopic Criterion for Traffic Phases in Congested Traffic

The spatiotemporal criteria for a wide moving jam [J], which distinguish the wide moving jam and synchronized flow phases, can be explained by a traffic flow discontinuity within a wide moving jam. Traffic flow is interrupted by the wide moving jam: there is no influence of the inflow into the jam on the jam outflow [52]. A difference between the jam inflow and the jam outflow changes the jam width only. This *traffic flow interruption effect* is a general effect for each wide moving jam.

The jam outflow becomes independent of the jam inflow, when the traffic flow interruption effect within

## A5-North, October 9, 1992



**Traffic Congestion, Modeling Approaches to, Figure 5**

Measured three traffic phases. **a,b** Vehicle speed (**a**) and flow rate averaged across all freeway lanes (**b**) as functions of time and space. **c,d** Flow rate and average vehicle speed at two different freeway locations **c** and **d** in different freeway lanes. In **b**, three intersections with other freeways are labeled by I1–I3 that show approximate locations of possible freeway bottlenecks. Taken from [52]

a moving jam occurs. In a hypothetical case, when all vehicles within a moving jam do not move, the criterion for the traffic flow interruption effect within the jam is

$$I = \frac{\tau_j}{\tau_{\text{del}}^{(a)}} \gg 1, \quad (1)$$

where  $\tau_j$  is the jam duration, i. e., the time interval between the upstream and downstream jam fronts passing a road location and  $\tau_{\text{del}}^{(a)}$  is the mean time delay in vehicle acceleration at the downstream jam front from a vehicle standstill;  $\tau_{\text{del}}^{(a)}$  determines the jam outflow;  $I$  is approximately equal to the vehicle number stopped within the jam. Note that corresponding to empirical results  $\tau_{\text{del}}^{(a)} \approx 1.5 - 2$  sec [52].

In real traffic, vehicles come to a stop at very different blank spaces to each other when they meet a wide moving jam. We call these blank spaces as *blanks* within the jam. Later the vehicles begin to cover these blanks within the

jam. This vehicle motion within the jam occurs at low vehicle speeds and it creates new blanks upstream. Thus low speed states that cover blanks within the jam are associated with upstream moving of the blanks within the jam (moving blanks for short; for more details see [67]). A sufficient criterion for flow interruption within the jam is

$$I_s = \frac{\tau_{\text{max}}}{\tau_{\text{del}}^{(a)}} \gg 1, \quad (2)$$

where  $\tau_{\text{max}}$  is the maximum time headway between two vehicles within the jam ( $\tau_{\text{max}} \leq \tau_j$ ).

For this and the following consideration, we should define the term *time headway* between vehicles. At a given time instant  $t = t_1$ , the time headway (time gap)  $\tau(t_1)$  is defined as a time it takes for a vehicle to reach a freeway location at which the bumper of the preceding vehicle is at the time instant  $t_1$ . In single vehicle measurements

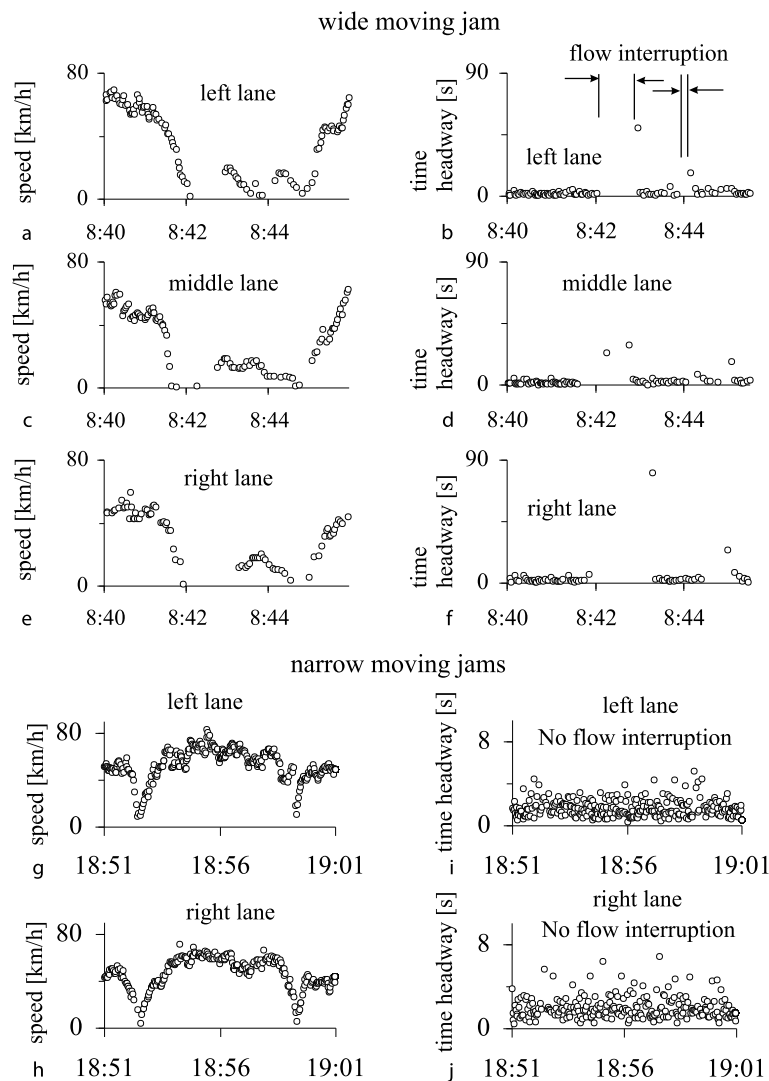


shown in Fig. 6,  $t_1$  is the time at which the preceding vehicle leaves the detector whose location is therefore related to the location of the bumper of the preceding vehicle in the time headway definition; the time headway is equal to  $\tau(t_1) = t_2 - t_1$ , where  $t_2$  is the time at which the vehicle front has been recorded at the detector.

Under condition (2), there are at least several vehicles within the jam that are in a standstill or if they are still moving, it is only with a negligible low speed in comparison with the speed in the jam inflow and outflow. These vehicles separate vehicles accelerating at the downstream jam front from vehicles decelerating at the upstream jam

front. For this reason, the jam outflow does not depend on the jam inflow. In other words, this jam propagates through a bottleneck while maintaining the mean downstream jam front velocity. In accordance with the macroscopic objective spatiotemporal criteria [J] and [S], this jam is a wide moving jam. Thus, we can assume that the traffic flow interruption effect can be used as a criterion to distinguish the synchronized flow and wide moving jam phases in single vehicle data. This is possible even if data is measured at a single freeway location.

The interruption of traffic flow within a moving jam can also be found in *empirical* single vehicle data. In an



**Traffic Congestion, Modeling Approaches to, Figure 6**

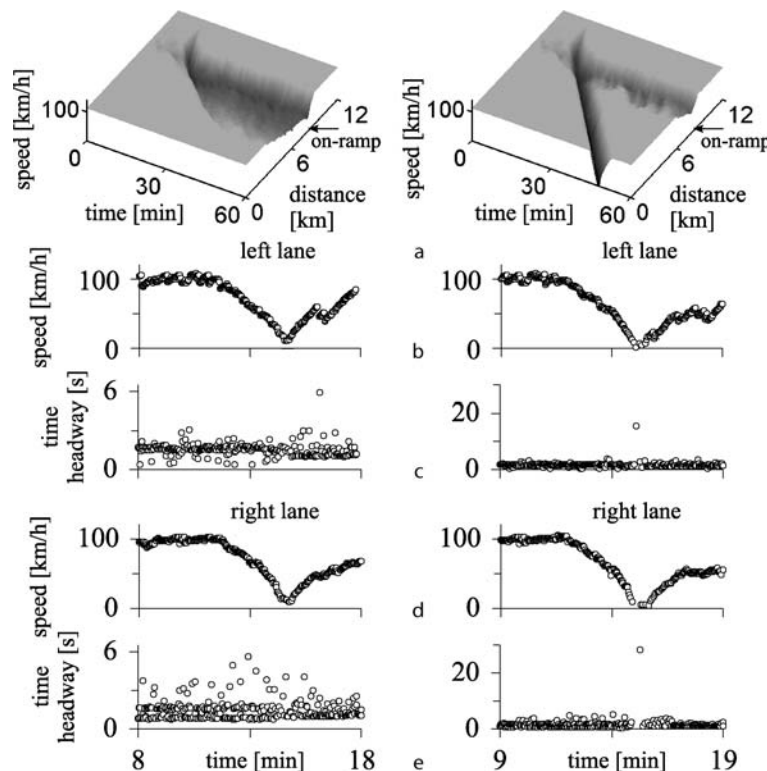
Measured single vehicle data analysis: Single vehicle data for speed in different freeway lanes for moving jams (left figures) and the related time headways (right figures). Taken from [67]

example shown in Fig. 6a–f, flow interruption effect occurs two times during upstream jam propagation through a road detector (these time intervals are labeled “flow interruption” in Fig. 6b). The values  $\tau_{\max}$  for the first flow interruption intervals within the wide moving jam are equal to approximately 50 s in the left lane, 24 s in the middle line, and 80 s in the right line. These values  $\tau_{\max}$  satisfy the criterion (2). This means that traffic flow is discontinuous within the moving jam, i. e., this moving jam can be associated with the wide moving jam phase. Later, some vehicles within the jam have time headways of about 4 s or longer associated with moving blanks [67,68].

In another example shown in Fig. 6g–j, there are two moving jams. However, rather than wide moving jams these moving jams can be classified as narrow moving jams. A narrow moving jam is a moving jam, which consists of the jam fronts only. Narrow moving jams are as-

sociated with the synchronized flow phase. This is because there are no traffic flow interruption intervals within a narrow moving jam. Indeed, in empirical observations upstream and downstream of the jam, as well as within the jam there are many vehicles that traverse the induction loop detector. There is no qualitative difference in time headways for different time intervals associated with these narrow moving jams and in traffic flow upstream or downstream of the jams (Fig. 6i, j).

This can be explained if it is assumed that each vehicle, which meets a narrow moving jam, which consists of the jam fronts only, can nevertheless accelerate later almost without any time delay within the jam. Within the upstream front vehicles must decelerate to a very low speed. However, the vehicles can accelerate almost immediately at the downstream jam front. These assumptions are confirmed by single vehicle data shown in Fig. 6i, j,



**Traffic Congestion, Modeling Approaches to, Figure 7**

Comparison of microscopic criterion with macroscopic spatiotemporal objective criteria [J] and [S] for the phases in congested traffic of identical vehicles (results of model simulations). *Left figures* are related to a narrow moving jam. *Right figures* are related to a wide moving jam. *a* The narrow moving jam is caught at an on-ramp bottleneck that is associated with the synchronized flow definition [S] (left); in contrast, the wide moving jam propagates through the bottleneck that is associated with the wide moving jam definition [J] (right). Vehicle speed on the main road in space and time that is averaged across two lanes. *b, d* Single vehicle data for speed in the left (*b*) and in right lanes (*d*) at location 50 m downstream of the end of the on-ramp merging region. *c, e* Time headways associated with *b* and *d*, respectively. Taken from [68]

in which time intervals between different measurements of time headways for different vehicles exhibit the same behavior away and within the jams. Even if within a narrow moving jam there are vehicles that are stopped, the condition

$$I_s \sim 1 \quad (3)$$

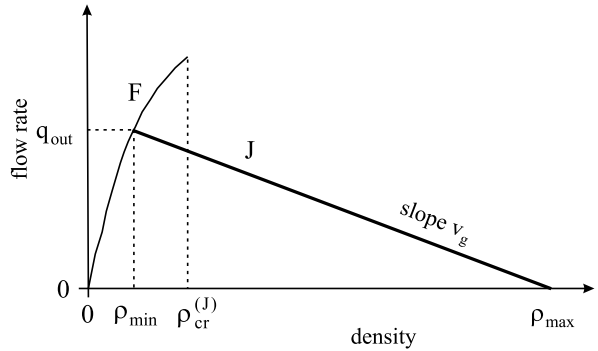
can be satisfied. Under this condition, there is no flow interruption within this jam. Thus regardless of these narrow moving jams, traffic flow is not discontinuous, i. e., the narrow moving jams are associated with the synchronized flow phase. This single vehicle analysis enables us to assume that congested traffic in Fig. 6a–f is associated with a wide moving jam. In contrast, congested traffic in Fig. 6g–j is associated with the synchronized flow phase.

It must be stressed that for a proof of the microscopic criterion for the traffic phases (2), we should compare results of phase identification based on this criterion in measured single vehicle data with the traffic phase definitions, i. e., with the macroscopic criteria for the traffic phases [J] and [S] (Sect. “Empirical Macroscopic Criteria for Defining Phases in Congested Traffic”). However, such an empirical proof can be possible, if the single vehicle data associated with a congested pattern in the neighborhood of a bottleneck is measured at many freeway locations including locations upstream and downstream of this bottleneck; this is because only in this case can we prove the behavior of the downstream front of the congested pattern at the bottleneck. Unfortunately, we do not have such measured single vehicle data. For this reason, we can make a proof of the microscopic criterion for the traffic phases based on numerical simulations in the context of three-phase traffic flow theory only. This proof, which is based on numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model (Sect. “Kerner–Klenov Stochastic Three-Phase Traffic Flow Model”), has been performed in [68] (Fig. 7).

The results of numerical simulations (Fig. 7) allow us to conclude that the microscopic criterion (2) enables us indeed to distinguish the wide moving jam and synchronized flow phases in single vehicle data measured at one road location.

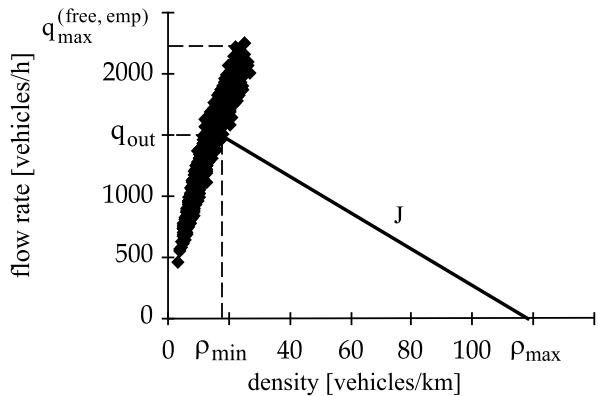
### Characteristic Parameters of Wide Moving Jam Propagation and Line J

The critical conclusion about earlier traffic flow models made in Sect. “Introduction” does not concern characteristic parameters of wide moving jam propagation, which was firstly discovered by Kerner and Konhäuser in 1994 [63] and later incorporated in a number of different



**Traffic Congestion, Modeling Approaches to, Figure 8**

Qualitative representation of the characteristic line for the downstream front of a wide moving jam (line J) whose slope is equal to the downstream jam front velocity  $v_g$ , and explanation of the metastability of states of free flow (curve F) with respect to moving jam formation [63]

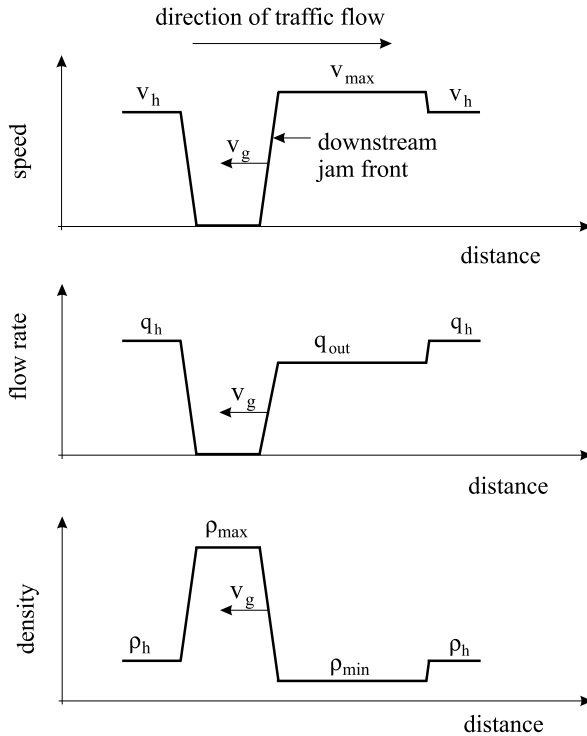


**Traffic Congestion, Modeling Approaches to, Figure 9**

Representation of the steady propagation of the downstream front of the downstream jam in the sequence of two wide moving jams in Fig. 5 by the line J in the flow-density plane and data for free flow (black diamonds). Taken from [52]

traffic flow models (see references in [40]). The Kerner–Konhäuser theory of wide moving jam propagation has been confirmed by empirical data in which wide moving jam propagation has been found (see references in [52]).

The characteristic feature of wide moving jam propagation [J], which defines the wide moving jam phase in congested traffic, i. e., a constancy of the mean velocity  $v_g$  of the downstream front of the jam while the jam propagates through bottlenecks can be presented by a line in the flow-density plane. This line is called the line J (Figs. 8 and 9). The slope of the line J is equal to the characteristic velocity  $v_g$ . If in the wide moving jam outflow free flow is formed, then the flow rate  $q_{out}$  in this jam outflow and the related density  $\rho_{min}$  give the left coordinates of



**Traffic Congestion, Modeling Approaches to, Figure 10**  
Characteristic parameters of wide moving jams. Schematic representation of a wide moving jam at a fixed time [63]. Spatial distribution of vehicle speed  $v$ , flow rate  $q$ , and vehicle density  $\rho$  in the wide moving jam, which propagates through a homogeneous state of free flow with speed  $v_h$ , flow rate  $q_h$ , and density  $\rho_h$

the line  $J$ ;  $q_{out}$ ,  $\rho_{min}$ , and the related mean vehicle speed  $v_{max}$  are characteristic parameters (Fig. 10). When control parameter of traffic (weather, percentage of long vehicles, etc.) do not change, the jam characteristic parameters do not depend on initial conditions, and they are the same for different wide moving jams. The right coordinates of the line  $J$  are related to the traffic variables within the jam, the density  $\rho_{max}$  and the average vehicle speed  $v_{min}$ . For simplicity we assume here that  $v_{min} = 0$  (low speed states within wide moving jams associated with so-called moving blanks have been studied in ► [Traffic Congestion, Spatiotemporal Features of](#) and [67,68]).

Concerning the characteristic jam velocity  $v_g$ , some general assumptions can be made [52]. Each driver standing within a wide moving jam can start to accelerate from the standstill inside the jam to free flow downstream after two conditions have been satisfied:

- (i) The preceding vehicle has already begun to move away from the jam.

- (ii) Due to the preceding vehicle motion, after some time the net distance (space gap) between the two vehicles has exceeded a space gap  $g_{del}$ .

The mean time delay  $\tau_{del}^{(a)}$  in vehicle acceleration (Sect. “[Microscopic Criterion for Traffic Phases in Congested Traffic](#)”) corresponds to the space gap  $g_{del}$  between two successive drivers that accelerate from the standstill within the wide moving jam to free flow downstream.

The motion of the downstream front of a wide moving jam results from acceleration of drivers from the standstill inside the jam to flow downstream of the jam. Because the average distance between vehicles inside the jam, including average length of each vehicle, equals  $1/\rho_{max}$ , the velocity of the downstream front of the wide moving jam is

$$v_g = -\frac{1}{\rho_{max}\tau_{del}^{(a)}}. \quad (4)$$

In empirical observations,  $v_g \sim -15$  km/h.

The line  $J$  and the characteristic parameters of wide moving jam propagation in free flow [63] are apparently the only results of earlier traffic flow theories to traffic congestion, which are confirmed in empirical observations of spatiotemporal features of congested patterns (see also Sect. “[Classic General Motors \(GM\) Model Approach: Free Flow Instability due to Driver Reaction Time](#)”).

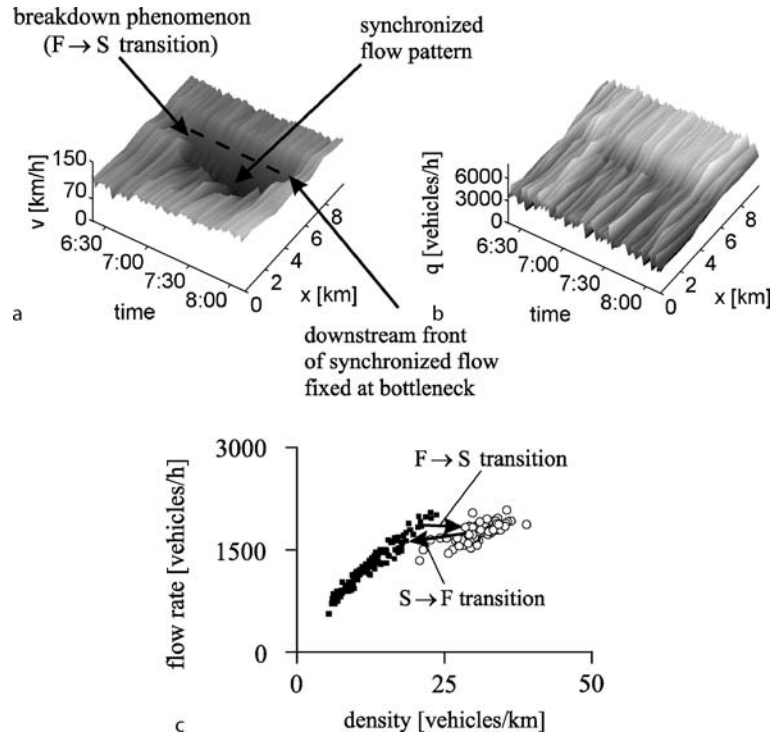
## Traffic Breakdown and Highway Capacity

### Spontaneous Traffic Breakdown (F $\rightarrow$ S Transition)

Traffic breakdown is accompanied by a sharp decrease in average vehicle speed in the free flow to a considerably lower speed in congested traffic (Figs. 11–13) (see, e.g., [30,38,39,93,105]).

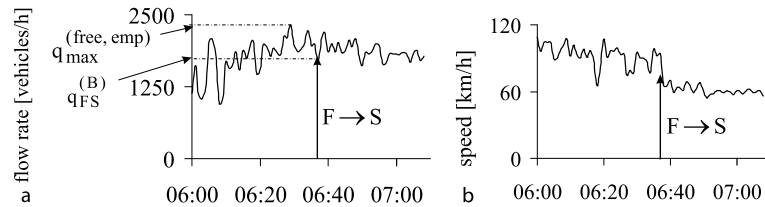
Before traffic breakdown occurs at the bottleneck shown in Fig. 11a, there is free flow at the bottleneck as well as upstream and downstream in a neighborhood of the bottleneck. In this case, the breakdown is caused by occurrence and subsequent growth of speed disturbances (fluctuations) within the free flow at the bottleneck. In accordance with definitions of phase transitions in complex systems, such traffic breakdown is called *spontaneous* traffic breakdown. There can be various sources of speed disturbances whose growth lead to traffic breakdown, e.g., unexpected braking of a vehicle, lane changing, fluctuations in flow rates upstream of the bottleneck, vehicle merging onto the main road from other roads (e.g., at on-ramp bottlenecks), etc.

Traffic breakdown occurs usually at the same freeway bottlenecks of a freeway section. These bottlenecks are



**Traffic Congestion, Modeling Approaches to, Figure 11**

Empirical example of traffic breakdown and hysteresis effect at on-ramp bottleneck: **a,b** Average speed (**a**) and flow rate (**b**) on the main road in space and time. **c** Hysteresis effect in the flow-density plane. 1-min average data. Taken from [52]. This example of traffic breakdown is qualitatively the same as many other examples observed in various countries (e. g., [30,38,39,105])



**Traffic Congestion, Modeling Approaches to, Figure 12**

Empirical example in which flow rate downstream of the bottleneck at which traffic breakdown occurs,  $q_{FS}^{(B)}$ , is smaller than  $q_{\max}^{(free, emp)}$ . Taken from [52]

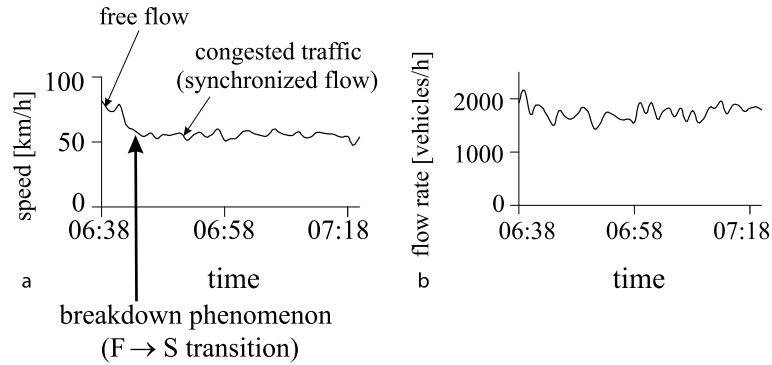
called *effectual bottlenecks*. An effectual bottleneck is a bottleneck where traffic breakdown most frequently occurs on many different days. Examples of effectual bottlenecks are adjacent bottlenecks  $B_1$  and  $B_2$  in Fig. 4b.

Elefteriadou et al. found that traffic breakdown has a probabilistic nature [30] that means the following: at a given flow rate traffic breakdown can occur but it should not necessarily occur. Thus on one day traffic breakdown occurs, however, on another day at the same flow rates and at the same traffic conditions traffic breakdown is not observed. In 1998, Persaud et al. found [105] that em-

pirical probability of traffic breakdown at a bottleneck is an increasing flow rate function (Fig. 14). Later such an empirical probability of traffic breakdown was also found on different freeways in various countries [10,11,12,87,94,131].

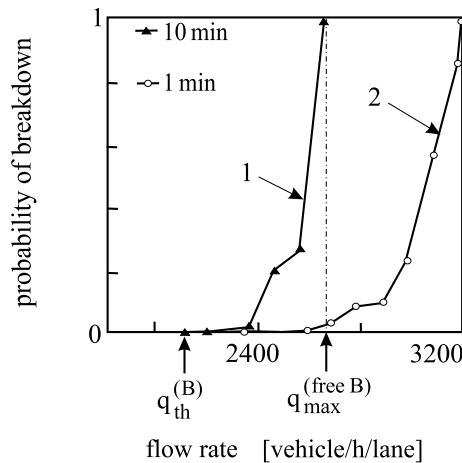
Another empirical probabilistic characteristic of traffic breakdown is as follows. At given traffic parameters (weather, etc.), the flow rate downstream of an on-ramp bottleneck associated with the empirical maximum flow rate in free flow  $q_{\max}^{(free, emp)}$ , which was measured on a specific day before congestion occurred, can be greater





**Traffic Congestion, Modeling Approaches to, Figure 13**

Traffic breakdown at an on-ramp bottleneck. Vehicle speed (a) and flow rate downstream of the bottleneck (b) as functions of time related to Fig. 11 (1-min average data). This example of traffic breakdown is qualitatively the same as many other examples observed in various countries (e. g., [30,38,39,105])



**Traffic Congestion, Modeling Approaches to, Figure 14**

Probability for traffic breakdown at an on-ramp bottleneck for two flow rate averaging time constants  $T_{av}$  ( $T_{av} = 1$  and 10 min). Data is measured on a freeway in Toronto, Canada. Taken from Persaud et al. [105] Note that the flow rates  $q_{max}^{(free B)}$  and  $q_{th}^{(B)}$  made on this figure are respectively the maximum and minimum free-flow capacities of free flow at the bottleneck [52] shown for the averaging time interval for empirical data  $T_{av} = 10$  min (see explanation in Sect. "Highway Capacity and First-Order F → S Transition")

than the flow rate  $q_{FS}^{(B)}$  at which traffic breakdown occurs (Fig. 12).

Congested traffic in Fig. 13 that has occurred due to traffic breakdown shows a further development in space and time (Fig. 11a, b), with the following effects observed in different countries (e. g., [30,38,39,105]):

(i) The upstream front of congested traffic propagates upstream. The upstream front separates free flow up-

stream from congested traffic downstream. Thus the region of congestion broadens in the upstream direction.

(ii) The downstream front of congested traffic is fixed at the bottleneck (dashed line in Fig. 11).

Corresponding to the objective criteria for traffic phases in congested traffic [J] and [S] considered previously, the result of item (ii) means that the congested traffic resulting from traffic breakdown belongs to the synchronized flow phase. In all known observations, traffic breakdown is associated with an F → S transition. Thus the terms "F → S transition," "breakdown phenomenon," "traffic breakdown," and "speed breakdown" are synonyms related to the same effect: the onset of congestion in free flow.

The congested pattern (Fig. 11a, b) exists for about one hour at the bottleneck: at 7:40 free flow occurs at the bottleneck. This restoration of free flow is related to a reverse phase transition from synchronized flow to the free flow (S → F transition for short) at the bottleneck. The F → S and S → F transitions are accompanied by a well-known *hysteresis effect* and hysteresis loop in the flow-density plane: a congested pattern emerges usually at a greater flow rate downstream of the bottleneck than this flow rate is at which the congested pattern dissolves (see references in [38,39]) (Fig. 11c).

### Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory

In three-phase traffic theory, traffic breakdown (F → S transition) is explained by a competition between two opposing tendencies occurring within a random local speed

disturbance in which the speed is lower and vehicle density is greater than in an initial free flow [52]:

- (i) A tendency towards the initial free flow due to vehicle acceleration associated with an *over-acceleration effect*;
- (ii) A tendency towards synchronized flow due to vehicle deceleration associated with a *speed adaptation effect*.

To explain this mechanism of traffic breakdown, we use the fundamental hypothesis of the three-phase traffic theory formulated as follows [49,50,51]:

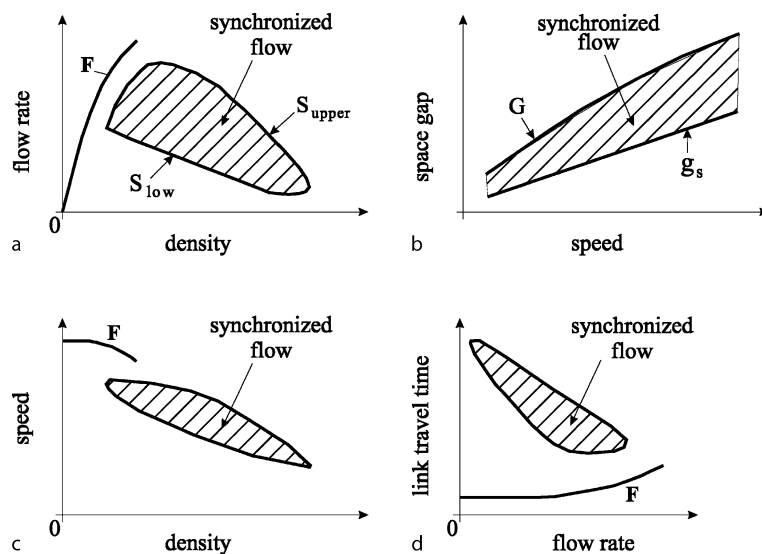
- Steady states of synchronized flow cover a two-dimensional (2D) region in the flow-density plane (Fig. 15a). The multitudes of free flow states overlap steady states of synchronized flow in the vehicle density. The free flow states on a multi-lane road and steady states of synchronized flow are separated by a gap in the flow rate and, therefore, by a gap in the speed at a given density: at each given density the synchronized flow speed is lower than the free flow speed.

This hypothesis and its consequences (Fig. 15b–d) contradict qualitatively with the fundamental diagram hypothesis and associated relationships of earlier traffic flow theories (Fig. 1). In particular, the fundamental hypothesis of the three-phase traffic theory states that for hypothetical steady states of congested traffic (synchronized flow) there

is *no* link travel time-flow relationship (Fig. 15d). This is in contrast with a link travel time-flow relationship (Fig. 1d) whose existence is the basic assumption of all known theories and models of congested traffic networks (see references in ► [Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment](#)).

It should be noted that a well-known wide scattering of empirical data for congested traffic in the flow-density plane *cannot* be considered a proof of the fundamental hypothesis of three-phase traffic theory: a 2D-region for steady states of synchronized flow (Fig. 15a) is a *theoretical* assumption only. Indeed, the wide scattering of measured data for congested traffic in the flow-density plane can be associated for example with a variety of driver characteristics and vehicle parameters as well as with speed fluctuations, which are present in real traffic flow. As a proof of the fundamental hypothesis, mathematical results of traffic flow models can be considered in which this and other hypotheses of three-phase traffic theory have been used. This is because these results explain and predict empirical features of traffic breakdown and resulting congested patterns [52].

The low boundary of the 2D region for steady states of synchronized flow in the flow-density plane (labeled  $S_{\text{low}}$  in Fig. 15a) is associated with a *synchronization space gap*  $G$  between vehicles; at  $g > G$  a vehicle accelerates. The upper boundary of the 2D region (labeled  $S_{\text{upper}}$ ) is associ-



**Traffic Congestion, Modeling Approaches to, Figure 15**

Fundamental hypothesis of three-phase traffic theory [49,50,51]: **a** Qualitative representation of free flow states (F) and 2D steady states of synchronized flow (*dashed region*) in the flow-density plane. **b** Qualitative representation of a part of the 2D steady states of synchronized flow shown in **a** in the space-gap-speed plane (*dashed region*). **c,d** Qualitative representation of free flow states (F) and 2D steady states of synchronized flow in the speed-density (**c**) and link-travel-time-flow (**d**) planes associated with a

ated with a *safe space gap*  $g_s$ , i. e., at  $g < g_s$  the vehicle decelerates. The synchronization gap is greater than the safe one at each given vehicle speed:  $G(v) > g_s(v)$  (Fig. 15b). Both the synchronization gap  $G$  and safe gap  $g_s$  are usually speed increasing functions.

In hypothetical steady states of synchronized flow, the synchronization gap is defined through the condition that at each given steady speed  $v$  in synchronized flow there are *infinite* space gaps  $g$  within the range

$$g_s \leq g \leq G \quad (5)$$

at which a driver can move with this steady speed  $v$ . In other words, through condition (5) the synchronization gap determines the existence of the 2D region for steady states of synchronized flow (Fig. 15). The condition (5) is consistent with a driver behavioral assumption that at relatively small space gaps in synchronized flow a driver can recognize whether the space gap to the preceding vehicle increases or decreases over time. This is true even if the speed difference between the vehicle speed and the speed of the preceding vehicle is negligible.

The condition (5) is also consistent with a driver behavioral assumption that for comfortable driving in car-following a driver tends to adapt the speed to the speed of the preceding vehicle within a greater space gap than the safe one. This driver behavior about speed adaptation with the synchronized gap explains the 2D region for steady states of synchronized flow (Fig. 15) introduced in three-phase traffic theory [49,50,51]. To understand the statement about a strong connection between the 2D region for steady states of synchronized flow and the dynamic driver behavior about speed adaptation, let us discuss a general case of car-following in which speeds of the vehicle and the preceding vehicle are not time-independent.

Then rather than the synchronization gap that determines the 2D region of steady states of synchronized flow via the condition (5), we should consider a *dynamic* synchronization space gap, which can depend on the vehicle speed *and* the speed of the preceding vehicle; if speeds of the vehicle and the preceding vehicle are equal to each other and time-independent, then the dynamic synchronization gap is equal to the synchronization gap in (5).

As above mentioned, at  $g > G$  a vehicle accelerates, whereas at  $g < g_s$  the vehicle decelerates. We define the dynamic synchronization gap as a space gap within which a vehicle tends to adapt the speed to the speed of the preceding vehicle without caring, what the precise space gap is, as long as this space gap is not smaller than the safe gap. The space gap at which this speed adaptation has finished

can be any space gap from the infinite number of space gaps within the space gap range

$$g_s \leq g \leq G(v, v_\ell) \quad (6)$$

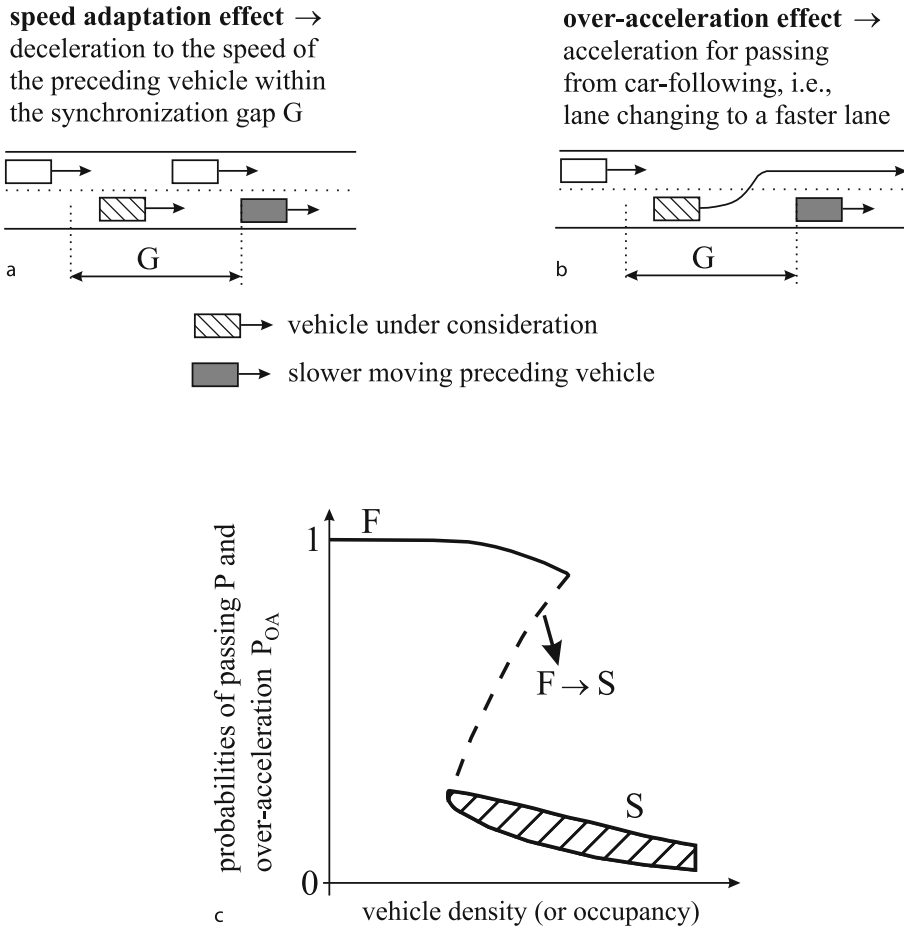
in which it is suggested that the dynamic synchronization gap  $G(v, v_\ell)$  can be a function of the vehicle speed  $v$  and the speed of the preceding vehicle  $v_\ell$ . This speed adaptation within the dynamic synchronization gap is called the *speed adaptation effect* (Fig. 16a) [52].

For hypothetical steady states of synchronized flow, the dynamic synchronization gap is equal to the synchronization gap in (5), i. e., for the steady states the synchronization gaps in the conditions (5) and (6) are identical ones. For this reason, we use the term “synchronization gap” for any case of car-following behavior. Thus the speed adaptation effect within the synchronization gap is indeed associated with the 2D-region of steady states of synchronized flow given by the condition (5) (Fig. 15).

The condition (5) is in contradiction with the fundamental diagram hypothesis of all the earlier traffic flow theories (see Sect. “Introduction”): for congested traffic there is no fundamental diagram in three-phase traffic theory. To explain this statement, let us consider a hypothetical case when a vehicle approaches the preceding vehicle that moves with a slower time-independent speed. Then the vehicle adapts the speed to the speed of the preceding vehicle at a space gap, which is only one possible space gap from the infinite number of space gaps between the synchronization gap  $G$  and the safe gap  $g_s$  within the above 2D-region at this speed, i. e., the gaps that satisfy (5). There is no desired (or optimal) space gap, when a vehicle follows the preceding vehicle that moves with at a time-independent speed. This feature is associated with a speed adaptation effect within the synchronization gap.

Under condition (6), a vehicle accelerates when it is slower than the preceding vehicle, and decelerates usually when it is faster than the preceding vehicle. However, under condition (6) the vehicle can also accelerate, even if it is not slower than the preceding vehicle. This vehicle acceleration is called *vehicle over-acceleration* (over-acceleration for short) [52]. In general, the over-acceleration is defined as vehicle acceleration from car-following, in particular, when the space gap  $g$  is satisfied to condition (6), i. e., the vehicle is within the synchronization gap.

To understand the sense of the term *over-acceleration*, we consider a scenario in which a vehicle that moves in free flow on a multi-lane road approaches a slower moving preceding vehicle. If firstly the vehicle cannot pass the preceding vehicle, then the vehicle decelerates within the synchronization gap to the speed of the preceding vehicle, i. e., the speed adaptation effect is realized leading to



**Traffic Congestion, Modeling Approaches to, Figure 16**

Qualitative explanation of competition of vehicle deceleration due to speed adaptation (a) and vehicle acceleration due to over-acceleration (b) [51,52]. c Qualitative Z-shaped density function for passing probability and probability of over-acceleration. In c,  $F$  – free flow,  $S$  – synchronized flow, dashed curves between states  $F$  and  $S$  are related to critical probabilities for passing and over-acceleration associated with a critical speed required for traffic breakdown: When the speed within a speed disturbance is lower than the critical speed, then probabilities for passing and over-acceleration are smaller than the critical probabilities; as a result, traffic breakdown occurs (arrows labeled by  $F \rightarrow S$ )

car-following of the slow preceding vehicle (Fig. 16a). We assume in this scenario that later the vehicle can pass this slow moving preceding vehicle. As well-known, to pass the preceding vehicle, the vehicle should change lane and accelerate. The vehicle acceleration takes place, even if the vehicle is *not* currently *slower* than the preceding vehicle; this explains why this vehicle acceleration is called *over-acceleration*. Thus in the case under consideration, over-acceleration is vehicle acceleration for passing from car-following, i. e., lane changing to a faster lane (Fig. 16b); as a result, the probability of the over-acceleration denoted by  $P_{OA}$  is equal to the passing probability denoted by  $P$ :

$$P_{OA} = P.$$

Passing in free flow is much more probable than in synchronized flow: passing probability  $P$  is greater in free flow than in synchronized flow. This obvious suggestion together with the fundamental hypothesis of three-phase traffic theory, specifically that at each given density the synchronized flow speed is lower than the free flow speed lead to the following hypothesis [51,52]:

- Passing probability  $P$  and, therefore, over-acceleration probability  $P_{OA}$  exhibit a *discontinuous* character; in particular, these probabilities are Z-shaped density functions (Fig. 16c): At a given density, there is a *finite drop* in passing and over-acceleration probabilities, when free flow transforms onto synchronized flow.

The Z-shape density function for probabilities of passing and over-acceleration (Fig. 16c) explain traffic breakdown as follows [52]. Free flow remains on a multi-lane road as long as the over-acceleration effect, which describes the tendency towards free flow, is stronger than the speed adaptation effect that describes the tendency towards synchronized flow. However, the greater the density in free flow, the smaller the probability of over-acceleration  $P_{OA}$ , i.e., the weaker the over-acceleration effect.

There is a critical probability of over-acceleration (dashed curve between states  $F$  and  $S$  in Fig. 16c) associated with a critical speed within a local speed disturbance required for traffic breakdown (explanation of the critical speed for traffic breakdown appears in Sects. “Highway Capacity and First-Order  $F \rightarrow S$  Transition” and “Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow”). This means that when probability of over-acceleration is equal to the critical probability of over-acceleration, then the tendency to free flow due to over-acceleration within the disturbance is on average exactly as strong as the tendency to synchronized flow due to speed adaptation.

When the speed within a local speed disturbance is lower than the critical speed, then probability of over-acceleration is smaller than the critical probability of over-acceleration. This means that the over-acceleration effect is weaker than the speed adaptation effect. In this case, the tendency to synchronized flow due to speed adaptation overcomes the tendency to free flow due to over-acceleration that results in traffic breakdown (arrow labeled by  $F \rightarrow S$  in Fig. 16c).

In free flow at a bottleneck, there is a permanent local speed decrease called a permanent speed disturbance. This speed disturbance is localized in a neighborhood of the bottleneck. For example, for an on-ramp bottleneck the permanent speed disturbance is due to vehicles that merge from the on-ramp onto the main road. This speed disturbance exists even if there were no random speed disturbances in free flow. The permanent speed decrease within the disturbance increases the probability of traffic breakdown at the bottleneck considerably in comparison to the traffic breakdown probability at other freeway locations away from bottlenecks at which a speed disturbance should initially occur in free flow for traffic breakdown. This explains why in empirical observations traffic breakdown occurs mostly at bottlenecks.

Due to traffic breakdown, a variety of congested traffic patterns can occur at a bottleneck. There are congested patterns that consist of the synchronized flow traffic phase only. These patterns are called *synchronized flow patterns*

(SP for short). The congested pattern, which emerges as a result of the  $F \rightarrow S$  transition discussed above (Fig. 11) consists of synchronized flow only, i.e., this is an example of an SP. In more detail, various SPs and other congested patterns are discussed in ► [Traffic Congestion, Spatiotemporal Features of](#).

### Induced Traffic Breakdown

In empirical observation, besides spontaneous traffic breakdown at a bottleneck (Figs. 11 and 13) (spontaneous  $F \rightarrow S$  transition) there can also be an induced traffic breakdown (induced  $F \rightarrow S$  transition) [52].

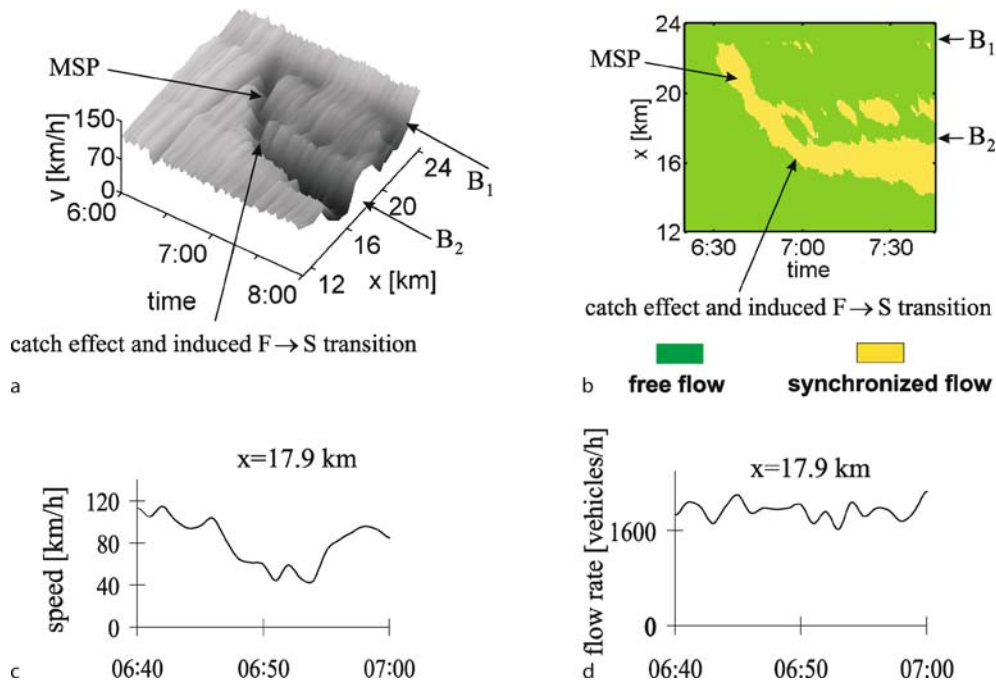
In contrast with spontaneous traffic breakdown, which occurs when before the breakdown free flow is in a neighborhood of the bottleneck (Fig. 11a), the induced  $F \rightarrow S$  transition at the bottleneck is caused by the propagation of a moving spatiotemporal *congested* traffic pattern, which has initially occurred at a *different* road location (e.g., at another bottleneck) than that of the bottleneck. When this congested pattern reaches the bottleneck, the pattern induces traffic breakdown at the bottleneck. The congested pattern whose propagation causes the breakdown at the bottleneck can be considered *external* disturbance at the bottleneck.

In particular, an  $F \rightarrow S$  transition can be induced by a wide moving jam propagating upstream through a bottleneck. This case is shown in Fig. 4a. Propagating through bottleneck  $B_2$ , the wide moving jam causes synchronized flow emergence at the bottleneck. Synchronized flow remains at bottleneck  $B_2$  even after the wide moving jam is far upstream of the bottleneck. The induced phase transition is labeled “induced  $F \rightarrow S$  transition” in this figure. In this case, after the wide moving jam has passed the bottleneck a *synchronized flow pattern* (SP for short) is formed. The SP remains at the bottleneck for a long time (this SP is labeled “synchronized flow” in Fig. 4a).

Induced  $F \rightarrow S$  transition at a bottleneck can also occur when a region of synchronized flow first occurs downstream of this bottleneck, and the region later reaches the bottleneck due to the upstream propagation of synchronized flow. However, in contrast to the above case of induced traffic breakdown caused by wide moving jam propagation, the initial synchronized flow is caught at the bottleneck [52].

To show this effect, let us consider propagation of a moving synchronized flow pattern (MSP) upstream of a bottleneck  $B_1$  (labeled MSP in Fig. 17). The MSP propagates upstream. After the MSP reaches the upstream bottleneck  $B_2$ , the MSP is caught at the bottleneck. This catch effect is inconsistent with the traffic phase definition [J]. In





Traffic Congestion, Modeling Approaches to, Figure 17

Empirical example of induced traffic breakdown at bottleneck. **a** Vehicle speed in space and time. **b** A graph of **a** with the free flow phase (green) and the synchronized flow phase (yellow). **c,d** Average (across the road) speed (**b**) and flow rate (**c**) within the MSP at the location  $x = 17.9$  km (about 0.8 km downstream of the bottleneck location  $B_2$ ). Bottlenecks  $B_2$  and  $B_1$  are the same as those in Fig. 4. Taken from [52]

other words, the MSP satisfies the traffic phase definition [S], i. e., this pattern is indeed an SP.

### Highway Capacity and First-Order F $\rightarrow$ S Transition

The possibility of both induced and spontaneous traffic breakdowns at the same bottleneck can be explained by the *nucleation* nature of traffic breakdown. This means that there should be some critical speed required for traffic breakdown. The breakdown occurs only, if due to a speed disturbance in free flow in the neighborhood of a bottleneck the speed decreases below the critical one. Otherwise, if free flow speed within the disturbance is greater than the critical one, breakdown does not occur.

The critical speed should depend on the flow rate downstream of the bottleneck; the smaller the flow rate, the lower the critical speed should be required for the breakdown in free flow. In contrast, when the flow rate increases, the critical speed at the bottleneck required for the breakdown increases. This means that the greater the flow rate downstream of the bottleneck, the smaller should be a local speed decrease in free flow at the bottleneck to cause the breakdown. Obviously, the lower this speed decrease

should be, the smaller the probability for the speed decrease. This explains the increasing character of the probability of traffic breakdown (Fig. 14).

Thus traffic breakdown at a bottleneck exhibits the following **fundamental empirical features**:

- Traffic breakdown is an F  $\rightarrow$  S transition.
- At the same bottleneck, traffic breakdown can be either *spontaneous* or *induced*.
- Onset and dissolution of congestion are accompanied by a *hysteresis* effect.
- Traffic breakdown exhibits the *probabilistic* nature.

In accordance with other systems of natural science, when the speed within a speed disturbance in free flow at the bottleneck is equal to or lower than the critical speed, this speed disturbance can be called critical speed disturbance or a *nucleus* for traffic breakdown. Thus the probabilistic feature of traffic breakdown is associated with the nucleation character of the breakdown. Only if a critical speed disturbance appears at the bottleneck, traffic breakdown occurs. In other words, the probability for traffic breakdown is the probability of critical speed disturbance appearance at the bottleneck.

The empirical features of traffic breakdown **A-D** mean that the breakdown is a first-order  $F \rightarrow S$  transition. In accordance with first-order  $F \rightarrow S$  transition features, there is the maximum flow rate (denoted  $q_{\max}^{(\text{free B})}$  and that is shown in Fig. 14 for the averaging time interval for empirical data  $T_{\text{av}} = 10$  min only) downstream of the bottleneck at which breakdown probability reaches one, i. e.,

$$P_{\text{FS}}^{(\text{B})} = 1. \quad (7)$$

There is also the threshold flow rate (denoted  $q_{\text{th}}^{(\text{B})}$ ). At the flow rate in free flow downstream of the bottleneck that is smaller than  $q_{\text{th}}^{(\text{B})}$ , breakdown probability

$$P_{\text{FS}}^{(\text{B})} = 0. \quad (8)$$

Highway (freeway) capacity of free flow at a bottleneck is limited by traffic breakdown at the bottleneck. Thus the fundamental empirical features of traffic breakdown **A-D** are also the fundamental empirical features for highway capacity.

For a given averaging time interval for traffic variables  $T_{\text{av}}$ , traffic breakdown can occur with probability  $P_{\text{FS}}^{(\text{B})}$ . In other words, an attribute of this *probabilistic highway capacity at the bottleneck* (bottleneck capacity) is the probability

$$P_{\text{C}}^{(\text{B})} = 1 - P_{\text{FS}}^{(\text{B})} \quad (9)$$

that free flow remains at the bottleneck during the time interval  $T_{\text{av}}$  [52]. Highway capacity is reached when

$$P_{\text{C}}^{(\text{B})} < 1. \quad (10)$$

Thus the definition of highway capacity made in three-phase traffic theory, which satisfies the fundamental empirical features of traffic breakdown **A-D**, reads as follows [52]:

- Highway capacity of free flow is equal to the flow rate downstream of the bottleneck at which free flow remains at the bottleneck with the probability  $P_{\text{C}}^{(\text{B})} < 1$  (10) during a given averaging time interval  $T_{\text{av}}$ .

Because there can be an infinite number of flow rates downstream of the bottleneck for which the condition (10) is satisfied, there can also be an infinite number of capacities of free flow at the bottleneck. Each of the freeway capacities has two attributes:

- (1) The probability  $P_{\text{C}}^{(\text{B})}$  (10) that free flow remains at the bottleneck during a given averaging time interval for traffic variables  $T_{\text{av}}$ .

- (2) The time interval  $T_{\text{av}}$ .

When for free flow at the bottleneck rather than the condition (10) the condition

$$P_{\text{C}}^{(\text{B})} = 1 \quad (11)$$

is satisfied, then the flow rate downstream of the bottleneck in this free flow is smaller than any of the freeway capacities. In accordance with (9), condition (11) is equivalent to (8). Thus the flow rate in free flow downstream of the bottleneck denoted by  $q_{\text{sum}}$  that is smaller than any of the capacities satisfy

$$q_{\text{sum}} < q_{\text{th}}^{(\text{B})}. \quad (12)$$

For this reason, the threshold flow rate  $q_{\text{th}}^{(\text{B})}$  is the minimum capacity.

In accordance with (9), condition (7) is equivalent to  $P_{\text{C}}^{(\text{B})} = 0$ . Thus the maximum flow rate  $q_{\max}^{(\text{free B})}$  determines the maximum capacity. Because  $q_{\max}^{(\text{free B})}$  is considerably greater than  $q_{\text{th}}^{(\text{B})}$ , there are infinite freeway capacities denoted by  $q_{\text{C}}^{(\text{B})}$ , which satisfy

$$q_{\text{th}}^{(\text{B})} \leq q_{\text{C}}^{(\text{B})} \leq q_{\max}^{(\text{free B})}. \quad (13)$$

Note that for an on-ramp bottleneck,  $q_{\max}^{(\text{free B})}$  and  $q_{\text{th}}^{(\text{B})}$  depend on  $q_{\text{on}}$  and  $q_{\text{in}}$ ; for this reason, there are also infinite maximum and infinite minimum capacities associated with different values  $q_{\max}^{(\text{free B})}$  and  $q_{\text{th}}^{(\text{B})}$ , respectively.

The maximum and minimum capacities can depend considerably on “control” parameters of traffic (weather, percentage of long vehicles, etc.). In empirical observations, they are found from a study of a flow rate dependence of breakdown probability  $P_{\text{FS}}^{(\text{B})}(q_{\text{sum}})$  (Fig. 14). For this reason,  $q_{\max}^{(\text{free B})}$  and  $q_{\text{th}}^{(\text{B})}$  are mean values found in many different realizations (days) at which traffic breakdowns have occurred.

To explain the meaning of the averaging time interval  $T_{\text{av}}$  in the capacity definition, we should note that flow rates are usually complicated time-functions in real traffic. For this reason, in empirical observations measured data is usually averaged for a time interval  $T_{\text{av}}$ . The longer  $T_{\text{av}}$ , the greater the probability of traffic breakdown at the same average flow rate. For this reason,  $T_{\text{av}}$  is an important attribute of highway capacity.

The sense of the infinite highway capacities (13) is as follows. At each flow rate  $q_{\text{sum}}$  downstream of a bottleneck, which satisfies the condition

$$q_{\text{th}}^{(\text{B})} \leq q_{\text{sum}} < q_{\max}^{(\text{free B})}, \quad (14)$$

traffic breakdown at the bottleneck is possible: traffic breakdown occurs spontaneously with a probability  $P_{FS}^{(B)}$ , which is within the range  $0 < P_{FS}^{(B)} < 1$ , or traffic breakdown can be induced. Thus each of these flow rates is highway capacity. Breakdown probability  $P_{FS}^{(B)}$  depends on  $q_{sum}$  strongly. Thus probability  $P_C^{(B)}$  (9) that traffic breakdown does not occur spontaneously, i. e., that free flow remains at the bottleneck is an attribute of highway capacity. This capacity attribute distinguishes different highway capacities. The probability function  $P_C^{(B)}(q_{sum})$  depends on  $T_{av}$ . Thus  $T_{av}$  is another attribute of highway capacity. Traffic breakdown cannot be possible at the bottleneck only then, when the flow rate  $q_{sum}$  satisfies the condition (12). Thus only the flow rates that satisfy the condition (12) are smaller than highway capacity.

In contrast with real traffic, in simulations of traffic with traffic flow models flow rates upstream of a bottleneck can be chosen as time-independent values. In this case,  $T_{av}$  in the above capacity definition should be replaced by a time interval for observing traffic flow  $T_{ob}$ . The related capacity definition and its relation to diagram of congested patterns at bottlenecks as well as other results of a probabilistic theory of traffic breakdown and highway capacity appears in ► [Traffic Breakdown, Probabilistic Theory of](#).

### Relation Between Speed and Flow Rate in Synchronized Flow Resulting from Traffic Breakdown

It can be seen in Fig. 13 that whereas there is a sharp decrease in average vehicle speed, the flow rate does not necessarily abruptly decrease during an  $F \rightarrow S$  transition and within an emergent SP. In most observations, during traffic breakdown the flow rate in the synchronized flow at the location of traffic breakdown is almost as great as in free flow. In some cases, the flow rate in synchronized flow can be even greater than the initial flow rate in free flow [52].

Another example of this feature is shown in Fig. 17c, d in which average speed within an MSP is considerably lower than the one in free flow, whereas the flow rate within the MSP is on average the same as that in free flow. In contrast with this very important feature of many other synchronized flow patterns observed in real traffic, the flow rate within a wide moving jam is considerably smaller than in free and synchronized flows (see Fig. 5c, d).

The empirical feature of synchronized flow, resulting from traffic breakdown that the flow rate within the SP can be as great as in free flow, is the important one for feedback on-ramp metering control applications (see Sect. 23.3 in [52] and a recent review [56]).

### Moving Jam Emergence in Synchronized Flow ( $S \rightarrow J$ Transition)

We have already mentioned that wide moving jams do not emerge spontaneously in free flow: no spontaneous phase transition from the free flow phase to the wide moving jam phase ( $F \rightarrow J$  transition for short) has been observed [52]. Wide moving jams can emerge *spontaneously only* in the synchronized flow phase ( $S \rightarrow J$  transition).

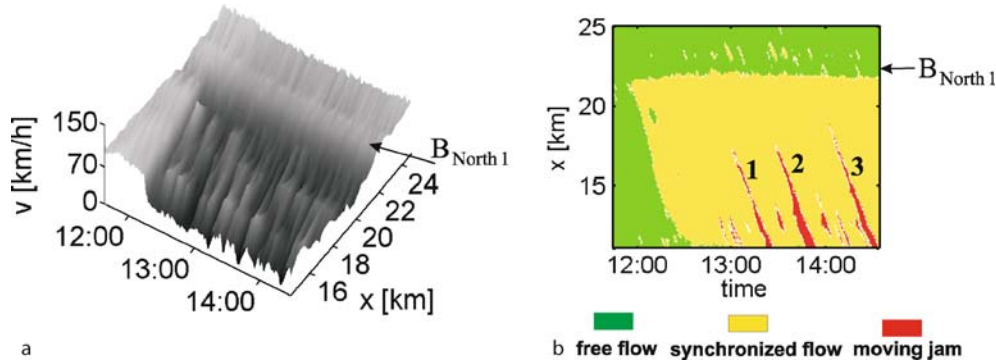
In synchronized flow at higher speeds, wide moving jams should not necessarily emerge spontaneously. Observations show that the greater the density in synchronized flow, the more likely is spontaneous moving jam emergence in that synchronized flow. Thus a wide moving jam emerges in an initial free flow due to a sequence of two phase transitions: Firstly, an  $F \rightarrow S$  transition occurs and synchronized flow emerges. Later and usually at other freeway locations than the location of the  $F \rightarrow S$  transition, an  $S \rightarrow J$  transition occurs spontaneously leading to wide moving jam emergence. This sequence of phase transitions is called the  $F \rightarrow S \rightarrow J$  transitions.

### Pinch Effect

In empirical observations,  $S \rightarrow J$  transition development is as follows. Narrow moving jams emerge spontaneously with subsequent self-compression of synchronized flow. This self-compression of synchronized flow with narrow moving jam emergence is called the pinch effect in synchronized flow. In the related pinch region of synchronized flow, the density is great and the speed is low, however, the flow rate can be great enough. Emergent narrow moving jams grow, propagating upstream within the pinch region. Some (or each) of these jams can transform into wide moving jams. Locations of these  $S \rightarrow J$  transitions are the upstream locations of the pinch region.

A congested pattern at a hypothetical isolated bottleneck, which occurs due to the  $F \rightarrow S \rightarrow J$  transitions, is called a *general pattern* (GP). Thus the GP is a congested pattern, which consists of synchronized flow upstream of the bottleneck and wide moving jams that emerge spontaneously in that synchronized flow [52]. Within the GP, there are two traffic phases of congested traffic, synchronized flow and wide moving jam(s).

A scenario of GP emergence can involve the transformation of an initial SP into a GP (Fig. 18). Firstly, the SP occurs at an off-ramp bottleneck (labeled  $B_{North 1}$ ). Over time, the density of synchronized flow increases. In the associated pinch region of synchronized flow, narrow moving jams emerge spontaneously. The narrow moving jams propagate upstream. Some of these narrow moving jams grow over time propagating upstream. Fi-



**Traffic Congestion, Modeling Approaches to, Figure 18**

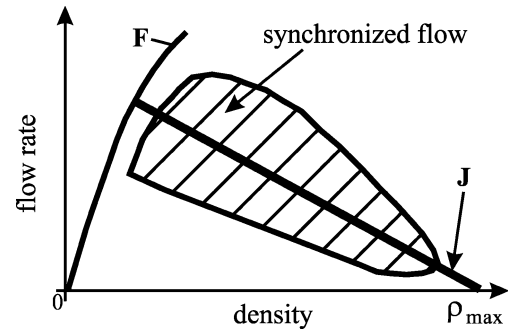
Empirical example of general pattern (GP) emergence at an off-ramp bottleneck (labeled  $B_{\text{North 1}}$ ). **a** Vehicle speed in space and time. **b** A graph of **a** with the free flow phase (green), the synchronized flow phase (yellow) and the wide moving jam phase (red) (the wide moving jams labeled 1, 2 and 3 in **b**). Taken from [52]

nally, the growing narrow moving jam can transform into a wide moving jam, i. e., an  $S \rightarrow J$  transition occurs spontaneously. In [► Traffic Congestion, Spatiotemporal Features of](#), it is shown that the moving jams 1, 2 and 3 shown in Fig. 18b propagate further through an upstream on-ramp bottleneck while maintaining their mean velocities of the downstream jam fronts. Thus in accordance with the traffic phase definition [J], these moving jams are indeed wide moving jams. A more detailed consideration of empirical GPs and other congested traffic patterns appears in [► Traffic Congestion, Spatiotemporal Features of](#).

### Explanations of $S \rightarrow J$ Transitions Through Three-Phase Traffic Theory

Empirical  $S \rightarrow J$  transitions are explained in three-phase traffic theory by the following hypotheses [49,50,51]:

- All infinite steady states of traffic flow in the flow-density plane that lie on the line  $J$  are threshold states for wide moving jam existence and emergence (Fig. 19).
- The line  $J$  intersects the 2D-region of steady states of synchronized flow in the flow-density plane, i. e., there are synchronized flow steady states below and above the line  $J$  (Fig. 19).
- The line  $J$  separates all steady states of synchronized flow in the flow-density plane into two different classes. All states below the line  $J$  are stable with respect to wide moving jam emergence, i. e., no  $S \rightarrow J$  transitions are possible within these states. All states on and above the line  $J$  are metastable states with respect to wide moving jam existence and emergence (Fig. 19).

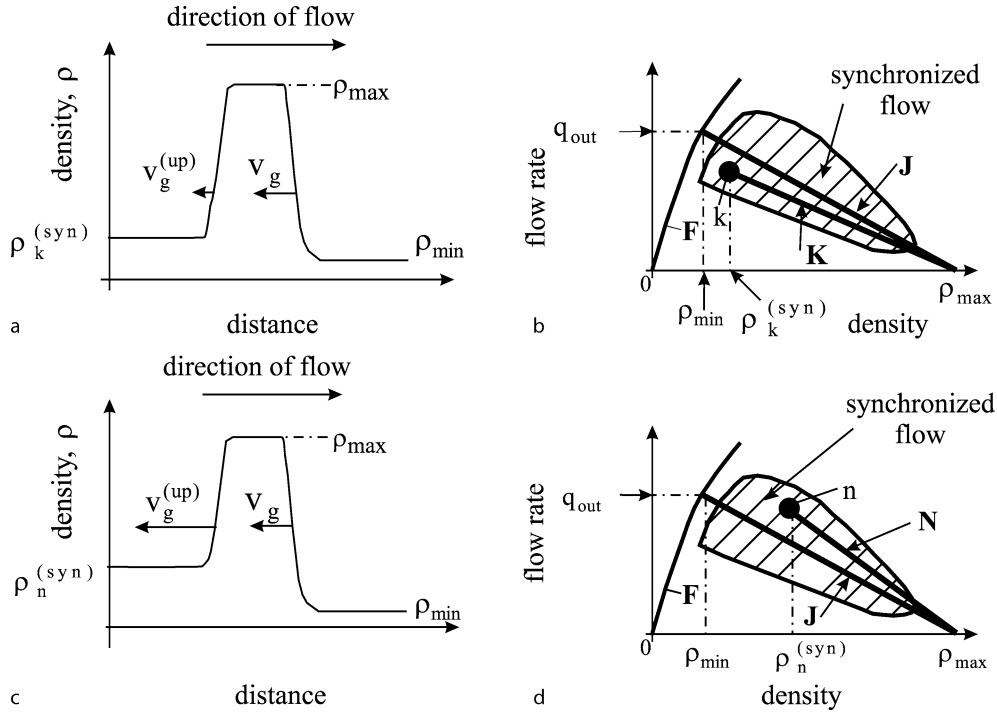


**Traffic Congestion, Modeling Approaches to, Figure 19**

Three-phase traffic theory in the flow-density plane [49,50,51]: Free flow (F), steady states of synchronized flow (2D dashed region) taken from Fig. 15a, and the line  $J$

- At a given synchronized flow speed, the greater the density in synchronized flow states that are above the line  $J$ , the greater the probability of an  $S \rightarrow J$  transition.

For a more detailed explanation of  $S \rightarrow J$  transitions (Fig. 20), let us assume that a state of synchronized flow directly upstream of a wide moving jam is associated with a point  $k$  in the flow-density plane. This point is below the line  $J$  (Fig. 20a, b). Because the velocity of the upstream front of the wide moving jam  $v_g^{(\text{up})}$  equals the slope of the line  $K$  (from a point  $k$  in free flow to the point  $(\rho_{\text{max}}, 0)$ ), the absolute value  $|v_g^{(\text{up})}|$  is always less than that of the downstream front  $|v_g|$  determined by the slope of the line  $J$ , i. e., the formula  $|v_g^{(\text{up})}| < |v_g|$  is valid. Therefore, the width of the wide moving jam gradually decreases and the jam dissolves. This means that no wide moving jams can persist continuously, i. e., all steady states of synchro-



**Traffic Congestion, Modeling Approaches to, Figure 20**

Explanations of moving jam emergence in synchronized flow in three-phase traffic theory [49,50,51]: a,c Qualitative forms of wide moving jams at two different densities in synchronized flow upstream of the jams,  $\rho^{(syn)} = \rho_k^{(syn)}$  (a) and  $\rho^{(syn)} = \rho_n^{(syn)}$  (c). b,d Representation of the line  $J$  and upstream wide moving jam fronts (lines  $K$  (b) and  $N$  (d)) in the flow-density plane for the wide moving jams in a and c, respectively. In b,d states for free flow (curve  $F$ ), steady states of synchronized flow (dashed region) are taken from Fig. 15a

nized flow below the line  $J$  are stable with respect to wide moving jam emergence.

In contrast, assume that a state of synchronized flow upstream of another wide moving jam is associated with a point  $n$  in the flow-density plane. This state is above the line  $J$  (Fig. 20c, d). In this case, the velocity of the upstream front of the wide moving jam  $v_g^{(up)}$  equals the slope of the line  $N$  (from a point  $n$  in synchronized flow to the point  $(\rho_{max}, 0)$ ), i. e., the absolute value  $|v_g^{(up)}|$  is always higher than that of the downstream front  $|v_g|$ , i. e., the formula  $|v_g^{(up)}| > |v_g|$  is valid. Therefore, the width of the wide moving jam in Fig. 20c should gradually increase. For these reasons, wide moving jams can be formed in states of synchronized flow that lie on or above the line  $J$ : these states are metastable with respect to  $S \rightarrow J$  transitions, i. e., an  $S \rightarrow J$  transition is a first-order phase transition. This analysis explains the hypothesis (c).

In synchronized flow states that are on the line  $J$ , the velocities of the downstream and upstream fronts of a wide moving jam are equal to each other; this explains the hy-

pothesis (a) that these states are threshold ones for  $S \rightarrow J$  transitions (see for more detail Sect. 6.3.1 in [52]).

Three-phase traffic theory explains  $S \rightarrow J$  transitions in the metastable synchronized flow states by a competition between the speed adaptation effect (Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”) and over-deceleration. Over-deceleration introduced by Herman et al. [43] (Sect. “Classic General Motors (GM) Model Approach: Free Flow Instability due to Driver Reaction Time”) occurs due to a finite driver reaction time: if the preceding vehicle begins to decelerate unexpectedly, then owing to the reaction time the vehicle starts deceleration with a delay; as a result, if the time delay is long enough the driver decelerates stronger than it is needed to avoid collisions. In earlier traffic flow models based on the fundamental diagram hypothesis, over-deceleration causes an instability of free flow beginning at a critical density [34,43]; the instability should explain traffic breakdown. As stressed in Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”, in



three-phase traffic theory rather than this instability associated with over-deceleration, a competition between the speed adaptation effect and over-acceleration explains empirical features of traffic breakdown.

To explain the competition between the speed adaptation effect and over-deceleration, we assume that a vehicle moving in a metastable synchronized flow state that is above the line  $J$  in Fig. 20d decelerates unexpectedly. If the following vehicle cannot pass this preceding vehicle, there are two possibilities: (i) The vehicle decelerates and due to the speed adaptation effect is able to adapt the speed to the speed of the preceding vehicle. Then rather than an  $S \rightarrow J$  transition, a new synchronized flow state of lower speed is formed. (ii) Due to over-deceleration, the vehicle decelerates stronger than it would be needed for speed adaptation, i. e., its speed becomes lower than the speed of the preceding vehicle. If each of the following vehicles decelerates also to lower speed than the associated preceding vehicle, then finally the speed upstream decreases up to zero leading to an  $S \rightarrow J$  transition. Thus if a local speed disturbance with a speed decrease occurs in synchronized flow, there is a tendency towards another state of synchronized flow associated with the vehicle adaptation effect. In contrast, due to over-deceleration there is also a tendency towards a wide moving jam. The greater the density, the more probable over-deceleration in comparison with speed adaptation. This explains the hypothesis (d) above.

Thus synchronized flow associated with a point in the flow-density plane that lies on or above the line  $J$  (Fig. 20d) is in a metastable state with respect to wide moving jam emergence. For this reason, there can be speed disturbances (fluctuations) within the synchronized flow whose growth lead to wide moving jam emergence. In accordance with definitions of phase transitions in complex systems, such wide moving jam emergence is called *spontaneous* wide moving jam emergence in synchronized flow ( $S \rightarrow J$  transition). This explains the term *spontaneous*  $S \rightarrow J$  transition used above [52].

A speed disturbance begins spontaneously to grow in a metastable state of synchronized flow [52], if within the initial speed disturbance the speed is equal or lower than a critical speed for an  $S \rightarrow J$  transition; otherwise the disturbance decays over time. A critical speed disturbance associated with the critical speed within the disturbance can be considered a nucleus for the  $S \rightarrow J$  transition.

There can be various sources for a speed disturbance whose occurrence and subsequent growth in metastable synchronized flow leads to a spontaneous  $S \rightarrow J$  transition [52]: unexpected braking of a vehicle in synchronized flow, lane changing within synchronized flow on the main road, fluctuations in flow rates, vehicle merging from

other roads (e. g., at bottlenecks), etc. Over time, a growing speed disturbance takes the form of a narrow moving jam (Sect. “Pinch Effect”), i. e., the growing speed disturbance of a great enough amplitude is synonym of a growing narrow moving jam in the pinch region of synchronized flow.

Note that lane changing as a source for the spontaneous occurrence of growing narrow moving jams in synchronized flow has recently been empirically studied in [1]. In numerical simulations of GP emergence at isolated bottlenecks based on three-phase traffic flow models, it has been found [62] that vehicle merging from other roads and lane changing in a neighborhood a bottleneck are the main sources for speed disturbances whose subsequent growth in metastable synchronized flow leads to spontaneous  $S \rightarrow J$  transitions (see also simulation results of [54] associated with the effect of lane changing on spontaneous  $S \rightarrow J$  transitions upstream of an off-ramp bottleneck).

It should be noted that in metastable synchronized flow there can also be an *induced*  $S \rightarrow J$  transition. In contrast with a spontaneous  $S \rightarrow J$  transition, induced wide moving jam emergence in the synchronized flow is caused by the upstream propagation of a moving jam, which has initially occurred within a *different* link of the road network connected with the road under consideration.

### Comparison of $F \rightarrow S$ and $S \rightarrow J$ Transitions

Both a spontaneous  $F \rightarrow S$  transition (Sect. “Spontaneous Traffic Breakdown ( $F \rightarrow S$  Transition)”) and a spontaneous  $S \rightarrow J$  transition are first-order phase transitions that occur in metastable states of free flow ( $F$ ) and synchronized flow ( $S$ ), respectively. In both cases, critical speed disturbances should appear whose subsequent growth lead to the related phase transition. However, there is a qualitative difference between these two phase transitions. This difference is associated with the kinetics of the growth of the critical disturbances.

When a critical speed disturbance appears in free flow at a bottleneck, the growth of this disturbance and resulting traffic breakdown occurs usually also at the bottleneck. This is explained by a permanent speed disturbance, which is on average motionless and localized in free flow at the bottleneck. This disturbance increases probability of synchronized flow emergence, i. e., breakdown probability at the bottleneck considerably.

In contrast, a growing narrow moving jam in metastable synchronized flow propagates upstream. It takes some time delay before the growth of the jam leads to wide moving jam, i. e., an  $S \rightarrow J$  transition occurs. For this reason, the  $S \rightarrow J$  transition occurs upstream of the

road location at which a critical speed disturbance has initially appeared and begun to grow in synchronized flow. Thus, even if a critical speed disturbance has appeared in metastable synchronized flow at the location of a bottleneck, the  $S \rightarrow J$  transition resulting from the growth of the disturbance occurs upstream of this bottleneck. The exclusion can be a very heavy bottleneck, which limits the flow rate in congested traffic considerably. A theory of traffic congestion at heavy bottlenecks appears in ► [Traffic Congestion, Spatiotemporal Features of](#).

Note that there can be also nucleation-interruption effects, which play an important role for both spontaneous  $F \rightarrow S$  and  $S \rightarrow J$  transitions. A consideration of these effects is, however, out of the scope of the article (see Sects. 6.5.3 and 10.2 in [52]).

### Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow

As mentioned above,  $F \rightarrow S$  and  $S \rightarrow J$  transitions are first-order phase transitions. Both of these first-order phase transitions can be illustrated by a double Z-characteristic for the phase transitions. An empirical double Z-characteristic for the phase transitions, which is measured during GP formation at an on-ramp bottleneck (Fig. 21a), is presented in Fig. 21b.

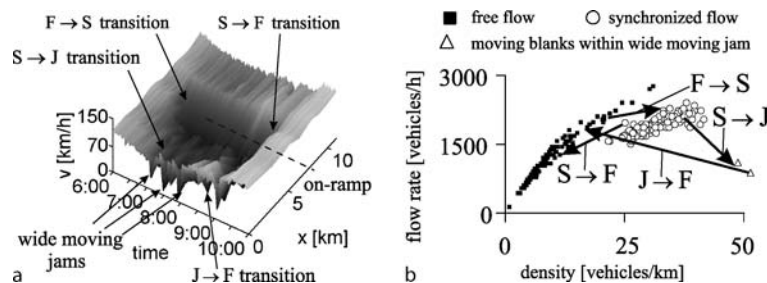
Firstly, an  $F \rightarrow S$  transition occurs at the bottleneck (arrows labeled by  $F \rightarrow S$  in Figs. 21a,b). Synchronized flow propagates upstream. Later, in the synchronized flow the pinch effect is realized upstream of the bottleneck. In the pinch region of synchronized flow, narrow moving jams emerge. The jams propagate upstream growing in the jam amplitude. Some of these narrow moving jams transform into wide moving jams, i.e.,  $S \rightarrow J$  transitions occur (one of these transitions at approximately  $x = 3.2$  km is labeled by arrow  $S \rightarrow J$  in Fig. 21a). Thus  $S \rightarrow J$  transitions occur later and upstream of the freeway location of

the  $F \rightarrow S$  transition. Within wide moving jams low speed states are observed that are probably associated with low speed within wide moving jams associated with so-called “moving blanks” that are discussed in ► [Traffic Congestion, Spatiotemporal Features of](#).

Later the GP begins to dissolve. In particular, some of wide moving jams associated with the GP dissolve and either synchronized flow or free flow occur ( $J \rightarrow S$  or  $J \rightarrow F$  transitions). An example of an  $J \rightarrow F$  transition associated with a wide moving jam that begins to dissolve between locations 1.5 and 0 km is shown in Fig. 21. Finally, an  $S \rightarrow F$  transition occurs at the bottleneck and the GP disappears (arrows labeled  $S \rightarrow F$  in Fig. 21).

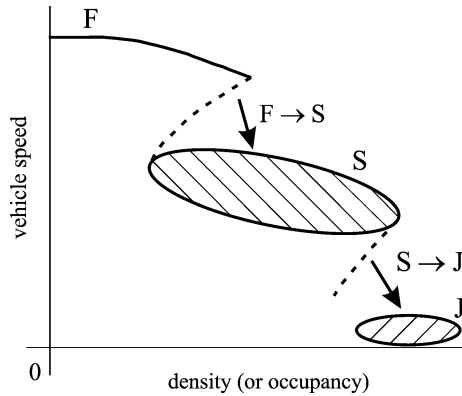
We see that there are many diverse transitions between the three traffic phases as well as between diverse traffic states within the phases associated with congested pattern emergence and dissolution. The phase transitions and transitions within the same traffic phase could exhibit a variety of hysteresis effects. The phase transitions in traffic and transitions within the same traffic phase cannot be distinguished each other without knowledge of the whole *spatiotemporal dynamics* of the congested pattern (Fig. 21a).

As discussed in Sect. “[Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory](#)”, for  $F \rightarrow S$  transition occurrence a local speed disturbance should occur in free flow at the bottleneck with the speed within the disturbance that is equal or lower than a critical speed. The greater the density (flow rate) downstream of the bottleneck, the higher this critical speed. Respectively, for  $S \rightarrow J$  transition occurrence, a local speed disturbance should appear in synchronized flow upstream of the bottleneck with the speed within the disturbance that is equal or lower than another critical speed. For this reason, in a theoretical double Z-characteristic (Fig. 22) between states of free flow, synchronized flow, and low speed states within wide moving jams there are states associated with the critical speed disturbances (nuclei for



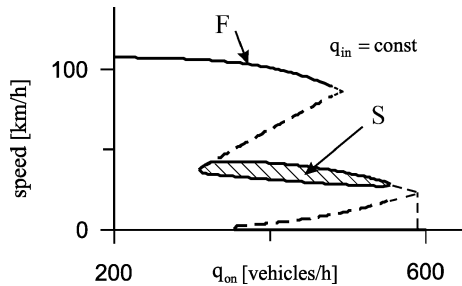
**Traffic Congestion, Modeling Approaches to, Figure 21**

Average speed in empirical GP at on-ramp bottleneck (a) and related empirical double Z-characteristic for phase transitions (b). Low speed states within wide moving jam are labeled as “moving blanks” in b



**Traffic Congestion, Modeling Approaches to, Figure 22**

Qualitative double Z-characteristic for  $F \rightarrow S \rightarrow J$  transitions at bottlenecks that includes states of free flow (F), synchronized flow (S) and low speed states (J) associated with moving blanks within wide moving jams [52]



**Traffic Congestion, Modeling Approaches to, Figure 23**

Simulated double Z-characteristic for  $F \rightarrow S \rightarrow J$  transitions at on-ramp bottlenecks. Taken from [52]

phase transitions) needed for phase transition occurrence between the related traffic phases.

It should be noted that in addition to the critical speed (density) within a critical disturbance, the critical disturbance is also characterized by a *critical spatial shape*, i. e., by the critical speed (density) spatial distribution within the disturbance. This conclusion of three-phase traffic theory is valid for  $F \rightarrow S$  and  $S \rightarrow J$  transitions. However, for simplicity we neglect the critical spatial shape of the disturbances in Fig. 22, in which critical speed disturbances (nuclei for the phase transitions) are qualitatively presented by dashed curves between states  $F$  and  $S$  associated with critical speeds within the nuclei for  $F \rightarrow S$  transitions and between states  $S$  and  $J$  associated with critical speeds within the nuclei for  $S \rightarrow J$  transitions.

The double Z-characteristic features can also be seen on simulated double Z-characteristic for the  $F \rightarrow S \rightarrow J$  transitions shown in Fig. 23. For simplicity, it has been suggested that the speed within wide moving jams is equal to zero, i. e., states  $J$  shown in Fig. 22 have been neglected.

As mentioned, the states  $J$  are associated with moving blanks within the jams (Fig. 21b) [52]; empirical and simulation results associated with moving blanks are discussed in more detail in ► [Traffic Congestion, Spatiotemporal Features of](#).

### Hypotheses of Three-Phase Traffic Theory as the Result of Traffic Phase Definitions

In Sects. “Free and Congested Traffic”–“Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow”, we have shown that empirical features of traffic breakdown and congested pattern formation are explained in the framework of three-phase traffic theory. Here we explain that and why hypotheses of this theory [49,50,51] are the *result* of the traffic phase definitions [J] and [S]:

1. The definition of wide moving jam [J] determines the line  $J$  in the flow-density plane. Any point on the line  $J$  can be a final steady traffic state for vehicles accelerating at the downstream front of a wide moving jam. If the speed in this state is lower than the minimum speed in free flow, the state is a synchronized flow steady state that lies on the line  $J$ . Thus there should be infinite synchronized flow states lying on the line  $J$  (Fig. 20b, d).
2. The definition of synchronized flow [S] means that downstream fronts of synchronized flow regions do *not* exhibit the jam characteristic feature [J]. For this reason, in addition to synchronized flow steady states lying on the line  $J$  (item 1) there should be synchronized flow steady states that are away from the line  $J$  in the flow-density plane. Indeed, only in this case the downstream fronts between different states of synchronized flow do not exhibit the characteristic jam velocity given by the slope of the line  $J$ . Thus there should be also synchronized flow steady states away from the line  $J$ , i. e., that there is a 2D-region of the steady states in the flow-density plane. This explains the fundamental hypothesis of three-phase traffic theory (Fig. 15).
3. The hypothesis of three-phase traffic theory that an  $F \rightarrow S$  transition is a first-order phase transition results from the 2D-region of synchronized flow steady states as already explained in Sect. “[Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory](#)”.
4. The definition [S] is associated with empirical results for moving jam behavior within synchronized flow [52]: depending on synchronized flow characteristics upstream of a moving jam, the jam can grow or dissolve propagating in the synchronized flow over time. We have noted (item 1) that there are infinite synchronized flow steady states that lie on the line  $J$ .

If these states were to build the upper boundary of the 2D-region of the synchronized flow steady states, i. e., there were no steady states above the line  $J$ , then, as explained in Sect. “Explanations of  $S \rightarrow J$  Transitions Through Three-Phase Traffic Theory”, all wide moving jams should dissolve propagating in the states of synchronized flow. This contradicts with mentioned empirical results about possible jam growth. In contrast, if the states on the line  $J$  were to build the low boundary of the 2D-region of the synchronized flow steady states, i. e., there were no steady states below the line  $J$ , then no wide moving jam can dissolve propagating in the states of synchronized flow. This also contradicts with mentioned empirical results about possible jam dissolution. Thus we should assume that there are synchronized flow steady states above *and* below the line  $J$ . This explains the hypothesis (b) of Sect. “Explanations of  $S \rightarrow J$  Transitions Through Three-Phase Traffic Theory”.

5. The hypotheses (a) and (c) of Sect. “Explanations of  $S \rightarrow J$  Transitions Through Three-Phase Traffic Theory” about  $S \rightarrow J$  transitions that result from item 2 and 4 have already been proven in Sect. “Explanations of  $S \rightarrow J$  Transitions Through Three-Phase Traffic Theory”.

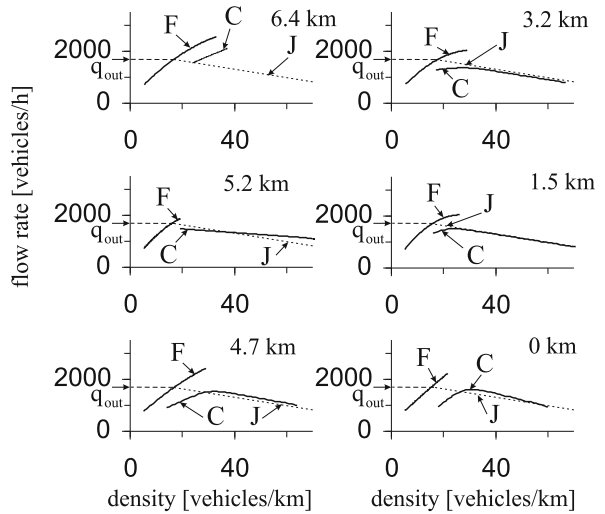
### Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion

In this section, we will explain and give reasons why all earlier traffic flow theories and models cannot show and explain *empirical* traffic breakdown discussed above.

#### Validation of Theoretical Fundamental Diagram for Traffic Flow

It has been mentioned in the introduction that the fundamental diagram (Fig. 1a) is the basis hypothesis for all earlier traffic flow theories models reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,102,126,127,128]. However, empirical observations show that whereas fundamental diagrams for free flow are qualitatively the same at different freeway locations, in contrast, the shape of empirical fundamental diagrams for congested traffic depends qualitatively on the location at which data measurements have been performed within a congested pattern. An example for different fundamental diagram types associated with a GP at an on-ramp bottleneck is shown in Fig. 24 (see explanations in Chap. 15 in [52]).

This means that in a general case the validation of a theoretical fundamental diagram of a traffic flow model, which is very often used by researchers as an empirical



**Traffic Congestion, Modeling Approaches to, Figure 24**

Dependence of empirical fundamental diagram types for congested traffic at different locations within the GP shown in Fig. 21a. F free flow. C congested traffic. J line  $J$  (dashed line). On-ramp bottleneck is at the location about 6.4 km. See explanations of these fundamental diagram types in Chap. 15 in [52]

basis for the fundamental diagram, cannot be performed. Depending on a road location various empirical fundamental diagram types exist. Moreover, if data, which are measured at many different locations within congested traffic, is used for averaging to one empirical fundamental diagram (this is often a case in the literature), then there is almost no possibility to find a relation of such a fundamental diagram to spatiotemporal features of real traffic patterns.

Thus there cannot be a choice of the theoretical fundamental diagram, which can alone give a correct relation between model solutions for spatiotemporal congested patterns and real measured spatiotemporal congested patterns. However, such a relation is very often suggested in a huge number of traffic flow models and theories within the fundamental diagram approach reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128] and briefly discussed in the reminder of this section. This may be one of the main reasons for failure of this well-accepted and widely used approach to traffic flow theory and modeling in describing of traffic breakdown and resulting spatiotemporal traffic congested patterns.

#### Classic Lighthill–Whitham–Richards (LWR) Theory of Onset of Traffic Congestion

The basic idea of the classic LWR traffic flow theory is as follows: the maximum flow rate  $q_0$  associated with

the maximum point  $(\rho_0, q_0)$  at the fundamental diagram (Fig. 1a) determines free flow capacity at a bottleneck (often called “bottleneck capacity”) denoted by  $q_{\text{cap}}$ :

$$q_{\text{cap}} = q_0. \quad (15)$$

It is further suggested that due to a road non-homogeneity introduced by the bottleneck the maximum flow rate  $q_0$  downstream of the bottleneck, which determines the capacity (15), is smaller than the capacity of a homogeneous road upstream of the bottleneck. For this reason, if the sum of the flow rates upstream of the bottleneck reaches the maximum flow rate  $q_0$  at the fundamental diagram (Fig. 1a), i. e., it reaches the bottleneck capacity (15), then a further increase in the flow rates must lead to congestion (queue) formation and upstream congestion propagation upstream of the bottleneck.

These hypotheses of the LWR theory mean that congested traffic occurs only then, when the upstream flow rate exceeds the bottleneck capacity  $q_{\text{cap}}$  determined by (15). This conclusion of the LWR traffic flow theory about the reason for the onset of congestion is very often used as the definition of congested traffic. As we will see below, this hypothesis of the LWR theory about the onset of congestion as well as the associated congested traffic definition are in a very deep contradiction with empirical results.

In accordance with the LWR theory, traffic flow phenomena should be explained based on the law of conservation of the number of vehicles on the road

$$\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial Q(\rho(x, t))}{\partial x} = 0, \quad (16)$$

in which is assumed that there is a relationship between the flow rate  $Q$  and density  $\rho$

$$Q = Q(\rho) \quad (17)$$

associated with the fundamental diagram for traffic flow. Here,  $x$  is a spatial coordinate in the direction of traffic flow, and  $t$  is time. The LWR-model (16), (17) has discontinuous solutions in the form of shock-waves (Fig. 25) with a shock-wave velocity

$$v_s = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1}, \quad (18)$$

where the flow rates  $Q(\rho_2)$ ,  $Q(\rho_1)$  associated with (17) correspond to points in the flow-density plane that lie on the fundamental diagram;  $\rho_1$  and  $\rho_2$  are the densities upstream and downstream of the shock wave, respectively (Fig. 25). The shock wave can be represented in the flow-

density plane by the line labeled “shock” whose slope is equal to the shock velocity  $v_s$  (Fig. 25b).

Because the LWR-model (16), (17) has discontinuous solutions in the form of shock waves, its numerical simulations are usually based on one of finite difference approximation methods for partial differential equations associated with the Godunov family methods. One of such discrete versions of the LWR-model (16) that is consistent with the LWR-hydrodynamic theory is the cell-transmission model of Daganzo [20].

In Daganzo’s cell transmission model, a rectangular lattice with time spacing  $\tau$  and space length (cell length)  $\Delta x$  is overlaid on the time-space plane; the  $x$ -coordinates represent the center of the cells into which a road has been discretized;  $t$ -coordinates are the times at which the cell vehicle densities are evaluated. Then the Daganzo model [20] for a single-lane road of length  $L$  with an on-ramp bottleneck can be written as follows:

$$\begin{aligned} \rho(x, t + \tau) = & \rho(x, t) - \left[ q \left( x + \frac{\Delta x}{2}, t + \frac{\tau}{2} \right) \right. \\ & \left. - q \left( x - \frac{\Delta x}{2}, t + \frac{\tau}{2} \right) \right] \left( \frac{\tau}{\Delta x} \right) \\ & + q_{\text{on}}(t) f(x) \left( \frac{\tau}{\Delta x} \right), \end{aligned} \quad (19)$$

$$\begin{aligned} q \left( x + \frac{\Delta x}{2}, t + \frac{\tau}{2} \right) \\ = \min(D(\rho(x, t)), R(\rho(x + \Delta x, t))) , \end{aligned} \quad (20)$$

$$f(x) = \begin{cases} \frac{1}{N_m} & \text{if } x_{\text{on}} \leq x < x_{\text{on}} + L_m, \\ 0 & \text{otherwise,} \end{cases} \quad (21)$$

with boundary conditions

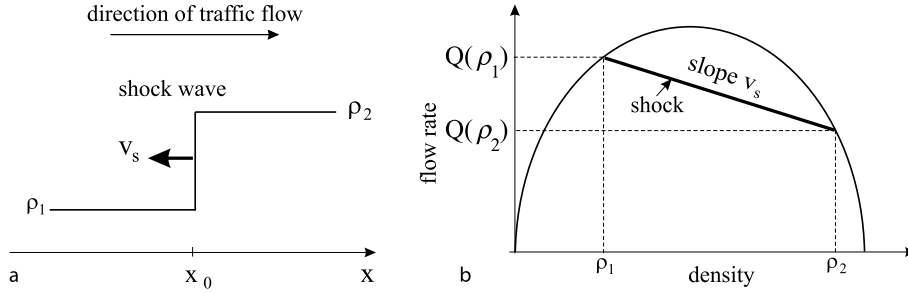
$$\rho(0, t) = \rho(\Delta x, t), \quad \rho(L, t) = \rho(L - \Delta x, t), \quad (22)$$

and initial conditions

$$\rho(x, 0) = \rho_{\text{in}}, \quad (23)$$

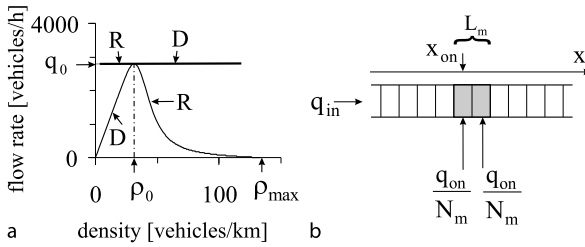
where  $R(\rho)$  and  $D(\rho)$  are non-increasing so-called *receiving* and non-decreasing so-called *sending* curves, respectively; these two monotonic curves  $R(\rho)$  and  $D(\rho)$  take values in the interval  $[0, q_0]$ , as shown in Fig. 26a:  $R(\rho) = Q(\rho)$  at  $\rho \geq \rho_0$ ;  $D(\rho) = Q(\rho)$  at  $\rho \leq \rho_0$ ;  $x = x_{\text{on}}$  is the location of the farthest upstream cell of the on-ramp merging region of length  $L_m = N_m \Delta x$  that includes the integer number  $N_m$  of cells (Fig. 26b); the flow rate  $q_{\text{on}}(t)$  is zero at  $t < t_0$  and  $q_{\text{on}}(t) = q_{\text{on}}$  is constant at  $t \geq t_0$ ;  $\rho_{\text{in}}$  is the density in an initial homogeneous free flow on the road,  $q_{\text{in}} = Q(\rho_{\text{in}})$  is the corresponding flow rate.





**Traffic Congestion, Modeling Approaches to, Figure 25**

Qualitative representation of a shock wave in space (a) and in the fundamental diagram with the line labeled “shock” whose slope is equal to the shock velocity  $v_s$  (b) [86]



**Traffic Congestion, Modeling Approaches to, Figure 26**

Parameters of the Daganzo cell-transmission model (19)–(23) used for numerical simulations of traffic breakdown in Fig. 27: **a** Receiving (R) and sending (D) curves. **b** Model of on-ramp bottleneck.  $Q(\rho) = \rho V(\rho)$ ,  $V(\rho) = V_g(g)$ , where  $g = \rho^{-1} - \rho_{\max}^{-1}$ ,  $V_g(g) = V_0 (\tanh((g - g_2)/g_1) + \tanh(g_2/g_1))$  at  $V_0 = 14$  m/s,  $g_2 = 17$  m,  $g_1 = 7$  m,  $\tau = 0.1$  s,  $\Delta x = 10$  m,  $L = 20$  km,  $N_m = 2$ ,  $t_0 = 7$  min,  $x_{on} = 16$  km,  $q_0 = 2795$  and  $q_{in} = 2676$  vehicles/h

Each of the simulations of congested traffic patterns with the Daganzo model (19)–(23) (Fig. 27) is performed during a limited time interval within which none of the patterns reach the road boundaries, i. e., free flow conditions remain at the boundaries. The simulations show common qualitative features of traffic breakdown and resulting congested traffic patterns that show the LWR theory at an isolated on-ramp bottleneck [62].

The onset of traffic congestion and resulting congested patterns within the LWR-theory are studied by an increase in the flow rate to the on-ramp  $q_{on}$ , while the flow rate in free flow upstream of the on-ramp bottleneck  $q_{in}$  is constant. Numerical simulations are made for various given values  $q_{on}$ . The difference in these values of  $q_{on}$  chosen for different simulations is small enough to ensure that all possible various congested patterns and their dynamic behavior characteristics can be found and studied.

If  $q_{on} = 0$ , the speed and density are spatially homogeneous. When the flow rate  $q_{on}$  increases beginning from zero (points 1–3 in Fig. 27a), the flow rate downstream of

the bottleneck  $q_{sum} = q_{on} + q_{in}$  increases too as long as the condition

$$q_{sum} = q_{on} + q_{in} < q_0 \quad (24)$$

is satisfied. In this case, a non-homogeneous motionless structure of traffic flow appears at the bottleneck related to different free flows downstream and upstream of the bottleneck (curves I in Fig. 27b, h). Speed decreases and density increases downstream of the bottleneck.

The differences in the speed and density upstream and downstream within this motionless free flow structure increase when  $q_{on}$  increases. However, this has a limit at a given large enough flow rate  $q_{in}$ . This limit is reached when the flow rate  $q_{on}$  reaches some critical value  $q_{on} = q_{on}^{(d)}$  at which the flow rate  $q_{sum}$  is equal to the maximum flow rate on the fundamental diagram:

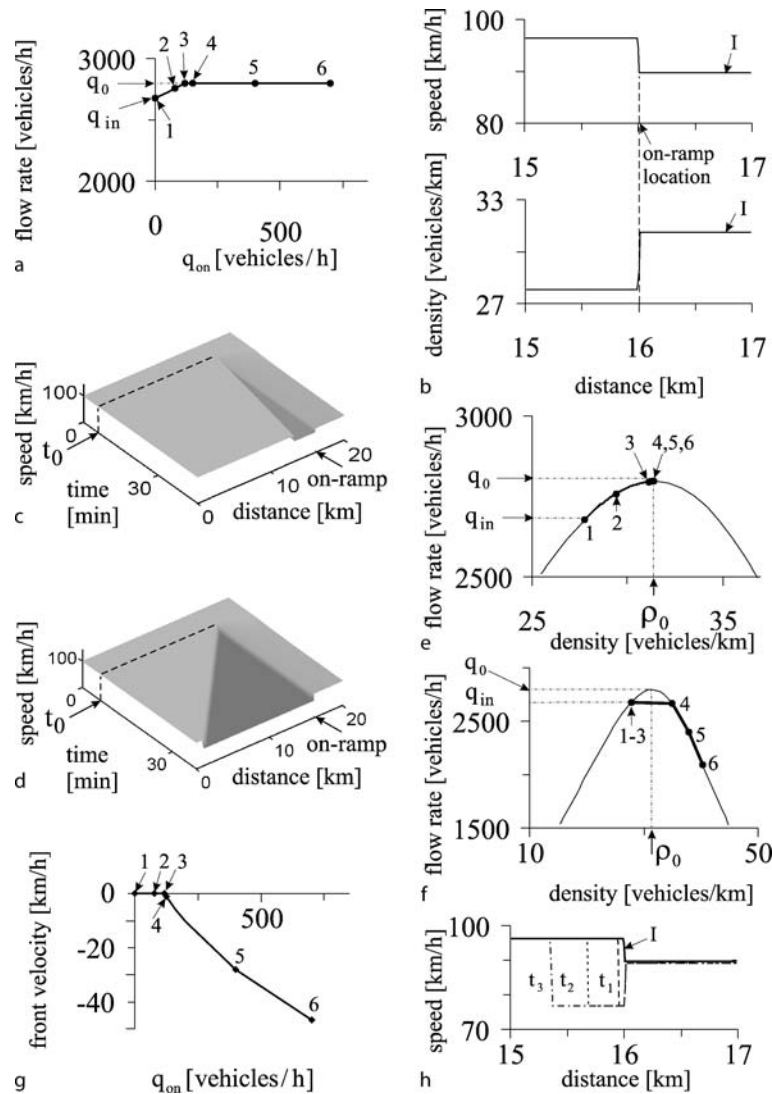
$$q_{sum} = q_{in} + q_{on}^{(d)} = q_0 \quad (25)$$

When the flow rate  $q_{on}$  increases further, i. e.,

$$\Delta q = q_{in} + q_{on} - q_0 > 0, \quad (26)$$

then the upstream front of the initial free flow structure, which is motionless at the condition  $q_{sum} = q_{in} + q_{on} \leq q_0$  (curve I in Fig. 27b), begins to move upstream of the bottleneck, i. e., a wave (shock wave) of lower speed and greater density propagating upstream appears (Fig. 27c, d). As a result, a dense flow associated with the branch of the diagram with a negative slope occurs upstream of the bottleneck (points 4–6 in Fig. 27e, f).

At the critical point (25), the flow rates on the main road downstream and upstream of the bottleneck are continuous functions of  $q_{on}$  (points 3, 4 in Fig. 27a, f). In contrast with the flow rates, under condition (26) already at  $\Delta q \rightarrow 0$  the minimum average speed  $v_{min}$  within the dense flow (point 4 in Fig. 27f) is lower than the speed



**Traffic Congestion, Modeling Approaches to, Figure 27**

Traffic breakdown in the LWR-theory [62]: **a** Flow rate downstream of bottleneck as a function of  $q_{on}$ . **b** Non-homogeneous motionless solution for free flow under condition (24). **c, d** Propagation of shock-waves under condition (26) for  $q_{on} = 160$  (c) and 400 vehicles/h (d). **e, f** On-ramp flow rate dependences of the flow rate and density on main road downstream (e) and upstream of bottleneck (f). **g**  $v_s(q_{on})$ . **h** Shock-wave propagation associated with point 4 in a, e-g. Simulations of the Daganzo model (19)–(23) with parameters given in caption of Fig. 26

upstream of the dense flow (point 3 in Fig. 27f); however, if we choose a greater  $q_{in}$ , then this speed difference decreases and it tends to zero, when  $q_{in} \rightarrow q_0$  (but  $q_{in} < q_0$ ).

Regardless of the choice in  $q_{in}$  (but  $q_{in} < q_0$ ), the absolute value of the shock-wave velocity  $|v_s|$  increases continuously beginning from zero, when  $q_{on}$  first reaches and then exceeds the critical flow rate  $q_{on}^{(d)}$  associated with (25). We find that if  $\Delta q \rightarrow 0$ , then  $|v_s| \rightarrow 0$ . For example, at

$\Delta q = 3$  vehicles/h ( $q_{on} = 120$  vehicles/h), the shock wave, within which the vehicle speed is associated with point 4 in Fig. 27f, has the velocity  $v_s \approx -0.7$  km/h only; this very slow shock wave propagation can be seen in Fig. 27h in which speed distributions within the wave related to the times  $t_1 = 25$ ,  $t_2 = 55$  and  $t_3 = 90$  min are shown.

The greater  $q_{on}$ , specifically, the greater  $\Delta q$  (26), the greater the absolute value of the shock-wave velocity  $|v_s|$  (Fig. 27g). In accordance with (18), in numerical simula-

tions for each of the flow rates  $q_{on}$  (points 4–6 in Fig. 27a) the front velocity  $v_s$  is determined by a slope of the line between the point related to the flow rate  $q_{in}$  in free flow upstream of the wave front (shock) (points 1–3 in Fig. 27f) and the related point at the fundamental diagram for congested traffic associated with the flow rate within dense flow downstream of the wave front (points 4–6 in Fig. 27f). In addition, the flow rate downstream of the bottleneck, which is equal to  $q_0$  under the condition (25), remains approximately to be equal to  $q_0$ , when  $q_{on}$  increases (points 4–6 in Fig. 27e).

The congested patterns in Fig. 27c, d at first glance resemble a widening SP. Indeed, in both cases a dense flow occurs upstream of the bottleneck whose downstream front is fixed at the bottleneck. Thus, this dense flow should satisfy the macroscopic spatiotemporal objective criteria for the synchronized flow phase [S] (Sect. “Introduction”). However, the criterion [S] defines the synchronized flow phase in *empirical* observations. This empirical synchronized flow occurs due to a first-order  $F \rightarrow S$  transition. This  $F \rightarrow S$  transition found in three-phase traffic theory explains the *fundamental empirical features* of traffic breakdown and bottleneck capacity (Sect. “Highway Capacity and First-Order  $F \rightarrow S$  Transition”). In contrast, the LWR-theory cannot explain these features.

Moreover, the classic formula (18) of the LWR-theory for shock waves, which is the fundamental formula of traffic science (see e.g., the books [21,84,93,126]), is not consistent with empirical features of synchronized flow; therefore, the LWR-formula (18) cannot be applied for an adequate calculation of front (shock) velocities of SPs as well as of the front velocities of wide moving jams when the jams propagate through synchronized flow.

To explain these critical conclusions, we summarize below features of congestion emergence in the LWR-theory, which are in deep contradiction with empirical results:

1. In the LWR-theory, there is *no* induced traffic breakdown at a bottleneck observed in empirical observations.
2. In the LWR-theory, there is *no* hysteresis effect, which is observed by congested pattern emergence and dissolution in empirical observations.
3. In the LWR-theory there is *no* discontinuous change in the velocity  $v_s$  when due to an increase in  $q_{on}$  the condition (26) is satisfied:  $|v_s|$  increases continuously beginning from zero, when  $q_{on}$  first reaches and then exceeds the critical flow rate  $q_{on}^{(d)}$  associated with the condition (25). This is qualitatively different from those for the upstream front of congested patterns at the bottleneck in empirical data. In the latter case, when the flow rate  $q_{on}$  increases and an  $F \rightarrow S$  transition occurs, a wave of synchronized flow occurs abruptly and propagates upstream with a *finite* velocity. This is associated with a first-order  $F \rightarrow S$  transition, which cannot be found and explained in the LWR-theory.
4. In empirical observations, the flow rate downstream of the bottleneck at which breakdown occurs in a realization (day) is often smaller than flow rates observed in free flow downstream of the bottleneck before breakdown occurrence (Fig. 12). This cannot be explained by the LWR-theory even if one suggests that the capacity  $q_{cap} = q_0$  (15) is a probabilistic value, which can be different in different realizations (days) at the same bottleneck (for a more detailed discussion of capacity definitions, see Sect. “Critical Discussion of Highway Capacity Definitions”). This is because for a given realization, within the framework of the LWR-theory there is a fixed bottleneck capacity at which traffic breakdown occurs. This contradicts with the above observations (Fig. 12).
5. In empirical data, after traffic breakdown has occurred at a bottleneck, wide moving jams can emerge spontaneously in synchronized flow upstream of the bottleneck, i.e., a GP emerges spontaneously. Regardless of the density, *no* spontaneous moving jam emergence can be found in the LWR-theory.
6. In empirical data, when a wide moving jam propagates through synchronized flow, there is no single relation between the density and flow rate in the synchronized flow. For this reason, in general case the classic LWR-formula for shock wave velocity (18) cannot be applied for a correct calculation of the jam front velocity while the jam propagates through synchronized flow.

Thus, both empirical features of traffic breakdown and resulting congested pattern formation discussed above are in a very deep contradiction with the LWR-theory. For these reasons, the LWR-theory and the associated traffic flow models (like cell-transmission models) cannot be used for reliable analysis of freeway traffic control and management strategies.

One of the reasons for drawbacks of the LWR-theory in explanation of empirical traffic breakdown is a single correspondence between the density and flow rate (17), i.e., the fundamental diagram for traffic flow used in this theory. In contrast, empirical measurements of synchronized flow show that there is no single correspondence between the density and flow rate in the synchronized flow. Thus instead of the LWR-formula (18), one should use the classic Stokes-formula for a shock-wave velocity derived

by Stokes in 1848 [114]:

$$v_s = \frac{q_2 - q_1}{\rho_2 - \rho_1}, \quad (27)$$

where  $\rho_1$ ,  $q_1$  and  $\rho_2$ ,  $q_2$  are the density and flow rate upstream and downstream of the shock wave, respectively. The fundamental difference between the LWR-formula (18) and the Stokes-formula (27) is that there is no given relationship between the density and flow rate in (27).

For these reasons, rather than the LWR-formula (18) in the model ASDA for tracking and prediction of wide moving jam propagation, which is successfully used in on-line applications on German freeways since 2000, the classic Stokes-formula (27) has been used. In the model ASDA, the density and flow rate  $\rho_1$ ,  $q_1$  upstream and  $\rho_2$ ,  $q_2$  downstream of the wide moving jam fronts in (27) are found based on measurements of these traffic flow variables (see a detailed consideration of the model ASDA and its application made in Chaps. 21 and 22 in [52]; for a brief review see ► [Traffic Prediction of Congested Patterns](#)).

#### Classic General Motors (GM) Model Approach: Free Flow Instability due to Driver Reaction Time

As long ago as 1958–1961, Herman, Montroll, Potts, Gazis, Rothery [34,43] and Komentani and Sasaki [73,74] suggested that traffic breakdown is associated with an instability of free flow. This instability is related to a finite reaction time of drivers. This driver reaction time is responsible for the over-deceleration effect discussed in Sect. “[Introduction](#)”; if the preceding vehicle begins to decelerate unexpectedly, then owing to the finite driver reaction time the following vehicle starts deceleration with a delay; as a result, if the time delay is long enough, the driver decelerates stronger than it is needed to avoid collisions. In this case, the driver speed can become lower than the speed of the preceding vehicle. If the over-deceleration effect is realized for next following drivers, the wave of vehicle speed reduction appears and increases in amplitude over time in traffic flow leading to a stop of some of the vehicles upstream. This instability should occur when the vehicle density on the fundamental diagram exceeds some critical density denoted  $\rho_{cr}^{(j)}$ .

In the car-following model of Herman, Montroll, Potts, Gazis, Rothery [34,43], which is often called the General Motors (GM) model, the driver reaction time denoted by  $\tau_0$  is explicitly used in the vehicle deceleration (acceleration) denoted by  $a(t + \tau_0)$ , i. e., the vehicle reacts with the time delay  $\tau_0$  on any changes in the space gap to the preceding vehicle  $g(t)$  and the speed difference

$\Delta v(t) = v_\ell(t) - v(t)$  (relative speed) between the speed of the preceding vehicle  $v_\ell(t)$  and the vehicle speed  $v(t)$ . The GM model reads as follows [34]:

$$a(t + \tau_0) = \frac{\Delta v(t)[v(t + \tau_0)]^{m_1}}{T_0[g(t) + d]^{m_2}}, \quad (28)$$

$$a(t) = \frac{dv(t)}{dt}, \quad (29)$$

where  $d$  is the vehicle length;  $T_0$ ,  $m_1$ ,  $m_2$  are constants. By integrating Eqs. (28), (29), one gets model solutions for steady states related to the fundamental diagram [34]

$$V(\rho) = V_0 \left[ 1 - (\rho/\rho_{\max})^{(m_2-1)} \right]^{(1-m_1)^{-1}}, \quad (30)$$

where  $V_0$  is constant,  $\rho_{\max} = 1/d$ .

The idea of the GM-model (28) with a driver reaction time is incorporated either explicitly or implicitly in many other traffic flow models reviewed in [16,18,33,35,37,40,84,92,93,95,98,102,126,127,128].

#### Newell Optimal Velocity (OV) Model and its Variants

At  $m_1 = 0$  the GM-model is associated with the Newell optimal velocity (OV) model [100]

$$v(t + \tau_0) = V(g(t) + d). \quad (31)$$

In this case, the OV-relation  $V(g + d)$  and the associated fundamental diagram  $Q(\rho) = \rho V(\rho)$  are determined from (31) at time-independent space gap  $g$  and speed  $v$ :

$$v = V(g + d). \quad (32)$$

In the Bando et al. OV model [3,4,5]

$$\frac{dv(t)}{dt} = \frac{V(g(t) + d) - v(t)}{\tau_0^{(0)}} \quad (33)$$

it is suggested that

$$V(g + d) = 0.5V_0[\tanh(g - g_0) + \tanh(g + d)], \quad (34)$$

$\tau_0^{(0)}$ ,  $g_0$  are model parameters.

There are many other traffic flow models associated with a generalization of OV-models, which can be written in the form (e. g., [40])

$$\frac{dv(t)}{dt} = \phi(v(t), \Delta v(t), g(t)), \quad (35)$$

where the OV-relation  $V(g)$  and the associated fundamental diagram  $Q(\rho) = \rho V(\rho)$  are determined from (35) at time-independent space gap  $g$  and speeds  $v$ ,  $v_\ell$ , as well as at  $v = v_\ell$ :

$$\phi(V, 0, g) = 0. \quad (36)$$

In the models (35), a driver reaction time is used only implicitly in a traffic flow model. To these model belongs, for example, the Treiber et al. Intelligent Driver Model (IDM) [119] with

$$\phi = c_3 \left[ 1 - \left( \frac{v}{v_0} \right)^4 - \left( \frac{g^*(v, \Delta v)}{g} \right)^2 \right], \quad (37)$$

$$g^*(v, \Delta v) = g_1 + \tau_1 v + \frac{v \Delta v}{2} (a_0 b_0)^{\frac{1}{2}}, \quad (38)$$

where  $c_3, g_1, v_0, \tau_1, a_0$  and  $b_0$  are model parameters.

These models exhibit qualitatively the same features of free flow instability and resulting wide moving jam formation as those firstly found in [63] from an analysis of a version of Payne's macroscopic traffic flow model.

### Payne-Macroscopic Traffic Flow Model and its Variants

The Payne-model reads as follows [103,104]:

$$\frac{\partial \rho(x, t)}{\partial t} + \frac{\partial (\rho(x, t) v(x, t))}{\partial x} = 0, \quad (39)$$

$$\frac{\partial v(x, t)}{\partial t} + v \frac{\partial v(x, t)}{\partial x} = \frac{V_0(\rho) - v(x, t)}{\tau^{(0)}} - \frac{c_0^2}{\rho} \frac{\partial \rho(x, t)}{\partial x}, \quad (40)$$

where the speed-density relationship  $V_0(\rho)$  determines the fundamental diagram  $Q(\rho) = \rho V_0(\rho)$ ;  $\tau^{(0)}, c_0$  are model parameters. To avoid solutions with speed discontinuities, Kühne introduced [78] the viscosity term  $\mu \partial^2 v(x, t) / \partial x^2$  in the Eq. (40) of the Payne-model. In [63], the Eq. (40) of the Payne-model has been rewritten as the Navier–Stokes equation:

$$\begin{aligned} & \frac{\partial v(x, t)}{\partial t} + v \frac{\partial v(x, t)}{\partial x} \\ &= \frac{V_0(\rho) - v(x, t)}{\tau^{(0)}} - \frac{c_0^2}{\rho} \frac{\partial \rho(x, t)}{\partial x} + \frac{\mu}{\rho} \frac{\partial^2 v(x, t)}{\partial x^2} \end{aligned} \quad (41)$$

with the speed-density relationship

$$V_0(\rho) = c_4 \left[ 1 + \left[ \frac{\exp[(\frac{\rho}{\rho_{\max}}) - c_5]}{c_6} \right]^{-1} - c_7 \right], \quad (42)$$

$c_i, i = 4, 5, 6, 7$  are model parameters. There are a variety of other Navier–Stokes-like and gas-kinetic non-local traffic flow models, which, as the Payne-model, consist of the vehicle balance and velocity equations (see [18,102] and Sect. D of [40]). We may call all these macroscopic traffic flow models Payne-like models. This is because all of these macroscopic models show qualitatively the same features of free flow instability and resulting wide moving jam formation in free flow as those firstly found in [63,64] for a version of the Payne-model (39), (41), (42).

In the Aw–Rascle macroscopic model [2], instead of the Payne-Eq. (40), the following velocity equation has been introduced:

$$\frac{\partial v(x, t)}{\partial t} + v \frac{\partial v(x, t)}{\partial x} = \frac{V_0(\rho) - v(x, t)}{\tau^{(0)}} + \rho \frac{dP}{d\rho} \frac{\partial v(x, t)}{\partial x}, \quad (43)$$

where  $P(\rho)$  is a density function. When the same fundamental diagram is chosen in this model as that in the Payne model, then the Aw–Rascle macroscopic model (39), (43) [2] exhibits also qualitatively the same features of free flow instability and resulting congested pattern formation [116] as those in Payne-like models and other GM-class models found earlier in [41,63,64,82].

**Wiedemann Psychophysical Traffic Flow Model** In Wiedemann's psychophysical model [127], a driver changes acceleration (deceleration) with a reaction time  $\tau_0$ , if some thresholds (action points) are crossed where the driver changes his or her behavior. These thresholds are usually presented in the relative-speed-space-gap plane for a pair of preceding and follower vehicles. When the relative speed  $\Delta v$  is a large enough negative value and the associated threshold for vehicle deceleration is crossed, the driver decelerates to have the relative speed  $\Delta v = 0$  at a desired space gap that depends on the vehicle speed. In the Wiedemann-model it is suggested that at small values of  $\Delta v$  a driver is not able to recognize whether he or she is slower and faster than the preceding vehicles. For this reason, when the speed of the preceding vehicle is a time-independent one, after reaching the desired space gap, the driver decelerates with a small deceleration before another threshold at a greater space gap is crossed. At this threshold the driver should recognize that he or she is slower than the preceding vehicle and so accelerates with a small acceleration to the desired space gap, then the driver decelerates with a small deceleration, and so on. Rather than this vehicle motion with small deceleration (acceleration), the dependence of the mentioned desired space gap on speed plays the role of the fundamental diagram for steady states in the GM-class models and together with a value of driver reaction time  $\tau_0$  determines the instability of model solutions such as in the GM-class models (see Sect. "Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence").

**Nagel–Schreckenberg Cellular Automata (CA) Traffic Flow Model** A different mathematical description for driver reaction time within the framework of the GM-class models has been introduced by Nagel and Schreckenberg [97]. In the Nagel–Schreckenberg (NaSch) stochastic



cellular automata (CA) model, in which the time and space are discrete values, the driver reaction time is described through the use of model driver fluctuations. The discrete time is  $t = n\tau$ ,  $n = 0, 1, 2, \dots$ ;  $\tau$  is the time step and the road is divided into cells of a finite length (see a history of CA traffic flow model development in review [88]). In the initial version of the NaSch CA model the cell length has been chosen to be equal the vehicle length  $d$  [97]. The update rules for vehicle motion in the NaSch CA model can be written as follows

$$v_{n+1} = \max(0, \min(v_{\text{free}}, v_n + 1, g_n) - \xi_n), \quad (44)$$

$$x_{n+1} = x_n + v_{n+1}, \quad (45)$$

$$\xi_n = \begin{cases} 1 & \text{for } r \leq p, \\ 0 & \text{otherwise,} \end{cases} \quad (46)$$

$r = \text{rand}(0, 1)$  denotes a random number uniformly distributed between 0 and 1;  $p < 1$ ; time and space are in the units of  $\tau$  and  $d$ , respectively.

To understand the NaSch-model, firstly note that  $g_n$  determines in (44) a safe speed: when  $v_n > g_n$ , then a driver decelerates at time step  $n + 1$  because from (44) it follows that

$$v_{n+1} \leq g_n. \quad (47)$$

This prevents vehicle collisions. Note that the safe speed determined through the space gap  $g$  was first introduced by Pipes [106].

Secondly, the condition

$$v_{n+1} = v_n + 1 \quad (48)$$

means that the vehicle accelerates at time step  $n + 1$ . However, in the NaSch-model due to model fluctuations with probability  $p$  this acceleration does not occur and instead of (48) from (44) it follows that the vehicle maintains its speed:

$$v_{n+1} = v_n. \quad (49)$$

Thus in this case, model fluctuations simulate a time delay in vehicle acceleration; this time delay is equal to

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p}. \quad (50)$$

Thirdly, let us assume that the vehicle should decelerate at time step  $n + 1$  that occurs if  $v_n > g_n$ . Then without taking fluctuations into account from (44) the condition  $v_{n+1} = g_n$  would be satisfied. However, due to model

fluctuations with probability  $p$  we find  $v_{n+1} = g_n - 1$ , i. e., the vehicle decelerates stronger than is needed for safety conditions. This is the main idea of the GM-model for the over-deceleration effect discussed above, which should explain traffic flow instability.

In car-following models like the GM-model and different kinds of OV-models, *each* of the vehicles exhibit the same microscopic time delays, which are determined by deterministic rules of model motion (for the models of identical vehicles that are considered here only). In contrast, in the NaSch-model these driver time delays are simulated through the use of random model fluctuations. As a result, in the NaSch-model these driver time delays are described as “collective effects” that occur *on average* in traffic flow.

One of the advantages of the NaSch CA model (44) is the very fast computer simulation times of traffic flow in large networks. A disadvantage of this model in comparison with deterministic traffic flow models is the very great (non-realistic) fluctuations of vehicle speed. We know now that these large model fluctuations are necessary to simulate driver delay times in vehicle acceleration and deceleration. The problem of large speed fluctuations has been solved in the Krauß-stochastic microscopic model [77], which uses the same ideas for mathematical description of driver delay times in vehicle acceleration and deceleration as those in the NaSch CA model. However, rather than a discrete space of CA models, in the Krauß-model the continuum space co-ordinate is used. The Krauß-model can be written as follows

$$v_{n+1} = \max\left(0, \min\left(v_{\text{free}}, v_n + a\tau, v_n^{(\text{safe})}\right) - \xi_n\right), \quad (51)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (52)$$

$\xi_n = a\tau r$ ,  $a$  is the maximum acceleration, the safe speed  $v_n^{(\text{safe})} = v^{(\text{safe})}(g_n, v_{\ell,n})$  in (51) is a solution of the Gipps-equation [36]

$$v^{(\text{safe})}\tau^{(\text{safe})} + X_d(v^{(\text{safe})}) = g_n + X_d(v_{\ell,n}), \quad (53)$$

where  $X_d(u)$  is the distance traveled by the vehicle with an initial speed  $u$  at a time-independent deceleration  $b$  until it comes to a stop;  $\tau^{(\text{safe})}$  is a safe time gap that can be individual for drivers; due to discrete model time  $t = n\tau$ ,  $n = 0, 1, 2, \dots$  with time step  $\tau = 1$  used in the model (51)

$$X_d(u) = b\left(\alpha\beta + \frac{\alpha(\alpha - 1)}{2}\right), \quad (54)$$

where  $\beta$  and  $\alpha$  are the fractional and integer parts of  $u/b$ , respectively;  $\tau^{(\text{safe})} = \tau$ .

### Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence

What types of congested traffic patterns should occur due to free flow instability in the GM-class models? This question has been answered by Kerner and Konhäuser in 1994 from their numerical study of a version of the Payne macroscopic model (39), (41), (42) [63]. As a result of the instability of free flow at a density that is greater than the critical one  $\rho_{cr}^{(j)}$ , wide moving jams emerge *spontaneously* in free flow ( $F \rightarrow J$  transition for short).

The  $F \rightarrow J$  transition is a first-order phase transition [63]. This means that at smaller densities  $\rho < \rho_{cr}^{(j)}$  associated with metastable states of free flow a wide moving jam can be excited by a speed (or density) disturbance in an initial homogeneous free flow (Fig. 28a). If an initial amplitude of the growing disturbance is small enough, the  $F \rightarrow J$  transition is associated with a “boomerang” behavior of the disturbance (Fig. 28a) [63]: firstly, the disturbance propagates downstream in free flow; then the disturbance comes to a stop strongly growing in its amplitude; as a result, the disturbance begins to propagate upstream; finally, a wide moving jam is forming that propagates upstream, i.e., the jam propagates through the location at which the initial disturbance has occurred.

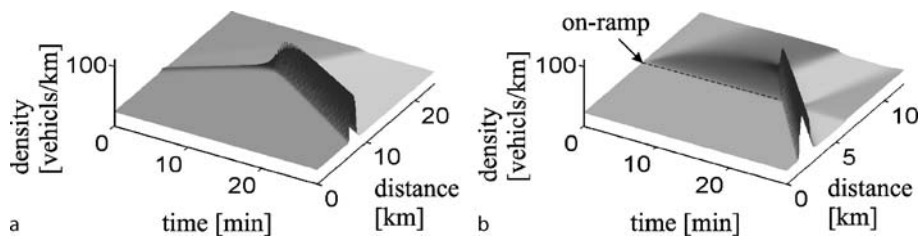
The same boomerang effect associated with the  $F \rightarrow J$  transition has been found, when small amplitude disturbances occur at an on-ramp bottleneck (Fig. 28b) [64]. Indeed, due to a speed disturbance that permanently exists at the bottleneck, a wide moving jam emerges spontaneously at the bottleneck in a metastable free flow [40, 41, 64, 82, 119] (Figs. 28b and 29). At the same flow rate on the main road upstream of the bottleneck and a greater flow rate to the on-ramp a sequence of wide moving jams emerges spontaneously (Fig. 29c).

The conclusion in [63] that instability of free flow, which should explain traffic breakdown, leads to  $F \rightarrow J$  transition is the *general* one for all models of the GM-class, which shows this instability (e.g., [40]). In partic-

ular,  $F \rightarrow J$  transitions at the bottleneck for the NaSch CA model (Fig. 29e, f) and for the Wiedemann-model (Fig. 29g, h) show qualitatively the same features of spontaneous wide moving jam emergence as those in a version of the macroscopic Payne-like model in [63, 64, 82] (Fig. 29b, c).

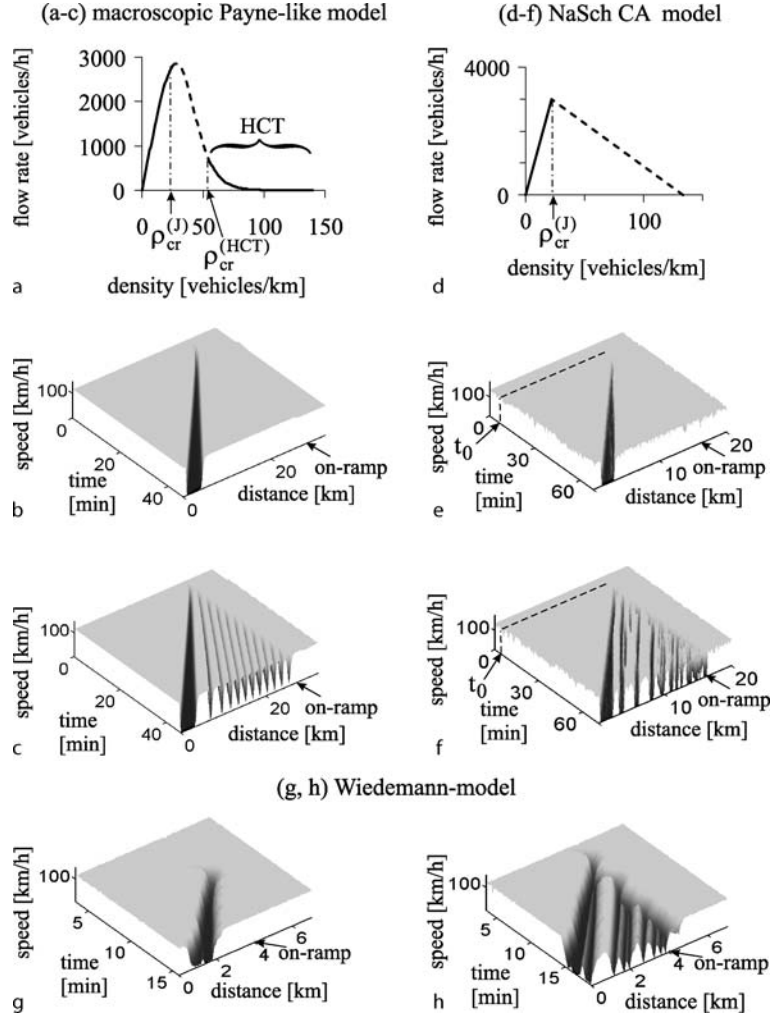
In the macroscopic model (39), (41), (42), the safety gap  $g_{del}$  has been simulated through a choice of the fundamental diagram (42) with a very small absolute value of the derivative  $|dV/d\rho|$  in the neighborhood of the jam density (Fig. 29a) [63]. Then due to this very small value  $|dV/d\rho|$  at the jam density a vehicle could not almost accelerate from the initial speed  $v = 0$  before the space gap to the preceding vehicle increases considerably. Thus only after a relatively long time delay in vehicle acceleration  $\tau_{del}^{(a)}$  associated with the gap  $g_{del}$ , which considerably exceeds the initial space gap within the jam, could the vehicle escape from the jam. Thus this choice of the fundamental diagram shape simulates also implicitly a much longer time delay in vehicle acceleration  $\tau_{del}^{(a)}$  at speeds that are close to zero  $v \approx 0$  within the jam in comparison with considerably shorter time delay in vehicle acceleration at higher speeds. Different time delays in vehicle acceleration  $\tau_{del}^{(a)}$  at  $v \approx 0$  and at higher speeds is known as a slow-to-start rule in traffic flow modeling.

By introducing of the slow-to-start rule in the NaSch-model [6], the initial NaSch-model has been further developed to show qualitatively the same features of wide moving jam propagation as those in the Kerner–Konhäuser theory of wide moving jams and in empirical data [52]. It must be noted that rather than the implicit simulation of the slow-to-start rule through the use of a special form of the fundamental diagram discussed above, another mathematical idea of simulation of slow-to-start rule firstly introduced in [115] has been used in the NaSch-model [6]. Recall that in the NaSch-model the probability of fluctuations  $p$  simulates the time delay in acceleration in accordance with (50). Thus to simulate the slow-to-start rule,



Traffic Congestion, Modeling Approaches to, Figure 28

“Boomerang” behavior of the growing disturbance in metastable free flow by wide moving jam formation. Vehicle density development in space and time. Simulations of the Payne-like model in [63]: a Homogeneous road [63]. b On-ramp bottleneck [64]



**Traffic Congestion, Modeling Approaches to, Figure 29**

Traffic breakdown at on-ramp bottleneck in the GM-model class with free flow instability [40,64,119]: Wide moving jam formation in free flow in a Payne-like model in [63] (a–c), in the NaSch CA model (d–f), and in the Wiedemann-model (g,h) (simulations with VISSIM 4.10-10 based on the Wiedemann-model; simulation parameters: flow rate in free flow on the main single-lane road upstream of the bottleneck  $q_{in} = 2750$ , flow rate to the on-ramp  $q_{on} = 200$  vehicles/h for g and  $q_{in} = 2500$ ,  $q_{on} = 400$  vehicles/h for h, length of on-ramp merging region is 300 m, maximum speed in free flow is within the range 119–120 km/h, passenger vehicles only). a,d The fundamental diagrams of the Payne-like model (a) with parameters of Fig. 28 and the NaSch CA model (d); *dashed parts* of the diagrams are related to unstable states. b,c,e,f–h Average speed in space and time

this probability has been taken considerably greater at  $v = 0$  than at  $v > 0$  [6]:

$$p(v_n) = \begin{cases} p_1 & \text{for } v_n = 0, \\ p_2 & \text{for } v_n > 0, \end{cases} \quad (55)$$

where  $p_1, p_2$  are constants,  $p_1 > p_2$ .

Thus, traffic breakdown is explained by an  $F \rightarrow J$  transition in all models with free flow instability discussed in

this subsection and a huge number of other traffic flow models of the GM-class models [16,40,95,98].

However, this general theoretical result [40,41,63,64, 82,92,95,98,119] is in contradiction with empirical results; rather than the  $F \rightarrow J$  transition, an  $F \rightarrow S$  transition governs traffic breakdown in real traffic flow (Sect. “[Highway Capacity and First-Order  \$F \rightarrow S\$  Transition](#)”). In other words, the traffic flow models with free flow instability of the GM-class [16,18,33,35,37,40,84,92,93,95,98,126,127,

128] cannot be used for a correct description of traffic breakdown.

The GM-class models reviewed in [16,18,33,35,37,40,84,92,93,95,98,102,126,127,128] cannot also be used for an adequate modeling of wide moving jam emergence; indeed, in the GM-class models wide moving jam emergence is explained by an  $F \rightarrow J$  transition. In contrast, in empirical observations wide moving jams occur due to the  $F \rightarrow S \rightarrow J$  transitions (Sect. “Moving Jam Emergence in Synchronized Flow ( $S \rightarrow J$  Transition)”).

**Why Wide Moving Jams do not Emerge Spontaneously in Empirical Free Flow** To explain the critical conclusions about the GM-class models made above, we consider the following hypothesis of three-phase traffic theory [49,50]:

- At any density of free flow at which an  $F \rightarrow J$  transition and an  $F \rightarrow S$  transition are possible, a critical local speed (density) disturbance in free flow needed for the  $F \rightarrow S$  transition is considerably smaller than that needed for the  $F \rightarrow J$  transition.

To understand this hypothesis, we should compare two driver behavioral effects in free flow: the speed adaptation effect, which is responsible for an  $F \rightarrow S$  transition (“Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”), and over-deceleration effect, which is responsible for an  $F \rightarrow J$  transition.

As the speed adaptation effect, the over-deceleration effect occurs also within a local speed disturbance within which the speed is lower and the density is greater than in an initial free flow. In accordance with the over-deceleration effect, a driver approaching the slower moving preceding vehicle decelerates stronger than is required to avoid collisions. As a result, the speed of the driver becomes *lower* than the speed of the preceding vehicle. This can occur with a great probability if an initial space gap is small enough. If all following vehicles move initially also at small enough space gaps, then due to over-deceleration they decelerate stronger than is required to avoid collisions. Then the speed of each next following vehicle decreases continuously up to zero – wide moving jam emerges. Thus, the nucleation of wide moving jam in free flow is possible if each (or many) of the drivers following each other decelerate stronger than the associated preceding vehicles.

In real traffic flow, however, a driver, which moves initially in free flow, while approaching a preceding vehicle moving with a synchronized flow speed, begins to decelerate at a synchronization gap. This gap is great enough. For this reason, the driver will not necessarily decelerate

to a lower speed than the synchronized flow speed of the preceding vehicle; this driver has enough time to compensate the driver’s reaction time and to adjust the speed to that of the preceding vehicle. Thus the necessary condition for an  $F \rightarrow J$  transition – the over-deceleration effect, which can occur only at small enough initial space gaps between vehicles following each other, is much harder to satisfy than the necessary condition for an  $F \rightarrow S$  transition – the speed adaptation effect, which occurs within the synchronization gap. This driver behavior – speed adaptation within the synchronization gap explains the above hypothesis of three-phase traffic theory and empirical observations in which the  $F \rightarrow S$  transition governs traffic breakdown.

Although a spontaneous  $F \rightarrow J$  transition is not observed in real traffic, however, an induced  $F \rightarrow J$  transition is possible. The induced  $F \rightarrow J$  transition can be observed, if free flow on a road in a metastable state with respect to wide moving jam emergence, i. e., the flow rate in the free flow is equal to or greater than the outflow from a wide moving jam  $q_{\text{out}}$  (Sects. “Characteristic Parameters of Wide Moving Jam Propagation and Line  $J$ ” and “Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence”). For example, a moving jam can initially occur within synchronized flow in an off-ramp lane of the road; if later this jam reaches the road, the jam can grow over time while propagating upstream on the road (see an empirical example of such an induced  $F \rightarrow J$  transition in Sect. 12.6 in [52]).

**Critical Discussion of Homogeneous Congested Traffic (HCT)** In the Payne-like macroscopic models (e. g., [63, 82]), the OV-models (e. g., [4,5]), the IDM [40,119] as well as some other traffic flow models (see references in [40]), the density region at the fundamental diagram, within which traffic flow is unstable, is limited at greater densities, i. e., flow states at the fundamental diagram are unstable *only* within the density range (dashed part of the fundamental diagram in Fig. 29a):

$$\rho_{\text{cr}}^{(J)} < \rho < \rho_{\text{cr}}^{(\text{HCT})}. \quad (56)$$

In other words, within the density range

$$\rho_{\text{cr}}^{(\text{HCT})} < \rho < \rho_{\text{max}} \quad (57)$$

homogeneous model states of congested traffic are stable with respect to small amplitude fluctuations. These states, in which the speed and flow rate are *homogeneous in space and time-independent*, have been called in [40,41] as *homogeneous congested traffic* (HCT) (Fig. 30a). The more the density exceeds  $\rho_{\text{cr}}^{(\text{HCT})}$ , the more stable is HCT with

respect to non-homogeneous speed disturbances in these models.

It must be noted that HCT is *not* a general result of traffic flow models of the GM-model class. No HCT appears regardless of the density, for example, in the NaSch CA model or in the Krauß-model: beginning from the critical density  $\rho_{cr}^{(l)}$ , all states of congested traffic at the fundamental diagram in these models are also unstable up to the jam density (Fig. 29d).

In simulations of traffic at an on-ramp bottleneck, HCT occurs at great enough flow rate to the on-ramp [8, 40, 41, 82, 119]. Thus, in simulations of the models satisfying (56) (Fig. 29a) the main features of congested patterns upstream of the on-ramp [8, 40, 41, 82, 95, 119] may be illustrated with the following simplified *theoretical schemes*:

- high flow rate on the main road upstream of the on-ramp *and* low flow rate to the on-ramp  $\rightarrow$  different kind of moving jams emerge spontaneously (Fig. 29b, c, e, f, g, h);
- high enough flow rate to the on-ramp  $\rightarrow$  homogeneous congested traffic (HCT) occurs in which the density is high, the speed is very low and no moving jams and no other non-homogeneous traffic states can emerge spontaneously (Fig. 30a).

In empirical observations [52], however, the following *empirical schemes* are observed:

- Low flow rate to the on-ramp  $\rightarrow$  synchronized flow in which the density is relatively low and the speed relatively high occurs; moving jams should not necessarily emerge;
- High enough flow rate to the on-ramp  $\rightarrow$  moving jams emerge spontaneously in synchronized flow upstream of the on-ramp.

It can be seen that the theoretical schemes [8, 40, 41, 82, 95, 119] and the empirical schemes [52] are inconsistent.

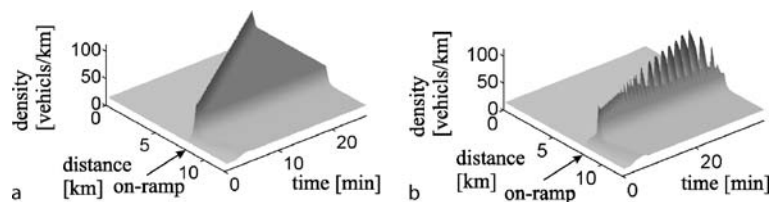
In contrast with this empirical scheme [52], in [111, 119] “empirical” congested traffic states have been published,

which should in particular prove the existence of HCT. Based on the measured data used in [111], let us show that this empirical proof is invalid.

In Fig. 10 in [111], spatiotemporal speed distributions within two congested patterns are shown to be homogeneous during congested pattern existence. The average speed within the associated patterns is very low, because these patterns have been caused by accidents. It must be stressed that these results in [111] have been derived with an adaptive smoothing method of data processing discussed in [111], i. e., with *processed data sets*. In contrast, our Fig. 31 shows *real unprocessed raw measured* data for one of these congested patterns related to Fig. 10a in [111].

To explain real measured data shown in Fig. 31, we should note that already in raw unprocessed data there is a large error in the average speed, when very low speeds are measured; if speeds,  $v$ , of all vehicles that have passed a detector during a 1-min interval are within the range  $0 < v < 20$  km/h, then the road computer sets the average speed to 10 km/h. Only if no vehicle passes a detector during a 1-min interval the speed (and flow rate) is zero. This explains why in the speed data shown in Fig. 31 there are mostly two speed values, zero and 10 km/h. Only when average speeds are higher than 30 km/h, the speed can be used in deciding whether the speed distribution is really a homogeneous one or not.

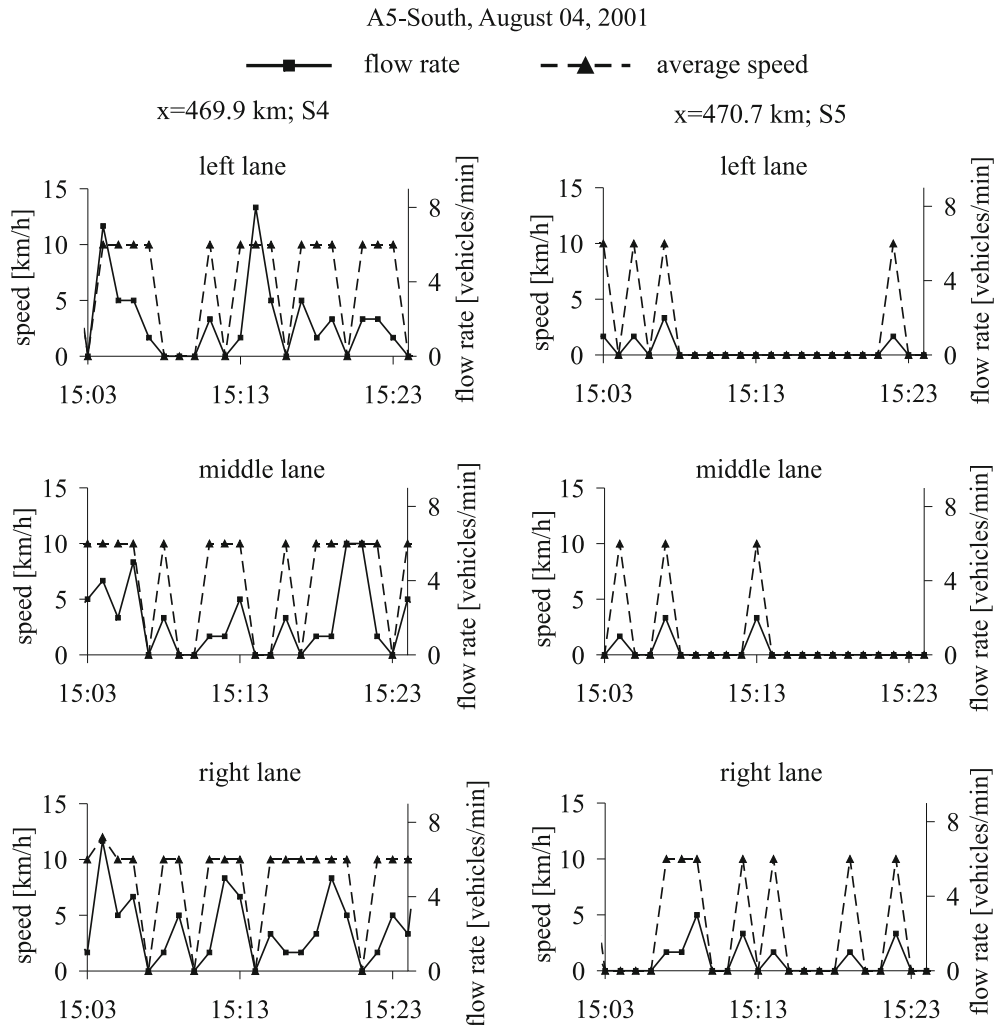
We see that for low speeds based on the data, regardless of a method of further speed data processing, *no* conclusion about features of a spatiotemporal pattern can be made from an analysis of speeds *only*, as made in [111]. Rather than the average speed, the flow rate is measured with a sufficient accuracy at any density. We can see from the flow rate distribution shown in Fig. 31 that there are extremely complex spatiotemporal flow rate changes both in space and time between zero and 8 vehicles/min. This explains that in contrast with the statement in [111], the congested pattern is extremely non-homogeneous in space and time. Note that features of such complex spatiotemporal congested patterns occurring at heavy bottlenecks caused for example by accidents or bad weather condition



Traffic Congestion, Modeling Approaches to, Figure 30

Simulations of HCT (a) and OCT (b) at on-ramp bottleneck [40, 41, 82, 119]





**Traffic Congestion, Modeling Approaches to, Figure 31**

Measured unprocessed flow rate (*squares*) and speed (*triangles*) at two detector locations within a congested pattern that has been presented in Fig. 10a in [111] as “homogeneous congested traffic” (HCT). The locations  $x = 469.9$  km for the detector S4 and  $x = 470.7$  km for the detector S5 are chosen in accordance with the freeway section sketch shown in Fig. 6 in [111]. The raw data is 1-minute averaged data

have recently been found in [55] and appears in ► **Traffic Congestion, Spatiotemporal Features of**.

Thus in reality the congested pattern shown in Fig. 10a in [111] has no relation with HCT. The same critical conclusion can be made about the congested pattern shown in Fig. 10b in [111], which should be another “empirical” example of HCT.

Due to existence of HCT model solutions, these traffic flow models exhibit also model solutions called oscillatory congested traffic (OCT) (Fig. 30b) [8,40,41,82,119]. Indeed, OCT appears in a neighborhood of the critical density  $\rho_{cr}^{(HCT)}$  for instability of HCT: when the density

in HCT decreases and it approaches the critical density  $\rho_{cr}^{(HCT)}$ , then due to instability of HCT, OCT occurs. Thus OCT model solutions result from the existence of HCT model solutions in these models, i. e., as HCT, OCT model solutions have no relation to real traffic flow.

In accordance with criticisms of the GM-class models made above, we should mention that congested pattern diagrams, which describe different types of congested traffic patterns in these models [8,40,41,64,65,82,95,119], have no relation to real traffic. This criticism includes also earlier results of the author et al. [64,65] made in the context of the fundamental diagram approach. The main point of

the criticism of these congested pattern diagrams is that at a great enough flow rate upstream of the bottleneck the onset of congestion in all these congested pattern diagrams is associated with  $F \rightarrow J$  transitions [8,40,41,64,65,82,95,119] (Fig. 29); in contrast with this model result, in real traffic the onset of congestion in all observations is governed by an  $F \rightarrow S$  transition (Sect. “Spontaneous Traffic Breakdown ( $F \rightarrow S$  Transition)”). In addition, at great flow rates to the on-ramp, OCT and HCT model solutions appear in these diagrams, which, as explained above, have no relation to real traffic flow.

### Models Combining LWR- and GM-Approaches

Considering wide moving jam emergence in the GM-class models, we have above suggested that the critical density  $\rho_{cr}^{(J)}$  for the instability leading to  $F \rightarrow J$  transitions is related to free flow, specifically, the condition

$$\rho_{cr}^{(J)} < \rho_0 \quad (58)$$

is satisfied, where the density  $\rho_0$  is associated with the maximum flow rate at the fundamental diagram.

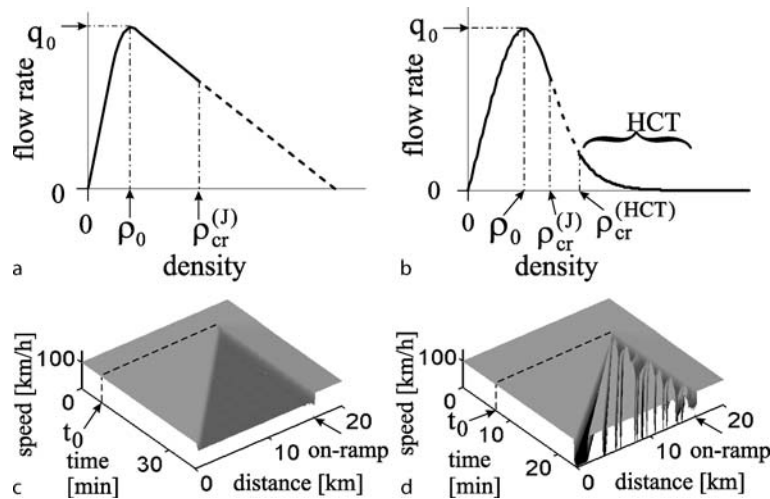
However, parameters of many traffic flow models of the GM-model class can also be chosen in the way that the critical density  $\rho_{cr}^{(J)}$ , at which steady speed states on the fundamental diagram becomes unstable as the density increases, is greater than the density  $\rho_0$  (Fig. 32a,b):

$$\rho_{cr}^{(J)} > \rho_0 \quad (59)$$

As in the LWR-model class, none of the fundamental empirical traffic breakdown features B–D of Sect. “Highway Capacity and First-Order  $F \rightarrow S$  Transition” can be shown in the models. Dependent on the shape of the fundamental diagram and model parameters, there are two types of such models. In the first model type, all steady speed states are unstable when the density satisfies the condition  $\rho > \rho_{cr}^{(J)}$  (Fig. 32a). In the second model type, when the density satisfies the condition  $\rho \geq \rho_{cr}^{(HCT)}$  (Fig. 32b), there are also HCT model solutions criticized in Sect. “Critical Discussion of Homogeneous Congested Traffic (HCT)”. Features of the GM-model class appear in these models when  $q_{on}$  increases and the density on the main road approaches the critical density  $\rho_{cr}^{(J)}$ . Then wide moving jams emerge spontaneously in the dense flow upstream of the bottleneck (Fig. 32d).

Thus by changing of model parameters in some of the models of the GM-class models, a model can exhibit qualitatively different conditions for congested traffic occurrence depending on whether the condition (58) (Fig. 29), or the condition (59) (Fig. 32) is satisfied, or else at any density there cannot be model instability leading to wide moving jam emergence at all, i. e., when the model exhibits qualitatively the same features as those of the LWR-model (Fig. 27).

In particular, when the shape of the fundamental diagram is chosen, then such transitions from condition (58) to condition (59) and, finally, to the LWR-model class occur in the OV model (33) by a gradual continuous decrease



**Traffic Congestion, Modeling Approaches to, Figure 32**

Traffic breakdown and congested patterns at on-ramp bottleneck in the models combining LWR- and GM-approaches under condition (59): a,b Two types of fundamental diagrams; dashed parts show unstable flow states. c Propagation of (shock) waves under condition (26) that is the same as that in Fig. 27d. d Spontaneous wide moving jam emergence in the dense traffic in a when the flow rate to the on-ramp increases and the density downstream of the bottleneck approaches the critical density  $\rho_{cr}^{(J)}$  [62]

in the model parameter  $\tau_0^{(0)}$ , in a version of the Payne-model (39), (41), (42) by a gradual continuous increase in the model parameter  $c_0$ , in the IDM (37), (38) by a gradual continuous increase in the model parameter  $c_3$ .

### Critical Discussion of Empirical Tests of Traffic Flow Models

Obviously, traffic flow theories must be based on real behavior of drivers in traffic, and their solutions should show phenomena observed in traffic flow. For this reason, an empirical test of a traffic flow theory is of great importance. There are several approaches to perform such a test [69]:

- (i) Vehicle speeds, time headways (net time gaps), and accelerations measured in a car-following experiment are compared with that, which a model shows (e. g., [84]).
- (ii) Spatiotemporal evolution of speed (or density) associated with a processing of real empirical data made for example through the use of adaptive smoothing methods are compared with results of numerical simulations of congested patterns [111,119].
- (iii) The empirical fundamental diagram for traffic flow (flow-density relationship) associated with measurements made at a freeway location are compared with a theoretical fundamental diagram for traffic flow used in a model or that results from the model (e. g., [15,31,42,44,71,72,83,99,112,117,129]).
- (iv) Time headway distributions, optimal velocity (OV) functions, and some other single vehicle characteristics that are measured at a freeway location or result from data aggregation are compared with the associated model results (e. g., [71,72,99]).
- (v) Traffic variables (e. g., flow rate and average speed) are measured on at least three different freeway locations. Measurements from two locations, which are the farthest upstream and downstream, are used as upstream and downstream boundary conditions for a model, respectively. The model calculates spatiotemporal distributions of traffic variables between these two locations. The distributions should correspond to empirical data measured at the locations, which have not been used as the boundary conditions for the modeling (e. g., [13,14,18,125]).

In all these cases, model parameters are chosen to have the best agreement with the associated empirical data.

As explained above in this section, past traffic flow theories and models reviewed in [16,33,40,95,98,128] cannot explain and reproduce empirical features of traffic breakdown. Nevertheless, these models can show a good corre-

spondence with empirical data in the above empirical test approaches (i)–(v).

To explain this “phenomenon”, we should note that empirical data used in the tests (i)–(iv) does not contain complete features of spatiotemporal traffic dynamics, which has been considered in part II of [52]. To study these fundamental features (traffic breakdown and resulting spatiotemporal congested patterns), real unprocessed measured data, which should include complete spatiotemporal distribution of synchronized flow within a congested pattern (see Sect. 2.4.11 in [52]), should be studied.

In particular, due to smoothing and/or selection of data in the empirical test (ii) [111,119], important characteristics of congested patterns get lost resulting in invalid analysis and classification of empirical congested traffic states. A characteristic example of such invalid analysis of empirical features of traffic congestion and the associated test of simulation results made in [111,119] has been considered in Sect. “Critical Discussion of Homogeneous Congested Traffic (HCT)” (Fig. 31).

The fundamental diagram, OV functions as classified in [71,72,99], time headway distributions, and other macroscopic and single vehicle (microscopic) characteristics of traffic flow used up to now in empirical tests (iii) and (iv) are associated with an averaging of spatiotemporal traffic pattern characteristics. For this reason, important features of spatiotemporal congested patterns and phase transitions in traffic flow are *lost* in these traffic flow characteristics. Thus in contrast with [71,72,99] it is not justified to use these macroscopic and microscopic characteristics as the *solely* empirical basis for a decision whether a traffic flow model can describe real traffic flow or not.

Moreover, as explained in ► [Traffic Congestion, Spatiotemporal Features of](#), a criterion for the definitions of synchronized flow and wide moving jam in [71,72,99], which is based on a comparison of the flow-density correlations within empirical data associated with congested traffic, is incorrect. As shown in ► [Traffic Congestion, Spatiotemporal Features of](#) and [67], this is associated with a *large systematic error* in calculations of vehicle density made in [71,72,99] from empirical data (for a more detailed consideration see ► [Traffic Congestion, Spatiotemporal Features of](#)). For this reason, empirical tests of the CA traffic flow models, which are related to these traffic phase definitions, made in [71,72,99] are also invalid.

These critical conclusions explain why the NaSch CA models [6,97] and their further developments including a CA model with “comfortable driving” [70] show empirical fundamental diagrams, empirical OV functions and

time headway distributions satisfactory [72], even though these CA models [6,70,97] as found in [66] (see explanations to simulation results of the NaSch CA model with “comfortable driving” [70] presented in Figs. 20–23 in [66]) cannot show and predict the main empirical spatiotemporal features of phase transitions and synchronized flow.

In addition, it should also be noted that in empirical tests of the CA traffic flow models made in [71,72,99], simulations of spatiotemporal congested patterns shown by the CA traffic flow models are made for a spatially homogeneous road, i.e., the road without bottlenecks. In contrast, measured data used for these tests are related to congested traffic occurring due to bottlenecks. Spatiotemporal congested patterns occurring in simulations made on the homogeneous road depend fundamentally on initial conditions and on inflow conditions given in a model at the upstream boundary of the road (e.g., on time interval distributions for vehicles entering the road). However, as shown in [52], the vehicle dynamics at a bottleneck makes the greatest influence on spatiotemporal congested patterns occurring in simulations performed at the bottleneck. This means that for an adequate empirical test of traffic flow models rather than simulations on the homogeneous road [71,72,99], model simulations of spatiotemporal congested patterns should be made at bottlenecks [68,69].

In empirical test (v), firstly congested traffic is measured at a road location. Then measured traffic variables associated with this congested traffic are used at the downstream boundary of a traffic flow model. In other words, at this boundary traffic variables are *given* as time functions associated with congested traffic measured at a freeway location related to this model boundary. As a result, in simulations congested traffic given at the model boundary propagates further upstream. If the characteristic features of wide moving jam propagation can be shown by the model (as explained above, this is the case for many models of the GM-model class [40,95,98]), then a wave of congested traffic can propagate in the model upstream with velocities, which can be chosen close to empirical ones. For this reason, an approximate correspondence between model and some empirical traffic variable functions is possible [13,14,125], although the models [40,95,98] cannot show and cannot reproduce empirical traffic breakdown and many resulting empirical spatiotemporal congested patterns.

This is because the test (v) is inconsistent with the “open character” of non-linear dynamic process, “traffic”. As mentioned, in this test *congested traffic* is often *given* at the farthest downstream boundary of a freeway network

model. This is not the case for real traffic, in which initial congestion occurs *spontaneously within* a real freeway network, mostly at a bottleneck. This bottleneck cannot be the *farthest* downstream boundary of this network. To simulate the open traffic process adequately with real traffic, vehicles should leave *freely* the farthest downstream boundary of a network model. This means that *free flow conditions* should be given at the farthest downstream boundary of the network model.

### Critical Discussion of Highway Capacity Definitions

There are two definitions of highway capacity, i.e., capacity of free flow at a bottleneck associated with the two model classes, the LWR-model and GM-model classes.

In the LWR-model, capacity of free flow at a bottleneck (bottleneck capacity) is equal to the flow rate  $q_{\text{cap}}$  downstream of the bottleneck (15): if the sum of flow rates upstream of the bottleneck exceeds  $q_{\text{cap}}$ , which is the maximum flow rate at the fundamental diagram, then traffic breakdown should occur leading to congested traffic upstream of the bottleneck.

To satisfy probabilistic features of traffic breakdown, in recent definitions of highway capacity is suggested that the capacity  $q_{\text{cap}}$  is a random value characterized by a capacity distribution function (e.g., [10]). However, even if it is assumed that the highway capacity  $q_{\text{cap}}$  (15) is a random value whose distribution function is defined through an empirical breakdown probability [10], the fundamental empirical features of the capacity A–C (Sect. “[Highway Capacity and First-Order F → S Transition](#)”) cannot be explained with the capacity determined by the maximum flow rate  $q_{\text{cap}}$  on the fundamental diagram (Fig. 24). This is because a random choice of  $q_{\text{cap}}$  changes *none* of qualitative features of the onset of congestion in the LWR-theory shown in Fig. 27; as explained in Sect. “[Classic Lighthill–Whitham–Richards \(LWR\) Theory of Onset of Traffic Congestion](#)”, these LWR-theory features are, however, in a deep contradiction with the fundamental empirical features of traffic breakdown (Sect. “[Highway Capacity and First-Order F → S Transition](#)”). Thus even under assumption about a random character of  $q_{\text{cap}}$  in the capacity definition (15), the definition is not associated with empirical results. Therefore, traffic management methods based on this capacity definition cannot be used for a reliable traffic management [53].

In the GM-model class, the maximum bottleneck capacity can be defined to be equal to the flow rate downstream of the bottleneck at which free flow instability occurs in a traffic flow model. However, as shown in Sect. “[Classic General Motors \(GM\) Model Approach](#)”

**Free Flow Instability due to Driver Reaction Time**” in this model class the instability leads to an  $F \rightarrow J$  transition. As mentioned above, this contradicts with the empirical fact that traffic breakdown is governed by an  $F \rightarrow S$  transition. For this reason, the definition of free flow capacity made within the framework of the GM-model class does not satisfy the empirical feature of the bottleneck capacity **A** (Sect. “**Highway Capacity and First-Order  $F \rightarrow S$  Transition**”).

The capacity definition made in three-phase traffic theory that satisfies all fundamental empirical features of the bottleneck capacity **A–D** has been considered in Sect. “**Highway Capacity and First-Order  $F \rightarrow S$  Transition**”.

### Common Features of Earlier Traffic Flow Modeling Approaches

The above consideration of earlier traffic flow models in the framework of the fundamental diagram hypothesis shows that stability features of steady state model solutions determine most of the qualitative features of the onset of congestion and resulting congested patterns. For this reason, it turns out that mathematically very different traffic flow models, which exhibit qualitatively the same stability features of steady speed states at the fundamental diagram, show also qualitatively the same features of traffic congestion. This leads to the traffic flow model classification made above and to the following conclusions about common features of earlier traffic flow modeling approaches:

- (i) Most of the traffic flow models used by traffic researches are based on fundamental diagram hypothesis. These models can be classified into two main classes. The first model class refers to the classic LWR-theory. The basic idea of the LWR-theory is that maximum flow rate associated with the maximum point at the fundamental diagram determines free flow capacity at a bottleneck. Thus if the flow rate upstream of the bottleneck exceeds the capacity, then traffic breakdown should occur. The second model class refers to the basic idea of the classic GM-model. The basic idea of the GM-model approach is that beginning at a critical density there is instability of free flow. This free flow model instability associated with a finite value of a driver reaction time should explain traffic breakdown.
- (a) The LWR-theory and all traffic flow models based on this theory cannot show and cannot predict the fundamental empirical features of traffic breakdown (Sect. “**Highway Capacity and**

**First-Order  $F \rightarrow S$  Transition**”) as well as empirical spontaneous moving jam emergence.

- (b) Free flow instability in models and theories based on the GM-model approach leads to an  $F \rightarrow J$  transition. In contrast, in empirical observations, traffic breakdown is associated with an  $F \rightarrow S$  transition. Thus, the models with free flow instability associated with a finite value of a driver reaction time cannot explain and cannot predict traffic breakdown observed in real traffic flow.
- (ii) These traffic flow models are basic traffic flow models for simulations of freeway control and management strategies. However, we have to conclude that the related simulations of the control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a management strategy.
- (iii) It must be stressed that the above criticism of all these earlier traffic flow modeling approaches does not diminish the following **achievements** of these approaches and associated models, which are also used in three-phase traffic theory and the related three-phase traffic flow models (see next Sect. “**Three-Phase Traffic Flow Models**”):
  - (a) There are characteristic parameters and features of wide moving jam propagation [63]. One of these characteristic jam features is that the downstream front of wide moving jams propagates steadily along a road.
  - (b) Classic ideas of traffic flow theories introduced and developed within the fundamental diagram approach about different driver time delays, various mathematical descriptions of driver acceleration and deceleration as well as safety conditions are also very important elements used in three-phase traffic flow models, which overcome the drawbacks of the earlier modeling approaches to traffic congestion (item (i) and (ii)). In particular, pertinent pioneering ideas, which were introduced in earlier models and theories by Herman, Montroll, Potts, Rothery, Gazis [34, 43], Komentani and Sasaki [73], Newell [100], Gipps [36], Payne [103,104], Nagel, Schreckenberg, Schadschneider, and co-workers [6,97], Bando, Sugiyama, and colleagues [4], Takayasu and Takayasu [115], Krauß et al. [77], Nagatani and Nakanishi [96] and by many other groups (see references in [16,18,21,33,35,37,40,84,88,92,93,95,98,102,126,127,128]), are also very important elements of three-phase traffic models discussed below.



### Three-Phase Traffic Flow Models

The first microscopic three-phase traffic theory that can reproduce empirical traffic breakdown and  $F \rightarrow S \rightarrow J$  transitions (Sect. “Characteristic Parameters of Wide Moving Jam Propagation and Line  $J$ ”) as well as all known resulting empirical spatiotemporal congested patterns was developed in 2002 by Kerner and Klenov [57] (see below and for more detail Sect. 16.3 in [52]). They proposed a microscopic spatial continuum and discrete-time traffic flow model in which general rules of vehicle motion are in accordance with the fundamental hypothesis (Fig. 15) as well as with other hypotheses of three-phase traffic theory. Based on this Kerner–Klenov model, numerical spatiotemporal features of traffic breakdown and the pinch effect have firstly been found and studied. Some months later, Kerner, Klenov, and Wolf developed a CA approach to three-phase traffic theory (the KKW CA models) [66], which can also explain traffic breakdown and the pinch effect as observed in real traffic (see Sect. 16.2 and Chaps. 17–19 in [52]).

Later other various traffic flow models in the framework of the three-phase traffic theory have been developed: Davis [22] as well as Kerner and Klenov [62] have proposed different three-phase microscopic deterministic models; Jiang and Wu [45], Lee et al. [81], and Gao et al. [32] have developed various three-phase CA models; Siebel and Mauser [113] and Laval [80] have suggested macroscopic models; Jiang et al. [48] have suggested a macroscopic version of the Kerner–Klenov speed adaptation model in [62]. Features of phase transitions and congested patterns that these models exhibit are similar to those found in 2002 in the Kerner–Klenov stochastic and KKW CA models. Recent simulation results in the framework of the author’s three-phase traffic theory can be found in [23,24,25,46,47,85,107,124].

It should be noted that the use of only some of the hypotheses of three-phase traffic theory in a traffic flow model does not ensure that the model can show empirical phase transitions in vehicular traffic. An example is Colombo’s continuum model [17] that consists of Eq. (16) in which dependence  $Q(\rho)$  is associated with the fundamental hypothesis of three-phase traffic theory (Fig. 19) [49,50,51]. However, solutions of Colombo’s model [17] can show *none* of the empirical features of  $F \rightarrow S$  and  $S \rightarrow J$  transitions because in this model other hypotheses of three-phase traffic theory needed for the description of phase transitions in real traffic flow have not been taken into account.

For an example of three-phase traffic flow model development, we consider basic assumptions and formulation

of the Kerner–Klenov stochastic three-phase traffic flow model for identical drivers and vehicles [57,58]. This are two reasons for this: (i) the model is the first three-phase traffic flow model and (ii) most of the simulation results discussed in Sects. “Free and Congested Traffic”–“Empirical Double Z-Characteristic for Phase Transitions in Traffic Flow” are based on this model.

#### Kerner–Klenov Stochastic Three-Phase Traffic Flow Model

Basic driver behavioral assumptions of the Kerner–Klenov model are as follows:

- (i) *Fundamental hypothesis of three-phase traffic theory* (Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”). In synchronized flow, a driver accepts a range of different hypothetical steady state speeds at the same space gap to the preceding vehicle. This means that hypothetical steady model states of synchronized flow cover a 2D-region in the flow-density plane (Fig. 33a). The boundaries of this 2D-region  $F$ ,  $L$ , and  $U$  are respectively associated with free flow, a synchronization space gap, and a safe space gap determined through a safe speed. The 2D-region of steady states is associated with a driver behavioral assumption that in synchronized flow a driver is able to recognize whether the space gap is increasing or decreasing regardless of the speed difference to the preceding vehicle.
- (ii) *Line  $J$  and 2D-region of steady states* (Sect. “Explanations of  $S \rightarrow J$  Transitions Through Three-Phase Traffic Theory”). In the model, the line  $J$  is between the boundaries  $L$ , and  $U$  (Fig. 33b), i. e., the line  $J$  divides the 2D-region of steady states of synchronized flow onto two classes: the states on and above the line  $J$  and the states below the line  $J$ , which are metastable and stable states with respect to wide moving jam formation, respectively. Thus at a given steady speed, a driver behavioral assumption is that the space gap in synchronized flow associated with the line  $J$ , i. e., in the jam outflows is greater than the safe one and it is smaller than the synchronization gap.
- (iii) *Speed adaptation effect in synchronized flow* (Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”). The speed adaptation effect takes place when the vehicle cannot pass the preceding vehicle, within the space gap range:

$$g_{s,n} \leq g_n \leq G_n, \quad (60)$$

where  $g_n = x_{\ell,n} - x_n - d$  is the space gap,  $x_n$  is the vehicle co-ordinate, the lower index  $\ell$  marks functions and values related to the preceding vehicle;  $g_{s,n}$  is the safe gap determined from the equation  $v_n = v_{s,n}$ , in which  $v_n$  is the vehicle speed,  $v_{s,n}$  is a safe speed;  $G_n$  is a synchronization gap; all vehicles have the same length  $d$ , which includes the minimum space gap between vehicles within a wide moving jam; index  $n$  corresponds to the discrete time  $t = n\tau$ ,  $n = 0, 1, 2, \dots$ ;  $\tau$  is time step. Under condition (60), the vehicle tends to adjust its speed to the preceding vehicle without caring, what the precise space gap is, as long as it is safe. For example, at a given time-independent speed of the preceding vehicle  $v_{\ell,n} = v_\ell = \text{const}$ , this speed adaptation leads to car following with  $v_n = v = v_\ell$  at a time-independent space gap  $g_n = g$ . There is an infinite number of these gaps associated with the same speed  $v = v_\ell$ . These gaps lie between the synchronization gap and safe gap, i. e., there is no desired (or optimal) space gap in synchronized flow.

- (iv) *Over-acceleration effect* (Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”). In synchronized flow of a lower density, a driver searches for the opportunity to accelerate and to pass. A competition between the speed adaptation (item (iii)) and over-acceleration effects simulates traffic breakdown ( $F \rightarrow S$  transition). The over-acceleration is simulated as a collective effect, which occurs on average in traffic flow, through the use of lane changing to a faster lane as well as through the use of random vehicle acceleration

$$\xi_a = a\tau\Theta(p_a - r), \quad (61)$$

where  $p_a$  is probability of random acceleration,  $a$  is the maximum acceleration,  $r = \text{rand}(0,1)$ ,  $\Theta(z) = 0$  at  $z < 0$  and  $\Theta(z) = 1$  at  $z \geq 0$ . The model fluctuations (61) are applied only if the vehicle should accelerate without model fluctuations.

- (v) *Pinch effect in synchronized flow* (Sect. “Pinch Effect”). Moving in synchronized flow, a driver comes on average closer to the preceding vehicle over time that should explain the pinch effect. A driver time delay in deceleration simulates this effect through model fluctuations in deceleration

$$b_n = a\Theta(P_1 - r_1), \quad (62)$$

applied under condition (60) only;  $P_1$  is probability of random deceleration,  $r_1 = \text{rand}(0,1)$ .

- (vi) *Over-deceleration effect*. As in the NaSch CA model, a well-known over-deceleration effect (human over-reaction) associated with driver reaction time [34,43] is also simulated as a collective effect through the use of random fluctuations in vehicle deceleration

$$\xi_b = a\tau\Theta(p_b - r), \quad (63)$$

which is applied only if the vehicle should decelerate without model fluctuations;  $p_b$  is the probability of random deceleration. In the Kerner–Klenov model, a competition between the over-deceleration and the speed adaptation effect (item iv) determines moving jam emergence in synchronized flow.

- (vii) *Driver time delay in acceleration*. In the model, this well-known effect should describe driver delay in acceleration at the downstream front of synchronized flow or wide moving jam (in the latter case, the driver delay in acceleration is known as a slow-to-start rule [6,115] (Sect. “Traffic Breakdown Resulting from Free Flow Instability in GM-class Models: Wide Moving Jam Emergence”)) after the preceding vehicle has begun to accelerate. A driver time delay in acceleration is simulated as a collective effect through the use of a random value of vehicle acceleration

$$a_n = a\Theta(P_0 - r_1) \quad (64)$$

applied under condition (60) and only then if the vehicle did not accelerate at the former time step; in the latter case  $P_0 = p_0 > 0$ , i. e., a vehicle accelerates with some probability  $p_0$  that depends on the speed  $v_n$ ; otherwise  $P_0 = 0$ . As in the NaSch CA model, the mean time in vehicle acceleration is

$$\tau_{\text{del}}^{(a)} = \frac{\tau}{1 - p_0}. \quad (65)$$

Thus, the basis of the Kerner–Klenov stochastic three-phase traffic flow model are driver behavioral assumptions made in three-phase traffic theory (items (i)–(v)). In addition, over-deceleration (item vi) and driver time delay in acceleration (item vii) introduced in earlier traffic flow models in the framework of the fundamental diagram approach have also been incorporated; in particular, as in the NaSch-model (44), these driver time delays appear through the use of model fluctuations. Finally, the Kerner–Klenov stochastic three-phase traffic flow model reads as follows:

$$v_{n+1} = \max(0, \min(v_{\text{free}}, \tilde{v}_{n+1} + \xi_n, v_n + a\tau, v_{s,n})), \quad (66)$$

$$x_{n+1} = x_n + v_{n+1}\tau, \quad (67)$$

$$\tilde{v}_{n+1} = \max(0, \min(v_{\text{free}}, v_{c,n}, v_{s,n})), \quad (68)$$

$$v_{c,n} = \begin{cases} v_n + \Delta_n & \text{at } g_n \leq G_n \\ v_n + a_n\tau & \text{at } g_n > G_n, \end{cases} \quad (69)$$

where

$$\Delta_n = \max(-b_n\tau, \min(a_n\tau, v_{\ell,n} - v_n)), \quad (70)$$

$v_{\text{free}}$  is the maximum speed in free flow that is constant.

The synchronization gap  $G_n$  depends on the vehicle speed  $v_n$  and on the speed of the preceding vehicle  $v_{\ell,n}$

$$G_n = G(v_n, v_{\ell,n}), \quad (71)$$

where the function  $G(u, w)$  is chosen as

$$G(u, w) = \max(0, k\tau u + \phi_0 a^{-1}u(u - w)), \quad (72)$$

$k > 1$  and  $\phi_0$  are constants. If  $v_n = v_{\ell,n}$ , the synchronization gap  $G_n$  is  $k v_n \tau$ ; this corresponds to a fixed time gap  $k\tau$  that determines the line  $L$  in Fig. 33. If  $v_n > v_{\ell,n}$ , the gap  $G_n$  increases and vice versa.

As explained in items (iv) and (vi) above, random deceleration and acceleration  $\xi_n$  in (66) are applied depending on whether the vehicle decelerates or accelerates, or else maintains its speed:

$$\xi_n = \begin{cases} -\xi_b & \text{if } S_{n+1} = -1 \\ \xi_a & \text{if } S_{n+1} = 1 \\ 0 & \text{if } S_{n+1} = 0, \end{cases} \quad (73)$$

where  $S$  in (73) denotes the state of motion ( $S_{n+1} = -1$  represents deceleration,  $S_{n+1} = 1$  acceleration, and  $S_{n+1} = 0$  motion at nearly constant speed)

$$S_{n+1} = \begin{cases} -1 & \text{if } \tilde{v}_{n+1} < v_n - \delta \\ 1 & \text{if } \tilde{v}_{n+1} > v_n + \delta \\ 0 & \text{otherwise,} \end{cases} \quad (74)$$

where  $\delta$  is constant ( $\delta \ll a\tau$ ).

In (64) and (62), the probabilities  $P_0$  and  $P_1$  are

$$P_0 = \begin{cases} p_0(v_n) & \text{if } S_n \neq 1 \\ 1 & \text{if } S_n = 1, \end{cases} \quad (75)$$

$$P_1 = \begin{cases} p_1 & \text{if } S_n \neq -1 \\ p_2(v_n) & \text{if } S_n = -1, \end{cases} \quad (76)$$

where speed functions for probabilities  $p_0(v_n)$  and  $p_2(v_n)$  are considered in [52];  $p_1$  is constant.

In the model, the safe speed  $v_{s,n}$  in (68) is chosen in the form

$$v_{s,n} = \min\left(v_n^{(\text{safe})}, \frac{g_n}{\tau} + v_{\ell}^{(a)}\right), \quad (77)$$

where  $v_n^{(\text{safe})} = v^{(\text{safe})}(g_n, v_{\ell,n})$  is taken from (51);  $v_{\ell}^{(a)}$  is an “anticipation” speed of the preceding vehicle at the next time step:

$$v_{\ell}^{(a)} = \max\left(0, \min\left(v_{\ell,n}^{(\text{safe})} - a\tau, v_{\ell,n} - a\tau, \frac{g_{\ell,n}}{\tau}\right)\right), \quad (78)$$

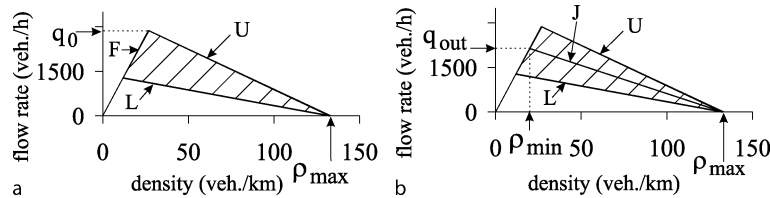
where  $v_{\ell,n}^{(\text{safe})}$  is the safe speed for the preceding vehicle,  $g_{\ell,n}$  is the space gap in front of the preceding vehicle.

Models for lane changing for a multi-lane freeway and of freeway bottlenecks associated with the Kerner–Klenov model as well as a further development of this model for heterogeneous flow consisting of various vehicles and drivers with different characteristics can be found in the book [52].

### Methodology of Empirical Test

For an adequate comparison of empirical and theoretical dynamic nonlinear features of spatiotemporal congested patterns in freeway traffic, the following methodology of the empirical test is used [52].

As in real traffic, in a model of a freeway with on-ramp and off-ramp bottlenecks traffic occurs at the upstream road boundary of the main road and of an on-ramp(s). In



Traffic Congestion, Modeling Approaches to, Figure 33

Steady speed states for the Kerner–Klenov stochastic three-phase traffic flow in the flow-density plane (a) and the line  $J$  (b). Taken from [52]

this open traffic process, all traffic variables downstream result from traffic demand at the upstream road boundary and the on-ramp(s), as well as drivers' destinations (whether a vehicle leaves the main road to the off-ramp or it further follows the main road). At downstream model boundary conditions for vehicle freely leaving a modeling freeway section(s) are given. Spatiotemporal congested patterns emerge, develop, and dissolve in this open freeway model with the same types of bottlenecks as those in empirical observations.

The spatiotemporal dynamics of traffic breakdown and resulting congested patterns found in simulations should be compared with the associated spatiotemporal dynamics found in empirical observations. Only after the fundamental spatiotemporal features of traffic have been simulated and compared with the empirical data, *secondary characteristics* associated with these spatiotemporal patterns, which include for example fundamental diagrams, time headway distributions, speed and flow-density correlation functions and OV functions, can be compared with the associated empirical characteristics.

Such an empirical test made in [67,68,69] shows that three-phase traffic flow models can reproduce and predict all known microscopic and macroscopic empirical features of traffic breakdown and resulting congested traffic patterns.

### Link Between Three-Phase Traffic Theory and Fundamental Diagram Approach

A link between three-phase traffic theory and the fundamental diagram approach to traffic flow modeling can be created through the use of the *averaging* of an infinite number of steady states of synchronized flow associated with a 2D region in the flow-density plane (Fig. 33) to one synchronized flow speed for each vehicle density. As a result, the averaged steady states of synchronized flow are related to 1D region, i. e., to a curve in the flow-density plane (curves *S* in Fig. 34). In this case, we should find a vehicle dynamics for a traffic flow model whose steady states are associated with such a fundamental diagram, however, the model is able to show and predict the existence of the free flow, synchronized flow, and wide moving jam phases, the  $F \rightarrow S \rightarrow J$  transitions between them including a double Z-characteristic for the phase transitions (Sects. "Traffic Breakdown and Highway Capacity" and "Moving Jam Emergence in Synchronized Flow ( $S \rightarrow J$  Transition)").

Recall that three-phase traffic theory is a qualitative theory. For this reason, the fundamental hypothesis of this theory is associated with features of hypothetical steady

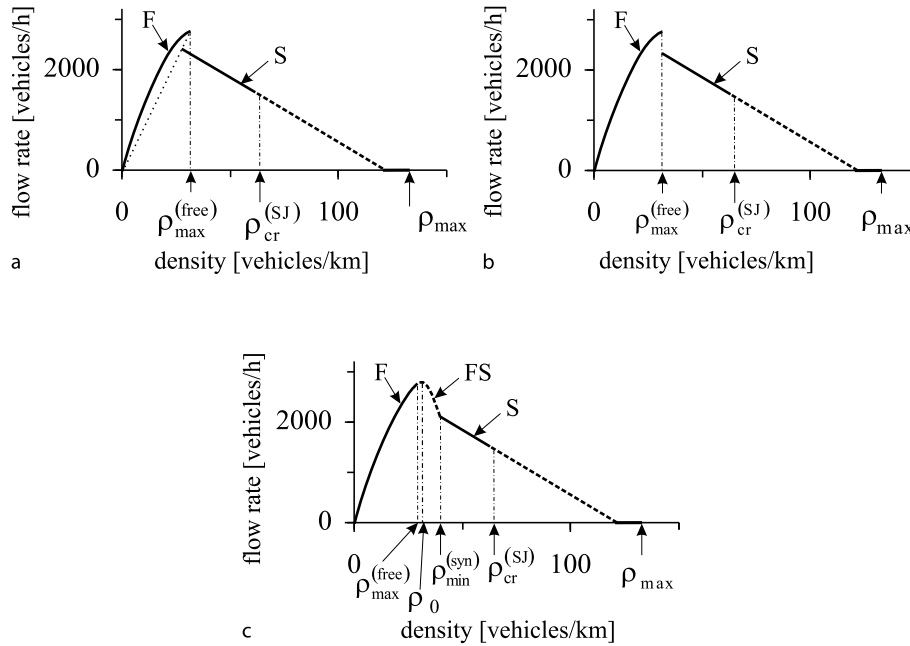
states of synchronized flow that should cover a 2D region in the flow-density plane (Fig. 15a). Rather than a hypothesis about model steady states of synchronized flow, the fundamental hypothesis of a mathematical three-phase traffic flow model can be a hypothesis about some specific dynamic model features (for more detail, see footnote 4 of Sect. 4.3.4 in [52]).

Although such three-phase traffic flow models are possible to develop, however, through the replacing of a 2D-region for steady states of synchronized flow by a 1D-region, i. e., by a curve in the flow-density plane, these three-phase traffic flow models lose a possibility of description of very important features of synchronized flow found in empirical observations. This emphasizes the sense and importance of 2D steady states of synchronized flow for the development of a three-phase traffic flow model. We illustrate these general conclusions through a brief discussion of speed adaptation (SA) three-phase traffic flow models [62].

In contrast with the Kerner–Klenov model discussed in Sect. "Kerner–Klenov Stochastic Three-Phase Traffic Flow Model", which incorporates the fundamental hypothesis of three-phase traffic theory (Fig. 33a), in the SA models hypothetical steady states of synchronized flow are associated with a curve (curve *S* in Fig. 34), i. e., they cover a 1D region in the flow-density plane. As mentioned, the basis hypothesis of the SA models is associated with the hypothesis of three-phase traffic theory about a double Z-characteristic for the sequence of  $F \rightarrow S \rightarrow J$  transitions. Based on a mathematical implementation of this hypothesis onto the SA models, the models can show and explain both traffic breakdown ( $F \rightarrow S$  transition) and moving jam emergence within synchronized flow (pinch effect and associated  $S \rightarrow J$  transitions) [62] as found in empirical observations [50,52].

An  $F \rightarrow S$  transition, i. e., traffic breakdown is simulated in the SA models through the use of one of the following features of steady speed states on the fundamental diagram:

- (i) A *discontinuity* of steady speed solutions between free flow and synchronized flow states (Fig. 34a, b). This discontinuity determines the maximum (limit) point of free flow ( $v_{\min}^{(\text{free})}, \rho_{\max}^{(\text{free})}$ ) on the fundamental diagram, i. e., at the densities  $\rho \geq \rho_{\max}^{(\text{free})}$  an  $F \rightarrow S$  transition should occur.
- (ii) An *instability* of steady speed solutions within a density range ( $\rho_{\max}^{(\text{free})}, \rho_{\min}^{(\text{syn})}$ ) (curve *FS* in Fig. 34c). This instability determines also the maximum (limit) point of free flow ( $v_{\min}^{(\text{free})}, \rho_{\max}^{(\text{free})}$ ) on the fundamental diagram.



**Traffic Congestion, Modeling Approaches to, Figure 34**

Steady states for three different SA three-phase traffic flow models with discontinuity (a,b) and instability (c) of steady speed solutions between free flow (curves F) and synchronized flow states (curves S). Taken from [62]

gram, i.e., at the densities  $\rho \geq \rho_{\max}^{(\text{free})}$  an  $F \rightarrow S$  transition should occur.

A first-order  $F \rightarrow S$  transition is simulated through the use of the above mentioned discontinuity or instability of steady speed solutions on the fundamental diagram (item (i) and (ii)) as well as a feature of the vehicle dynamics incorporated into the SA models that a vehicle tries to adjust the speed to the speed of the preceding vehicle while approaching synchronized flow states; this simulates the speed adaptation effect in synchronized flow (Sect. “Traffic Breakdown Explanation Through Fundamental Hypothesis of Three-Phase Traffic Theory”).

The pinch effect and  $S \rightarrow J$  transitions are simulated in the SA models through an instability of those synchronized flow model steady states whose speeds are lower than some critical speed  $v_{\text{cr}}^{(\text{SJ})}$ , i.e., when synchronized flow density is greater than some critical density (denoted by  $\rho_{\text{cr}}^{(\text{SJ})}$  in Fig. 34): at each density  $\rho > \rho_{\text{cr}}^{(\text{SJ})}$  the synchronized flow state is unstable with respect to infinitesimal disturbances (dashed parts of curves S in Fig. 34).

It must be noted that rather than the fundamental diagram choice is associated with some empirical fundamental diagrams, in the SA models, to simulate the  $F \rightarrow S$  and  $S \rightarrow J$  transitions, which occur at *different* freeway lo-

cations, we chose some artificial features of steady states on fundamental diagrams (Fig. 34). Thus rather than validate these fundamental diagrams with empirical data measured at a single freeway location, the characteristic variables on these diagrams like  $\rho_{\max}^{(\text{free})}$  and  $\rho_{\text{cr}}^{(\text{SJ})}$  should be validated based on spatiotemporal empirical features of the  $F \rightarrow S \rightarrow J$  transitions.

The SA models allow us to overcome drawbacks of earlier traffic flow models in the framework of the fundamental diagram approach in description of traffic breakdown and the pinch effect discussed in Sect. “Critical Discussion of Fundamental Diagram Modeling Approach to Traffic Congestion”. This is because of the formulation of such a vehicle dynamics in the SA-models that leads to two *separated regions for phase transitions* and associated *two different critical densities*  $\rho_{\max}^{(\text{free})}$  for traffic breakdown and  $\rho_{\text{cr}}^{(\text{SJ})}$  for  $S \rightarrow J$  transitions on the fundamental diagram (Fig. 34). For these reasons, the SA models are able to show and predict the sequence of  $F \rightarrow S \rightarrow J$  transitions and the associated double Z-characteristic.

On the other hand, due to the averaging of an infinite number of steady states of synchronized flow to one synchronized flow speed, the SA models are not able to show important features of synchronized flow as well as many features of coexistence of the free flow, synchronized flow,



and wide moving jam traffic phases found in empirical observations [62]. These drawbacks of the SA models are associated with the ignoring of the fundamental hypothesis of three-phase traffic theory.

One of these drawbacks of the SA models is illustrated in Fig. 35. In accordance with empirical results [52], three-phase traffic flow models with 2D-region of steady states for synchronized flow can show and predict *both* free flow *and* synchronized flow between wide moving jams (Fig. 35c, d). In contrast, for the SA models, whose fundamental diagrams are shown in Fig. 34, *only* free flow can occur between wide moving jams, when the jams are moving in large enough distances each other (Fig. 35a, b).

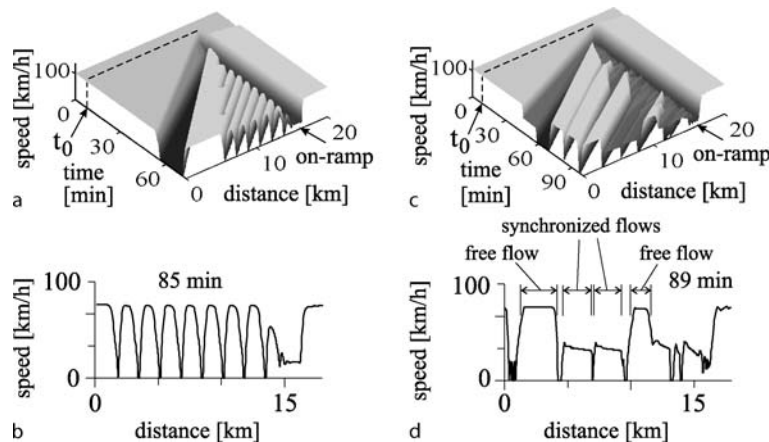
To explain this, note that if a three-phase model exhibits a 2D region of steady states for synchronized flow (Fig. 36a), in the wide moving jam outflow either a state of free flow with the density  $\rho_{\min}$  or any state of synchronized flow from infinite synchronized flow states that lie on the line  $J$  is possible. In contrast, there is only an unstable steady state of synchronized flow that lies on the line  $J$  in the case of the SA-models (Fig. 36b). For this rea-

son, when distances between wide moving jams are great enough, free flow can be formed in the jam outflows only (Fig. 35a,b).

We can conclude that the fundamental hypothesis of three-phase traffic theory about a 2D-region for steady states of synchronized flow permits infinite synchronized flow states on, above, and below the line  $J$ . This allows us to simulate different synchronized flow states in a diverse variety of combinations with wide moving jams and free flows, as observed in empirical data. This is not possible to do based on the SA-models with a fundamental diagram for steady states of synchronized flow.

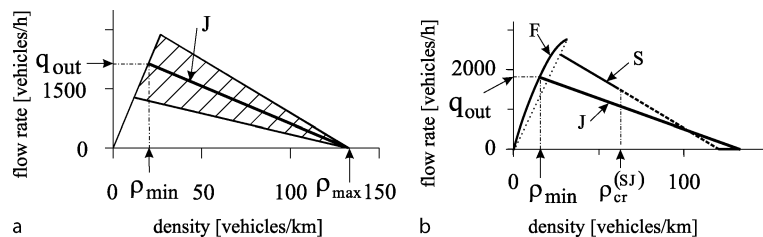
## Conclusions

1. Freeway traffic flow can be free or congested. There are two phases in congested traffic, synchronized flow and wide moving jams. Thus there are three traffic phases: 1. Free flow. 2. Synchronized flow. 3. Wide moving jam. The synchronized flow and wide moving jam traffic phases are defined through spatiotemporal empiri-



**Traffic Congestion, Modeling Approaches to, Figure 35**

Explanation of drawbacks of SA models: a,b In the SA models (Fig. 34), only free flow is formed between wide moving jams. c,d In a three-phase traffic flow model with 2D-region of steady states for synchronized flow both free and synchronized flows are possible between wide moving jams. Taken from [62]



**Traffic Congestion, Modeling Approaches to, Figure 36**

Steady states for synchronized flow and line  $J$  for two classes of three-phase traffic flow models

cal criteria associated with propagation characteristics of the downstream fronts of these phases.

2. Empirical traffic breakdown at a freeway bottleneck, i. e., the onset of congestion in an initial free flow at the bottleneck is associated with a phase transition from free flow to synchronized flow ( $F \rightarrow S$  transition). An  $F \rightarrow S$  transition at the bottleneck can be spontaneous or induced.
3. Wide moving jams can emerge in synchronized flow ( $S \rightarrow J$  transition). Thus wide moving jams emerge spontaneously due to a sequence of the  $F \rightarrow S \rightarrow J$  transitions associated with a double Z-characteristic for these phase transitions. Wide moving jams do not emerge spontaneously in free flow, i. e., spontaneous  $F \rightarrow J$  transitions are not observed in real traffic flow.
4. The most of traffic flow models used by traffic researches are based on the fundamental diagram hypothesis. These models can be classified into two main classes referred to the classic LWR-theory and to the classic GM-model. None of the traffic flow models reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128] can explain and predict traffic breakdown at the bottleneck as is observed in empirical data. These traffic flow models are basic traffic flow models for simulations of freeway control and management strategies. However, we have to conclude that the related simulations of the control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated management strategy.
5. In contrast with the traffic flow models and theories reviewed in [16,18,21,33,35,37,40,84,92,93,95,98,126,127,128], three-phase traffic theory can explain the empirical features of traffic breakdown and resulting congested patterns. Thus traffic flow models in the framework of this theory can be used for simulations of control and management strategies in freeway traffic before they are introduced to the market.

### Future Directions

There are still many “black areas” in understanding empirical *microscopic* features of synchronized flow and phase transitions in freeway traffic. Traffic flow models in the framework of three-phase traffic theory are only at the beginning of their development. There are almost no analytical results within the context of this theory (the exception is a probabilistic theory of traffic breakdown at a bottleneck (see ► [Traffic Breakdown, Probabilistic Theory of](#))). These and many other unsolved problems are an interest-

ing and important field of the future empirical and theoretical investigations.

Although the FOTO and ASDA models for tracking of spatiotemporal congested patterns based on three-phase traffic theory are already successfully used in on-line installations and there are ideas about applications of the theory for feedback on-ramp metering, speed limit control and advance driver assistant systems in vehicles, many other applications for control and management of traffic networks could be expected.

### Cross References

► [Freeway Traffic Management and Control](#)

### Acknowledgments

I would like to thank Sergey Klenov, Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald-Hiller and Olivia Brickley for help and useful suggestions.

### Bibliography

1. Ahn S, Cassidy MJ (2007) Freeway traffic oscillations and vehicle lane-change maneuvers. In: Allsop RE, Bell MGH, Hydecker BG (eds) *Transportation and Traffic Theory 2007*. Elsevier, Amsterdam, pp 691–710
2. Aw A, Rascle M (2000) Resurrection of “Second Order” models of traffic flow. *SIAM J Appl Math* 60:916–938
3. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1994) Structure stability of congestion in traffic dynamics. *Jpn J Appl Math* 11:203–223
4. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51:1035–1042
5. Bando M, Hasebe K, Nakayama A, Shibata A, Sugiyama Y (1995) Phenomenological study of dynamical model of traffic flow. *J Phys I France* 5:1389–1399
6. Barlović R, Santen L, Schadschneider A, Schreckenberg M (1998) Metastable states in cellular automata for traffic flow. *Eur Phys J B* 5:793–800
7. Bellomo N, Coscia V, Delitala M (2002) On the mathematical theory of vehicular traffic flow I. Fluid dynamic and kinetic modelling. *Math Mod Meth App Sci* 12:1801–1843
8. Berg P, Woods A (2001) On-ramp simulations and solitary waves of a car-following model. *Phys Rev E* 64:035602(R)
9. Bovy PHL (ed) (1998) *Motorway analysis: new methodologies and recent empirical findings*. Delft University Press, Delft
10. Brilon W, Geistefeld J, Regler M (2005) Reliability of freeway traffic flow: a stochastic concept of capacity. In: Mahamassani HS (ed) *Proc of the 16th inter sym on transportation and traffic theory*. Elsevier, Amsterdam, pp 125–144
11. Brilon W, Zurlinden H (2004) Kapazität von Straßen als Zufallsgröße. *Straßenverkehrstechnik* 4:164
12. Brilon W, Regler M, Geistefeld J (2005) Zufallscharakter der Kapazität von Autobahnen und praktische Konsequenzen – Teil 1. *Straßenverkehrstechnik* 3:136

13. Brockfeld E, Kühne RD, Skabardonis A, Wagner P (2003) Toward benchmarking of microscopic traffic flow models. *Trans Res Rec* 1852:124–129
14. Brockfeld E, Kühne RD, Wagner P (2005) Calibration and validation of simulation models. In: *Proc of the transportation research board 84th annual meeting*, TRB Paper No. 05-2152. TRB, Washington DC
15. Ceder A (ed) (1999) *Transportation and traffic theory*. Proc of the 14th international symposium on transportation and traffic theory, Elsevier, Oxford
16. Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
17. Colombo RM (2003) Hyperbolic Phase Transitions in Traffic Flow. *SIAM J Appl Math* 63:708–721
18. Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
19. Cowan RJ (1976) Useful headway models. *Trans Res* 9:371–375
20. Daganzo CF (1993) The cell-transmission model: a dynamic representation of highway traffic consistent with the hydrodynamic theory. *Trans Res B* 28:269–287
21. Daganzo CF (1997) *Fundamentals of transportation and traffic operations*. Elsevier, New York
22. Davis LC (2004) Multilane simulations of traffic phases. *Phys Rev E* 69:016108
23. Davis LC (2006) Controlling traffic flow near the transition to the synchronous flow phase. *Physica A* 368:541–550
24. Davis LC (2006) Effect of cooperative merging on the synchronous flow phase of traffic. *Physica A* 361:606–618
25. Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Physica A* 379:274–290
26. Edie LC, Foote RS (1958) Traffic flow in tunnels. *Highw Res Board Proc Ann Meet* 37:334–344
27. Edie LC, Foote RS (1960) Effect of shock waves on tunnel traffic flow. In: *Highway Research Board Proceedings*, vol 39. National Research Council, Washington DC, pp 492–505
28. Edie LC (1961) Car-following and steady state theory for non-congested traffic. *Oper Res* 9:66–77
29. Edie LC, Herman R, Lam TN (1980) Observed multilane speed distribution and the kinetic theory of vehicular traffic. *Trans Sci* 14:55–76
30. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic nature of breakdown at freeway merge junctions. *Trans Res Rec* 1484:80–89
31. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and Granular Flow' 01*. Springer, Heidelberg
32. Gao K, Jiang R, Hu S-X, Wang B-H, Wu Q-S (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76:026105
33. Gartner NH, Messer CJ, Rathi A (eds) (1997) *Special report 165: revised monograph on traffic flow theory*. Transportation Research Board, Washington DC
34. Gazis DC, Herman R, Rothery RW (1961) Nonlinear follow-the-leader models of traffic flow. *Oper Res* 9:545–567
35. Gazis DC (2002) *Traffic theory*. Springer, Berlin
36. Gipps PG (1981) Behavioral car-following model for computer simulation. *Trans Res B* 15:105–111
37. Haight FA (1963) *Mathematical theories of traffic flow*. Academic Press, New York
38. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Trans Res Rec* 1320:91–98
39. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Trans Res Rec* 1365:12–18
40. Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
41. Helbing D, Hennecke A, Treiber M (1999) Phase diagram of traffic states in the presence of inhomogeneities. *Phys Rev Lett* 82:4360–4363
42. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and Granular Flow' 99*. Springer, Heidelberg
43. Herman R, Montroll EW, Potts RB, Rothery RW (1959) Traffic dynamics: analysis of stability in car following. *Oper Res* 7: 86–106
44. Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) *Traffic and Granular Flow' 03*. Springer, Heidelberg
45. Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. *J Phys A Math Gen* 37:8197–8213
46. Jiang R, Wu QS (2005) Toward an improvement over Kerner–Klenov–Wolf three-phase cellular automaton model. *Phys Rev E* 72:067103
47. Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. *Physica A* 377:633–640
48. Jiang R, Hua M-B, Wang R, Wu Q-S (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner–Klenov speed adaptation model. *Phys Lett A* 365:6–9
49. Kerner BS (1998) Theory of congested traffic flow. In: Rysgaard R (ed) *Proc of the 3rd symposium on highway capacity and level of service*, vol 2. Road Directorate, Ministry of Transport, Denmark, pp 621–642
50. Kerner BS (1998) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
51. Kerner BS (1999) Congested traffic flow: observations and theory. *Trans Res Rec* 1678:160–167
52. Kerner BS (2004) *The physics of traffic*. Springer, Berlin
53. Kerner BS (2007) On-ramp metering based on three-phase traffic theory I. *Traffic Eng Control* 48:28–35
54. Kerner BS (2007) Study of freeway speed limit control based on three-phase traffic theory. *Trans Res Rec* 1999:30–39
55. Kerner BS (2008) A theory of traffic congestion at heavy bottlenecks. *J Phys A Math Theor* 41:215101
56. Kerner BS (2008) Three-phase traffic theory and its applications for freeway traffic control. In: Inweldi PO (ed) *Transportation research trends*. Nova Science Publishers, New York, pp 1–93
57. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A Math Gen* 35:L31–L43
58. Kerner BS, Klenov SL (2003) Microscopic theory of spatio-temporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
59. Kerner BS, Klenov SL (2005) Probabilistic breakdown phenomenon at on-ramps bottlenecks in three-phase traffic theory. cond-mat/0502281. e-print in <http://arxiv.org/abs/cond-mat/0502281>
60. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: congestion nucleation in spatially non-homogeneous traffic. *Physica A* 364:473–492

61. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory. *Trans Res Rec* 1965:70–78
62. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39:1775–1809
63. Kerner BS, Konhäuser P (1994) Structure and parameters of clusters in traffic flow. *Phys Rev E* 50:54–83
64. Kerner BS, Konhäuser P, Schilke M (1995) Deterministic spontaneous appearance of traffic jams in slightly inhomogeneous traffic flow. *Phys Rev E* 51:6243–6246
65. Kerner BS, Konhäuser P, Schilke M (1996) “Dipole-layer” effect in dense traffic flow. *Phys Lett A* 215:45–56
66. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A Math Gen* 35:9971–10013
67. Kerner BS, Klenov SL, Hiller A, Rehborn H (2006) Microscopic features of moving traffic jams. *Phys Rev E* 73:046107
68. Kerner BS, Klenov SL, Hiller A (2006) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. *J Phys A Math Gen* 39:2001–2020
69. Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. *Non Dyn* 49:525–553
70. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Towards a realistic microscopic description of highway traffic. *J Phys A Math Gen* 33:L477–L485
71. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: microscopic description of traffic phases. *Phys Rev E* 65:056133
72. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. *Phys Rev E* 70:016115
73. Kometani E, Sasaki T (1958) *J Oper Res Soc Jap* 2:11
74. Kometani E, Sasaki T (1959) A safety index for traffic with linear spacing. *Oper Res* 7:704–720
75. Koshi M (2003) An interpretation of a traffic engineer on vehicular traffic flow. In: Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow* 01. Springer, Heidelberg, pp 199–210
76. Koshi M, Iwasaki M, Ohkura I (1983) Some findings and an overview on vehicular flow characteristics. In: Hurdle VF (ed) *Proc 8th international symposium on transportation and traffic theory*. University of Toronto Press, Toronto, pp 403
77. Krauß S, Wagner P, Gawron C (1997) Metastable states in a microscopic model of traffic flow. *Phys Rev E* 55:5597–5602
78. Kühne R (1991) In: Brannolte U (ed) *Highway capacity and level of service*. A.A. Balkema, Rotterdam, pp 211
79. Kühne R, Mahnke R, Lubashevsky I, Kaupužs J (2002) Probabilistic description of traffic breakdown. *Phys Rev E* 65:066125
80. Laval JA (2007) Linking synchronized flow and kinematic waves. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Proc of the international workshop on traffic and granular flow*. Springer, Berlin, pp 521–526
81. Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical restriction versus human overreaction triggering congested traffic states. *Phys Rev Lett* 92:238702
82. Lee HY, Lee H-W, Kim D (1999) Dynamic states of a continuum traffic equation with on-ramp. *Phys Rev E* 59:5101–5111
83. Lesort J-B (ed) (1996) *Transportation and traffic theory*. Proc of the 13th international symposium on transportation and traffic theory. Elsevier, Oxford
84. Leutzbach W (1988) *Introduction to the theory of traffic flow*. Springer, Berlin
85. Li XG, Gao ZY, Li KP, Zhao XM (2007) Relationship between microscopic dynamics in traffic flow and complexity in networks. *Phys Rev E* 76:016110
86. Lighthill MJ, Whitham GB (1955) On kinematic waves. I Flow movement in long rivers. II A theory of traffic flow on long crowded roads. *Proc Roy Soc A* 229:281–345
87. Lorenz M, Elefteriadou L (2000) A probabilistic approach to defining freeway capacity and breakdown. *Trans Res Cir E-C* 018:84–95
88. Maerivoet S, De Moor B (2005) Cellular automata models of road traffic. *Phys Rep* 419:1–64
89. Mahmassani HS (ed) (2005) *Transportation and traffic theory*. Proc of the 16th inter sym on transportation and traffic theory. Elsevier, Amsterdam
90. Mahnke R, Kaupužs J (1999) Stochastic theory of freeway traffic. *Phys Rev E* 59:117–125
91. Mahnke R, Pieret N (1997) Stochastic master-equation approach to aggregation in freeway traffic. *Phys Rev E* 56:2666–2671
92. Mahnke R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
93. May AD (1990) *Traffic flow fundamentals*. Prentice-Hall, New Jersey
94. Okamura H, Watanabe S, Watanabe T (2000) An empirical study of the capacity of bottlenecks on the basic suburban Expressway sections in Japan. TRB Circular EC 018, Transportation Research Board, Washington DC
95. Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
96. Nagatani T, Nakanishi K (1998) Delay effect on phase transitions in traffic dynamics. *Phys Rev E* 57:6415–6421
97. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys (France) I* 2:2221–2229
98. Nagel K, Wagner P, Woessler R (2003) Still flowing: approaches to traffic flow and traffic jam modeling. *Oper Res* 51:681–716
99. Neubert L, Santen L, Schadschneider A, Schreckenberg M (1999) Single-vehicle data of highway traffic: a statistical analysis. *Phys Rev E* 60:6480–6490
100. Newell GF (1961) Nonlinear effects in the dynamics of car following. *Oper Res* 9:209–229
101. Newell GF (1982) *Applications of queuing theory*. Chapman Hall, London
102. Papageorgiou M (1983) *Application of automatic control concepts in traffic flow modeling and control*. Springer, Berlin
103. Payne HJ (1971) Models of freeway traffic and control. In: Bekey GA (ed) *Mathematical models of public systems*, vol 1. Simulation Council, La Jolla
104. Payne HJ (1979) *Trans Res Rec* 772:68
105. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the breakdown phenomenon in freeway traffic. *Trans Res Rec* 1634:64–69
106. Pipes LA (1953) An operational analysis of traffic dynamics. *J Appl Phys* 24:274–287
107. Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical restriction versus human overreaction: accident avoidance and two-lane simulations. In: Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) *Proc of the international workshop on traffic and granular flow*. Springer, Berlin, pp 503–508



108. Prigogine I, Herman R (1971) Kinetic theory of vehicular traffic. Elsevier, New York
109. Richards PI (1956) Shockwaves on the highway. *Oper Res* 4:42–51
110. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) Traffic and Granular Flow' 05. In: Proc of the international workshop on traffic and granular flow. Springer, Berlin
111. Schönhof M, Helbing D (2007) Empirical features of congested traffic states and their implications for traffic modelling. *Trans Sci* 41:135–166
112. Schreckenberg M, Wolf DE (eds) (1998) Traffic and Granular Flow' 97. In: Proc of the international workshop on traffic and granular flow. Springer, Singapore
113. Siebel F, Mauser W (2006) Synchronized flow and wide moving jams from balanced vehicular traffic. *Phys Rev E* 73:066108
114. Stokes EE (1848) On a difficulty in the theory of sound. *Phil Mag* 33:349–356
115. Takayasu M, Takayasu H (1993) Phase transition and 1/f type noise in one dimensional asymmetric particle dynamics. *Fractals* 1:860–866
116. Tanga CF, Jiang R, Wu QS (2007) Phase diagram of speed gradient model with an on-ramp. *Physica A* 377:641–650
117. Taylor MAP (ed) (2002) Transportation and traffic theory in the 21st century. Proc of the 15th international symposium on transportation and traffic theory. Elsevier, Amsterdam
118. Tilch B, Helbing D (2000) Evaluation of single vehicle data in dependence of the vehicle-type, lane, and site. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow' 99. Springer, Heidelberg, pp 333–338
119. Treiber M, Hennecke A, Helbing D (2000) Congested traffic states in empirical observations and microscopic simulations. *Phys Rev E* 62:1805–1824
120. Treiterer J (1967) Improvement of traffic flow and safety by longitudinal control. *Trans Res* 1:231–251
121. Treiterer J (1975) Investigation of traffic dynamics by aerial photogrammetry techniques. Ohio State University Technical Report PB 246 094. Columbus, Ohio
122. Treiterer J, Taylor JI (1966) Traffic flow investigations by photogrammetric techniques. *Highw Res Rec* 142:1–12
123. Treiterer J, Myers JA (1974) The hysteresis phenomenon in traffic flow. In: Buckley DJ (ed) Proc 6th international symposium on transportation and traffic theory. A.H. & AW Reed, London, pp 13–38
124. Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. *Physica A* 378:475–484
125. Wang Y, Papageorgiou M, Messmer A (2004) Predictive feedback routing control strategy for freeway network traffic [E-text type]. In: Proc of the 83rd Annual Transportation Research Board Meeting, TRB Paper No. 04-3429. TRB, Washington DC
126. Whitham GB (1974) Linear and nonlinear waves. Wiley, New York
127. Wiedemann R (1974) Simulation des Verkehrsflusses. University of Karlsruhe, Karlsruhe
128. Wolf DE (1999) Cellular automata for traffic simulations. *Physica A* 263:438–451
129. Wolf DE, Schreckenberg M, Bachem A (eds) (1995) Traffic and Granular Flow. Proc of the international workshop on traffic and granular flow. World Scientific, Singapore
130. Zhang P, Wong SC (2006) Essence of conservation forms in the traveling wave solutions of higher-order traffic flow models. *Phys Rev E* 74:026109
131. Zurlinden H (2003) Ganzjahresanalyse des Verkehrsflusses auf Straßen. In: Schriftenreihe des Lehrstuhls für Verkehrswesen der Ruhr-Universität Bochum, vol 26. Ruhr-Universität Bochum, Bochum

## Traffic Congestion, Spatiotemporal Features of

BORIS S. KERNER

GR/ETI, HPC: G021, Daimler AG,  
Sindelfingen, Germany

### Article Outline

Glossary

Definition of the Subject

Introduction

Congested Patterns on Homogeneous Road

Congested Patterns at Isolated Bottlenecks

Diagram of Congested Patterns at Isolated Bottlenecks

Complex Congested Patterns and Pattern Interaction

Reproducible and Predictable Congested Patterns

Microscopic Features of Traffic Phases

Congested Patterns at Heavy Bottlenecks

Conclusions. Fundamental Empirical Features

of Spatiotemporal Congested Freeway Traffic Patterns

Future Directions

Acknowledgments

Bibliography

### Glossary

**Free flow** Free flow is usually observed, when the vehicle density in traffic is small enough. The flow rate increases in free flow with increase in vehicle density, whereas the average vehicle speed is a decreasing density function.

**Congested traffic** Congested traffic is defined as a state of traffic in which the average speed is *lower* than the minimum average speed that is possible in free flow. In empirical observations, congested traffic occurs mostly at a freeway bottleneck; the average speed decreases sharply in an initial free flow to a lower speed at the bottleneck. This speed breakdown observed during the onset of congestion is called the breakdown phenomenon or traffic breakdown. The bottleneck can be a result of road works, on- and off-ramps, a decrease



in the number of freeway lanes, road curves and road gradients, bad weather conditions, accidents, etc.

**Effectual bottleneck** An effectual bottleneck is a bottleneck at which traffic breakdown is observed.

**Three traffic phases** In three-phase traffic theory, there are the three traffic phases:

1. Free flow.
2. Synchronized flow.
3. Wide moving jam.

The wide moving jam and synchronized flow traffic phases associated with congested traffic are defined through the spatiotemporal macroscopic empirical (objective) criteria [J] and [S]. The definition [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks. The definition [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

Traffic breakdown is associated with a local phase transition from free flow to synchronized flow. Wide moving jams can emerge spontaneously in synchronized flow only; the jams result from growth of narrow moving jams that occur in the synchronized flow.

**Synchronized flow patterns** A synchronized flow pattern (SP) is a congested traffic pattern that consists of the synchronized flow phase only.

**Pinch effect in synchronized flow** The pinch effect in synchronized flow is the effect of spontaneous occurrence and growth of narrow moving jams in synchronized flow. Within the associated pinch region of synchronized flow a self-compression of synchronized flow is usually observed.

**General congested patterns** A general pattern (GP) is a congested traffic pattern at a hypothetical isolated bottleneck that consists of both traffic phases in congested traffic – synchronized flow and wide moving jam. Firstly, synchronized flow occurs at the bottleneck. Due to the pinch effect in this synchronized flow, later and upstream of the bottleneck narrow moving jams emerge spontaneously; jam amplitudes grow while the jams propagate upstream within the pinch region. Some of the jams (or each of them) transform into wide moving jams at the upstream boundary of the pinch region of synchronized flow. Thus GP structure consists of the pinch region of synchronized flow and a sequence of wide moving jams upstream.

**Diagram of congested patterns** The diagram of congested traffic patterns at a hypothetical isolated bottleneck shows regions of congested pattern existence and excitation depending on traffic demand and bottleneck characteristics.

**Expanded congested patterns** An expanded congested pattern (EP) is a congested traffic pattern whose synchronized flow affects two or more adjacent freeway bottlenecks.

**Microscopic structure of wide moving jam** The wide moving jam structure consists of alternations of regions in which traffic flow is interrupted and flow states of low speeds associated with moving blanks within the jam. The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move.

**Moving blanks within wide moving jam** When vehicles meet a wide moving jam, they come to a stop at very different blank spaces (blanks for short) to each other. Later vehicles begin to move within the jam covering these blanks. Due to this low speed vehicle motion new blanks between vehicles occur upstream within the jam. A moving blank is a blank between vehicles within a wide moving jam, which moves upstream due to low speed vehicle motion within the jam.

**Heavy bottleneck** A heavy bottleneck is a bottleneck with a great bottleneck strength, i. e., a great bottleneck influence on traffic flow that limits the average flow rate within a congested pattern occurring upstream of the bottleneck to very small values. A heavy bottleneck can occur due to e. g., bad weather conditions, accidents, or heavy roadworks.

**Traffic congestion at heavy bottleneck** Traffic congestion at a heavy bottleneck can be associated with the phenomena of random disappearance and appearance of the pinch region over time and a complex non-regular dynamics of wide moving jams upstream of the pinch region as well as with the phenomenon of the merger of wide moving jams into a mega-wide moving jam (mega-jam). When the bottleneck strength increases strongly, the pinch region disappears and only the mega-jam survives and synchronized flow remains only within its downstream front separating free flow and congested traffic.

## Definition of the Subject

Empirical observations show that traffic congestion exhibits extremely complex spatiotemporal features. Spatiotemporal features of traffic congestion are as follows:

- (i) Diverse variety of spatiotemporal congested traffic patterns measured at a freeway bottleneck.
- (ii) Complex evolution of these congested patterns in space and time that occurs when traffic demand or/and bottleneck characteristics change.
- (iii) Complex spatiotemporal phenomena associated with congested patterns occurring at two or more adjacent freeway bottlenecks.
- (iv) Transformations between various congested patterns that occurs due to phase transitions between different traffic phases within the pattern.
- (v) Diverse microscopic characteristics of traffic congestion associated with complex driver behavior within congested patterns.
- (vi) Complex non-regular dynamics of wide moving jams that can occur when a heavy bottleneck appears on a road.

Traffic congestion is a fact of life for many car drivers. For this reason, one of the aims of transportation and traffic research is to provide an understanding of freeway traffic congestion that can be used for effective traffic management, traffic control, organization, traffic assignment and other engineering applications, which should improve traffic safety and result in high-quality mobility.

Moreover, due to the very complex non-linear character of freeway traffic observed in empirical data, analysis of complexity in spatiotemporal traffic phenomena is needed for development of modeling approaches, which can explain these traffic phenomena and, therefore, are applicable for traffic prediction, traffic control, organization, traffic assignment and other traffic engineering applications.

## Introduction

There are a huge number of publications devoted to empirical and experimental investigations of spatiotemporal features of vehicular traffic congestion (e. g., [9,10,11,12,51,68,69,70] and references in [2,17,20,38,54,57,58,59]). However, the puzzle of empirical spatiotemporal features of traffic congestion has been solved only recently [30,36]. Consequently, as has already explained in ► [Traffic Congestion, Modeling Approaches to](#), earlier traffic flow theories and models reviewed in [1,2,3,4,16,17,20,54,56,57,58,59,72,73,74] cannot explain and predict most empirical features of the onset of congestion (traffic breakdown) and many important empirical features of resulting congested patterns (with the exclusion of explanation of wide moving jam propagation; see ► [Traffic Congestion, Modeling Approaches to](#)). These traffic flow models are standard ones for validation of freeway traffic control, management and assignment. Thus the related simulations of

freeway control and management strategies cannot predict many of the freeway traffic phenomena that would occur through the use of a simulated control strategy.

Therefore, in 1996–1999 the author introduced an alternative traffic flow theory called three-phase traffic theory, which overcomes the disadvantages of all other traffic flow models and theories in the explanation of spatiotemporal empirical features of traffic congestion. For this reason, the three-phase traffic theory is the basis for an explanation of empirical spatiotemporal features of traffic congestion made in this article.

In three-phase traffic theory, besides the free flow traffic phase there are two traffic phases in congested traffic: synchronized flow and wide moving jam. Thus there are three traffic phases in this theory:

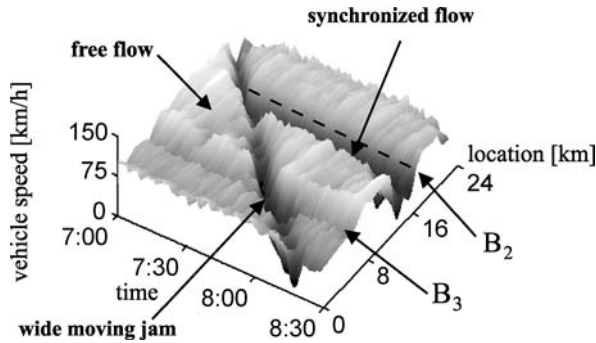
1. Free flow.
2. Synchronized flow.
3. Wide moving jam.

These phases are defined through spatiotemporal empirical criteria for these phases. Before we consider the phase definitions, we define the term “moving jam”. A moving jam is a localized structure of great vehicle density spatially limited by two jam fronts. Within the downstream jam front vehicles accelerate escaping from the jam; within the upstream jam front, vehicles slow down approaching the jam. Both jam fronts (and, therefore, the jam as a whole structure) propagate upstream in traffic flow. Within the jam (i. e., between the jam fronts) vehicle density is great and speed is very low (sometimes as low as zero).

The definition of the wide moving jam phase [J]: A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or bottlenecks.

The definition of the synchronized flow phase [S]: In contrast with the wide moving jam phase, the downstream front of the synchronized flow phase does not exhibit the wide moving jam characteristic feature; in particular, the downstream front of the synchronized flow phase is often fixed at a bottleneck.

Note that the downstream front of synchronized flow separates synchronized flow upstream from free flow downstream. Within the downstream front of synchronized flow vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow downstream of the synchronized flow. To apply the traffic phase definitions [J] and [S], spatiotemporal features of congested traffic should be known. The criteria [J] and [S] also mean that if a congested traffic state is not related to the wide moving jam phase, then with certainty the state is asso-



**Traffic Congestion, Spatiotemporal Features of, Figure 1**  
Measured three traffic phases: Average vehicle speed in space and time. Taken from [38]

ciated with the synchronized flow phase. This is because in three-phase traffic theory congested traffic can be either within the synchronized flow phase or within the wide moving jam phase. In other words, if in measured data congested traffic states associated with the wide moving jam phase have been identified, then with certainty all remaining congested states in the data set are related to the synchronized flow phase. For example, the definition [S] means that moving jams, which are caught at the bottleneck, are related to the synchronized flow phase rather than to the wide moving jam phase.

The traffic phase definitions [J] and [S] are illustrated by measured data shown in Fig. 1. There are two qualitatively different types of congested patterns. The first congested pattern propagates through both on-ramp bottlenecks (labeled  $B_2$  and  $B_3$  in Fig. 1) while maintaining the mean velocity of the downstream front of the pattern. In accordance with the definition [J], this congested pattern is associated with the wide moving jam traffic phase in congested traffic. The downstream fronts of another congested pattern is fixed at bottleneck  $B_2$ . In accordance with the definition [S], these congested patterns are associated with the synchronized flow traffic phase.

In measured data, the onset of congestion in free flow is accompanied by a sharp decrease in the vehicle speed from a free flow speed to a considerably lower speed in congested traffic and it is called traffic breakdown, or speed breakdown, or else the breakdown phenomenon (see, e.g., [13,18,19,57,61]). As explained in ► [Traffic Congestion, Modeling Approaches to](#), traffic breakdown is a first-order local phase transition from free flow to synchronized flow traffic phase ( $F \rightarrow S$  transition) at the bottleneck, i.e., an  $F \rightarrow S$  transition and traffic breakdown are synonyms associated with the same phenomenon: the onset of congestion in an initial free flow.

In synchronized flow resulting from an  $F \rightarrow S$  transition, narrow moving jams can emerge spontaneously whose subsequent growth can lead to wide moving jam occurrence ( $S \rightarrow J$  transition). The spontaneous emergence of narrow moving jams in synchronized flow with the subsequent jam growth, while the jams propagate upstream within the synchronized flow, is called the pinch effect in synchronized flow.  $S \rightarrow J$  transitions occur usually at other road locations than the road location of the  $F \rightarrow S$  transition resulting in the emergence of synchronized flow. Thus in real free flow, wide moving jams emerge due to a sequence of two phase transitions called  $F \rightarrow S \rightarrow J$  transitions: firstly, an  $F \rightarrow S$  transition occurs; later and usually at another road location within the emergent synchronized flow an  $S \rightarrow J$  transition is realized ([38], ► [Traffic Congestion, Modeling Approaches to](#)).

In ► [Traffic Congestion, Modeling Approaches to](#) of this Encyclopedia, the fundamental empirical features of traffic breakdown, a critical discussion of earlier traffic flow modeling approaches to traffic congestion, which cannot explain these features of traffic breakdown, a theory of the pinch effect in synchronized flow, as well as a stochastic traffic flow model in the framework of three-phase traffic theory have been reviewed. The latter model explains and predicts empirical features of traffic breakdown and wide moving jam emergence.

The main focus of this article is a review of empirical *spatiotemporal features* of traffic congested patterns resulting from traffic breakdown as well as an explanation of these empirical traffic patterns. The article is organized as follows. Congested patterns occurring on homogeneous roads (outside or else without highway bottlenecks) are discussed in Sect. “[Congested Patterns on Homogeneous Road](#)”. In Sect. “[Congested Patterns at Isolated Bottlenecks](#)”, we consider the main types of congested patterns that occur at isolated highway bottlenecks. A diagram of these patterns at an on-ramp bottleneck, i.e., regions of spontaneous occurrence of various types of congested patterns in the flow–flow plane whose coordinates are the flow rate to the on-ramp and the flow rate in free flow upstream of the bottleneck is discussed in Sect. “[Diagram of Congested Patterns at Isolated Bottlenecks](#)”. In this section, we will also briefly discuss another congested pattern diagram associated with an off-ramp bottleneck. Complex congested patterns and their interaction, which are observed on real freeways with several adjacent bottlenecks, will be reviewed in Sect. “[Complex Congested Patterns and Pattern Interaction](#)”. Reproducible and predictable features of spatiotemporal congested traffic patterns are briefly reviewed in Sect. “[Reproducible and Pre-](#)

dictable Congested Patterns”. In Sect. “Microscopic Features of Traffic Phases”, we consider some microscopic features of traffic congestion. In Sect. “Congested Patterns at Heavy Bottlenecks” we discuss features of congested patterns at heavy bottlenecks caused for example by bad weather conditions, accidents, or heavy roadworks. Finally, in Sect. “Conclusions. Fundamental Empirical Features of Spatiotemporal Congested Freeway Traffic Patterns” we make a summary of fundamental empirical spatiotemporal features of traffic congested patterns.

### Congested Patterns on Homogeneous Road

As follows from empirical results and explained in three-phase traffic theory ([30,31,38], ► [Traffic Congestion, Modeling Approaches to](#)), a bottleneck increases the probability of traffic breakdown ( $F \rightarrow S$  transition) in a neighborhood of this bottleneck considerably in comparison to other road locations. This explains why in real free flow the breakdown ( $F \rightarrow S$  transition) occurs mostly at the bottleneck, whereas outside highway bottlenecks traffic breakdown is observed extremely seldom.

Although in real free flow the breakdown occurs mostly at a highway bottleneck, in this section we discuss spatiotemporal features of congested patterns occurring on homogeneous roads (outside or else without highway bottlenecks). This is because empirical, experimental, and theoretical studies of congested patterns occurring on homogeneous roads have a great *methodological* importance: these studies show that phase transitions and congested patterns in vehicular traffic result from complex driver interactions in traffic, rather than just as a result of highway bottlenecks [32,38].

### Moving Synchronized Flow Patterns

An empirical example of the spontaneous emergence of synchronized flow in an initially free flow is shown in Fig. 2 [32]. The observation of this  $F \rightarrow S$  transition outside bottlenecks was possible, because the distance between two adjacent bottlenecks (downstream off-ramp bottleneck at location  $x = 23.4$  km, detectors D23 and upstream on-ramp bottleneck at  $x = 17.1$  km, detectors D16) is great enough. We see that during the  $F \rightarrow S$  transition at  $x = 19$  km (detectors D18; up-arrow  $F \rightarrow S$  at 06:36 in Fig. 2b) free flow was both at the off-ramp- and on-ramp bottlenecks (D23 and D16, D15).

After the  $F \rightarrow S$  transition has occurred (labeled  $F \rightarrow S$  in Figs. 2 and 3), the region of the synchronized flow propagates both downstream (up-arrow at location D20 in Fig. 3) and upstream (up-arrow at D17). A synchronized flow pattern, which propagates on a road as

a whole localized structure is called a moving synchronized flow pattern (MSP).

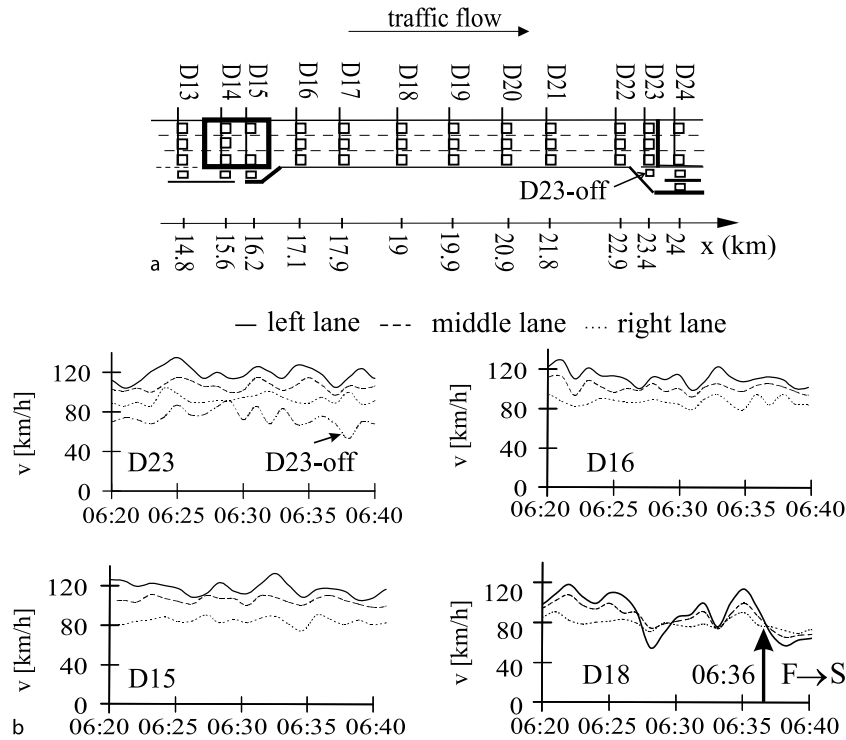
A numerical simulation of an MSP on a homogeneous road without bottlenecks is shown in Fig. 4a. Due to a small speed disturbance, an  $F \rightarrow S$  transition occurs spontaneously in an initially homogeneous free flow. This phase transition is a first-order one as this can be seen from Fig. 4b: there should be a speed disturbance of *finite* amplitude for the occurrence of the  $F \rightarrow S$  transition. Thus there is a nucleation effect that is the usual one for other various first-order phase transitions in complex systems of different nature: there is a critical speed (density) disturbance in free flow, i. e., a nucleus required for the nucleation of an  $F \rightarrow S$  transition. If the nucleus appears in free flow, the disturbance grows leading to the  $F \rightarrow S$  transition. If in contrast the disturbance is smaller than the critical one, the disturbance decays. We see that the greater the density, the smaller the nucleus required for the  $F \rightarrow S$  transition (curve  $F_S$  in Fig. 4b) [41].

Numerical simulations of  $F \rightarrow S$  transitions (labeled  $F \rightarrow S$ ) with spontaneous emergence of many MSPs (labeled S) on a homogeneous circle road without bottlenecks is shown in Fig. 5. In this case, a KKW cellular automata three-phase traffic flow model has been studied [45] discussed in ► [Traffic Breakdown, Probabilistic Theory of](#). There are model fluctuations of great amplitude in this model. For this reason, there are many local disturbances, which are nuclei for  $F \rightarrow S$  transitions. As a result, many MSPs occur almost simultaneously on the road.

### Pinch Effect and Moving Jam Emergence

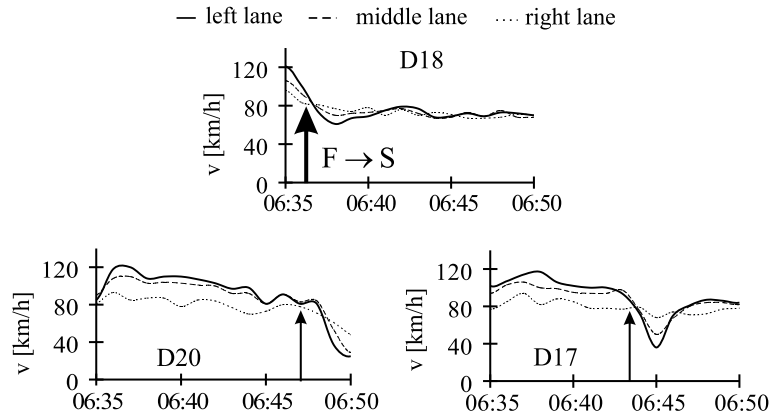
Pinch effect conditions can be realized in synchronized flow outside bottlenecks [32]. An empirical example is shown in Fig. 6. In the synchronized flow that has initially occurred outside bottlenecks discussed in Subsect. “[Moving Synchronized Flow Patterns](#)” (Figs. 2 and 3), the region of synchronized flow with a greater density ( $x = 20.9$  km, D20) occurs. Within this synchronized flow a narrow moving jam appears, which grows in its amplitude propagating upstream (down-arrows in Fig. 6). Finally, the jam transforms into a wide moving one. Both the spontaneous emergence of the jam and the resulting  $S \rightarrow J$  transition occur outside the bottlenecks. The fact that the resulting jam is a wide moving one is confirmed by the empirical evidence that the jam propagates upstream through the on-ramp bottleneck (down-arrows at D16 and D15) while maintaining the characteristic velocity of the downstream front of the jam.

A numerical simulation of a spontaneous emergence of a moving jam within synchronized flow of an MSP with



#### Traffic Congestion, Spatiotemporal Features of, Figure 2

Measured  $F \rightarrow S$  transition outside bottlenecks: **a** Scheme of a part of freeway section of the freeway A5 South in Germany with road detectors labeled D13–D24 (see Fig. 2.1 of [38]). **b** Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]



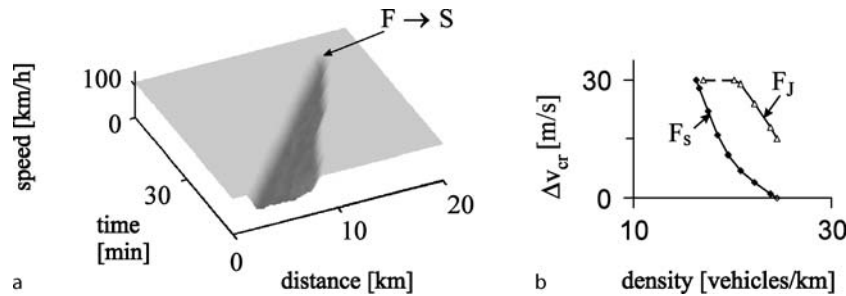
#### Traffic Congestion, Spatiotemporal Features of, Figure 3

Empirical propagation of synchronized flow outside bottlenecks from location of  $F \rightarrow S$  transition (D18) downstream (D20) and upstream (D17) in data measured on the freeway section shown in Fig. 2a. Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]

the subsequent  $S \rightarrow J$  transition on a homogeneous road without bottlenecks is shown in Fig. 7 (labeled  $S \rightarrow J$ ) [41]. Another scenario of a sequence of  $F \rightarrow S \rightarrow J$  transitions, which occurs on a circular road is shown in Fig. 5: firstly,

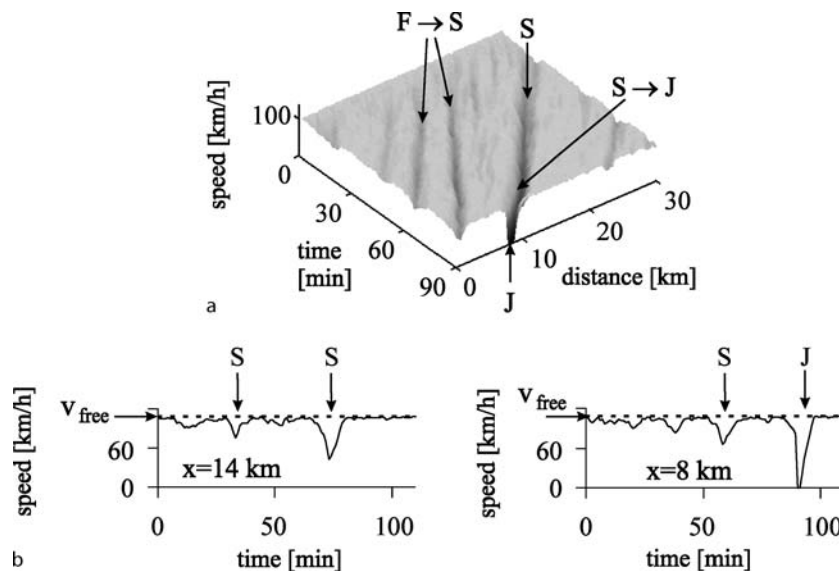
an  $F \rightarrow S$  transition occurs and an MSP appears. Later and upstream of the location of the  $F \rightarrow S$  transition a narrow moving jam emerges in synchronized flow of the MSP; the jam grows propagating upstream. Finally, a wide moving





**Traffic Congestion, Spatiotemporal Features of, Figure 4**

Simulations of  $F \rightarrow S$  transition on homogeneous road without bottlenecks: **a** MSP formation. **b** Critical amplitude (denoted by  $\Delta v_{cr}$ ) of local speed disturbance, i. e., speed nucleus required for  $F \rightarrow S$  transition (curve  $F_S$ ) and  $F \rightarrow J$  transition (curve  $F_J$ ) as functions of vehicle density in initially homogeneous free flow. Taken from [41]



**Traffic Congestion, Spatiotemporal Features of, Figure 5**

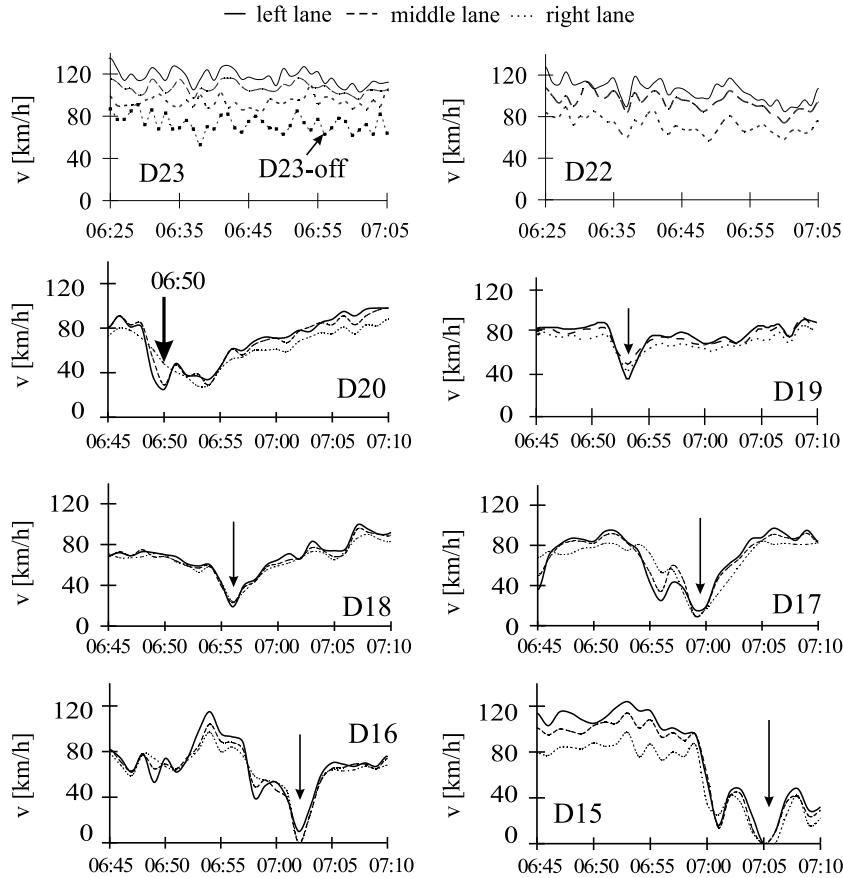
Simulations of the spontaneous emergence of synchronized flow ( $F \rightarrow S$  transition) in initially homogeneous free flow with the subsequent moving jam emergence in the emergent synchronized flow ( $S \rightarrow J$  transition) on circle road without bottlenecks. Taken from [45]

jam emerges (labeled J). Results of three-phase traffic theory about  $F \rightarrow S$  and  $S \rightarrow J$  transitions on a homogeneous road [30,32,41,45] discussed above have also been confirmed in simulations of a cellular automata three-phase traffic flow model [24].

### Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road

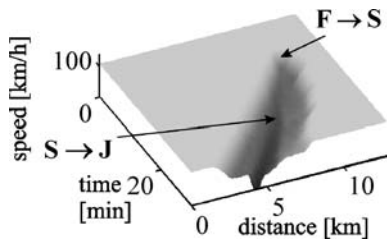
In three-phase traffic theory has been found that an  $S \rightarrow J$  transition is a first-order phase transition occurring in metastable states of synchronized flow. A detailed consideration of the multitude of metastable synchronized flow

appears in Sect. “Moving Jam Emergence in Synchronized Flow ( $S \rightarrow J$  Transition)” of [Traffic Congestion, Modeling Approaches to](#) in this Encyclopedia. In an initial homogeneous metastable synchronized flow, the  $S \rightarrow J$  transition occurs, if a critical local speed (density) disturbance appears in this flow. This critical disturbance can be considered a nucleus required for the nucleation of an  $S \rightarrow J$  transition. If the nucleus appears in the synchronized flow, the disturbance grows leading to the  $S \rightarrow J$  transition. If in contrast the disturbance is smaller than the critical one, the disturbance decays. At a given speed, the greater the density in synchronized flow, the smaller the critical disturbance, i. e., the smaller the nucleus required for the



**Traffic Congestion, Spatiotemporal Features of, Figure 6**

Empirical spontaneous emergence of wide moving jam in synchronized flow (labeled by down-arrows) outside bottlenecks in data measured on the freeway section shown in Fig. 2a. Time-dependences of 1 min average speed in three road lanes at different locations. Taken from [32]

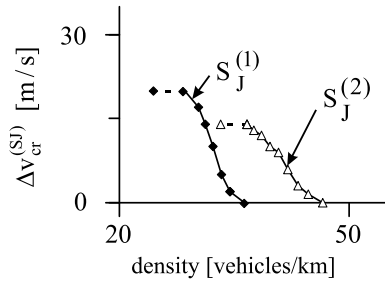


**Traffic Congestion, Spatiotemporal Features of, Figure 7**

Simulations of sequence of  $F \rightarrow S \rightarrow J$  transitions on homogeneous road without bottlenecks. Taken from [41]

$S \rightarrow J$  transition. Thus pinch effect conditions, i.e., synchronized flow in which speed disturbances grow spontaneously leading to the  $S \rightarrow J$  transition can be realized on homogeneous road without bottlenecks ([30,31,32,38], ► [Traffic Congestion, Modeling Approaches to](#)).

Results of simulations of an  $S \rightarrow J$  transition in synchronized flow on a circle road without bottlenecks under pinch effect conditions are presented in Fig. 8 [41]. In contrast with simulations of a sequence of  $F \rightarrow S \rightarrow J$  transitions shown in Fig. 5, in Fig. 8 rather than free flow, the initial traffic flow is associated with homogeneous synchronized flow that is in a metastable state with respect to an  $S \rightarrow J$  transition. The initial synchronized flow speeds (72 km/h for the curve  $S_f^{(1)}$  and 51 km/h for  $S_f^{(2)}$ ) have been considerably lower than free flow speed ( $v_{\text{free}} = 108$  km/h in Fig. 5). During a short time interval a local speed disturbance of an initial amplitude  $\Delta v$  has been applied in the synchronized flow: within the disturbance, the decrease in speed is equal to  $\Delta v$ . Then, the subsequent dynamic evolution of the initial speed disturbance in the initially homogeneous synchronized flow on the circle road is studied. It is found that there is a critical amplitude of the speed disturbance denoted by  $\Delta v_{\text{cr}}^{(SJ)}$ , i.e., there is a speed



**Traffic Congestion, Spatiotemporal Features of, Figure 8**

Simulations of the spontaneous emergence of wide moving jams ( $S \rightarrow J$  transition) in homogeneous synchronized flow on circle road without bottlenecks: Critical amplitude of local speed disturbance denoted by  $\Delta v_{cr}^{(S/J)}$ , i. e., speed nucleus required for  $S \rightarrow J$  transition (curves  $S_J^{(1)}$  and  $S_J^{(2)}$  are related to the synchronized flow speeds 72 and 51 km/h, respectively). Taken from [41]

nucleus of a “size”  $\Delta v_{cr}^{(S/J)}$  required for an  $S \rightarrow J$  transition: if  $\Delta v \geq \Delta v_{cr}^{(S/J)}$ , then the disturbance grows leading to the spontaneous emergence of a wide moving jam in synchronized flow. If in contrast,  $\Delta v < \Delta v_{cr}^{(S/J)}$ , then the disturbance decays over time resulting in homogeneous synchronized flow. At a given synchronized flow speed, the study has been made for various initial densities of synchronized flow. We found that at each of the given speeds, the greater the density in synchronized flow, the smaller the nucleus required for the  $S \rightarrow J$  transition (curves  $S_J^{(i)}$ ,  $i = 1, 2$  in Fig. 8). If the density of synchronized flow is great enough, then  $\Delta v_{cr}^{(S/J)} \rightarrow 0$  (Fig. 8), i. e., a very small nucleus is required for the spontaneous emergence of a wide moving jam in the dense synchronized flow.

This means that in such an initially homogeneous dense synchronized flow on a homogeneous circle road without bottlenecks pinch effect conditions are fully satisfied: already very small speed (density) disturbances in this synchronized flow grow propagating upstream with the subsequent spontaneous moving jam emergence [32,41].

#### Pinch Effect Conditions in Driver Experiment: Spontaneous Emergence of Moving Jams in Synchronized Flow of Great Density on Circle Road

The result of three-phase traffic theory discussed in Subsect. “Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road” [32,38,41,45] has recently been confirmed in a driver experiment [66]. In this experiment, wide moving jams emerge spontaneously within an initially homogeneous synchronized flow of great density. In the study,  $N_{veh} = 22$  vehicles have initially homogeneously moved along a circle road of a length  $L_{circle} = 230$  m with a speed of about 30 km/h. Over time,

growing local speed (density) disturbances have occurred in this vehicle motion. The growth of these disturbances has led to wide moving jam formation, which propagate upstream with the velocity about 20 km/h that is associated with empirical observations of moving jams that emerge in the pinch region of synchronized flow [38].

The experiment [66] is a very interesting and important one for a deeper understanding of moving jam emergence in synchronized flow. However, we should note that both the statement of the authors that the spontaneous moving jam emergence occurs in *free flow* as well as the associated physical explanation of the jam emergence in traffic are *invalid*. Indeed, in the experiment [66] vehicles could not overtake each other and the vehicle speed (30 km/h) has been considerably lower than the speed in free flow (usually higher than 80 km/h for passenger cars). Moreover, the initial density of the initially homogeneous traffic flow of [66] has been  $\rho_{hom} = N_{veh}/L_{circle} \approx 95$  vehicles/km that is considerably greater than the maximal density observed in free flow, which is usually appreciably smaller than 40 vehicles/km. The vehicle density 95 vehicles/km in the initially homogeneous flow in the experiment [66] is even considerably greater than the usual density in the pinch region of synchronized flow occurring upstream of on- and off-ramp bottlenecks; this density is usually within the range 35–60 vehicles/km [38].

Thus rather than free flow, the initially homogeneous flow in the experiment of [66] is associated with synchronized flow of a very great density. Rather than the moving jam emergence in free flow, in the experiment [66] the spontaneous moving jam emergence in an initially homogeneous synchronized flow of a very great density has been shown discussed in Subsect. “Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road”. This is in accordance with all other observations of wide moving jam emergence in synchronized flow occurring on homogeneous roads (Subsect. “Pinch Effect and Moving Jam Emergence”) as well as upstream of bottlenecks (see for example Fig. 16 discussed in Subsect. “General Congested Patterns”) [32,38].

In contrast with the invalid physical explanation for the spontaneous moving jam emergence in traffic made in [66], as follows from all known empirical observations of real traffic as well as from the characteristics of an initial traffic flow in the driver experiment [66] discussed above, moving jams do not emerge spontaneously in free flow ([32,38], ► [Traffic Congestion, Modeling Approaches to](#)). Three-phase traffic theory explains this result of empirical observations as follows. At the same density of free flow, the nucleus required for the synchronized flow emer-

gence in this free flow (curve  $F_S$  in Fig. 4) is considerably smaller than the nucleus required for the emergence of a moving jam in this free flow (curve  $F_J$  in Fig. 4). This can explain why rather than the moving jam emergence, the synchronized flow emergence governs the onset of congestion in real free flow. For this reason, all empirical observations show that moving jams spontaneously emerge in *free flow* due to a sequence of  $F \rightarrow S \rightarrow J$  transitions *only*. This is regardless of whether there are bottlenecks on a road or not. A more detailed explanation of this important empirical result can be found in [38] as well as in ► [Traffic Congestion, Modeling Approaches to](#).

### Congested Patterns at Isolated Bottlenecks

As mentioned above, the onset of congestion ( $F \rightarrow S$  transition) in an initial free flow occurs mostly at a freeway bottleneck. A bottleneck at which traffic breakdown occurs is called the *effectual bottleneck*. After traffic breakdown has occurred, synchronized flow is formed at the effectual bottleneck. In many cases, the downstream front of the synchronized flow is fixed at the bottleneck (the exclusion is moving synchronized flow patterns). The location in a neighborhood of the bottleneck at which this downstream front of synchronized flow is fixed is called an *effective location of the effectual bottleneck* (the effective location of the bottleneck for short). It must be noted that the effective location of the bottleneck can be different from the location at which traffic breakdown ( $F \rightarrow S$  transition) has occurred leading to congested pattern emergence. Moreover, both the location of traffic breakdown and the effective location of the bottleneck are probabilistic values in real traffic. Even for the same type of congested pattern the effective location of the bottleneck can randomly change over time [38].

On real freeways there are many effectual bottlenecks. When a congested pattern has occurred at a downstream effectual bottleneck, the pattern can propagate further upstream reaching an upstream effectual bottleneck. In this case, if free flow at the upstream bottleneck is in a metastable state with respect to an  $F \rightarrow S$  transition (i. e., with respect to synchronized flow emergence), then congestion pattern propagation from the downstream bottleneck causes an induced  $F \rightarrow S$  transition at the upstream bottleneck. This can lead to complex congested pattern transformation effects considered below.

Before we discuss these complex effects of congested pattern transformation, we would like to discuss congested patterns that emerge at a hypothetical isolated effectual bottleneck. As we have already mentioned, on real freeways there are many effectual bottlenecks, therefore,

a consideration of the isolated bottleneck is an idealization of real traffic. Obviously, this idealization is easy to perform with a traffic flow model rather than with real traffic. However, on real freeways there can be freeway sections with adjacent effectual bottlenecks that are located far enough from each other. On such freeway sections during long enough time intervals before a congested pattern, which has emerged at a downstream bottleneck, reaches an upstream bottleneck, the congested pattern can be considered the pattern at an isolated bottleneck. This bottleneck is the downstream bottleneck at which the pattern has initially occurred. Such a simplification of the reality allows us to understand some important common spatiotemporal features of congested patterns. Nevertheless, although in this section only spatiotemporal features of congested patterns at isolated bottlenecks are discussed, all illustrations of *empirical* examples of congested patterns represent the real development of measured congested patterns, which sometimes cause induced traffic breakdowns and complex pattern transformations at upstream bottlenecks when the patterns reach them. To prevent confusions, these real induced pattern formation and complex congested pattern transformation will be further explained in Sect. “Complex Congested Patterns and Pattern Interaction” of this article. There are two main types of congested patterns at an isolated bottleneck: 1. Synchronized flow patterns (SP for short). 2. General patterns (GP for short).

### Synchronized Flow Patterns

As a result of an  $F \rightarrow S$  transition at an isolated bottleneck, various SPs can occur at the bottleneck. There are three main types of SPs [38]:

- (1) Localized SP (LSP for short): The downstream front of an LSP is fixed at the bottleneck. The upstream front of the LSP does not continuously propagate upstream over time: this front is localized at some distance upstream of the bottleneck. In other words, the region of synchronized flow in the LSP is localized in space. However, the location of the upstream front of synchronized flow in the LSP and therefore the LSP width (in the longitudinal direction) can exhibit complex oscillations over time. Note that the upstream front of synchronized flow separates free flow upstream from synchronized flow downstream of this front. Vehicles must slow down within the upstream front of synchronized flow.
- (2) Widening SP (WSP for short): As in the LSP, the downstream front of an WSP is fixed at the bottleneck. In contrast to the LSP, the upstream front of the WSP

propagates upstream continuously over time. In other words, the width of synchronized flow (in the longitudinal direction) in the WSP widens in the upstream direction. It must be noted that due to this widening, the upstream front of synchronized flow must eventually reach an upstream adjacent effectual bottleneck. At this upstream bottleneck another congested pattern either already exists or can be induced. In both cases, the WSP can be caught at this upstream bottleneck (catch effect; see below). Thus, in real traffic the WSP can usually exist only for a finite time, before the WSP reaches the upstream effectual bottleneck.

- (3) Moving SP (MSP for short): In contrast to the LSP and WSP, an MSP is a localized pattern on a freeway that propagates within free flow over time between freeway bottlenecks; both downstream and upstream MSP fronts, which separate synchronized flow within the MSP and free flows away from the MSP, propagate on the freeway. Depending on flow rate distribution within the MSP and in free flows away from the MSP, each of the MSP fronts can propagate either upstream or downstream. In particular, downstream MSP propagation (i. e., when both MSP fronts propagate downstream) is possible; this occurs if the flow rate within the MSP is greater than the flow rates in free flows away from the MSP. Otherwise, upstream MSP propagation is realized. Due to MSP propagation, the MSP (or one of the MSP fronts) can reach an adjacent effectual bottleneck at which synchronized flow can be excited; then the MSP (or the associated MSP front) is caught at the bottleneck (catch effect). Thus, the MSP that propagates between bottlenecks can exist only for a finite time, before the MSP (or one of the MSP fronts) reaches a nearest adjacent effectual bottleneck at which synchronized flow can be excited.

An empirical example of an LSP is shown in Fig. 9. It can be seen that there are no wide moving jams in the LSP. Moreover, we can see that whereas the speed is considerably lower within the LSP (Fig. 9a), the flow rate is close to the flow rate in free flow (see Fig. 9b, where the increase in flow rate downstream of the bottleneck is related to the on-ramp inflow).

The upstream front of synchronized flow in the LSP (Fig. 9a) does not continuously propagate upstream of the bottleneck over time: this upstream front is localized at some finite distance upstream of the bottleneck. This distance is a function of time. In other words, this SP is indeed an empirical example of an LSP.

In numerical simulations made in the framework of three-phase traffic theory, LSPs exhibit qualitatively the

same spatiotemporal features as those found in empirical data (Fig. 10) [38]:

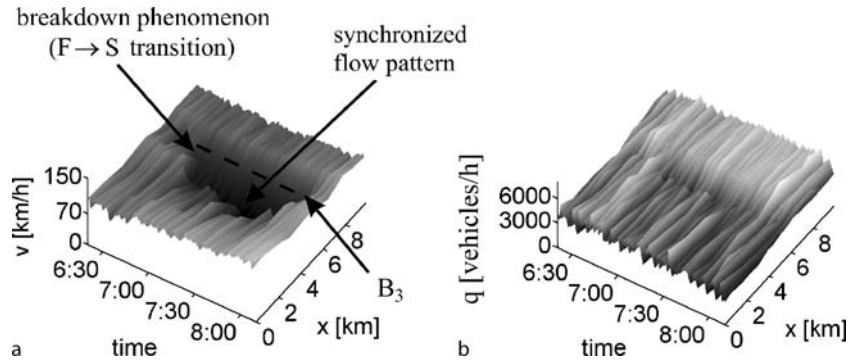
- (i) The downstream front of the LSP is fixed at the on-ramp bottleneck. However, the upstream front of the LSP is localized on the main road at some distance  $L_{\text{LSP}}$  upstream of the on-ramp bottleneck (Fig. 10a,f).
- (ii) Whereas the speed within the LSP is considerably lower than in free flow, the flow rate is approximately the same as that in free flow (Fig. 10b–e).
- (iii) The LSP width  $L_{\text{LSP}}$  can be a complicated time function (Fig. 10f). The mean width of the LSP  $L_{\text{LSP}}^{(\text{mean})}$  can depend strongly on  $q_{\text{in}}$  and  $q_{\text{on}}$ . In numerical simulations,  $L_{\text{LSP}}^{(\text{mean})}$  varies between 0.5 km and 10 km.

An empirical example of an MSP is shown in Fig. 11. We consider two adjacent bottlenecks, a downstream bottleneck  $B_1$  and an upstream bottleneck  $B_2$ . Synchronized flow first occurs at the downstream bottleneck  $B_1$ . The synchronized flow subsequently propagates upstream as a distinct localized structure. This is an example of an MSP (MSP is labeled “MSP” in Fig. 11a,c). When the MSP reaches the upstream bottleneck  $B_2$ , it does not propagate through the bottleneck as a localized structure: the MSP is caught at the bottleneck (catch effect). This is in contrast to wide moving jam propagation through an upstream bottleneck, as shown in Fig. 1 and for a wide moving jam labeled “wide moving jam” in Fig. 11b,c. It should be also noted that whereas the flow rate is very small within the latter wide moving jam, we cannot almost distinguish the MSP in the flow rate distribution in space and time shown in Fig. 11b. As already mentioned above, this is a very important features of many empirical (measured) SPs: The flow rate within the SP can be as great as in an initial free flow.

Numerical simulations of MSPs made in the framework of three-phase traffic theory show qualitatively the same spatiotemporal features as those found in empirical data [38]:

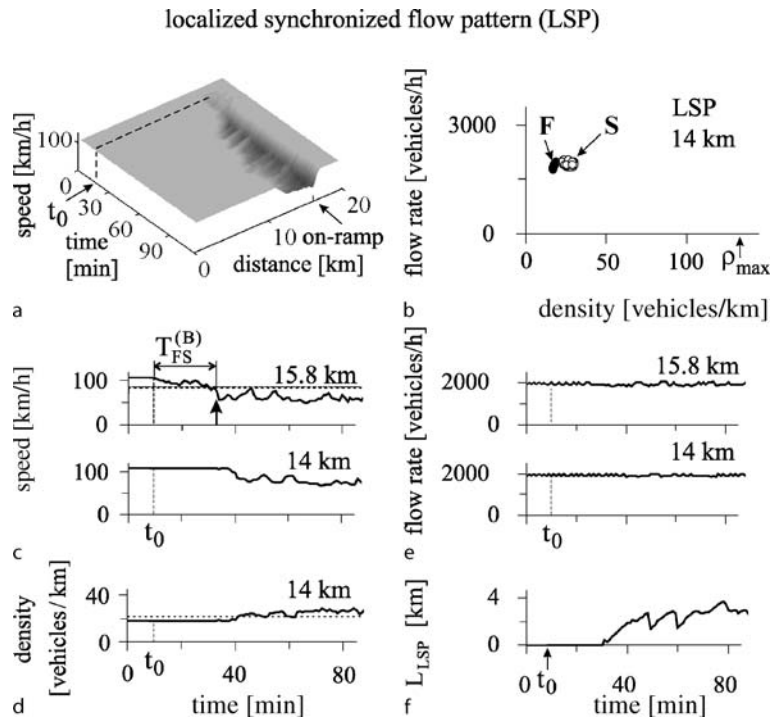
- (i) The MSP occurs spontaneously on the main road at the off-ramp (Fig. 12a) or on-ramp (Fig. 12b) bottlenecks because the downstream front of synchronized flow can depart from the bottleneck. As a result, the MSP begins to propagate on the main road as an independent localized structure on a homogeneous road. As in empirical observations (Fig. 11b), whereas the speed within the MSP is considerably lower than in free flow, the flow rate is approximately the same as that in free flow (Fig. 12c,d).
- (ii) When two adjacent bottlenecks exist on a freeway and an MSP appears at the downstream one, then the MSP





**Traffic Congestion, Spatiotemporal Features of, Figure 9**

Empirical example of a localized synchronized flow pattern (LSP) at on-ramp bottleneck  $B_3$ . **a** Averaged vehicle speed in the LSP in space and time. **b** Total flow rate across the freeway in space and time. The *dashed line* in **a** shows the location of the bottleneck. Taken from [38]



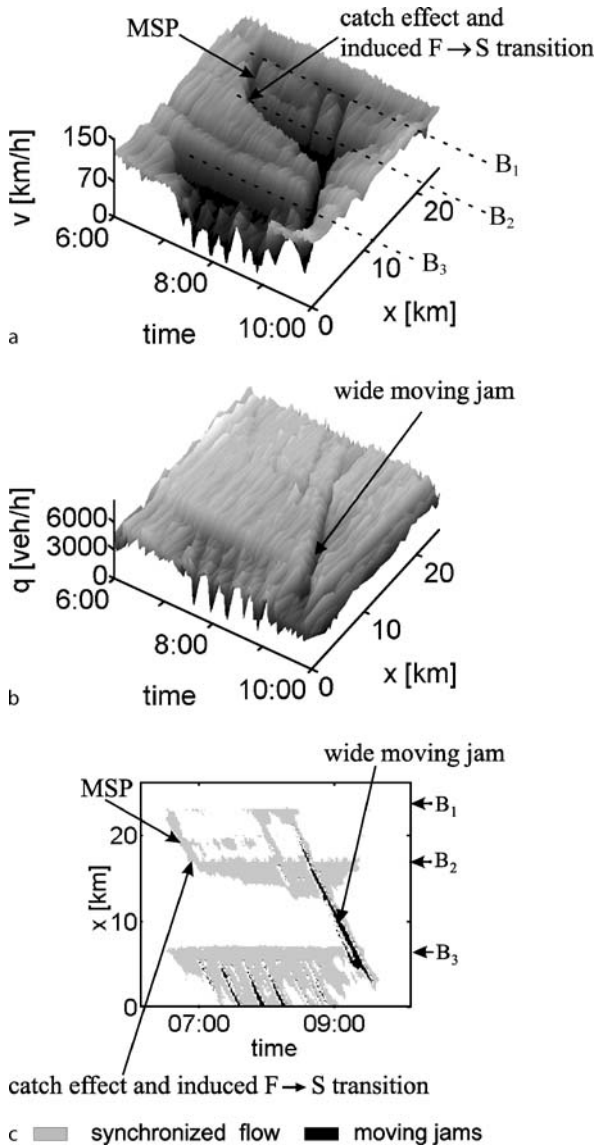
**Traffic Congestion, Spatiotemporal Features of, Figure 10**

Localized synchronized flow pattern (LSP) at on-ramp bottleneck in the framework of three-phase traffic theory (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► [Traffic Congestion, Modeling Approaches to](#)): **a** Averaged vehicle speed in the LSP in space and time. **b** Free flow (F) and synchronized flow within the LSP (S) in the flow–density plane. **c–e** Speed (**c**), density (**d**) and flow rate (**e**) as time-functions at fixed freeway locations. **f** The LSP width  $L_{LSP}$  as a time-function. Taken from [38]

can be caught at the downstream bottleneck with the subsequent SP formation.

An empirical example of an WSP is shown in Fig. 13. In this case, synchronized flow due to a spontaneous traffic breakdown is formed at an off-ramp bottleneck  $B_{North1}$ .

The downstream front of this synchronized flow is fixed at this bottleneck. The upstream front of synchronized flow propagates continuously upstream over time. When this front reaches an upstream on-ramp bottleneck  $B_{North2}$ , the WSP is caught at the bottleneck (catch effect) causing an induced traffic breakdown at the bottleneck with the sub-



**Traffic Congestion, Spatiotemporal Features of, Figure 11**

Empirical example of MSP with subsequent catch effect and induced  $F \rightarrow S$  transition at bottleneck  $B_2$ : a, b Average speed (a) and flow rate (b) in space and time. c Graph of a with overview of free flow (white), synchronized flow (gray), and moving jams (black). Bottlenecks  $B_2$  and  $B_3$  are the same as those in Fig. 1. Taken from [38]

sequent formation of another congested pattern upstream of the bottleneck.

Numerical simulations of WSPs made in the framework of three-phase traffic theory show qualitatively the same spatiotemporal features as those found in empirical data (Fig. 14) [38]. The WSP occurs spontaneously on the

main road at the off-ramp (Fig. 14a) or on-ramp (Fig. 14b) bottlenecks. The downstream front of an WSP is fixed at the on-ramp bottleneck. The upstream front of the WSP is continuously widening upstream. As within other SPs, the speed within the WSP is considerably lower than in free flow and the flow rate is approximately the same as that in free flow (Figs. 13b and 14c,d).

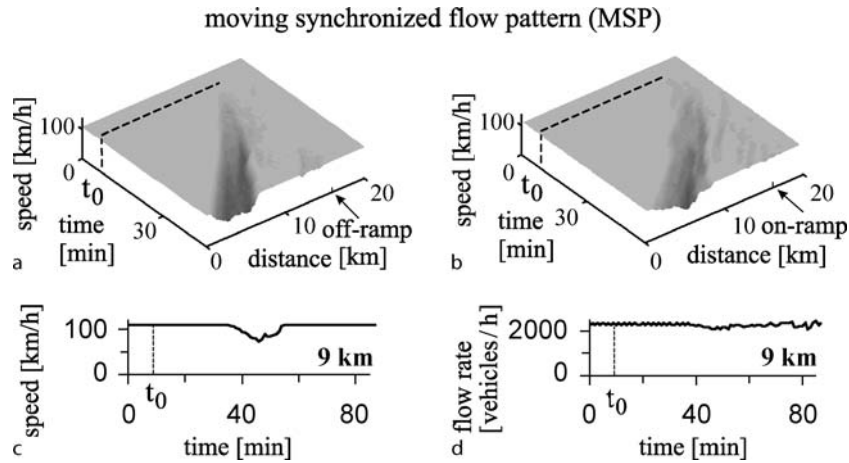
### General Congested Patterns

It must be noted that over time the average speed in synchronized flow of an WSP can decrease, and wide moving jams can begin to emerge spontaneously. In other words, the WSP can transform into a general pattern (GP). A GP is a congested pattern at an isolated bottleneck, which consists of synchronized flow whose downstream front is fixed at the bottleneck and wide moving jams that emerge spontaneously in this synchronized flow upstream of the bottleneck (Fig. 15a,b). Thus the GP consists of the both traffic phases of congested traffic, synchronized flow and wide moving jams.

In the empirical example shown in Fig. 15, WSP transformation into a GP occurs downstream of an upstream on-ramp bottleneck  $B_{\text{North}2}$ . For this reason, we can qualitatively consider the congested pattern that is upstream of the downstream off-ramp bottleneck  $B_{\text{North}1}$  and downstream of the upstream on-ramp bottleneck  $B_{\text{North}2}$  as an empirical example of a GP at the off-ramp bottleneck. Because the jams labeled 1, 2, and 3 propagate through the upstream bottleneck while maintaining their mean velocities of the jam downstream front (Fig. 15b), in accordance with the phase definition [J], these jams are indeed wide moving jams.

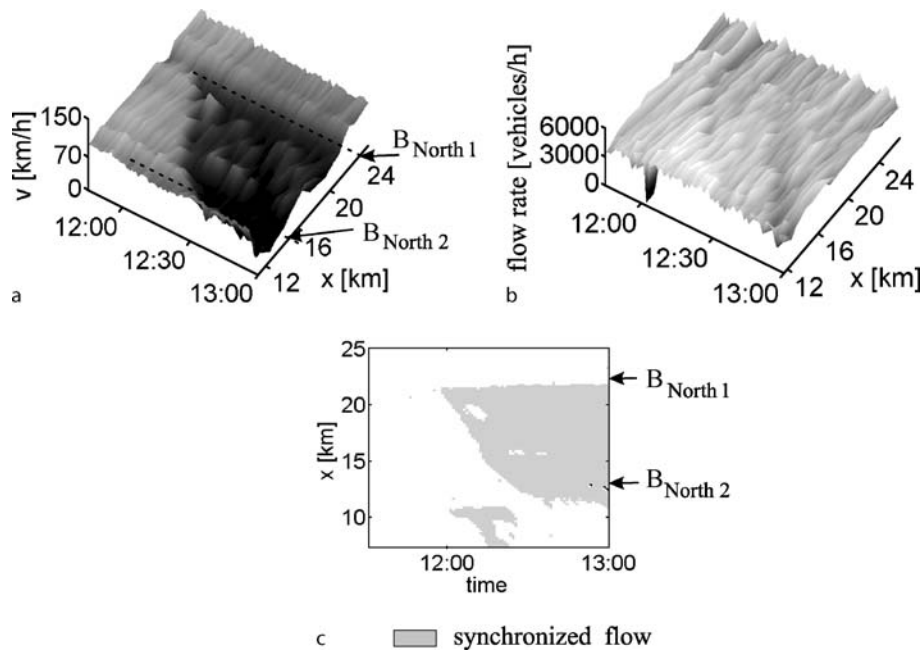
However, we must note that in many cases, upstream congestion (i. e., synchronized flow and wide moving jams in Fig. 15a,b upstream of the on-ramp bottleneck  $B_{\text{North}2}$ ) can influence downstream congested pattern parameters considerably (Sect. “Complex Congested Patterns and Pattern Interaction”). For this reason, the above consideration of the empirical congested pattern between the adjacent bottlenecks  $B_{\text{North}1}$  and  $B_{\text{North}2}$  as a GP can only show some qualitative features of GPs at an isolated bottleneck.

Another empirical example of a GP, which is formed at an on-ramp bottleneck before wide moving jams or/and synchronized flow of this GP reach an eventual upstream bottleneck on the freeway, is shown in Fig. 16. We see that firstly, synchronized flow is formed upstream of the bottleneck (labeled by arrows  $F \rightarrow S$  and  $S$ ). Later, and at other freeway locations wide moving jams emerge in that synchronized flow. These wide moving jams propa-



**Traffic Congestion, Spatiotemporal Features of, Figure 12**

Moving synchronized flow patterns (MSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of three-phase traffic theory [38] (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► [Traffic Congestion, Modeling Approaches to](#)): a, b Averaged vehicle speed in the MSP in space and time. c, d Speed (c) and flow rate (d) as time-functions at a fixed freeway location. Taken from [38]



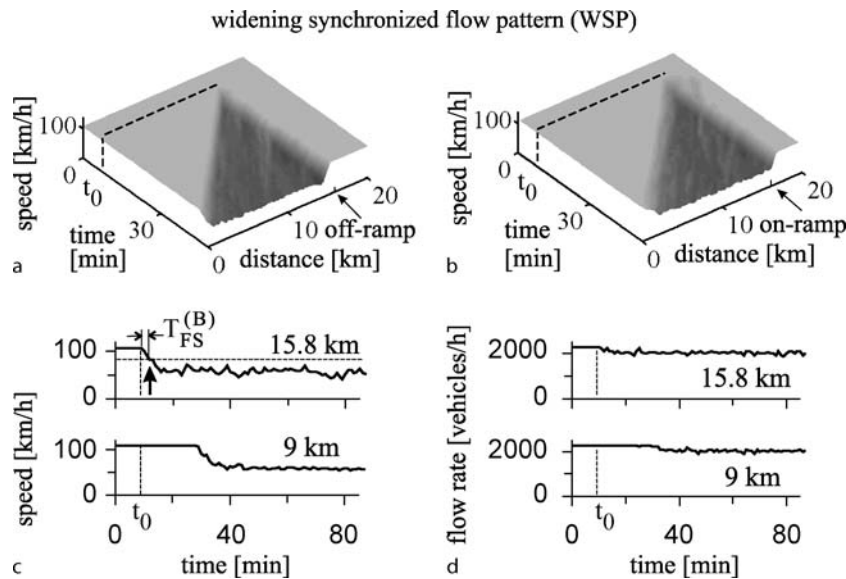
**Traffic Congestion, Spatiotemporal Features of, Figure 13**

Empirical example of an WSP and the catch effect: a, b Average speed (a) and flow rate (b) in space and time. c Graph of a with overview of free flow (white) and synchronized flow (gray). Taken from [38]

gate further upstream while maintaining the velocity of their downstream fronts. Free flow remains downstream of the bottleneck (location 7.8 km).

We can see that qualitative features of this GP are the same as those in Fig. 15. In both cases, firstly an  $F \rightarrow S$  transition occurs at the bottleneck (arrow S in Figs. 15c

and 16b). Synchronized flow propagates upstream. Later, the pinch effect is realized: moving jams emerge spontaneously in the synchronized flow upstream of the location of the initial traffic breakdown. These moving jams propagate upstream growing in amplitude. Some of the jams transform into wide moving jams, i. e.,  $S \rightarrow J$  transitions



**Traffic Congestion, Spatiotemporal Features of, Figure 14**

Widening synchronized flow patterns (WSP) at off-ramp (a) and on-ramp (b–d) bottlenecks in the framework of three-phase traffic theory [38] (numerical simulations of the Kerner–Klenov stochastic three-phase traffic flow model considered in ► [Traffic Congestion, Modeling Approaches to](#)): a, b Averaged vehicle speed in the WSP in space and time. c, d Speed (c) and flow rate (d) as time-functions at a fixed freeway location. Taken from [38]

occur. In the example, the mean time interval between the emergence of a narrow moving jam and the jam transformation into a wide moving one is about 10 min, i. e., about ten times longer than the mean duration of  $F \rightarrow S$  transitions, which are about 1 min in duration.

However, there are also some differences in GP formation in Figs. 15 and 16. In Fig. 15, WSP exists during a long enough time interval before the pinch effect occurs within the WSP, i. e., the WSP transforms into the GP. In contrast, in Fig. 16 the pinch effect occurs in synchronized flow after a short time interval. Secondly, in Fig. 15 the frequency of wide moving jam emergence in synchronized flow is considerably smaller than the one for the GP shown in Fig. 16.

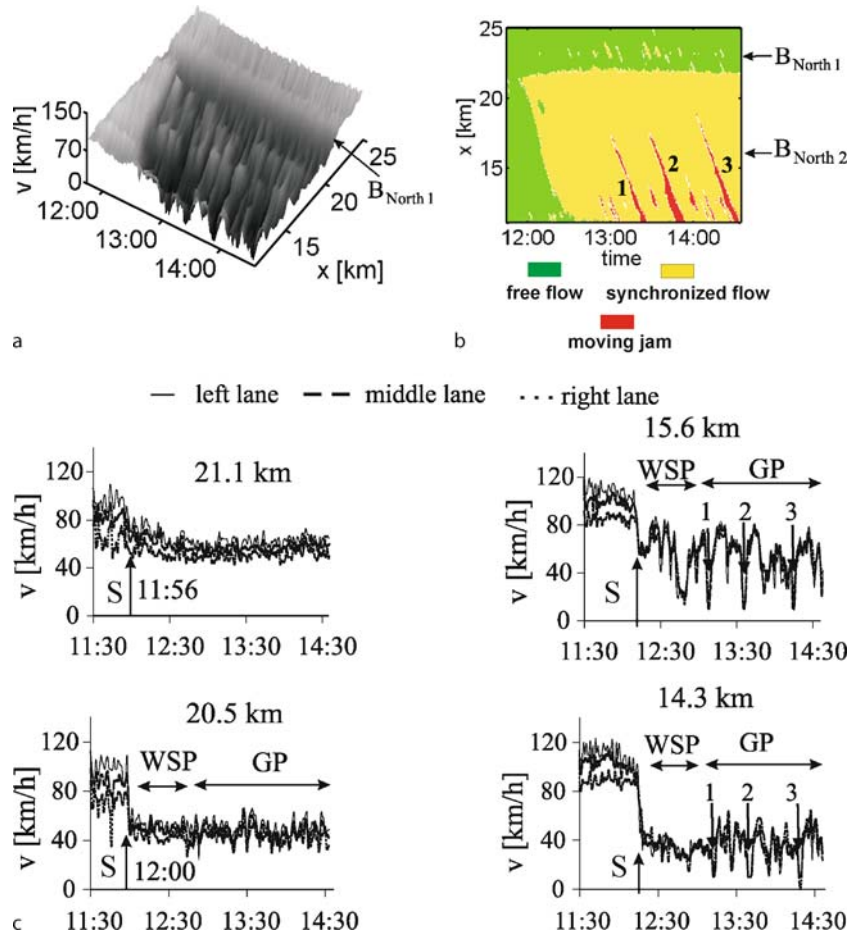
The differences in GP formation that have been mentioned above result from different pinch effect features in these two cases (Figs. 15 and 16). It has been found that the stronger the self-compression of synchronized flow in the pinch region is (i. e., the more the increase in synchronized flow density), the more likely moving jams emerge and the greater the frequency of this moving jam emergence is. This is in accordance with pinch effect features discussed in Subsect. “Pinch Effect Conditions in Homogeneous Synchronized Flow on Circle Road”. In the first case of the off-ramp bottleneck shown in Fig. 15, the pinch effect is weak (so-called “weak congestion” case), i. e., a rel-

atively small increase in density of the synchronized flow within the initial WSP occurs. As a result, the frequency of wide moving emergence is small; the frequency depends on the percentage of vehicle leaving the main road to the off-ramp.

In contrast, in Fig. 16 the pinch effect is relatively strong, i. e., density increases considerably in the pinch region of synchronized flow; the frequency of wide moving jam emergence is greater than in the former case. It has been found that if the flow rates upstream of the on-ramp bottleneck increase, then a case of so-called “strong congestion” within the pinch region can be reached. In this case, regardless of further on-ramp flow rate increase, the flow rate within the pinch region on the main road reaches a limit (minimum) flow rate and wide moving jam frequency reaches the maximum value. This is valid, if bottleneck characteristics and traffic parameters like weather, percentage of long vehicles, and other road conditions do not change considerably.

These and other empirical spatiotemporal features of GP formation have been explained in three-phase traffic theory (Fig. 17). In particular, the following features of the pinch effect have been found:

- (i) Firstly, an  $F \rightarrow S$  transition occurs at the bottleneck. The emergent synchronized flow propagates

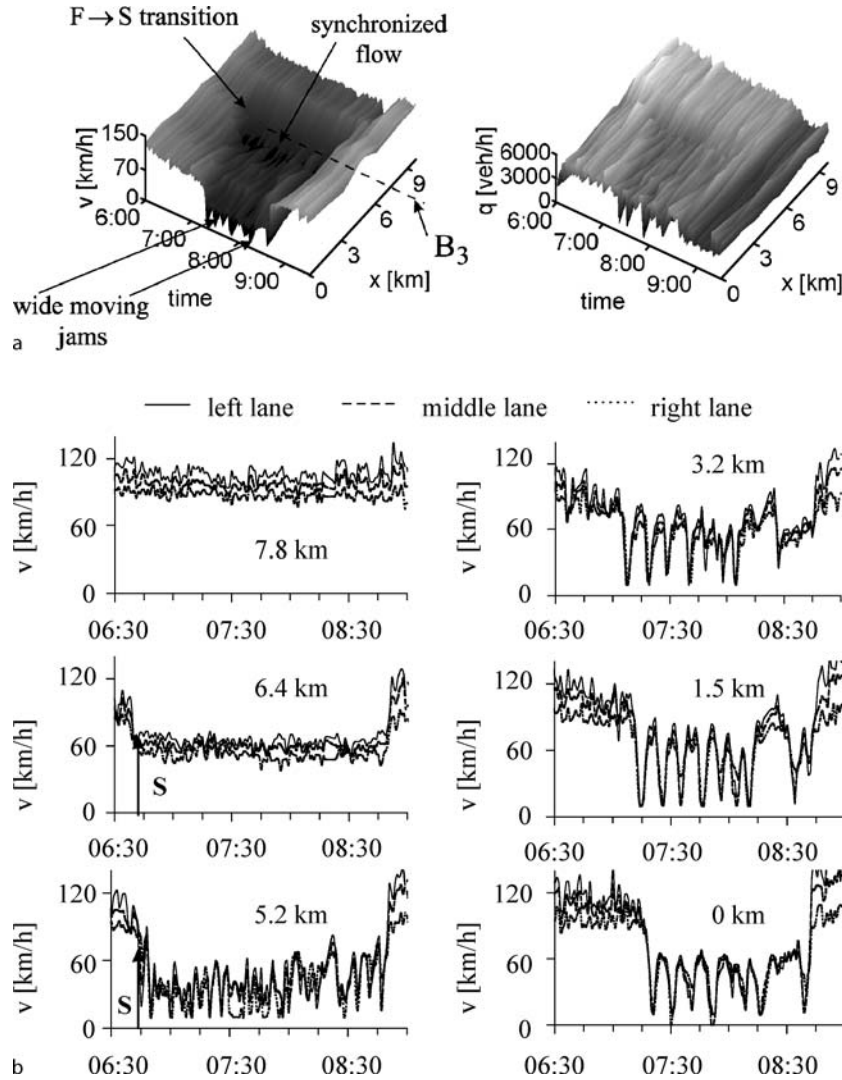


**Traffic Congestion, Spatiotemporal Features of, Figure 15**

Empirical example of WSP transformation into a GP over time: **a** Average speed in space and time. **b** Graph of **a** with overview of free flow (green), synchronized flow (yellow) and wide moving jams labeled 1, 2, and 3 (red). **c** Time dependence of speed at some fixed freeway locations. Taken from [38]

- upstream. In the examples of GPs shown in Fig. 17, the upstream front of the congested patterns is the upstream front of the synchronized flow that has initially emerged at the bottleneck. There are also other GP examples, when the upstream front of the GP is associated with the upstream front of the farthest upstream wide moving jam of a GP (such GPs are not shown here; see the book [38]).
- (ii) Later, the pinch effect occurs in synchronized flow upstream of the bottleneck. However, the flow rate does not necessarily decrease within the synchronized flow. Even if the flow rate decreases due to the pinch effect, the relative value of this decrease is considerably smaller than the decrease in average speed; this is explained by considerable increase in density, i. e., by self-compression of the synchronized flow. In the pinch region of a GP, narrow moving jams emerge spontaneously and grow propagating upstream (Fig. 17c),  $x = 15.8$  km and  $x = 14.5$  km.
  - (iii) Points related to the pinch region lie above the line J in the flow–density plane, i. e., these points are associated with metastable synchronized flow states with respect to wide moving jam emergence. This explains narrow moving jams emergence and growth in the pinch region; see explanations of metastable states of synchronized flow and of growing narrow moving jams, which are associated with the synchronized flow phase, in ([38], ► [Traffic Congestion, Modeling Approaches to](#)).
  - (iv) The location of an  $S \rightarrow J$  transition, at which a narrow moving jam transforms into a wide moving jam, is related to the upstream boundary of the pinch re-





**Traffic Congestion, Spatiotemporal Features of, Figure 16**

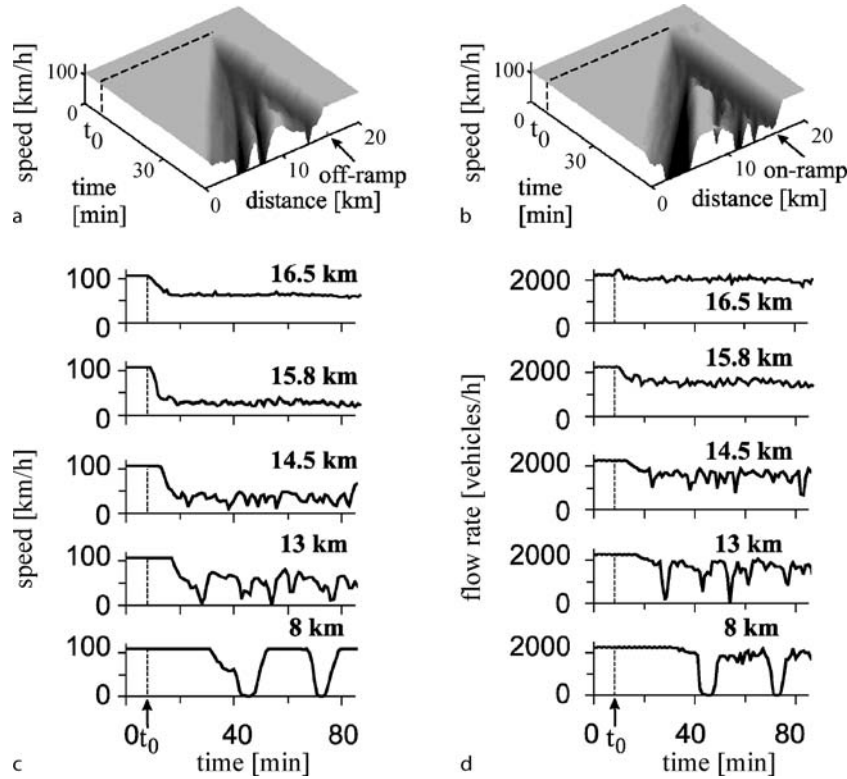
Empirical example of GP at on-ramp bottleneck  $B_3$ . **a** Average vehicle speed (*left*) and total flow rate (*right*) across the freeway in space and time. **b** Speed in the three freeway lanes at different locations (1 min data). Taken from [38]

gion of the GP. Different narrow moving jams can transform into wide moving jams at different locations. For this reason, the upstream boundary of the pinch region can exhibit complicated oscillations over time.

- (v) After a narrow moving jam has transformed into a wide moving jam, this wide moving jam can suppress the growth of the downstream narrow moving jam (Fig. 17c,  $x = 13$  km). This jam suppression effect occurs only if the narrow moving jam is close to the downstream wide moving jam front.

- (vi) Distances between different narrow moving jams in the pinch region can be very different to one another.

It must be noted that the well-known and very old term “stop-and-go” traffic, which is related to a sequence of moving traffic jams, will not be used in this article. For a traffic observer, both a sequence of narrow moving jams and a sequence of wide moving jams constitutes “stop-and-go” traffic. However, as already mentioned, narrow moving jams belong to the synchronized



**Traffic Congestion, Spatiotemporal Features of, Figure 17**

Simulations of GP at off-ramp (a) and on-ramp bottlenecks (b–d). a, b Average vehicle speed in space and time. c, d Speed (c) and flow rate (d) at different locations (1 min data averaging) for the GP in b. Taken from [38]

flow phase, whereas wide moving jams belong to the qualitative different wide moving jam phase.

### Diagram of Congested Patterns at Isolated Bottlenecks

A diagram of different congested patterns at an on-ramp bottleneck gives regions of congested pattern occurrence in the flow–flow plane whose coordinates are the flow rate in free flow on the main road upstream of the bottleneck  $q_{in}$  and the flow rate to the on-ramp  $q_{on}$  (Fig. 18a) [41,42,45]. There are two main boundaries in this diagram,  $F_S^{(B)}$  and  $S_J^{(B)}$ .

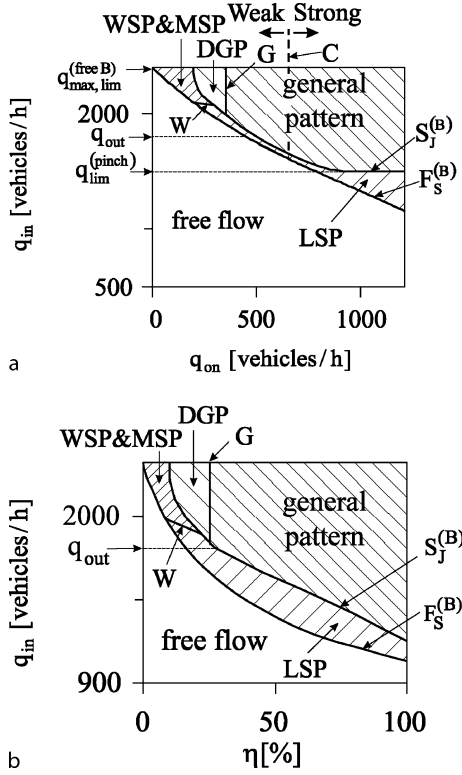
Below and left of the boundary  $F_S^{(B)}$  free flow occurs. At the boundary  $F_S^{(B)}$  within a given time interval for observing free flow at the bottleneck an  $F \rightarrow S$  transition occurs that leads to synchronized flow emergence.

Between the boundaries  $F_S^{(B)}$  and  $S_J^{(B)}$  different SPs (LSP, MSP, and WSP) emerge upstream of the on-ramp (Figs. 10, 12, and 14). Right of the boundary  $S_J^{(B)}$  wide

moving jams occur spontaneously in synchronized flow, i. e., different GPs appear (Figs. 17 and 19). At the boundary  $S_J^{(B)}$  within a given time interval for observing synchronized flow an  $S \rightarrow J$  transition occurs that leads to wide moving jam emergence.

Considering regions of SP emergence in the diagram, the following conclusions can be made. The WSP occurs above the boundary  $W$  in the diagram in Fig. 18a. Below the boundary  $W$  an LSP occurs. At greater  $q_{in}$  and a very small  $q_{on}$  an MSP can occur rather than the WSP. The MSP occurs spontaneously on the main road at the on-ramp bottleneck because at very low  $q_{on}$  the downstream front of synchronized flow can depart from the on-ramp bottleneck. The MSP begins to propagate on the main road as an independent localized structure (Fig. 12) on a homogeneous road. A sequence of MSPs can also occur on the main road at the on-ramp bottleneck.

When right of the boundary  $S_J^{(B)}$  at a given  $q_{in}$  the flow rate  $q_{on}$  increases, the average density within the pinch region of a GP increases too. There is a boundary labeled  $C$  in the diagram, which separates weak and strong conges-



**Traffic Congestion, Spatiotemporal Features of, Figure 18**  
Diagrams of congested patterns at on-ramp (a) and off-ramp bottlenecks (b). Results of numerical simulations. Taken from [38]

tion condition within the pinch region of GPs. Right of this boundary weak congestion at which the pinch region characteristic depends on  $q_{on}$  transforms into strong congestion at which the pinch region characteristics like the flow rate (this limit flow rate is denoted by  $q_{lim}^{(pinch)}$  in the diagram) and the frequency of wide moving jam emergence reach their limit values as observed in empirical data. As a result of strong congestion, there is saturation of the boundary  $S_j^{(B)}$  at greater  $q_{on}$  associated with the limit flow rate  $q_{lim}^{(pinch)}$  (Fig. 18a).

There is a region in the diagram that is between the boundary  $S_j^{(B)}$  and the line  $G$  within which a dissolving GP (DGP) occurs (Fig. 18a). In an DGP, after a wide moving jam has been formed this wide moving jam suppresses the pinch region in the initial GP. The suppression effect can occur when the initial flow rate  $q_{in}$  is high enough and the flow rate to the on-ramp  $q_{on}$  is lower than some characteristic value that determines the position of the boundary  $G$  in the diagram of congested patterns in Fig. 18a. To understand the physics of the DGP, let us consider flow rates  $q_{in}$  that satisfy the condition  $q_{in} > q_{out}$  ( $q_{out}$  is the flow rate in

the wide moving jam outflow under condition that in this jam outflow free flow is formed). After the first wide moving jam has been formed in the GP, the flow rate in the jam outflow cannot be greater than  $q_{out}$ . This means that rather than an initial high flow rate  $q_{in}$  a smaller flow rate related to the jam outflow determines the inflow into the pinch region of the GP. At a low enough flow rate to the on-ramp  $q_{on}$  (left of the boundary  $G$  in Fig. 18a), the pinch region cannot exist at this lower inflow into the pinch region: the pinch region dissolves.

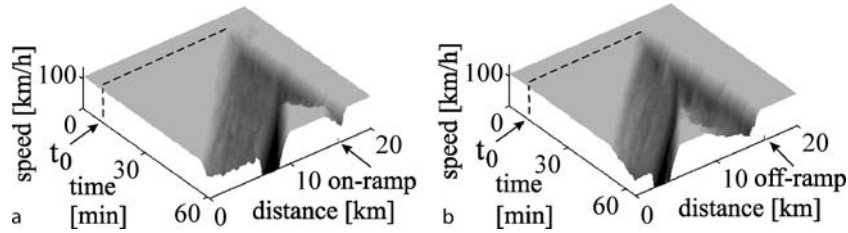
As a result of this dissolution effect, either free flow or an LSP appears at the on-ramp while the wide moving jam moves on the main road upstream of the on-ramp (Fig. 19). From the above consideration of the physics of a DGP, one can conclude that the boundary  $G$ , which separates DGPs and GPs in the diagram of congested patterns, should intersect the boundary  $S_j^{(B)}$  at the point  $q_{in} = q_{out}$ . This is correct if hysteresis effects are not taken into account. However, the hysteresis effect of GP formation leads to the result that the line  $G$  does not necessarily intersect the boundary  $S_j^{(B)}$  at the point  $q_{in} = q_{out}$  as is shown in Fig. 18a.

In contrast to the on-ramp bottleneck, at an off-ramp bottleneck (Fig. 17a) a GP *only* under weak congestion can be formed. This is true for all values  $\eta < 100\%$ , where  $\eta$  is the percentage of vehicles that want to leave the main road to the off-ramp in the flow rate  $q_{in}$ .

The pattern diagram at the off-ramp bottleneck, i.e., the regions of pattern formation in the plane  $(\eta, q_{in})$  (Fig. 18b) qualitatively resembles the diagram at the on-ramp bottleneck (Fig. 18a). However, there is no saturation of the boundary  $S_j^{(B)}$  at a higher  $\eta$ . This is associated with the mentioned fact that in GPs only the weak congestion condition occurs at all  $\eta < 100\%$ . Indeed, all pinch region characteristics in GPs do depend on  $\eta$ .

Note that GP parameters depend on the average flow rate  $q^{(cong)}$  within a GP at a bottleneck. The time interval of flow rate averaging of  $q^{(cong)}$  is suggested to be considerably longer than time distances between any moving jams within the GP. For the GP, the average flow rate  $q^{(cong)}$  is equal to the average flow rate within the pinch region of synchronized flow:  $q^{(cong)} = q^{(pinch)}$ . In the theory of GPs and the pinch effect [41,42,45] was found that the greater the bottleneck influence (called the bottleneck strength) on traffic is, the smaller  $q^{(pinch)}$ . In turn, the smaller  $q^{(pinch)}$  is, the greater the mean frequency of wide moving jams in a GP, and the lower the average speed between the wide moving jams.

In particular, for an on-ramp bottleneck at a given flow rate  $q_{in}$  the bottleneck strength is the greater, the greater



**Traffic Congestion, Spatiotemporal Features of, Figure 19**

Dissolving GP at on-ramp (a) and off-ramp (b) bottlenecks. Results of numerical simulations. Taken from [38]

the flow rate to the on-ramp  $q_{on}$ . For an Off-ramp bottleneck, at a given flow rate  $q_{in}$  the bottleneck strength is the greater, the greater the percentage of vehicles that leave the main road to the off-ramp  $\eta$ . However, as already mentioned for given on-ramp bottleneck (e. g., characteristics of vehicle merging from the on-ramp onto the main road) and traffic parameters (weather, percentage of long vehicles, etc.) and a given flow rate  $q_{in}$ , there is a saturation in decrease of  $q^{(pinch)}$  when  $q_{on}$  increases: beginning at a great enough  $q_{on}$ , the flow rate  $q^{(pinch)}$  reaches a limit (minimum) value  $q_{lim}^{(pinch)}$  [38,42]. In contrast, for the off-ramp bottleneck  $q^{(pinch)}$  decreases continuously when  $\eta$  increases up to  $\eta \rightarrow 100\%$ . At  $\eta = 100\%$  all vehicles leave the main road to the off-ramp. Thus if the lane number on the main road is greater than the lane number of the off-ramp, then at  $\eta = 100\%$  the off-ramp bottleneck can be considered a merge bottleneck caused by the lane reduction.

It can be seen from the diagrams of congested patterns (Fig. 18) that when the flow rates (and/or  $\eta$  for an off-ramp bottleneck) change, complex transformations between various SPs as well as between SPs and GPs occur. The latter is more probable under weak congestion conditions within the GPs. Moreover, there are also regions in these diagrams in which SPs and GPs are metastable patterns. This means that pattern transformation can also occur at given flow rates (and a given  $\eta$ ) as a result of one of the local phase transitions (e. g.,  $S \rightarrow F$ ,  $S \rightarrow J$ , or  $J \rightarrow S$  transitions) within an initial congested pattern (for more detail, see [38]).

Metastable free flow states with respect to SP formation as well as some of the metastable regions of SPs existence are considered in ► **Traffic Breakdown, Probabilistic Theory of** in connection with freeway capacity of free flow at a bottleneck, which is a very important traffic flow characteristic.

It should be noted that theoretical results discussed in Sects. “**Congested Patterns at Isolated Bottlenecks**” and “**Diagram of Congested Patterns at Isolated Bottle-**

**necks**” are taken from simulations of three-phase traffic flow models of [41,42,45]. Recently other various traffic flow models in the framework of the three-phase traffic theory have been developed [5,15,23,27,44,52,53]. Features of congested patterns that these models exhibit [6,7,8,15,23,25,26,27,44,52,53,62,71] are similar to those found earlier in [41,42,45].

### Complex Congested Patterns and Pattern Interaction

When there are two or more adjacent bottlenecks, there can be many new spatiotemporal phenomena. In particular, we consider the following empirical spatiotemporal features of congested patterns, which have been found out in empirical data [36,38]:

- (i) Induced  $F \rightarrow S$  transitions caused by upstream congested pattern propagation. A congested pattern, which occurs at the downstream bottleneck, can induce another congested pattern at the upstream bottleneck where free flow is realized before.
- (ii) The catch effect. The catch of synchronized flow at an effectual bottleneck with the subsequent formation of a new congested pattern at this bottleneck.
- (iii) An occurrence of expanded congested patterns (EP) where synchronized flow, which has initially appeared at the downstream bottleneck, affects the upstream bottleneck.
- (iv) The influence of foreign wide moving jam propagation on a congested pattern at the upstream bottleneck. These foreign wide moving jams occur downstream of the congested pattern within a GP at a downstream bottleneck.
- (v) An intensification of downstream congestion due to the onset of upstream congestion.

*Induced  $F \rightarrow S$  transition with catch effect.* In an empirical example shown in Fig. 11, when the MSP has not yet reached bottleneck  $B_2$ , free flow is there. After the MSP is caught at the bottleneck, synchronized flow occurs at

bottleneck  $B_2$ . This synchronized flow continues for a long time (more than 2 hours) at bottleneck  $B_2$ . The upstream front of synchronized flow is now localized at some finite distance upstream of bottleneck  $B_2$ . This means that after the MSP has induced another SP at the bottleneck, instead of MSP propagation through the bottleneck, the MSP is caught at the bottleneck. Thus, the catch of the MSP at the bottleneck causes an induced speed breakdown at the bottleneck (these two effects are labeled “catch effect and induced  $F \rightarrow S$  transition” in Fig. 11).

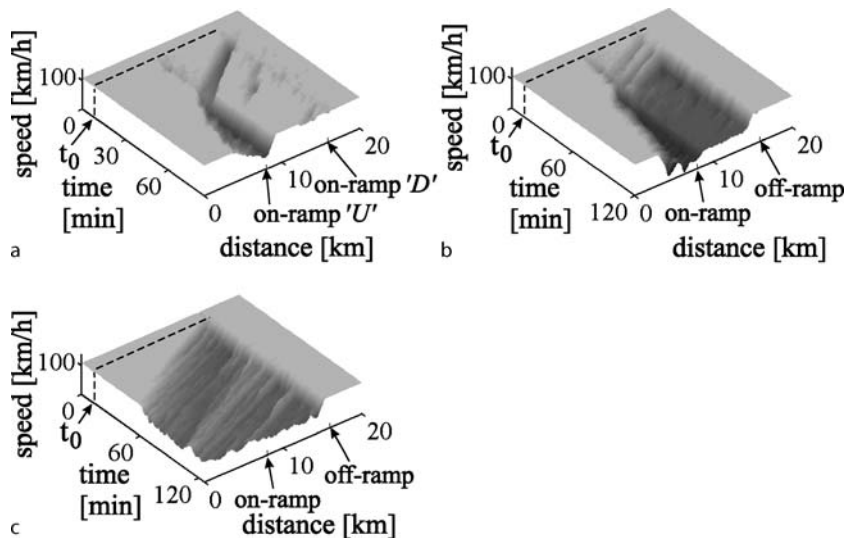
In the context of three-phase traffic theory, both induced traffic breakdown and catch effect can be found in accordance with empirical results. When two adjacent bottlenecks exist on a freeway and either an WSP or MSP appears at the downstream one, then the associated SP can be caught at the upstream bottleneck with the subsequent formation of a new congested pattern upstream of the upstream bottleneck (Fig. 20). For example, when the MSP shown in Fig. 20a reaches the upstream bottleneck, this synchronized flow causes an induced traffic breakdown at this upstream bottleneck. Synchronized flow is localized at some distance from the bottleneck, rather than the upstream front of this synchronized flow propagating further upstream continuously over time. This means that the initial WSP is caught at the upstream bottleneck. This catch effect causes the induced speed breakdown at the bottleneck. In another example, the upstream front of the WSP shown in Fig. 20b causes both induced traffic breakdown and catch effect after the upstream front of the WSP

reaches the upstream bottleneck. Indeed, we can see that the initial WSP is caught at the upstream bottleneck. This catch effect causes also the induced speed breakdown at the bottleneck.

*Expanded congested patterns (EP for short).* An EP is defined as a congested traffic pattern whose synchronized flow affects at least two adjacent bottlenecks. A typical empirical example of EP is shown in Fig. 13a,c: the WSP that has occurred initially at the downstream off-ramp bottleneck  $B_{\text{North1}}$  exists only for a relatively short time, which is determined by the propagation of the upstream front of the WSP to the upstream on-ramp bottleneck  $B_{\text{North2}}$ . In this case, the lifetime of the WSP is about 20 min. After synchronized flow of the WSP is caught at the upstream bottleneck, a new congested pattern is formed. Synchronized flow in this pattern covers both bottlenecks  $B_{\text{North1}}$  and  $B_{\text{North2}}$ . Thus in accordance with EP definition, this congested pattern is an EP.

If the above empirical congested pattern (Fig. 13) is considered during a considerably longer time interval and on the whole freeway section on which measurements are available, we find a typical complex EP with several wide moving jams propagating through the EP and two upstream bottlenecks  $B_{\text{North2}}$  and  $B_{\text{North3}}$  while maintaining the mean velocity for the downstream jam front that are the same for these different jams (jams labeled 1,...,6 in Fig. 21).

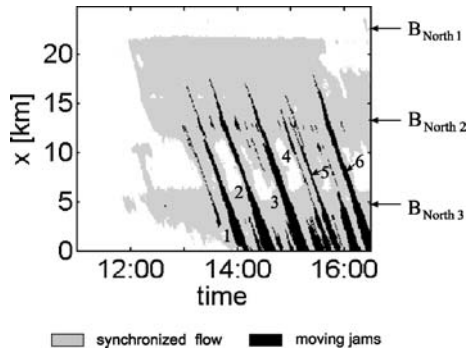
What is typical in this EP? Firstly, synchronized flow affects several effectual bottlenecks. Secondly, additionally



Traffic Congestion, Spatiotemporal Features of, Figure 20

Numerical simulations of induced pattern formation with the catch effect and expanded congested patterns (EPs). a, b Catch effect of MSP (a) and WSP (b) with induced formation of a new congested pattern at the upstream bottleneck. Taken from [38]





**Traffic Congestion, Spatiotemporal Features of, Figure 21**

Empirical example of an expanded congested pattern (EP). Free flow (white), synchronized flow (gray), moving jams (black). Taken from [38]

to wide moving jams that have emerged between bottlenecks  $B_{\text{North}1}$  and  $B_{\text{North}2}$ , new pinch regions in synchronized flow upstream of bottlenecks  $B_{\text{North}2}$  and  $B_{\text{North}3}$  are formed within which new moving jams emerge spontaneously. Many of these moving jams, however, are often suppressed by the initially wide moving jams propagating through the pinch region. An explanation of the wide moving jam suppression effect can be found in Sect. 7.6.3 of [38]. Thirdly, upstream and downstream of wide moving jams either synchronized flow or free flow can be formed within the same EP. Note that much more complex EPs can be observed, when effectual bottlenecks are located at short distances to each other (see Chap. 2 in [38]).

Simulations of a part of this EP (Fig. 22) made with the use of empirical road characteristics and flow rates are both qualitative and quantitative similar to the empirical EP (Figs. 15 and 21).

**Foreign wide moving jams.** A foreign wide moving jam is a wide moving jam, which has initially occurred downstream of an effectual bottleneck and propagates through synchronized flow of another congested pattern upstream of this bottleneck.

We mentioned above that wide moving jams in Fig. 21, which initially emerged upstream of the downstream bottleneck  $B_{\text{North}1}$ , propagate through the upstream bottleneck  $B_{\text{North}2}$  while maintaining the velocity of the downstream jam front. These wide moving jams subsequently propagate through the next upstream bottleneck  $B_{\text{North}3}$  as well.

Upstream of the bottlenecks  $B_{\text{North}2}$  and  $B_{\text{North}3}$  there are synchronized flows of other congested patterns formed. Thus the wide moving jams labeled 1, 2, ..., 6 in Fig. 21 that initially emerged downstream of the bottleneck  $B_{\text{North}2}$  are *foreign* wide moving jams when they propagate through synchronized flows upstream of the bottle-

neck  $B_{\text{North}2}$  or the upstream of bottleneck  $B_{\text{North}3}$  [38]. An example of numerical simulations of foreign wide moving jam propagation is shown in Fig. 22.

The identification of foreign wide moving jams is associated with the jam interaction effect; a foreign wide moving jam can exert a considerable influence on other moving jams that are just emerging in synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.

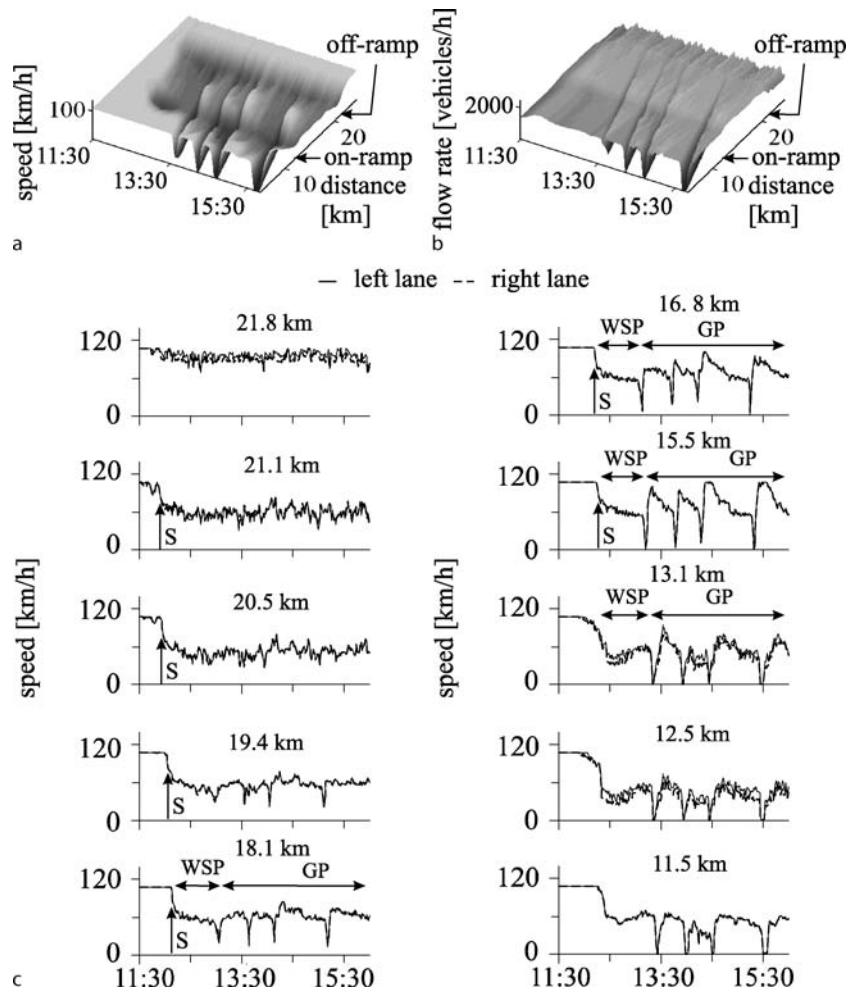
It can turn out that in addition to foreign wide moving jams, new wide moving jams can emerge in the pinch region of synchronized flow upstream of the upstream bottleneck. As a result, a “united” sequence of wide moving jams is formed from the joining of the foreign wide moving jams and the wide moving jams that have emerged in synchronized flow upstream of the upstream bottleneck.

It should be noted that when the distances between adjacent effectual bottlenecks are great enough, spatially separated regions of synchronized flow can occur between adjacent effectual bottlenecks where the onset of congestion is realized. Often pinch regions are formed in these synchronized flows, leading to GP emergence. These GPs at different adjacent bottlenecks have spatially separated synchronized flow regions. We call these GPs “spatially separated” GPs [38].

An empirical example is shown in Fig. 23, where two spatially separated GPs occur at adjacent effectual bottlenecks  $B_2$  and  $B_3$ . Synchronized flow upstream of the GP at bottleneck  $B_2$  does not cover bottleneck  $B_3$ , i. e., no EP occurs between them. In other words, a congested pattern can consist of several spatially separated GPs, if there is enough distance between adjacent effectual bottlenecks (distances should usually be more than 4–5 km).

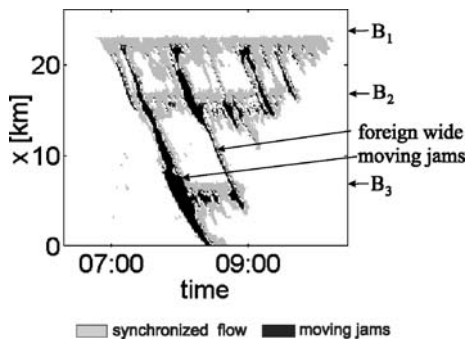
There can also be an intermediate case. An example can be seen between downstream bottleneck  $B_1$  and upstream bottleneck  $B_2$  in Fig. 23. We find that during some time intervals synchronized flow from the downstream GP at bottleneck  $B_1$  covers bottleneck  $B_2$ , i. e., EPs appear. During other time intervals two spatially separated GPs at adjacent bottlenecks  $B_1$  and  $B_2$  are realized.

However, wide moving jams have emerged upstream of the downstream bottleneck  $B_1$ . The wide moving jams propagate through adjacent bottlenecks  $B_2$  and  $B_3$  while maintaining the velocity of the downstream jam fronts. These wide moving jams are foreign wide moving jams when they are upstream of adjacent bottlenecks  $B_2$  and  $B_3$ , where other moving jams are emerging. Thus, these GPs have only spatially separated regions of synchronized flow. In contrast to the synchronized flow regions, wide moving jams of the GP at the downstream bottleneck propagate



**Traffic Congestion, Spatiotemporal Features of, Figure 22**

Simulations of a part of empirical EP shown in Figs. 15 and 21: a, b Speed (a) and flow rate (b) in space and time. c Time-functions of speed at different locations. Taken from [46]



**Traffic Congestion, Spatiotemporal Features of, Figure 23**

Empirical example of foreign wide moving jam propagation. Free flow (white), synchronized flow (gray), and moving jams (black) in space and time. Taken from [38]

through the GP at the upstream bottlenecks. As already mentioned, this foreign jam propagation can considerably influence moving jam emergence at the upstream bottleneck. This usually occurs when spatially separated GPs appear at the freeway. Thus, due to foreign wide moving jam propagation, spatially separated GPs can form a complex congested pattern on the freeway (Fig. 23).

*Intensification of downstream congestion due to upstream congestion.* It can be expected that upstream congestion (congested pattern formation at an upstream bottleneck) should tend to a reduction in downstream congestion. This is because at the same traffic demand the flow rate in free flow downstream of the congested bottleneck (discharge flow rate) is usually lower than the initial flow rate in free flow at the same freeway location before

a congested pattern at the upstream bottleneck has been formed.

Nevertheless in some cases, we will see that rather than a *reduction* in downstream congestion, an *intensification* of downstream congestion can occur due to upstream congestion. This is the result of complex nonlinear interactions among congested patterns occurring at *different spatially separated* adjacent effectual bottlenecks.

To discuss an example of this effect, we consider a road with two on-ramp bottlenecks (Fig. 24). There are downstream and upstream on-ramp bottlenecks labeled 'D' and 'U', respectively. The initial flow rate in free flow upstream of the on-ramp 'D'

$$q_{in}^{(down)} = q_{in} + q_{on}^{(up)} \quad (1)$$

in Fig. 24a are equal to each other in Fig. 24a,b. However, in Fig. 24a the flow rate

$$q_{on}^{(up)} = 0, \quad (2)$$

whereas in Fig. 24b the flow rate

$$q_{on}^{(up)} > 0. \quad (3)$$

The equality of the flow rate  $q_{in}^{(down)}$  (1) in Fig. 24a and b is realized due to greater flow rate  $q_{in}$  in free flow upstream of the upstream bottleneck in Fig. 24a in comparison with the flow rate  $q_{in}$  in Fig. 24b.

An initial DGP at the on-ramp 'D', where *only one* wide moving jam emerges (Fig. 24a), transforms into a GP at this bottleneck where an *uninterrupted sequence* of wide moving jams emerges (Fig. 24b).

To explain this intensification of downstream congestion, note that after the wide moving jam in the DGP (Fig. 24a) emerges, the current flow rate upstream of the on-ramp 'D',  $q_{in}^{(down)}$ , has dropped to  $q_{out}$ :

$$q_{in}^{(down)} = q_{out}. \quad (4)$$

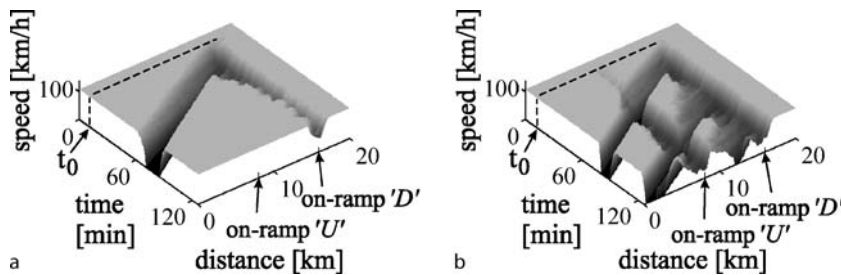
This decrease in the flow rate  $q_{in}^{(down)}$  occurs because vehicles merging onto the main road from the on-ramp 'U' must slow down when they reach the wide moving jam. In other words, when the wide moving jam is between the on-ramps 'D' and 'U' the flow rate  $q_{on}^{(up)}$  does not lead to an increase in the flow rate  $q_{in}^{(down)}$ : due to the flow rate  $q_{on}^{(up)}$  only the jam inflow increases. In the example, the jam outflow  $q_{out} = 1810$  vehicles/h. At the flow rate  $q_{in}^{(down)} = 1810$  vehicles/h the GP cannot exist upstream of the on-ramp 'D' at the chosen flow rate to the on-ramp 'D',  $q_{on}^{(down)}$  (Fig. 24a).

In Fig. 24b, an LSP is induced at the on-ramp 'U' after the wide moving jam from a DGP (downstream congestion) has reached the on-ramp 'U'. After the wide moving jam has passed the on-ramp 'U', there are no wide moving jams between the on-ramps 'D' and 'U'. For this reason, vehicles merging onto the main road from the on-ramp 'U' cause an increase in the flow rate  $q_{in}^{(down)}$  to the value  $q_{in}^{(down)} = 2150$  vehicles/h. This flow rate is much higher than 1810 vehicles/h. This leads to subsequent wide moving jam emergence at the on-ramp 'D', and so on. As a result, a GP is formed at the on-ramp 'D' (Fig. 24b).

The mean period of moving jam emergence in the GP is determined by the time of wide moving jam propagation to the on-ramp 'U'. Before a wide moving jam has passed the on-ramp 'U' the condition (4) is satisfied, and no moving jam can emerge at the given flow rate  $q_{on}^{(down)}$  at the on-ramp 'D'.

### Reproducible and Predictable Congested Patterns

It has been found that for each effectual bottleneck or each set of several adjacent effectual bottlenecks, where congested patterns occur, the spatiotemporal structure of congested patterns exhibits predictable, i. e., characteristic, unique, and reproducible features [38]. These predictable and reproducible pattern features can be almost *the same*

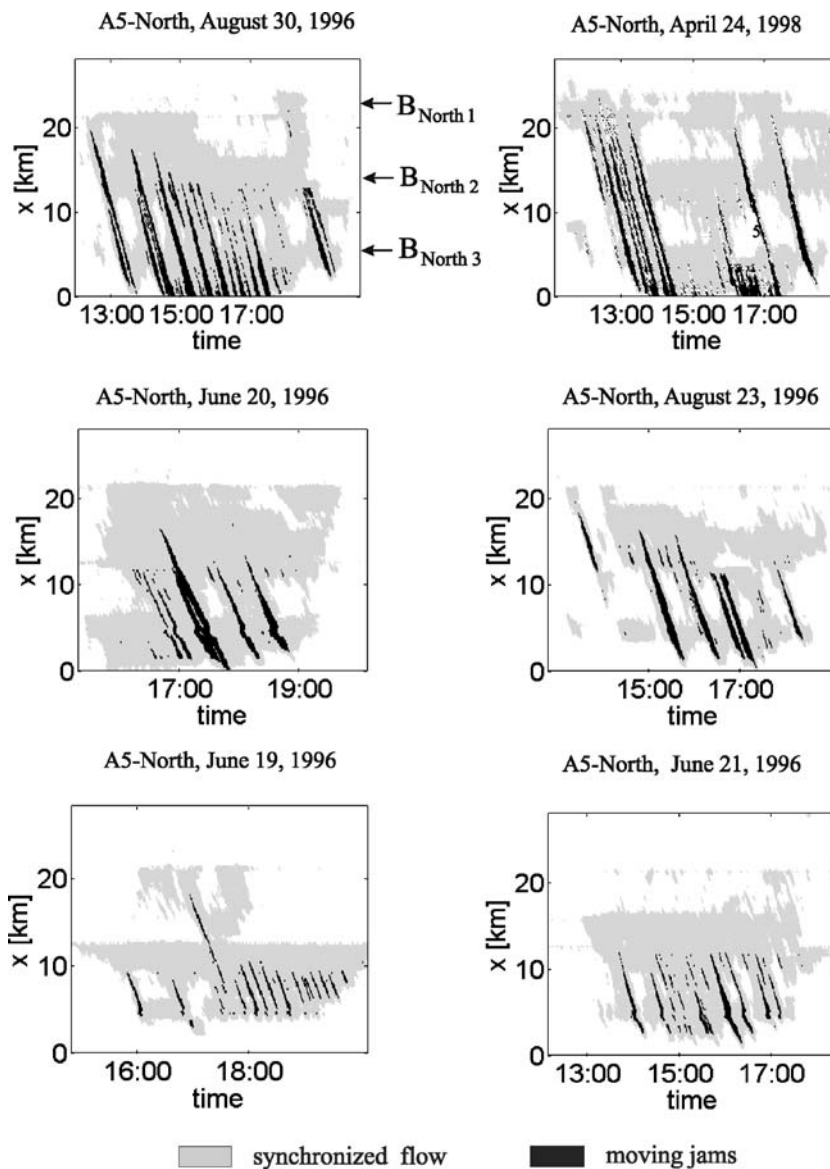


Traffic Congestion, Spatiotemporal Features of, Figure 24

Simulations of intensification of upstream congestion due to downstream congestion. Speed on the main road in space and time: a, b DGP (a) transforms into a GP (b). Taken from [38]

for different days and years. They can also persist over a large range of flow rates (traffic demand) at which the patterns exist. The predictable and reproducible pattern features remain in traffic flows with very different driver behavioral characteristics and different vehicle parameters. The predictable and reproducible pattern features can be used in the forecast of congested patterns at freeway bottlenecks. Overviews of some of the reproducible congested pattern features can be seen in Fig. 25.

There are pattern features that have a probabilistic nature. This means that in different realizations (days) congested pattern features can be different for the same traffic control parameters (the percentage of long vehicles, weather, other road conditions), traffic demand, and initial conditions. However, we can identify different “degrees of predictability” for different pattern features. Some “deterministic” pattern features and pattern features with a high “degree of predictability” are as follows:



**Traffic Congestion, Spatiotemporal Features of, Figure 25**

Empirical examples of predictable and reproducible patterns. Overviews of free flow (white), synchronized flow (gray), and moving jams (black) in space and time. The effectual bottlenecks are the same as those in Fig. 21 [38]

- (1) Traffic breakdown ( $F \rightarrow S$  transition) occurs mostly at the same effectual freeway bottlenecks on different days and years.
- (2) In contrast to synchronized flow, which occurs mostly at bottlenecks, wide moving jam emergence is usually observed in synchronized flow at a finite distance upstream of bottlenecks, if distances between the bottlenecks are large enough.
- (3) Wide moving jams propagate through any complex state of traffic and through bottlenecks at the mean velocity of the downstream jam front.
- (4) If free flow emerges in the wide moving jam outflow, then the flow rate in this jam outflow is considerably lower than the maximum possible flow rate in free flow.
- (5) The catch effect: a local synchronized flow region reaching a bottleneck is caught at the bottleneck, when free flow at the bottleneck is in a metastable state to traffic breakdown.
- (6) If a narrow moving jam is very close to the downstream front of a foreign wide moving jam, the foreign jam suppresses the growth of this narrow jam.

Some traffic pattern features with a middle “degree of predictability” are as follows:

- (i) The instant of spontaneous traffic breakdown at an effectual bottleneck at the same time-dependences of upstream flow rates. This reveals the probabilistic nature of an  $F \rightarrow S$  transition.
- (ii) For a bottleneck or for a set of adjacent effectual bottlenecks that are close to one another there is a certain type of spatiotemporal congested pattern that appears with the highest probability, given the same traffic demand.

Some pattern features with a low “degree of predictability” are as follows:

- (a) Whether a spontaneous or an induced traffic breakdown occurs, leading to synchronized flow at an effectual bottleneck.
- (b) The instant of wide moving jam emergence in synchronized flow.

### Microscopic Features of Traffic Phases

#### Moving Blanks Within Wide Moving Jams

The spatiotemporal criterion for a wide moving jam [J] of Sect. “[Introduction](#)” that defines the wide moving jam phase in congested traffic can be explained by a traffic flow

interruption that occurs when vehicles are in a standstill within the jam. A sufficient criterion for this flow interruption within the jam is [47]

$$I_s = \frac{\tau_{\max}^{(a)}}{\tau_{\text{del}}^{(a)}} \gg 1, \quad (5)$$

where  $\tau_{\max}$  is the maximum time headway between two vehicles within a wide moving jam and  $\tau_{\text{del}}^{(a)}$  is the mean time delay in vehicle acceleration at the downstream jam front from a vehicle standstill;  $\tau_{\text{del}}^{(a)}$  determines the jam outflow; corresponding to empirical results  $\tau_{\text{del}}^{(a)} \approx 1.5\text{--}2$  sec [38].

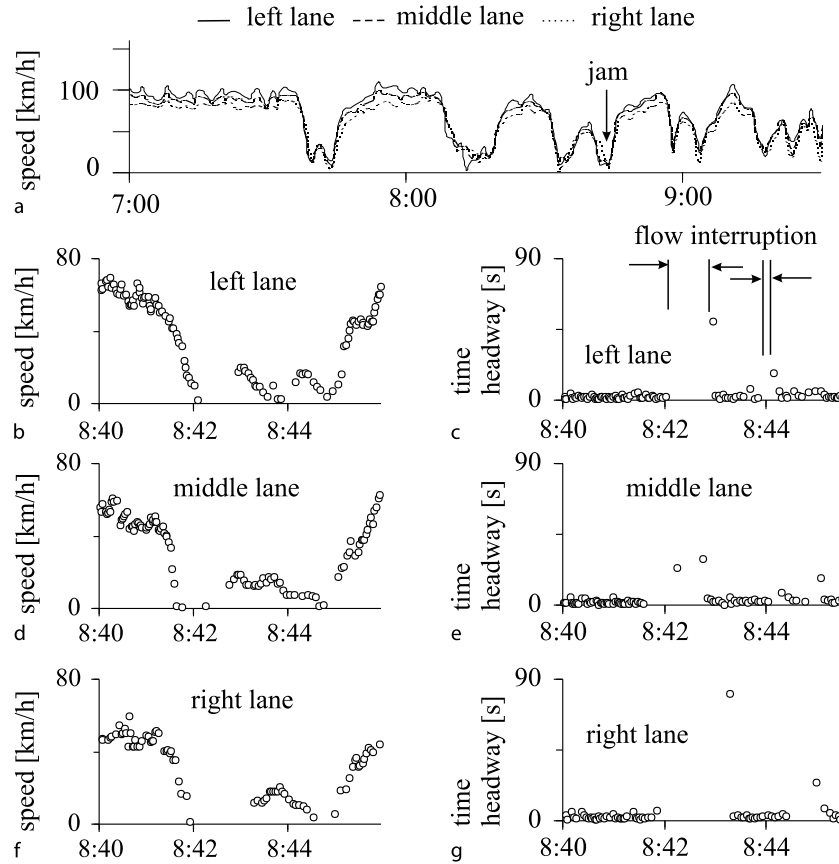
The flow interruption effect is a general effect for each wide moving jam (see examples of flow interruption intervals labeled by “flow interruption” in Figs. 26c and 27a). For this reason, the criterion (5) can be considered as a microscopic criterion for the wide moving jam phase (more detailed explanations of flow interruption intervals and the criterion (5) for wide moving jams appear in ► [Traffic Congestion, Modeling Approaches to](#) and Sect. “[Congested Patterns at Heavy Bottlenecks](#)”). Between flow interruption intervals within the wide moving jam shown in Fig. 26b–g, vehicles within the jam exhibit time headways about 1.5–7 sec (Fig. 27b). These time headways are considerably shorter than flow interruption intervals (about 20 sec and longer) in Fig. 26c,e,g. The time headways are related to low speed states measured at detectors within the jam (Fig. 26b,d,f).

To understand the effect of these low speed states, note that when vehicles meet the wide moving jam, firstly they decelerate to a standstill at the upstream jam front. As a result, the first flow interruption interval in all lanes appears (Figs. 26c,e,g and 27a). Net distances (space gaps) between these vehicles can be very different and the mean space gap can exceed a minimum (safe) space gap considerably, i. e., regions with no vehicles can appear within the jam. These regions with no vehicles are called *blanks* within the wide moving jam.

Later vehicles move covering these blanks. This low speed vehicle motion is responsible for low speed states mentioned above (Fig. 26b,d,f). Consequently, due to this vehicle motion new blanks between vehicles occur upstream, i. e., the blanks move upstream within the jam. Then other vehicles within the jam that are upstream of these blanks begin also to move covering the latter blanks. This leads to moving blanks that propagate upstream within the jam.

Thus we define a *moving blank* as a blank between vehicles, which moves upstream due to vehicle motion within the jam. The vehicle motion occurs at a low vehi-





**Traffic Congestion, Spatiotemporal Features of, Figure 26**

Measured single vehicle data analysis: **a** Overview of measured data (1 min average data). **b–g** Single vehicle data for speed in three freeway lanes for a wide moving jam (left figures) labeled by “jam” in **a** and the related time headways (right figures). Taken from [47]

cle speed, i. e., the vehicle motion creating moving blanks within the jam is related to low speed states discussed above (Fig. 26b,d,f); these low speed states are the reason for moving blank emergence within the jam. This explains why moving blanks are associated with low speed states within the jam.

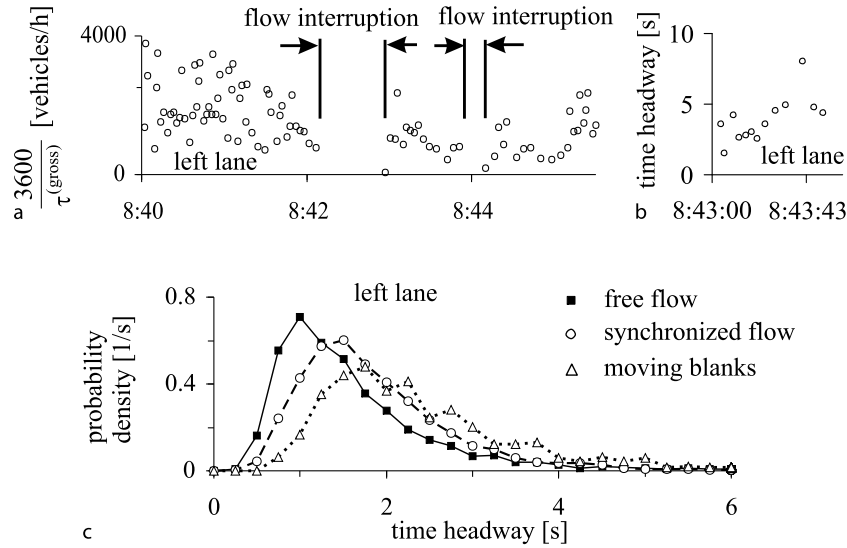
Time headways  $\tau$  between vehicles in these low speed states with moving blanks are about 1.5–7 sec (Fig. 27b). Because these low speed states are the reason for moving blanks, we will refer the time headways in the low speed states to moving blanks. Thus the term *time headways for moving blanks* means the time headways in the low speed states causing upstream motion of blanks.

To explain possible short time headways associated with moving blanks observed in empirical data (about 1.5 sec), note that when vehicles begin to cover a blank within a wide moving jam, they start to move from a standstill within the jam. Thus, some of the vehicles can move with time headways between each other that are close to

a time delay in vehicle acceleration from a standstill within the jam whose average value is  $\tau_{\text{del}}^{(a)} \approx 1.5\text{--}2$  sec.

Time headways in low speed states within the jam, i. e., time headways associated with moving blanks (about 1.5–7 sec in Fig. 27b) must be distinguished from long time intervals of traffic flow interruption within the jam (about 20 sec and longer) (Figs. 26c,e,g and 27a). The flow interruption intervals within the jam occur when some of the vehicles within the jam do not move. This vehicle stoppage does not indicate whether there are some blanks between these vehicles or not. This is because during these long flow interruption intervals space gaps between vehicles that are stopped within the jam might be quite small.

Thus at a fixed freeway location, within a wide moving jam, long time intervals, in which no vehicle passes the location (flow interruption intervals), alternate with time intervals in which vehicles pass the location with low speeds and time headways of about 1.5–7 sec associated with moving blanks within the jam. This empiri-



**Traffic Congestion, Spatiotemporal Features of, Figure 27**

Measured microscopic characteristics of traffic phases associated with single vehicle data shown in Fig. 26b–g: **a** Time distributions of the value  $3600/\tau^{(\text{gross})}$  in the left lane ( $\tau^{(\text{gross})}$  is the gross time headway). **b** Time headways associated with moving blanks in the left lane related to the jam shown in Fig. 26b. **c** Probability density of time headways in the left lane for free flow (solid curve), synchronized flow (dashed curve), and moving blanks within wide moving jams (dotted curve). Taken from [47]

cal result could be associated with the following microscopic jam space structure: freeway regions in which vehicles are in a standstill (traffic flow interruption) alternate with freeway regions in which blanks moving upstream (moving blanks) occur resulting from vehicle motion with low speeds.

Empirical data show that low speed states within wide moving jams associated with moving blanks are not necessarily synchronized between different lanes (Fig. 26b,d,f).

Moving blanks within a wide moving jam resemble electron holes in the valence band of semiconductors. As the moving blanks that propagate upstream appear due to downstream vehicle motion within the jam, so appearance of electron holes moving with the electric field results from electron motion against the electric field in the valence band of semiconductors. However, an electron hole is associated with a single vacancy for one electron in the valence band. In contrast, there can be wide (in the longitudinal direction) moving blanks associated with a “vacancy” for two or more vehicles within a wide moving jam. This moving blank can be split into two or more moving blanks during upstream blanks motion. On the other hand, several adjacent moving blanks can merge into one moving blank (a more detailed consideration of the spatiotemporal dynamics of moving blanks and flow interruption intervals appear in Sect. “Congested Patterns at Heavy Bottlenecks”).

One of the microscopic characteristics of the traffic phases is time headway distributions. In particular, it has been found that the mean time headway for free flow is shorter than for congested traffic (see references in [20]). This can be explained at least by the following peculiarities of free flow and congested traffic:

- (i) A vehicle can come very close to the preceding vehicle just before the vehicle passes it. However, a road detector cannot recognize whether a very short time headway that has been measured is related to car-following or to this vehicle passing effect.
- (ii) A vehicle can easily change the lane, if the preceding vehicle would decelerate unexpectedly; thus, the vehicle can choose a shorter time headway in car-following than in congested traffic.
- (iii) There is a frustration effect in congested traffic that can be responsible for some long time headways [20].

There can also be another reason for shorter mean time headways in free flow in comparison with mean time headways in congested traffic. It has been found that there are different distributions of time headways in synchronized flow and moving blanks associated with wide moving jams (Fig. 27c). It can be assumed that as long as vehicles are within a wide moving jam, the vehicles have no hurry while covering blanks with the jam.

### Traffic Phases in Flow-Density Plane

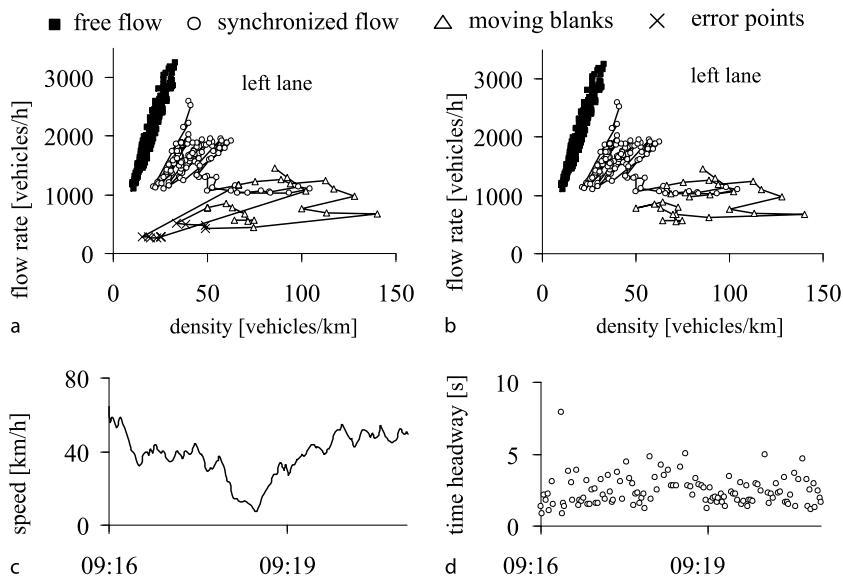
There are a huge number of publications associated with an analysis of measured traffic data in the flow-density plane (see references in [2,20,57]). One of the aims of this analysis is to find criteria for a classification of qualitatively different traffic states and phases associated with congested traffic (see, e.g., [49,50,60]). However, as mentioned above traffic is a complex process in space and time, i.e., rather than a data analysis in the flow-density plane, to find qualitatively different traffic states and phases one should firstly study spatiotemporal features of traffic congested patterns. Otherwise, invalid conclusions about features of traffic states and phases can be made very easily.

To illustrate this, we consider single vehicle characteristics of low vehicle speeds within a wide moving jam associated with moving blanks (cross and triangle points) with free flow (black squares) and synchronized flow (circles) in the flow-density plane (Fig. 28a). Note that in contrast with single-vehicle data for a wide moving jam shown in Fig. 26b–g, which are used for a calculation of data for moving blanks shown in Fig. 28a, there are no flow interruption intervals in low speed data shown in Fig. 28c,d, i.e., the latter data are associated with the synchronized flow phase.

As well-known [2,20,57], in free flow the flow rate is on average an increasing density function in the flow-density plane (black squares in Fig. 28a). Due to small density states associated with low vehicle speeds within wide moving jams (cross points in Fig. 28a), the flow rate might also be considered on average an increasing density function in the flow-density plane for wide moving jams. For synchronized flow (Fig. 28a), measured points for synchronized flow exhibit the great scattering in the flow-density plane. These results have been used in [49,50,60] for the definition of a criterion for the traffic phases in measured data of congested traffic based on the flow-density correlation function

$$cc_{\rho,q}(\tau) = \frac{\langle \rho(t)q(t+\tau) \rangle - \langle \rho(t) \rangle \langle q(t+\tau) \rangle}{\sqrt{\langle \rho^2(t) \rangle - \langle \rho(t) \rangle^2} \sqrt{\langle q^2(t) \rangle - \langle q(t) \rangle^2}}. \quad (6)$$

In accordance with [49,50,60], small flow-density correlations in the data should be associated with synchronized flow, while large flow-density correlations in the data for congested traffic – with wide moving jams. Indeed, for data shown in Fig. 28 we find that the flow-density correlation coefficient (6) for synchronized flow is a very small value  $cc_{\rho,q}^{(\text{syn})}(0) \approx 0.06$ , whereas for the wide moving jam  $cc_{\rho,q}^{(\text{jam})}(0) \approx 0.67$ , i.e., it is comparable with the correlation coefficient for free flow  $cc_{\rho,q}^{(\text{free})}(0) \approx 0.97$ .



**Traffic Congestion, Spatiotemporal Features of, Figure 28**

Empirical characteristics of traffic phases: **a** Traffic phases in the flow-density plane for data in the left lane. **b** The same traffic phases as those in **a**, however, with improved data for moving blanks. Microscopic data for moving blanks, for synchronized flow and free flow used in **a**, **b** are associated with single-vehicle data, respectively, for wide moving jam shown in Fig. 26b,c, for synchronized flow shown in **c**, **d**, and for free flow shown in Fig. 26a. **c**, **d** Single-vehicle speeds (**c**) and time headways (**d**) for synchronized flow states used in **a**, **b** that are associated with data shown in Fig. 26a. Moving averaging over five vehicles is used. Taken from [47]

However, these conclusions of the definition for the traffic phases of [49,50,60] based on the flow-density correlation coefficient (6) are *invalid*. As explained in [47], the increasing density function (Fig. 28a) for wide moving jams in the flow-density plane as well as the associated great value of flow-density correlations are associated with a *large systematic error* in the data measured at road detectors. This error is due to incorrect density estimation within moving jams made in [49,50,60]. To understand this critical conclusion, recall that there are traffic flow interruption intervals within the wide moving jam (Fig. 26b–g). This leads to an incorrect calculation of the average speed  $v$  within the jam. The average speed measured by a detector can be related to a time interval, which includes at least one traffic flow interruption interval. In this case, the measured average speed is considerably higher than the real average speed within the jam. This is because vehicles do not move during traffic flow interruption intervals. As a result, the subsequent density estimation through the formula  $\rho = q/v$  ( $q$  the flow rate) leads to very small densities within the jam. However, the real density of vehicles stopped within the jam during traffic flow interruption is very large. Thus the measured data associated with the traffic flow interruption effect within wide moving jams cannot be used for density estimation within the jams. To estimate real vehicle density within a wide moving jam during time intervals that include flow interruption intervals, the density definition (number of vehicles per freeway length at a given time moment), i. e., the *spatial averaging* should be used, rather than density estimation through the use of the formula  $\rho = q/v$  in which the flow rate and average speed are related to the *time averaging* of vehicles passing a detector during a given time interval. The conclusion about the large systematic error in density estimation within wide moving jams is confirmed by a numerical analysis discussed in Subsect. “[Numerical Simulations of Moving Blanks Within Wide Moving Jams](#)”.

This analysis shows also that within wide moving jams density estimation  $\rho = q/v$  exhibits a relative small error during vehicle motion associated with moving blanks within the jams. This is because in this case the associated time headways are related to vehicles moving within the jams. If only these single vehicle data is used for density estimation, then points for moving blanks within the jam appear that are associated with random transformations in different directions in the flow-density plane (triangle points in Fig. 28b) rather than an increasing function of the flow rate with density (cross and triangle points in Fig. 28a). This is explained by low vehicle speed states that are associated with non-regular vehicle motion cov-

ering blanks within the wide moving jam. If error measurement points are removed (Fig. 28b), then the flow-density correlation coefficient for the wide moving jam  $cc_{\rho,q}^{(\text{jam})}(0) \approx 0.18$ , i. e., it is small. This means that in contrast with [49,50,60] no valid traffic phase definition based on the flow-density correlations in congested traffic can be made.

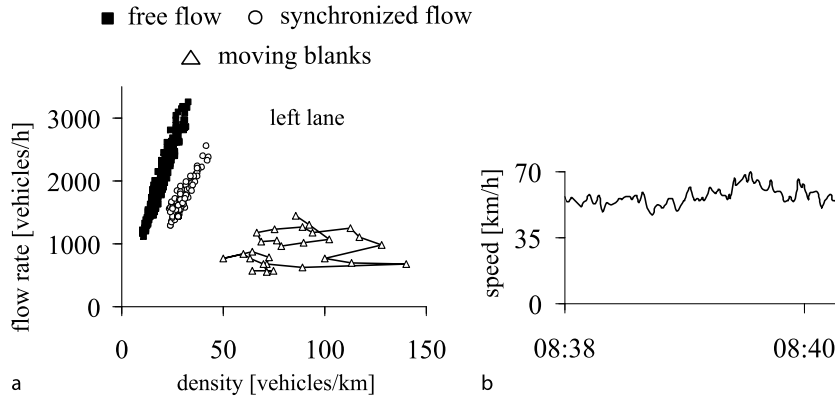
In addition, we should mention that synchronized flow states can overlap states associated with moving blanks within a wide moving jam as this is shown in Fig. 28b. This is because the speed in synchronized flow states can be as low as the speed in states associated with moving blanks within the jam. Thus in a general case low speed states associated with moving blanks within wide moving jams cannot also be used for clearly distinguishing the synchronized flow and wide moving jam phases in congested traffic.

Moreover, the result mentioned above that synchronized flow exhibits a small flow-density correlation, as this is the case for measured data shown in Fig. 28b,c, is *not* a general one. An empirical example of synchronized flow, for which the flow-density correlation coefficient  $cc_{\rho,q}^{(\text{syn})}(0) \approx 0.91$  is as great as the one for free flow, is shown in Fig. 29. Thus we see that in contrast with [49, 50,60] a great value of the flow-density correlation cannot be used as the criterion for distinguishing free flow from synchronized flow.

### Numerical Simulations of Moving Blanks Within Wide Moving Jams

To understand features of propagation and evolution of moving blanks within a wide moving jam, we consider a moving jam that is induced downstream of an on-ramp bottleneck by applying of a short-time and high amplitude local speed disturbance in heterogeneous free flow consisting of fast and long vehicles (Fig. 30a,b). There are two time intervals in which flow interruption within the jam occurs. The maximum time headways (Fig. 30c) associated with these flow interruption effects satisfy the microscopic criterion (5) for wide moving jams ([47], ► [Traffic Congestion, Modeling Approaches to](#)) (time delays used in (5) for fast and long vehicles are  $\tau_{\text{del}}^{(a)} \approx 1.77$  and 3.33 sec, respectively). As a result, in accordance with the macroscopic spatiotemporal criterion [J] for the wide moving jam phase, this jam propagates through the bottleneck while maintaining the mean velocity of the downstream jam front.

Within the wide moving jam, between time intervals of flow interruptions there are intervals in which low speed states appear associated with moving blanks (Fig. 30b–d).



#### Traffic Congestion, Spatiotemporal Features of, Figure 29

Empirical characteristics of traffic phases: **a** Traffic phases in the flow-density plane for data in the left lane with improved data for moving blanks. Data for free flow and wide moving jam are the same as those in Fig. 28b. **b** Single-vehicle speeds for synchronized flow used in **a** are associated with data shown in Fig. 26a. Moving averaging over five vehicles is used

Simulated time headways for moving blanks that are about 2.5–7.5 sec in the left lane (Fig. 30d, left) correspond to empirical observed values (Fig. 27). As can be seen from vehicle trajectories within the jam, these low speed states associated with moving blanks are related to non-homogeneous vehicle motions that cover blanks between vehicles within the jam (Fig. 31). These moving blanks propagate upstream with a negative velocity that is on average equal to the characteristic speed of the downstream jam front. Spatial and temporal speed distributions in low speed states associated with moving blanks exhibit complex variations within the jam, which are different in the left and right lanes (Fig. 31b,c); these variations seem to correspond to a random vehicle speed behavior within the jam.

In accordance with empirical results of Subsect. “Traffic Phases in Flow-Density Plane”, due to flow interruption within the wide moving jam shown in Fig. 30a–c, there is a large error in density distributions calculated through the formula  $\rho = q/v$  at greater density (curves 2 in Fig. 30e) in comparison with density distributions found based on the density definition (vehicles per freeway length) (curves 1). Indeed, curves 2 in Fig. 30e show a significant density underestimation within moving jams. These error points (crosses in Fig. 32a) explain the systematic error in the empirical studies of states within wide moving jams made in [49,50,60] that has been illustrated in Fig. 28a (compare error points in Figs. 28a and 32a).

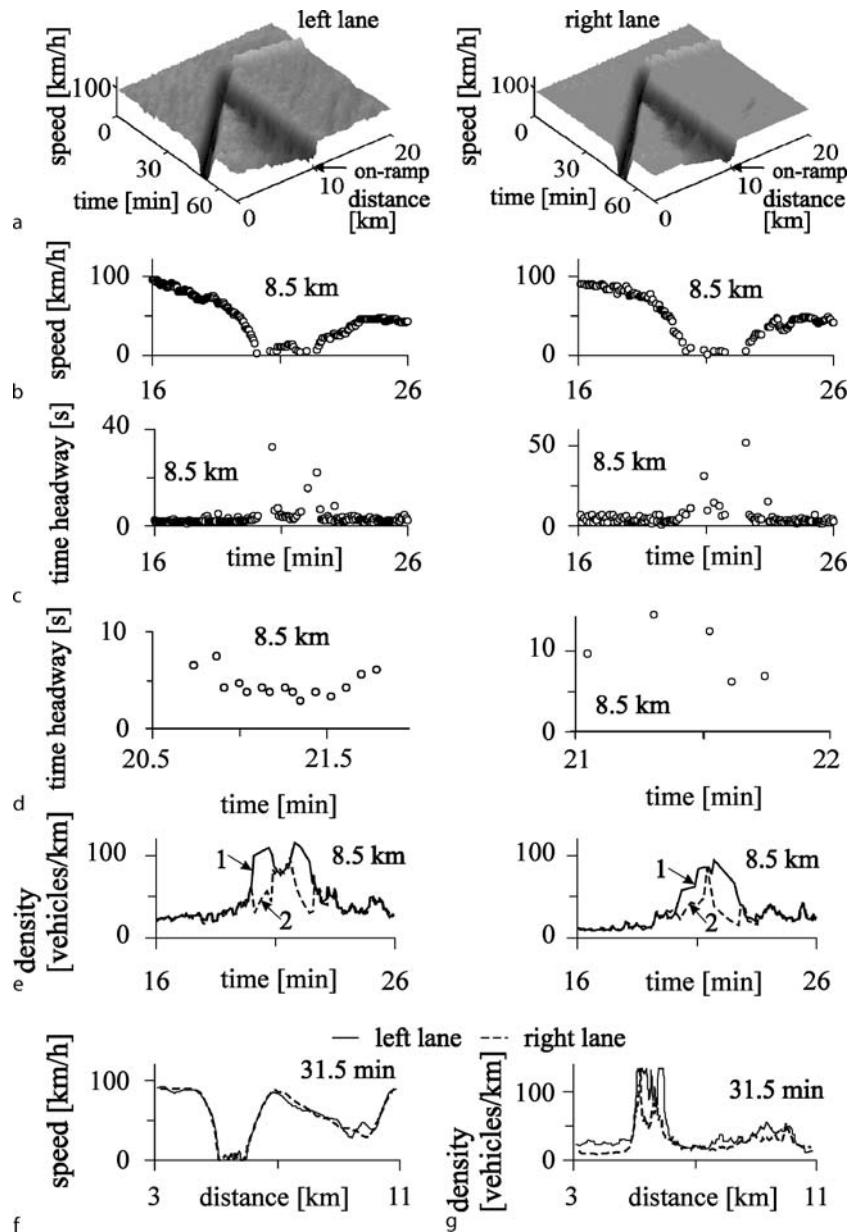
Within the regions of moving blanks, the difference between density calculated through the formula  $\rho = q/v$  (curves 2 in Fig. 30e) and density found based on the density definition (curves 1) decreases considerably. For

this reason, if error points associated with flow interruption within the jam are removed, then remaining points in the flow-density (Fig. 32b) exhibit a qualitative correspondence with the states within the jam calculated through the density definition (Fig. 32c). The latter states are related to the speed and density found at a fixed time moment 31.5 min (Fig. 30f,g). Nevertheless, there are some quantitative differences between speed and density distributions found through detector measurements (Fig. 32b) and through the use of the density definition in Fig. 32c. In particular, there are points in the flow-density plane found in the density calculation through the density definition (Fig. 32c), which are related to greater density up to the maximum jam density. These points cannot usually be found based on traffic measurements at a detector location (Fig. 32b).

To understand possible scenarios of emergence of moving blanks within a wide moving jam, spontaneous jam emergence in synchronized flow of a general pattern (GP) at an on-ramp bottleneck has been simulated in heterogeneous flow consisting of passenger cars and long vehicles (Fig. 33). Firstly, an  $F \rightarrow S$  transition occurs spontaneously at the bottleneck. Synchronized flow that appears due to this  $F \rightarrow S$  transition propagates upstream. Moving jams emerge spontaneously in that synchronized flow propagating upstream.

Simulations show that there are the following mechanisms for emergence of moving blanks within the jams: (i) Vehicle lane changing at the upstream jam front. (ii) Vehicles come to a stop at the upstream front of the wide moving jam at different space gaps to the related preceding vehicles. (iii) When vehicles move with low speeds cover-





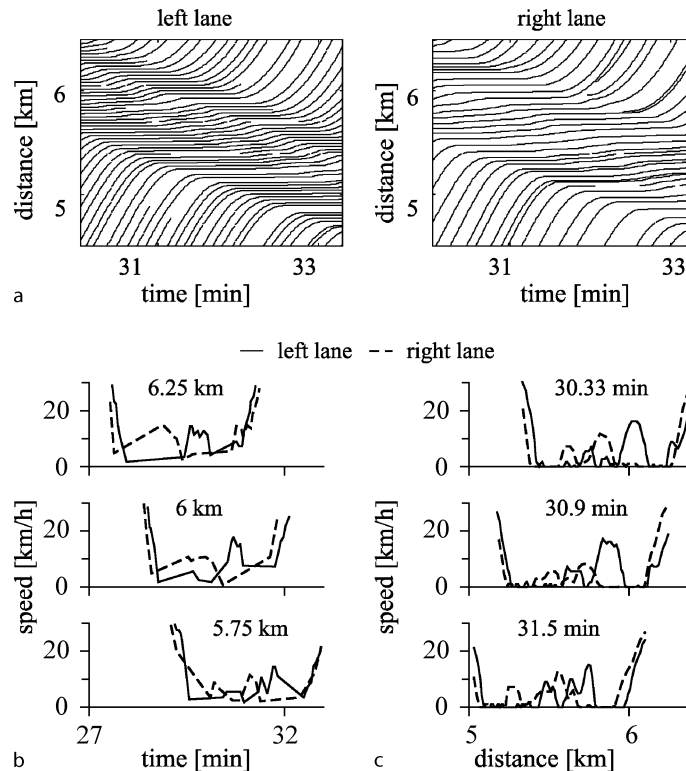
**Traffic Congestion, Spatiotemporal Features of, Figure 30**

Simulated characteristics of moving blanks within a wide moving jam propagating through an on-ramp bottleneck. **a** Speed in space and time on the main road. **b–e** Single vehicle data for time dependences of speed (**b**), time headways (**c**, **d**), and the associated density distributions calculated through the density definition (*solid curves 1*) and, as in empirical Fig. 28a, via the formula  $\rho = q/v$  (*dashed curves 2*) (**e**) at location  $x = 8.5$  km. *Left and right figures (a–e) are related to the left and right lanes, respectively.* **f**, **g** Speed (**f**) and density (**g**) related to **a** as functions of distance at time 31.5 min in the left and right lanes. Taken from [47]

ing downstream blanks within the jam, vehicle lane changing can result in new moving blanks.

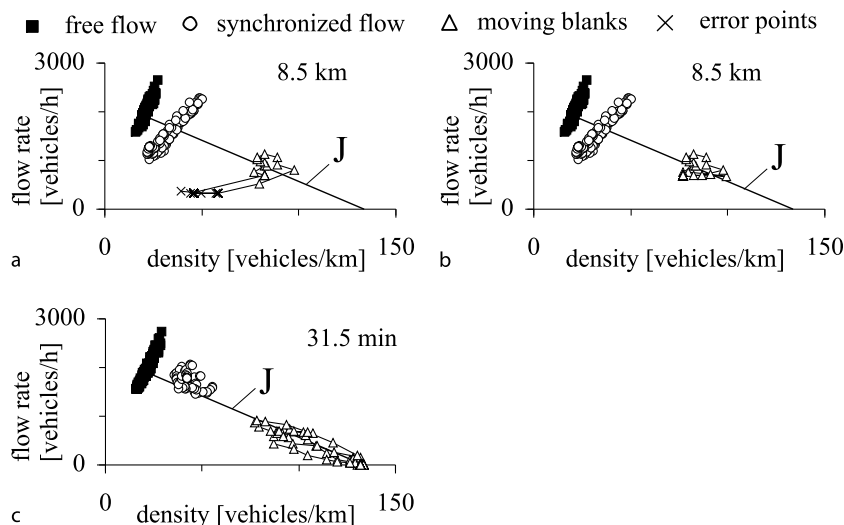
In heterogeneous flow, moving blanks emerge most frequently due to vehicle lane changing (item (i) and (iii)). Specifically, a moving jam emerges firstly in the right lane

only ( $t = 23$  min in Figs. 33b,c and 34a). Then some vehicles change from the right lane to the left lane (arrow 1 in Fig. 34a). On the one hand, this lane changing decreases speed in the left lane. This leads to narrow jam formation in this lane. On the other hand, blanks between vehicles



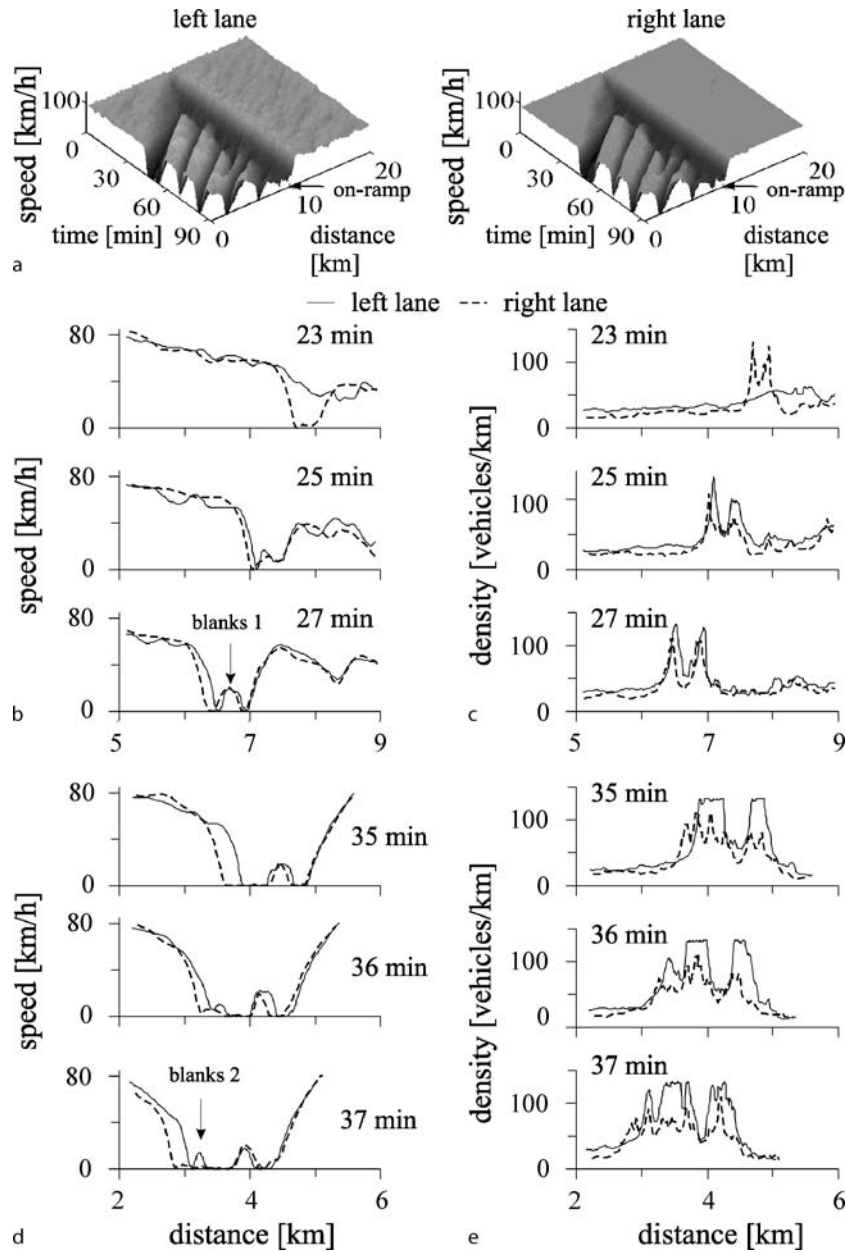
#### Traffic Congestion, Spatiotemporal Features of, Figure 31

Simulated microscopic characteristics of moving blanks within the wide moving jam shown in Fig. 30a: a Vehicle trajectories for each 4th vehicle in the left (*left*) and right (*right*) lanes. b, c Low speed states associated with moving blanks at three different locations (b) and three different time moments (c) in the left and right lanes. Taken from [47]



#### Traffic Congestion, Spatiotemporal Features of, Figure 32

Simulated characteristics of moving blanks within the wide moving jam in the left lane shown in Fig. 30a together with states of free flow and synchronized flow in the flow-density plane. a, b States within the jam that are determined at detector at  $x = 8.5$  km through the formula  $\rho = q/v$  in two cases in which all states within the jam measured at the detector are shown (a) and error states are removed (b). c States within the jam that are determined through the density definition (vehicles per freeway length) at time 31.5 min. J is the line J for the downstream jam front. Taken from [47]

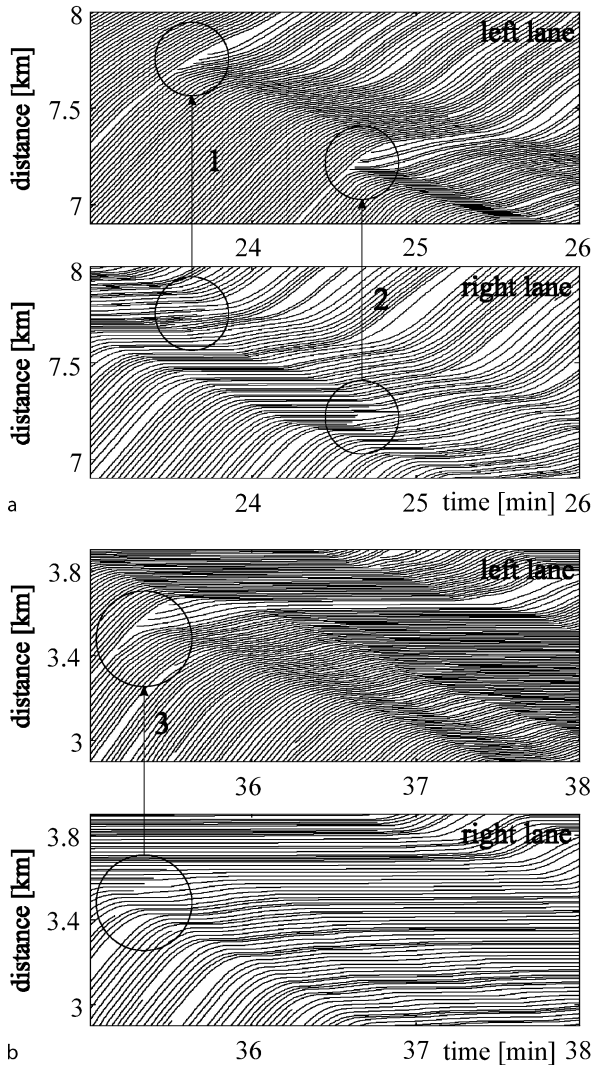


**Traffic Congestion, Spatiotemporal Features of, Figure 33**

Simulated characteristics of moving blanks emergence: a Speed in space and time within a general pattern (GP) at an on-ramp bottleneck in the left (left) and right lanes (right). b–e Speed (b, d) and density (c, e) at different time moments associated with the GP in a in the left and right lanes. In c, e, for density calculation the density definition (vehicles per freeway length) is used. Taken from [47]

in the right lane increase due to lane changing. Vehicles begin to cover these blanks ( $t \approx 24$  min in Fig. 34a). As a result, low speed states within the jam appear associated with these moving blanks. Later, subsequent lane changing of vehicles from the right to the left lane results in an

abrupt decrease in speed in the left lane upstream of the narrow moving jam ( $t \approx 25$  min in Figs. 33b,c and 34a). Then this complex spatiotemporal speed distribution in the left lane transforms into a wide moving jam with moving blanks within the jam ( $t \approx 25$  min in Fig. 33b,c). Thus



**Traffic Congestion, Spatiotemporal Features of, Figure 34**  
 Simulated characteristics of emergence of moving blanks in the GP shown in Fig. 33a: a, b Vehicle trajectories for different time intervals in the left and right lanes. Taken from [47]

in this scenario, vehicle lane changing from the right to the left lane at the upstream jam front is the main reason for emergence of moving blanks within the jam (moving blanks labeled “blanks 1” in Fig. 33b).

During subsequent jam propagation new moving blanks emerge (Fig. 33d,e). Firstly, the upstream jam front in the right lane is upstream of the upstream jam front in the left lane ( $t = 35$  min, Fig. 33d,e). Due to lane changing of vehicles from the right to the left lane, the upstream jam front locations are synchronized with each other (arrow 3 in Fig. 34b). This lane changing causes emergence of

moving blanks within the jam ( $t = 37$  min, moving blanks labeled “blanks 2” in Fig. 33d).

### Diverse Transitions Between Traffic Phases and States

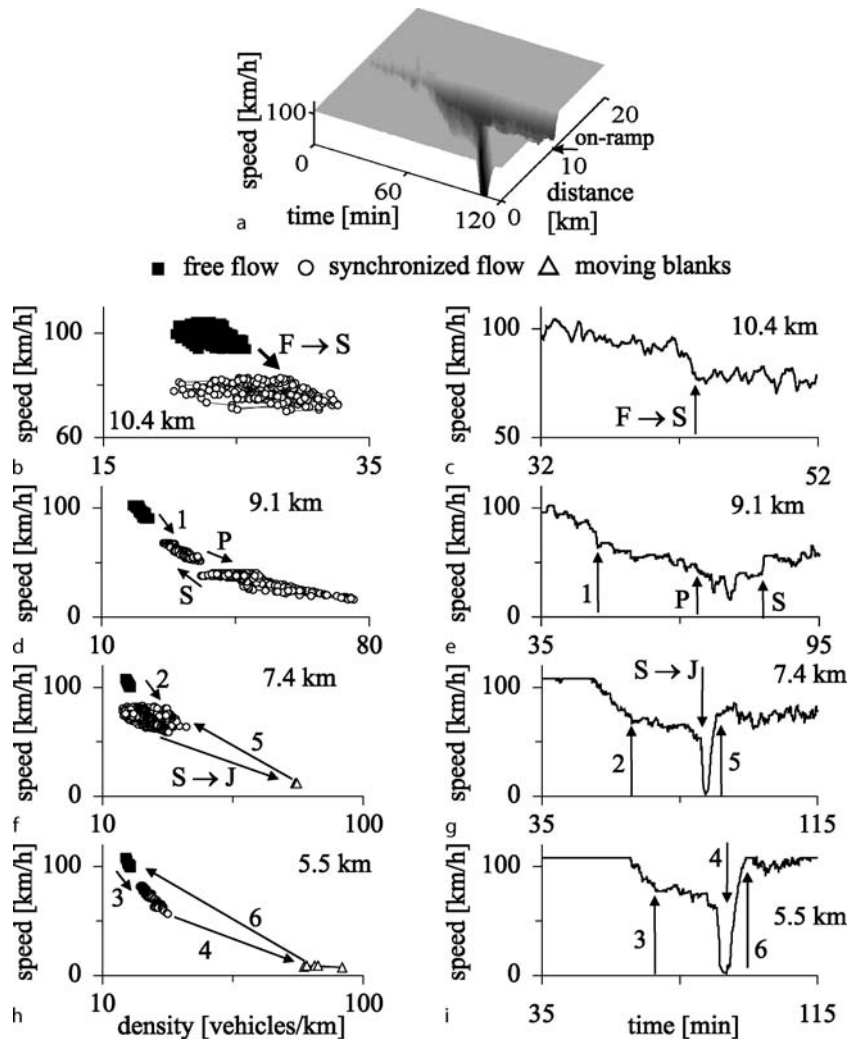
Microscopic effects of moving jam emergence, propagation, and dissolution lead to diverse *Transitions* between traffic phases and traffic states. In a microscopic analysis of these transitions, we should distinguish the following three basic types of qualitatively different transitions measured at fixed freeway locations.

- Local phase transitions (denoted by  $F \rightarrow S$ ,  $S \rightarrow F$ ,  $S \rightarrow J$ , and  $J \rightarrow S$  transitions). Within a local freeway region a new traffic phase emerges instead of an initial traffic phase.
- Transitions between existing traffic phases that are caused by propagation of a front separating free and synchronized flows or by wide moving jam propagation.
- Transitions between different traffic states of the same traffic phase.

We consider the associated microscopic effects on an example of moving jam emergence within the DGP shown in Fig. 35a.

Firstly, an  $F \rightarrow S$  transition occurs in a neighborhood of the bottleneck (location  $x = 10.4$  km; transition of type (i) shown by arrows labeled “ $F \rightarrow S$ ” in Fig. 35b,c). As a result, synchronized flow appears at the bottleneck. Secondly, this synchronized flow propagates on the main road upstream of the bottleneck, i.e., a wave of upstream transitions from free flow to synchronized flow occurs (e.g.,  $x = 9.1, 7.4$ , and  $5.5$  km; transitions of type (ii) shown by arrows 1–3 in Fig. 35d–i). Then a self-compression of this synchronized flow (pinch effect) within a localized region upstream of the bottleneck occurs. The pinch effect causes an abrupt transition between different states within the synchronized flow phase ( $x = 9.1$  km; transition of type (iii) shown by arrows labeled “P” in Fig. 35d,e).

Within the related pinch region of synchronized flow, a narrow moving jam emerges spontaneously. During upstream jam propagation, the jam amplitude grows and the narrow moving jam transforms into a wide moving jam, i.e., an  $S \rightarrow J$  transition occurs in the synchronized flow (location  $x = 7.4$  km; transition of type (i) shown by arrows labeled “ $S \rightarrow J$ ” in Fig. 35f,g). The flow rate in the jam outflow  $q_{out}$  is smaller than the flow rate in the pinch region. As a result, the pinch region dissolves after the wide moving jam has been formed; synchronized flow of higher speed remains ( $x = 9.1$  km; transition of type (iii) shown by arrows labeled “S” in Fig. 35d,e). For this reason,



**Traffic Congestion, Spatiotemporal Features of, Figure 35**

Simulated characteristics of microscopic transitions between different traffic phases and states measured at fixed freeway locations as well as the associated hysteresis phenomena in traffic flow: **a** Speed in space and time within DGPs at on-ramp bottleneck. **b–i** Transitions in the speed-density plane (**b, d, f, h**) and in the related time-dependences of speed (**c, g, e, i**) at different road locations during wide moving jam emergence and propagation shown in **a**. Taken from [47]

in contrast with the GP in Fig. 33a the wide moving jam in Fig. 35a prevents subsequent moving jam emergence in synchronized flow upstream of the bottleneck, i. e., the DGP appears.

The resulting DGP consists of synchronized flow upstream of the wide moving jam, the jam propagating upstream, and synchronized flow that is downstream of the jam and upstream of the bottleneck (Fig. 35a). At each of the virtual detectors, jam propagation causes different transitions of type (ii):

1) From an initial traffic phase at a detector (free flow

or synchronized flow) to low speed states associated with moving blanks within the wide moving jam (e. g.,  $x = 5.5$  km, transition from synchronized flow to the jam shown by arrows “4” in Fig. 35h,i).

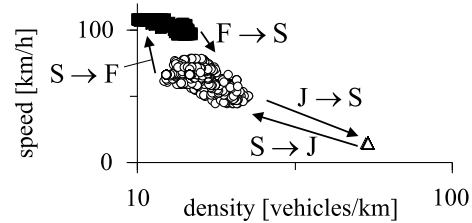
2) From the jam to either free flow or synchronized flow after the jam has passed the detector ( $x = 7.4$  km, transition from the jam to synchronized flow shown by arrows “5” in Fig. 35f,g and  $x = 5.5$  km, transition from the jam to free flow shown by arrows “6” in Fig. 35h,i).

In Fig. 35, the flow rates to the on-ramp  $q_{on}$  and the flow rate in free flow upstream of the DGP  $q_{in}$  do not vary



over time. In contrast, if the flow rate  $q_{in}$  begins to decrease, then the DGP dissolves (Fig. 36). In this case, DGP dissolution starts upstream and propagates downstream. Firstly, a wave of synchronized flow dissolution, which is upstream of the jam, propagates downstream leading to transitions of type (ii) from synchronized flow to free flow (Fig. 36a). When this wave reaches the jam, the latter begins to dissolve. As a result of wide moving jam dissolution synchronized flow remains, i. e., an  $J \rightarrow S$  transition occurs ( $x = 7$  km; arrows labeled “ $J \rightarrow S$ ” in Fig. 36b). Then a new wave of synchronized flow dissolution propagating downstream is formed, which leads to transitions of type (ii) from synchronized flow to free flow ( $x = 7$  km; arrow 7 in Fig. 36b). Due to the latter transitions, finally, an  $S \rightarrow F$  transition (transition of type (i)) occurs at the bottleneck ( $x = 10.4$  km; arrows labeled “ $S \rightarrow F$ ” in Fig. 36c).

These numerical simulations enable us to find states of free flow and synchronized flow as well as states within a wide moving jam on a double Z-characteristic for phase transitions (Fig. 37). Arrows between these states labeled “ $F \rightarrow S$ ” and “ $S \rightarrow F$ ” are associated with  $F \rightarrow S$  and  $S \rightarrow F$  transitions at the bottleneck ( $x = 10.4$  km), whereas arrows labeled “ $S \rightarrow J$ ” and “ $J \rightarrow S$ ” are associated with  $S \rightarrow J$  and  $J \rightarrow S$  transitions, which have occurred at different locations,  $x = 7.4$  km and 7 km, respectively (Figs. 35g and 36b). Thus, wide moving jam emergence ( $S \rightarrow J$  transition) and dissolution ( $J \rightarrow S$  transition) occur upstream of the bottleneck, whereas the  $F \rightarrow S$  and  $S \rightarrow F$  transitions occur at the bottleneck. As a result, states for the three traffic phases on the double Z-characteristic cannot be found through measurements of traffic variables at a single freeway location. Consequently, traffic



**Traffic Congestion, Spatiotemporal Features of, Figure 37**

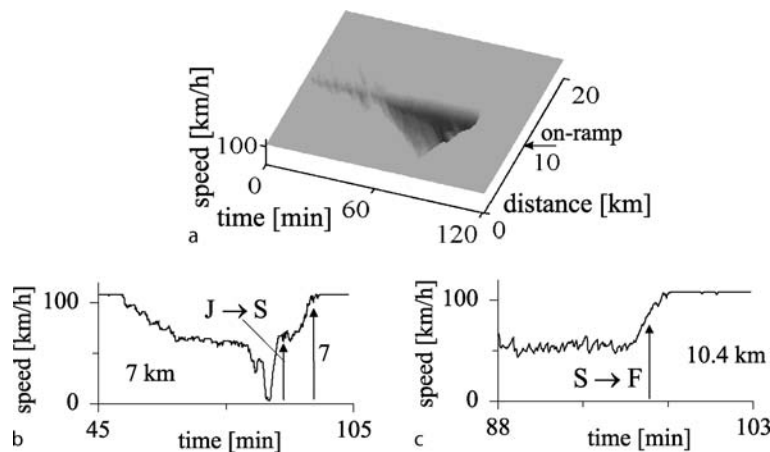
States of the three traffic phases associated with double Z-characteristic in the speed-density plane for local phase transitions associated with Figs. 35 and 36

states and transitions shown by arrows in Fig. 37 are taken from different locations (i. e., from Fig. 35) at which the related local phase transitions have occurred.

We can conclude that during synchronized flow and wide moving jam emergence, propagation, and dissolution there are all three types of abrupt transitions (i), (ii), and (iii) between traffic variables. These transitions can cause a variety of “hysteresis” phenomena in the speed-density plane (Figs. 35, 36, and 37).

### Congested Patterns at Heavy Bottlenecks

In general, the average flow rate  $q^{(cong)}$  within a congested traffic pattern upstream of a bottleneck is the smaller, the greater the bottleneck influence of traffic (called the bottleneck strength). As in Sect. “Diagram of Congested Patterns at Isolated Bottlenecks”, the time interval of the averaging of the flow rate  $q^{(cong)}$  is suggested to be considerably longer than time distances between any moving jams



**Traffic Congestion, Spatiotemporal Features of, Figure 36**

Speed in time and space during a dissolution of the DGP in Fig. 35a occurring when the flow rate in free flow upstream of the DGP decreases beginning at  $t = 75$  min (a) and the associated transitions in time-dependences of speed (b, c) at different road locations

within a congested pattern. Empirical data show that the flow rate  $q^{(\text{cong})}$  within GPs occurring at usual bottlenecks like on- and off-ramp bottlenecks discussed above in this article, which is equal to the average flow rate in the pinch region  $q^{(\text{pinch})}$ , is approximately within a range

$$q^{(\text{cong})} = q^{(\text{pinch})} = 1100\text{--}1700 \text{ vehicles/h/lane} . \quad (7)$$

In contrast with the usual bottlenecks, due to for example bad weather conditions or accidents heavy bottlenecks can occur, which exhibit a much greater influence on traffic (greater bottleneck strength) that limits  $q^{(\text{cong})}$  to very small values, sometimes as low as zero. Results of a theory of traffic congestion at heavy bottlenecks developed recently [39,40] are discussed in this Sect. “Congested Patterns at Heavy Bottlenecks”.

In this case, we can expect new interesting physical phenomena associated with complexity of traffic congestion at heavy bottlenecks. This expectation follows already from a qualitative consideration of the influence of the decrease in  $q^{(\text{cong})}$  on a sequence of wide moving jams. Indeed, from empirical single vehicle data studied in [46,47], we can conclude that the flow rate  $q^{(\text{blanks})}$  of low speed states associated with moving blanks within the jams satisfies an approximate condition

$$q^{(\text{blanks})} \lesssim 600 \text{ vehicles/h/lane} . \quad (8)$$

The average flow rate within wide moving jams  $q^{(\text{blanks})}$  is defined as a number of vehicles  $N^{(\text{blanks})}$ , which have passed a virtual road detector during time intervals of the propagation of  $n_j$  wide moving jams (it is assumed that  $n_j \gg 1$ ) through the detector, divided by the sum of all jam durations:  $q^{(\text{blanks})} = N^{(\text{blanks})} / \sum_{i=1}^{n_j} \tau_j^{(i)}$ , where  $\tau_j^{(i)}$  is a duration of a  $i$ th wide moving jam, i. e., the time interval between the downstream and upstream fronts of the jam  $i$ , while this jam propagates through the detector.

Between the wide moving jams, the average flow rate associated with non-interrupted flows in the jam outflows  $q_{\text{out}}^{(j)}$  is greater than  $q^{(\text{pinch})}$  (7), i. e.,  $q_{\text{out}}^{(j)}$  is considerably greater than  $q^{(\text{blanks})}$  (8). Now we assume that due to bad weather conditions or an accident a heavy bottleneck occurs with a great strength for which

$$q^{(\text{cong})} = q^{(\text{blanks})} . \quad (9)$$

In this case,  $q_{\text{out}}^{(j)}$  must reduce to  $q^{(\text{blanks})}$  (8), i. e., the difference between flows within and outside wide moving jams disappears. As a result, at

$$q^{(\text{cong})} \leq q^{(\text{blanks})} \quad (10)$$

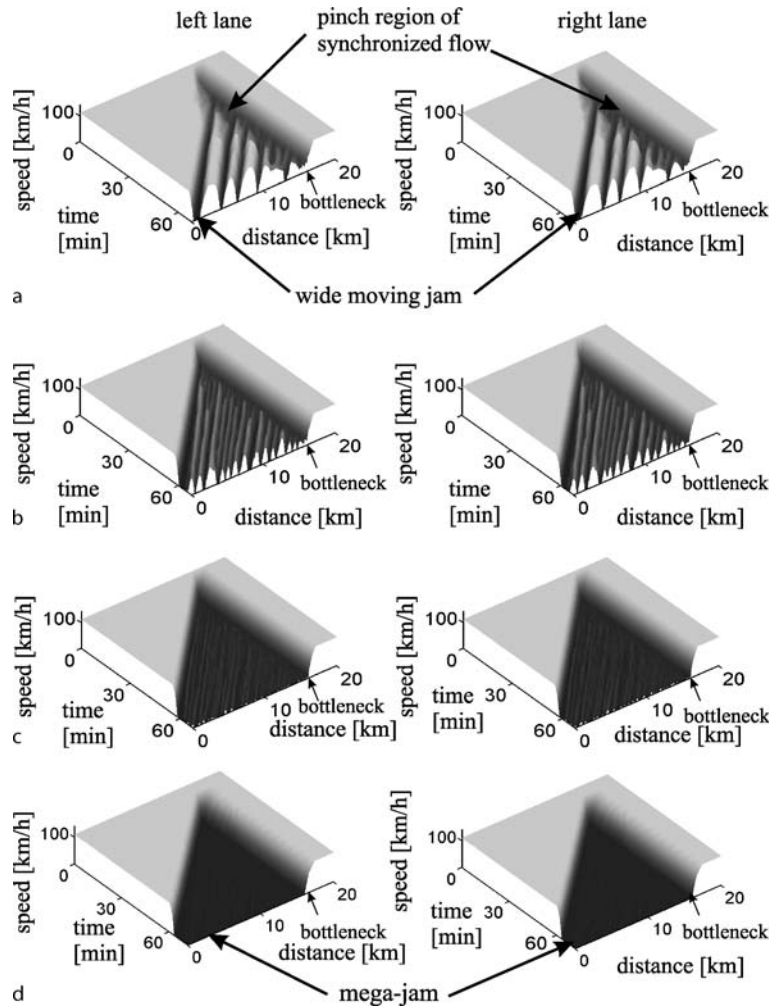
wide moving jams should merge into a mega-wide moving jam (mega-jam for short).

For an analysis of traffic congestion at heavy bottlenecks, the Kerner–Klenov stochastic three-phase traffic flow model of vehicular traffic on a two-lane road with identical drivers and vehicles [41,42] considered in ► [Traffic Congestion, Modeling Approaches to](#) has been used with the following model for a heavy bottleneck.

We suggest that there is a section of the road of a length  $L_B$  within which due to an accident or bad weather conditions drivers should increase a safety time headway  $\tau^{(\text{safe})}$  to the preceding vehicle as well as decrease the maximum speed to some  $v_B$  that is lower than the maximum speed in free flow: within the section, the safety time headway is equal to  $\tau^{(\text{safe})} = T_B > 1$  sec that is longer than  $\tau^{(\text{safe})} = 1$  sec used in the model for vehicles moving outside this section. In according with the model,  $\tau^{(\text{safe})}$  determines a safe speed, which should not be exceeded by a driver; otherwise, the driver decelerates. As a result, within this section drivers move with the mean time headway that is longer than the mean time headway outside the bottleneck section and it is the longer, the longer  $\tau^{(\text{safe})} = T_B$ . Therefore the section with longer  $\tau^{(\text{safe})} = T_B$  acts as a bottleneck on the road. In the bottleneck model, each chosen value  $T_B$  defines a specific bottleneck: the strength of this bottleneck is the greater, the longer  $T_B$ ; in turn, the longer  $T_B$ , the smaller  $q^{(\text{cong})}$ , i. e., the greater the flow rate limitation within congestion caused by the bottleneck.

This model feature allows us to simulate a heavy bottleneck caused by bad weather conditions or accidents leading to a long enough  $T_B$  within the bottleneck. Under bad weather conditions, a road section with a long value  $T_B$  can be caused, e. g., by a poor view due to fog on the section or a much longer deceleration way needed by snow and ice on the section. If an accident occurs on a road, a road section with a long value  $T_B$  can be caused, e. g., by much narrower lane widths allowed for driving on the road section; the same effect can occur under heavy roadworks.

In simulations discussed in below, we consider traffic phenomena occurring when the bottleneck strength increases gradually through the increase in  $T_B$ , when other bottleneck parameters are given constants (Fig. 38). However as simulations show, at a given  $T_B$  the increase in  $L_B$  or/and decrease in  $v_B$  lead also to an increase in the bottleneck strength with the subsequent decrease in the average flow rate  $q^{(\text{cong})}$  in congested traffic upstream of the bottleneck. For this reason, numerical values of  $T_B$  mentioned below, at which characteristic traffic phenomena occur at a heavy bottleneck, depend on  $L_B$  and  $v_B$ . In contrast, we find that numerical values of the flow rate  $q^{(\text{cong})}$  at which the traffic phenomena occur at the bottleneck do not change appreciably, when a greater value of  $L_B$



**Traffic Congestion, Spatiotemporal Features of, Figure 38**

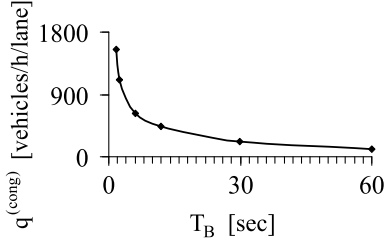
Simulated speed in space and time in the left (*left*) and right (*right*) road lanes at different  $T_B$ :  $T_B = 1.8$  (a), 2.4 (b), 12 (c), 60 sec (d).  $q_{in} = 1946$  vehicles/h/lane. The upstream boundary of bottleneck region of the length  $L_B = 300$  m is at  $x = 16$  km; the maximum speed within this bottleneck region is  $v_B = 60$  km/h. Resulting values of  $q^{(cong)} = 1546$  (a), 1114 (b), 440 (c), 127 vehicles/h/lane (d)

or/and another value of  $v_B$  (in the range 50–80 km/h) are chosen. For this reason, the flow rates  $q^{(cong)}$  can be considered characteristic values representing the bottleneck strength.

### Evolution of Traffic Phases at Heavy Bottlenecks

When the bottleneck strength is not great, i. e.,  $T_B$  is chosen to be not very long ( $T_B = 1.8$  sec), then at a great enough flow rate  $q_{in}$  in free flow upstream of the bottleneck, firstly synchronized flow occurs spontaneously at the bottleneck. Then the pinch region with a relatively great flow rate  $q^{(cong)} = q^{(pinch)}$  is formed (Figs. 38a and 39). At the pinch region upstream boundary wide moving

jams emerge. Thus we found the phenomena of regular GP formation (Figs. 38a and 40a,b) discussed briefly in Subsect. “General Congested Patterns”. The pinch region width  $L^{(pinch)}(t)$  changes over time between about 1 and 2 km (Fig. 41a).  $L^{(pinch)}$  is defined as the distance between the upstream boundary of the bottleneck ( $x = 16$  km) and the road location upstream at which a wide moving jam has just been identified through the use of the jam microscopic criterion (5). There is also a nearly constant frequency of  $L^{(pinch)}(t)$  oscillations associated with the maximum in the Fourier spectrum (Fig. 41b). Speed autocorrelation functions and associated Fourier spectra of speed time-dependences at shorter  $T_B$  show regular character of wide moving jam propagation (Fig. 40c,d).



**Traffic Congestion, Spatiotemporal Features of, Figure 39**  
Simulated average flow rate  $q^{(\text{congr})}$  within congested patterns shown in Fig. 38 as a function of  $T_B$ . The averaging time interval for  $q^{(\text{congr})}$  is 60 min

When  $T_B$  becomes longer and therefore the bottleneck strength increases, the flow rate  $q^{(\text{congr})} = q^{(\text{pinch})}$  decreases (Fig. 39). However, if  $T_B$  remains a relatively small value ( $1.8 < T_B < 3$  sec), then as for other bottleneck types, we found the following GP features some of them have already been discussed in Sect. “Diagram of Congested Patterns at Isolated Bottlenecks”: the smaller the flow rate  $q^{(\text{pinch})}$ , the greater the frequency of narrow moving jam emergence within the pinch region, the lower the maximum (and the average) speed between wide moving jams upstream of the pinch region, the smaller the mean pinch region width  $L_{\text{mean}}^{(\text{pinch})}$  (Figs. 38b and 41c,d,k,l). This can also be seen from a comparison of time-dependences of average speed within the region of wide moving jams for different  $T_B$  (Fig. 40a,e).

Qualitatively other phenomena are found when  $T_B$  further increases ( $T_B > 3$  sec) and the average flow rate  $q^{(\text{congr})}$  decreases considerably (Figs. 38c,d and 39).

We find that there is a critical strength of a heavy bottleneck associated with  $T_B = T_B^{(\text{break})}$  that results in the critical flow rate  $q^{(\text{congr})} = q_{\text{break}}^{(\text{congr})}$ . When

$$q^{(\text{congr})} \leq q_{\text{break}}^{(\text{congr})} \quad (11)$$

related to  $T_B \geq T_B^{(\text{break})}$ , then there are random time intervals when the pinch region disappears, i. e.,  $L^{(\text{pinch})} = 0$  (Fig. 41e,g). This means that there are time instants at which there is no pinch region and wide moving jams emerge directly at the upstream boundary of the bottleneck, whereas for other time intervals the pinch region appears again (Fig. 41e,g). We found that  $q_{\text{break}}^{(\text{congr})} \approx 920$  vehicles/h/lane ( $T_B^{(\text{break})} \approx 3$  sec at the chosen bottleneck parameters  $L_B$  and  $v_B$ ). Under condition (11),  $L^{(\text{pinch})}(t)$  becomes a non-regular time-function (Fig. 41e,g,i) whose Fourier spectrum is broader, the longer  $T_B$ , i. e., the smaller  $q^{(\text{congr})}$  (Fig. 41f,h,j). Such a GP at an isolated

heavy bottleneck can be considered the GP with a non-regular pinch region.

The more the bottleneck strength exceeds the critical bottleneck strength, the greater the difference  $q_{\text{break}}^{(\text{congr})} - q^{(\text{congr})}$  and the longer the mean duration of time intervals within which the regular structure of the GP breaks and the pinch region disappears, and therefore, the smaller the mean length  $L_{\text{mean}}^{(\text{pinch})}$  of the pinch region (Fig. 41k,l). In contrast with regular time-dependences of the average speed within a sequence of wide moving jams (Fig. 40a–h), the time-functions of 1 min average speed at a greater bottleneck strength exhibit a non-regular behavior (Fig. 40i–l). This can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time-dependences (Fig. 40k,l).

When the bottleneck strength increases further,  $L_{\text{mean}}^{(\text{pinch})}$  decreases continuously (Fig. 41l).  $L_{\text{mean}}^{(\text{pinch})}$  reaches zero at a threshold bottleneck strength for the pinch region existence associated with  $T_B = T_B^{(\text{th})}$  that results in a threshold flow rate  $q^{(\text{congr})} = q_{\text{th}}^{(\text{congr})}$ . When

$$q^{(\text{congr})} \leq q_{\text{th}}^{(\text{congr})} \quad (12)$$

related to  $T_B \geq T_B^{(\text{th})}$ , then the pinch region of GPs does not exist. At model parameters, the pinch region of a GP disappears fully at  $q_{\text{th}}^{(\text{congr})} \approx 220$  vehicles/h/lane ( $T_B^{(\text{th})} \approx 30$  sec at the chosen  $L_B$  and  $v_B$ ).

There is also another critical bottleneck strength, which we call the critical bottleneck strength for the mega-jam formation associated with  $T_B = T_B^{(\text{mega})}$  that results in the critical flow rate  $q^{(\text{congr})} = q_{\text{mega}}^{(\text{congr})}$ . When

$$q^{(\text{congr})} \leq q_{\text{mega}}^{(\text{congr})} \quad (13)$$

related to  $T_B \geq T_B^{(\text{mega})}$ , then wide moving jams merge onto a mega-jam. Numerical simulations show that (13) is equivalent to (10), i. e.,

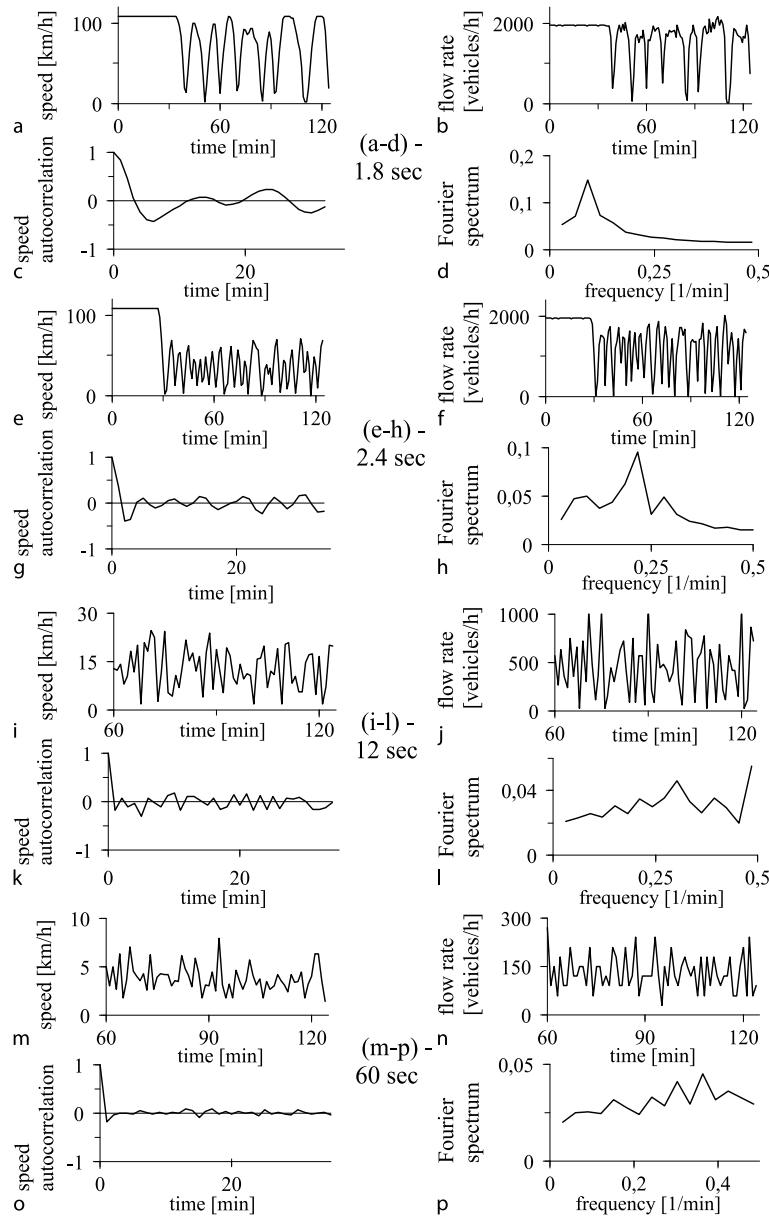
$$q_{\text{mega}}^{(\text{congr})} = q^{(\text{blanks})}. \quad (14)$$

At model parameters,  $q_{\text{mega}}^{(\text{congr})} \approx 130$  vehicles/h/lane ( $T_B^{(\text{mega})} \approx 60$  sec at the chosen  $L_B$  and  $v_B$ ).

We found also that the critical bottleneck strength for the mega-jam formation is greater than the threshold bottleneck strength for the pinch region existence that results in the condition

$$q_{\text{mega}}^{(\text{congr})} < q_{\text{th}}^{(\text{congr})} \quad (15)$$

related to  $T_B^{(\text{mega})} > T_B^{(\text{th})}$ . Under condition (13), i. e., when all wide moving jams merge onto a mega-jam, traffic congestion upstream of an isolated heavy bottleneck cannot



**Traffic Congestion, Spatiotemporal Features of, Figure 40**

Simulated characteristics of congested patterns shown in Fig. 38 related to location 10 km. Time-functions of speed (a, e, i, m), speed correlations (c, g, k, o), associated Fourier spectra (d, h, l, p), and flow rate (b, f, j, n) for different  $T_B = 1.8$  (a-d), 2.4 (e-h), 12 (i-l), 60 sec (m-p). 1 min average data in the left lane

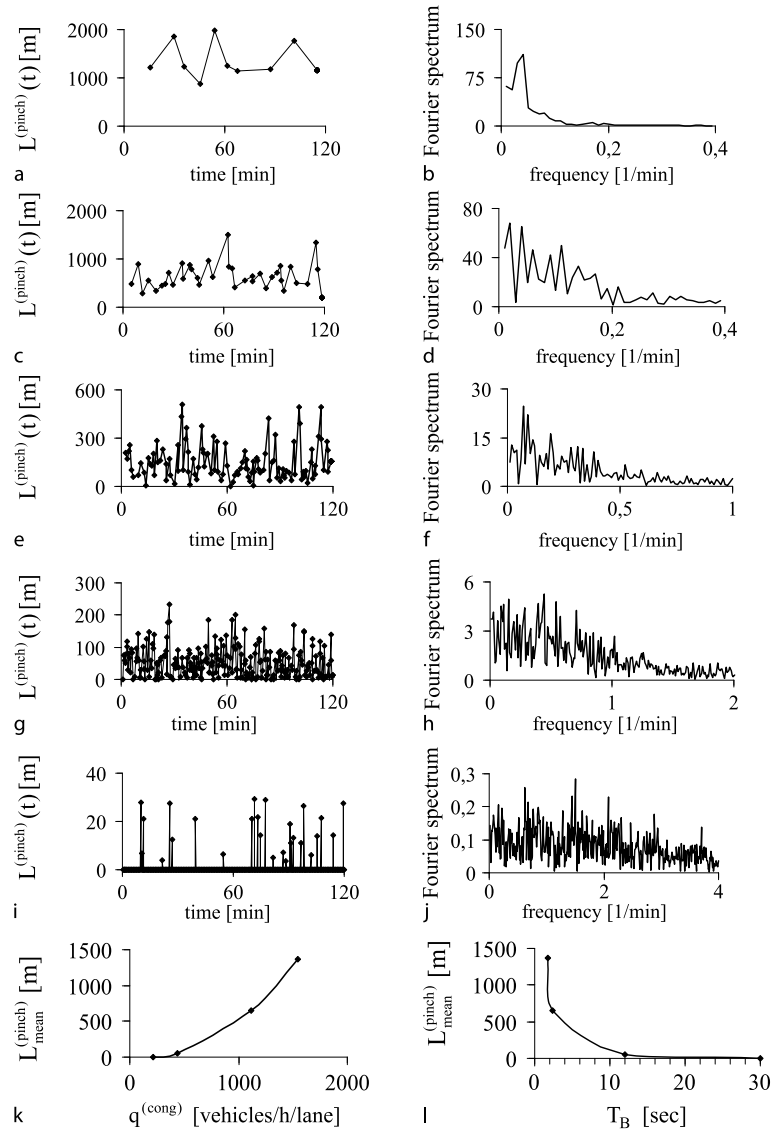
be considered as a GP any more. Thus in accordance with (13), (14), if traffic breakdown has occurred at a bottleneck, then the condition

$$q^{(\text{cong})} > q^{(\text{blanks})} \quad (16)$$

is a *necessary* condition for GP existence at this bottleneck.

As for GPs with a very non-regular pinch region (Fig. 40i-l), the time-functions of 1 min average speed within the mega-jam exhibit a non-regular behavior (Fig. 40m); this can be seen from speed autocorrelation functions and associated Fourier spectra of the average speed time-dependences (Fig. 40o,p).





**Traffic Congestion, Spatiotemporal Features of, Figure 41**

Simulations of pinch effect:  $L^{(\text{pinch})}(t)$  (a, c, e, g, i), associated Fourier spectra (b, d, f, h, j) for congested patterns related to  $T_B = 1.8$  (a, b), 2.4 (c, d), 6 (e, f), 12 (g, h), and 30 sec (i, j). k, l  $L_{\text{mean}}^{(\text{pinch})}(q^{(\text{congl})})$  (k) and  $L_{\text{mean}}^{(\text{pinch})}(T_B)$  (l)

When  $q^{(\text{congl})}$  becomes zero, because behind a road location the road is closed, the mega-jam transforms into a queue of *motionless* vehicles, which therefore is not associated with vehicular traffic. Nevertheless, there is a link between the queue and the mega-jam. If at a time instant we allow several vehicles to escape from this queue, then simulations show that motion of these vehicles results in wide moving jam occurrence: the downstream front of the jam separates moving vehicles escaping from the queue and vehicles standing within the queue upstream of the

front. When the number of vehicles escaping from the initial queue decreases, the downstream jam front transforms into a moving blank(s) subsequently covered by vehicles standing in the queue. When a vehicle per a long enough time interval is allowed to escape from the mega-jam, as simulations show, a sequence of such moving blanks within this jam occur; these moving blanks exhibit, however, the dynamics specifically associated with the mega-jam (Subsect. “[Microscopic Spatiotemporal Structure of Mega-Jam](#)”).

Thus based on a study of three-phase traffic flow model we found that the complexity of traffic congestion at a heavy bottleneck caused for example by bad weather conditions or accidents is associated with the phenomenon of random disappearance and appearance of the pinch region over time as well as with the phenomenon of the merger of wide moving jams into a mega-jam. When the bottleneck strength increases continuously, the mean length of the pinch region decreases also continuously up to zero; at such a heavy bottleneck, the pinch region does not exist any more. Beginning from a greater bottleneck strength only the mega-jam survives in congested traffic upstream of the bottleneck and synchronized flow remains only within its downstream front separating free flow and congested traffic. In other words, in congested traffic occurring at such a very heavy bottleneck there are no continuous flows within the congested pattern any more with the one exception of the downstream front of synchronized flow that separates free flow downstream and the mega-jam upstream of the front. Synchronized flow remains *only* within this front. These phenomena reveal the evolution of the traffic phases in congested traffic when heavy bottlenecks occur in highway traffic.

### Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics

For GPs with a regular pinch region ((11) is not satisfied), we find results of Subsect. “General Congested Patterns” that at a small enough distance upstream of the bottleneck there is the pinch region of non-interrupted synchronized flow (Fig. 42a,b) with short enough time headways for which criterion (5) is satisfied for *none* of time headways (Fig. 42c,d) [38].

In contrast, under condition (11), when GPs with a non-regular pinch region occur, at the same small distance upstream of the bottleneck some long time headways appear in the pinch region of the GP for which criterion (5) is satisfied (Fig. 42g,h). This means that there are wide moving jams that emerge almost directly upstream of the bottleneck. This is related to the previous result that the pinch region width  $L^{(\text{pinch})}$  of such a GP is nearly as small as zero at some random time instants (Fig. 41e,g). During other time intervals, there is synchronized flow with short enough time headways for which criterion (5) is not satisfied (Fig. 42g,h). Thus at small enough distance upstream of the bottleneck, such GPs consists of a random alternation between synchronized flow and wide moving jams. This is related to the conclusion of Subsect. “Evolution of Traffic Phases at Heavy Bottlenecks” that for these GPs at

some random time instants the pinch region disappears and appears again.

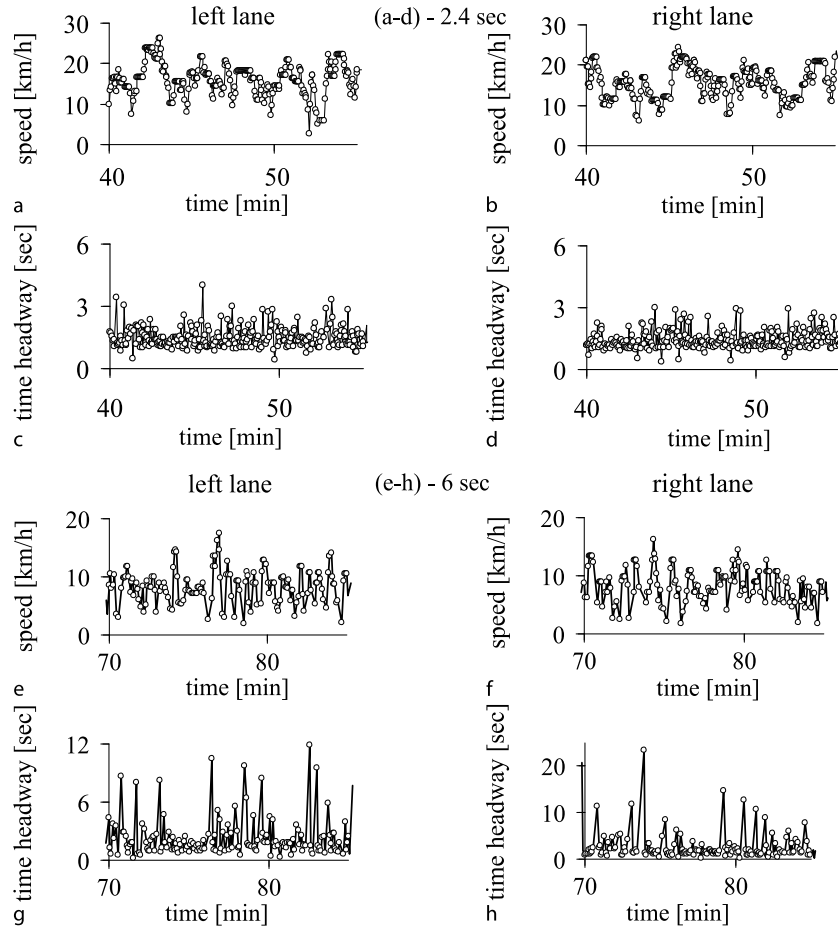
Single-vehicle data measured by a virtual road detector 6 km upstream of the bottleneck within the GPs with regular (Fig. 43a–d) and non-regular pinch regions (Fig. 43e–h) exhibits qualitatively the same and typical time-dependences of the speed and time headway for a sequence of wide moving jams of a GP [47]. We see that there are moving jams within which the maximum time headway  $\tau_{\max}$  between two vehicles within the jams is very long, i. e., traffic flow is interrupted and criterion (5) is satisfied (Fig. 43c,d,g,h). Therefore, these jams are wide moving ones. There are time intervals between these jams within which the speed is high (Fig. 43a,b,e,f) and vehicle time headways are small (Fig. 43c,d,g,h). These regions between the wide moving jams are related to non-interrupted traffic flow. Thus at some distance upstream of the bottleneck there is a sequence of wide moving jams that is characteristic for GPs.

A wide moving jam consists of alternations of flow interruption intervals and moving blanks within the jam (Sect. “Microscopic Features of Traffic Phases”). In a general case, for a wide moving jam we can observe two or more flow interruption intervals, i. e., two or more long time headways  $\tau^{(i)}$ ,  $i = 1, 2, \dots$  each of them satisfies criterion (5):

$$I_s^{(i)} = \tau^{(i)} / \tau_{\text{del}}^{(a)} \gg 1, \quad i = 1, 2, \dots \quad (17)$$

Between these flow interruption intervals, we find one or more vehicles that have passed the detector with speeds, which are considerably lower than the average speed both upstream and downstream of the wide moving jam. This low speed vehicle motion is associated with moving blanks (Subsect. “Numerical Simulations of Moving Blanks Within Wide Moving Jams”): the flow interruption intervals  $\tau^{(i)}$ ,  $i = 1, 2, \dots$  are separated by one or more moving blanks within the wide moving jam.

At low scales in time and space, both GPs with a regular (Fig. 44a,b) and non-regular pinch regions (Fig. 44c,d) exhibit *spatiotemporal microscopic structures* in which regions of lower speed alternate with regions of higher speed; in both cases, we find that these regions propagate upstream. At the first glance, we can see only that the mean frequency of low speed regions increases with the increase in  $T_B$ , i. e., when the bottleneck strength increases; this result has already been mentioned in Subsect. “Evolution of Traffic Phases at Heavy Bottlenecks”. However, if we consider the microscopic structures of GPs with a non-regular pinch region in larger scales in space and time (Figs. 45 and 46), we can see that there is a crucial qualitative differ-



**Traffic Congestion, Spatiotemporal Features of, Figure 42**

Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 15.8 km within congested traffic in the left figures (left) and right lanes (right) related to different  $T_B = 2.4$  (a–d) and 6 sec (e–h). For e–h  $q^{(\text{cong})} = 626$  vehicles/h/lane

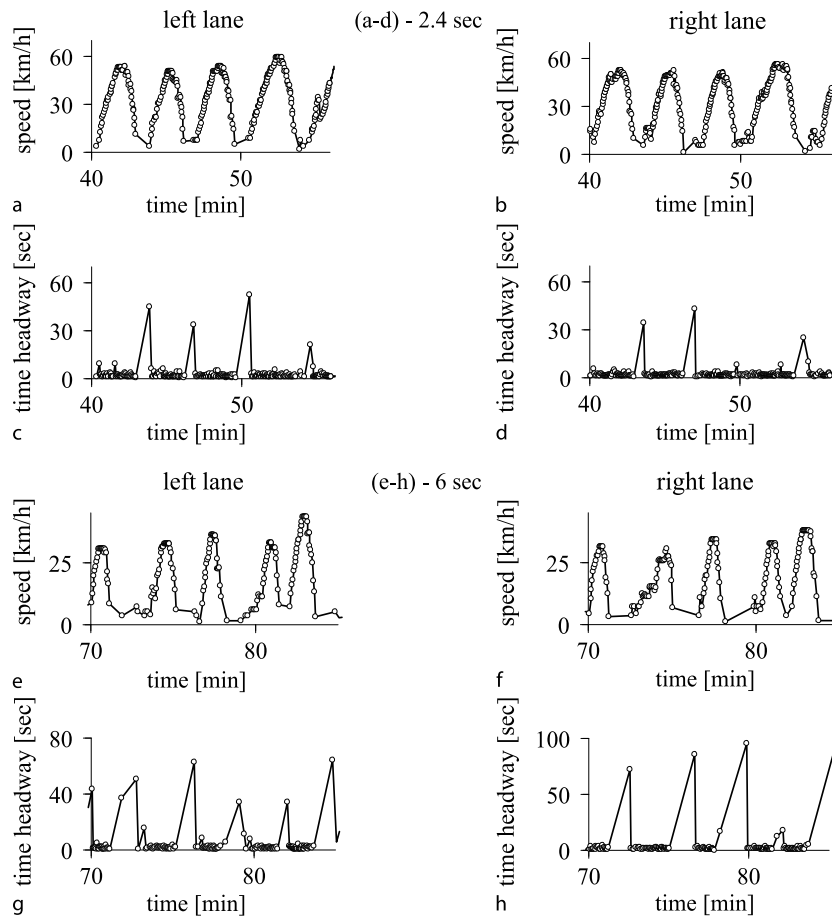
ence between them and the microscopic structures of GPs with a regular pinch region (Fig. 44a,b).

GPs with a regular pinch region exhibits a regular dynamics of wide moving jams discussed in Subsect. “General Congested Patterns”, when the jams propagate upstream of the pinch region on a homogeneous road (Fig. 44a,b): some of the wide moving jams can dissolve during their upstream propagation (a so-called jam suppression effect) [38]. An example of the effect of the dissolution of wide moving jams can be seen in Fig. 44b (labeled “jam dissolution”).

However, under condition (11), i. e., when a GP with a non-regular pinch region is realized, wide moving jams of the GP can exhibit complex and non-regular spatiotemporal dynamic behavior (Figs. 45 and 46). This non-regular jam dynamics is associated with the following effects:

- (1) The effect of the *splitting* of a flow interruption interval onto two (or more) flow interruption intervals (Figs. 45a and 46c).
- (2) The effect of the *emergence* of a new flow interruption interval (Figs. 45a and 46d).
- (3) The effect of the *merging* of two (or more) traffic flow interruption intervals (Figs. 45a and 46a).
- (4) The effect of the *dissolution* of a flow interruption interval (Figs. 45b,c and 46).

The dynamic effects (1)–(4) occur spontaneously during upstream propagation of wide moving jams. In different lanes, these effects occur often independently of each other and at different road locations. A spatiotemporal competition of these effects as well as a diverse variety of the *initiating* and *resulting* dynamic effects determine a non-regular

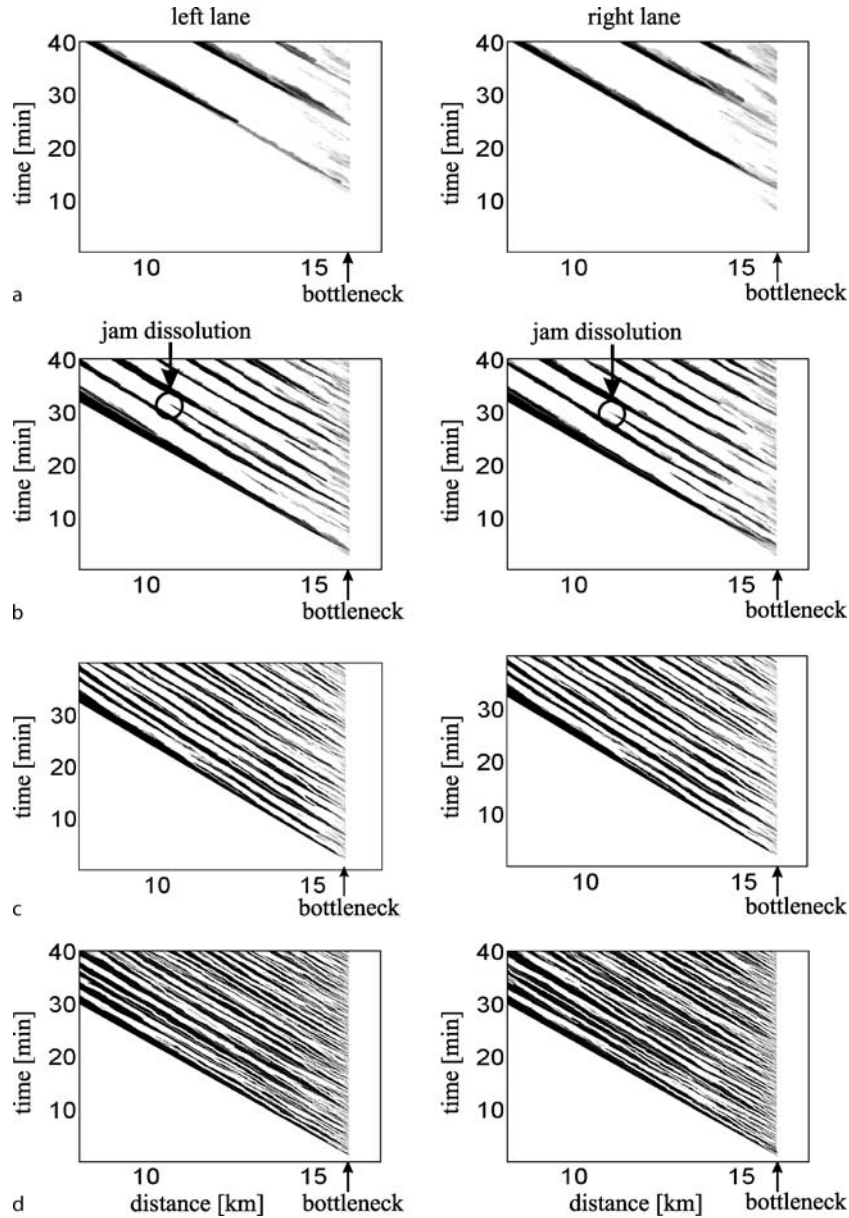


**Traffic Congestion, Spatiotemporal Features of, Figure 43**

Simulated single-vehicle data for speed (a, b, e, f) and time headway (c, d, g, h) measured by a virtual detector at location 10 km within congested traffic in the left lanes (*left*) and right lanes (*right*) related to different  $T_B = 2.4$  (a–d) and 6 sec (e–h)

dynamic behavior of wide moving jams. The variety of the initiating and resulting effects associated with the effects (1)–(4) is as follows:

- (i) The splitting of a flow interruption interval within a wide moving jam can result from the emergence of a new region of moving blanks within the jam.
- (ii) The splitting of a flow interruption interval within an initial wide moving jam can result in the splitting of the jam onto two (or more) wide moving jams; in the latter case, synchronized flow(s) (or free flow(s)) occurs that separates the emergent wide moving jams. This is related to an  $J \rightarrow S$  (or  $J \rightarrow F$ ) transition(s) occurring within the initial jam.
- (iii) The effect of the emergence of a new flow interruption interval within a wide moving jam can result from the splitting of a region of moving blanks onto two (or more) regions of moving blanks separated by the emergent flow interruption interval.
- (iv) The effect of the emergence of a new flow interruption interval occurring within synchronized flow between two wide moving jams can result in the emergence of a new wide moving jam. This is related to an  $S \rightarrow J$  transition occurring in metastable synchronized flow between the jams.
- (v) The effect of the merging of two (or more) traffic flow interruption intervals can result from the dissolution of two (or more) regions of moving blanks within a wide moving jam without wide moving jam dissolution.
- (vi) The effect of the merging of two (or more) traffic flow interruption intervals can result in the merging of two (or more) wide moving jams.



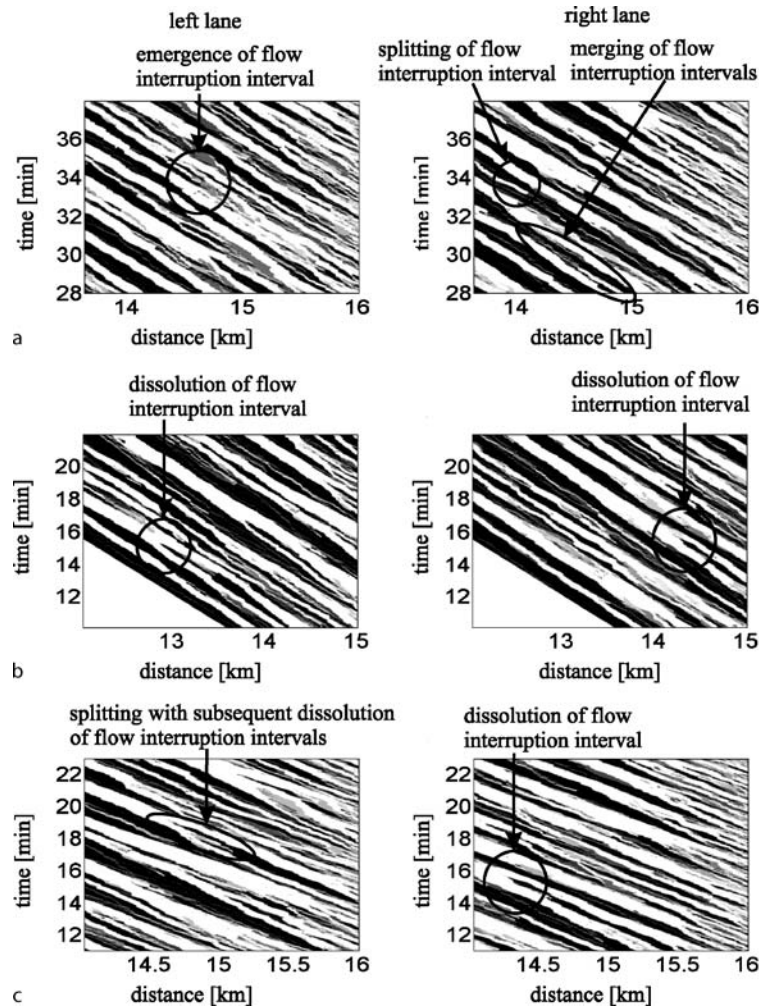
**Traffic Congestion, Spatiotemporal Features of, Figure 44**

Simulated spatiotemporal microscopic structures of GPs with regular (a, b) and non-regular pinch regions (c, d): Single-vehicle speeds within GPs presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (*left*) and right (*right*) road lanes associated with speed distributions at different  $T_B = 1.8$  (a), 2.4 (b), 4 (c), 12 sec (d). For  $c q^{(\text{cong})} = 776$  vehicles/h/lane. The upstream boundary of the bottleneck region is at location  $x = 16$  km (labeled "bottleneck")

- (vii) The effect of the dissolution of a traffic flow interruption interval within a wide moving jam can result from the merging of two (or more) regions of moving blanks within the jam.
- (viii) The effect of the dissolution of a traffic flow interruption interval can result in the dissolution of a wide moving jam.

The initiating and resulting effects (v)–(viii) decrease the mean frequency of regions of flow interruption intervals and regions of moving blanks within wide moving jams as well as the mean frequency of wide moving jams of a GP. In particular, the effects (vii) and (viii) occur very frequently close to the upstream boundary of the pinch region of GPs (see road locations between 16 and 15 km in





#### Traffic Congestion, Spatiotemporal Features of, Figure 45

Fragments of Fig. 44d in larger scales in time and space in the left lanes (left) and right lanes (right). White regions (single-vehicle speeds are equal to or higher than 5.4 km/h) are related to synchronized flows and moving blanks. Black regions (single-vehicle speeds are equal to zero) are related to flow interruption intervals.  $T_B = 12$  sec

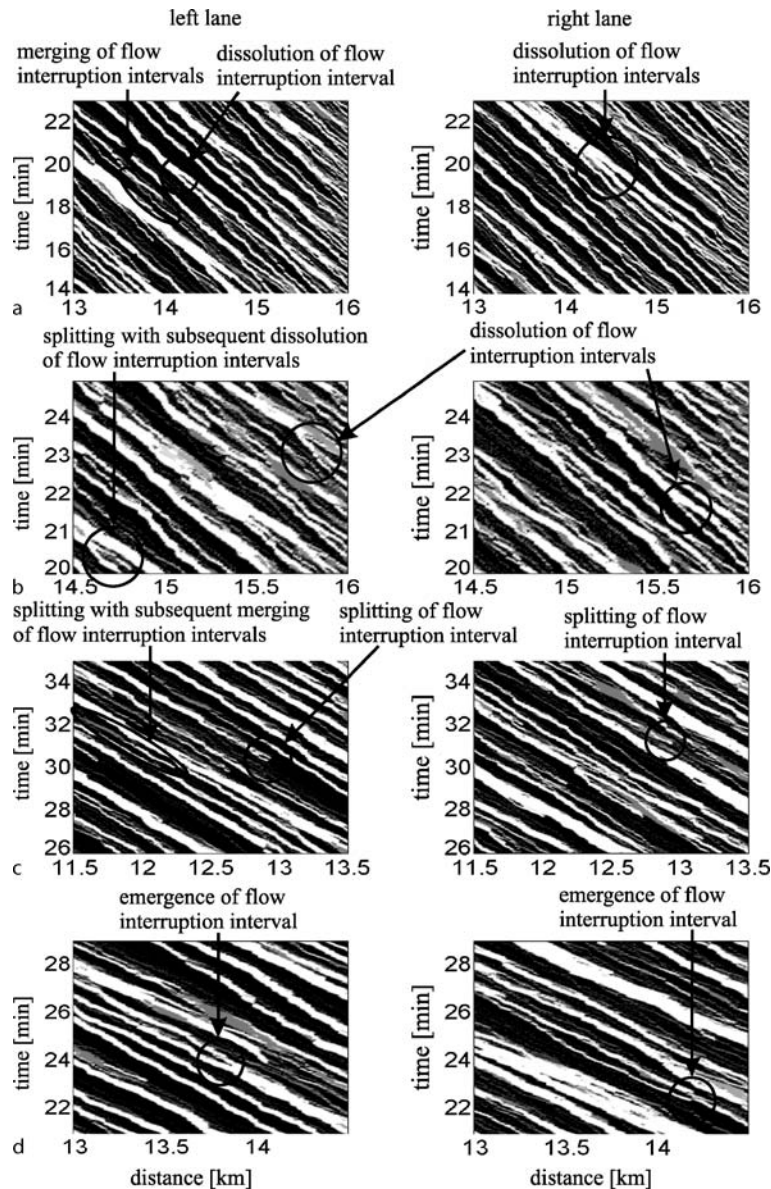
Figs. 45a,c and 46a,b). This leads to a decrease in the mean frequency of wide moving jams, when the jams propagate upstream.

There can be diverse sequences of the effects (1)–(4) over time with a random and arbitrarily sequence of the effects (i)–(viii). For example, the splitting of a flow interruption interval onto two intervals can be followed by the subsequent dissolution of these two intervals (Figs. 45c and 46b); the splitting of a flow interruption interval onto two intervals can be followed by the subsequent merging of these two intervals (Fig. 46c). The random and arbitrarily spatiotemporal sequence of the effects discussed can explain the non-regular spatiotemporal jam dynamics found in GPs with a non-regular pinch region. In turn,

this non-regular spatiotemporal jam behavior can also explain the result of Subsect. “[Evolution of Traffic Phases at Heavy Bottlenecks](#)” that we cannot distinguish a sequence of wide moving jams in 1 min average data (Fig. 40i–l): the microscopic non-regular jam dynamics is averaged in this 1 min average data leading to non-regular and non-homogeneous speed distributions in which this real fine spatiotemporal jam dynamics *cannot be found*.

#### Microscopic Spatiotemporal Structure of Mega-Jam

The crucial difference between GPs and a mega-jam, which occurs under condition (13), is as follows: far enough upstream of the bottleneck, single-vehicle data of



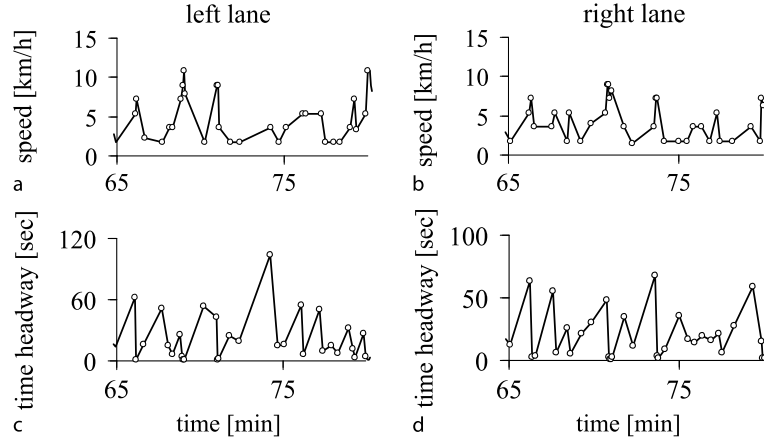
**Traffic Congestion, Spatiotemporal Features of, Figure 46**

Simulated single-vehicle speeds within GP presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (left) and right lanes (right). White regions (single-vehicle speeds higher than 3.6 km/h) are related to synchronized flows and moving blanks. Black regions (single vehicle speeds are equal to zero) are related to flow interruption intervals.  $T_B = 30$  sec.  $q^{(cong)} = 218$  vehicles/h/lane

a GP shows a sequence of wide moving jams separated by non-interrupted flows (Fig. 43); in contrast, within the mega-jam we can distinguish *no* sequence of wide moving jams (Fig. 47a,b). Rather than a sequence of wide moving jams, within the mega-jam there is a complex sequence of upstream moving flow interruption intervals (black regions in Fig. 48) for which criterion (17) is satis-

fied (Fig. 47c,d). These flow interruption intervals are separated by short time intervals within which vehicles move with low speeds (Fig. 47c,d). The latter time intervals are associated with moving blanks (white regions in Fig. 48).

Thus as the microscopic spatiotemporal structure of a wide moving jam, the microscopic structure of a mega-jam consists of an alternation of flow interruption inter-



**Traffic Congestion, Spatiotemporal Features of, Figure 47**

Simulated single-vehicle data for speed (a, b) and time headway (c, d) measured by a virtual detector at location  $x = 5$  km related to  $T_B = 60$  sec. Left and right figures are related to the left and right lanes, respectively. The upstream boundary of the bottleneck is at location  $x = 16$  km

vals and moving blanks (Fig. 48). This explain the term a *mega-jam* that is also a wide moving jam, however, with an extremely great width continuously growing over time. The continuous growth of the mega-jam width occurs as long as the flow rate upstream of the mega-jam  $q_{in}$  exceeds  $q^{(cong)}$  (Fig. 38d). As for GPs with a non-regular pinch region, within the mega-jam we found a very complex and non-regular spatiotemporal dynamics of flow interruption intervals and moving blanks. However, because within the mega-jam there are no wide moving jams separated by non-interrupted flows, the non-regular mega-jam dynamics is associated with the dynamic effects (i), (iii), (v), and (vii) of Subsect. “Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics” only.

### Physics of Non-regular Spatiotemporal Dynamics of Traffic Congestion at Heavy Bottlenecks

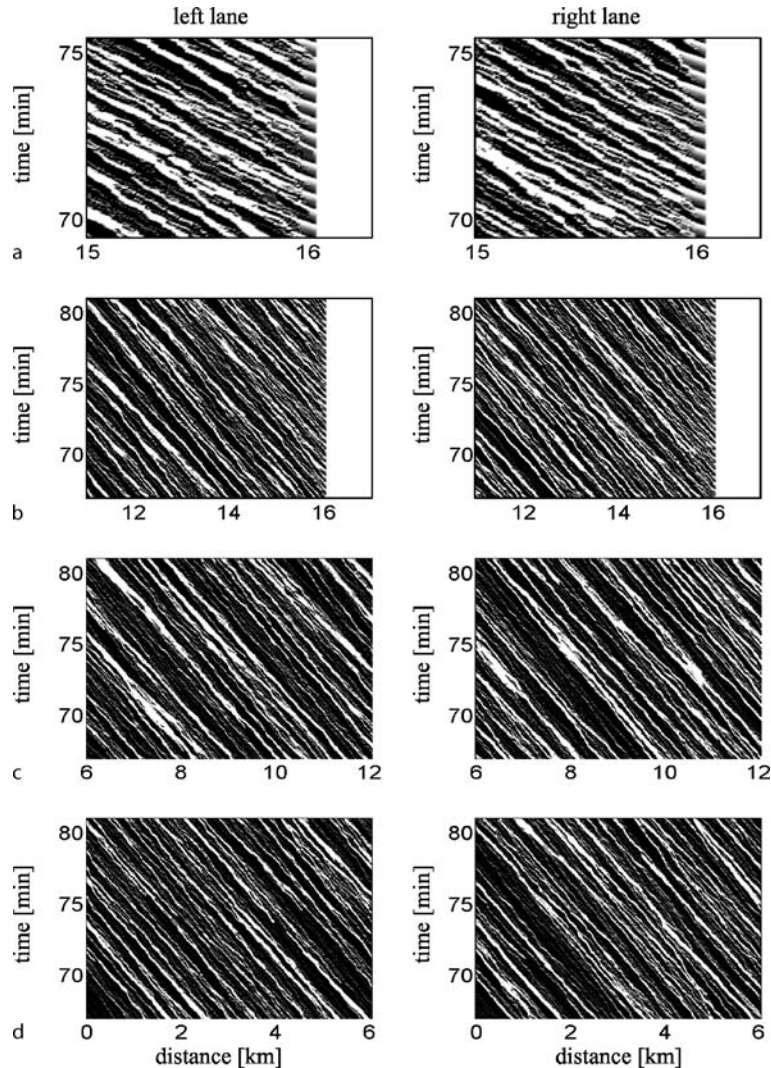
A phase transition from synchronized flow to a wide moving jam is a first-order phase transition, which is characterized by a random time delay  $T_{SJ}$  [38].  $T_{SJ}$  includes a random time delay of spontaneous nucleation of a narrow moving jam and a random time of the jam growth within the pinch region until the jam transforms into a wide moving jam at the upstream boundary of the pinch region. Thus the smaller  $T_{SJ}$ , the shorter  $L^{(pinch)}$ . The random character of  $T_{SJ}$  explains a time-dependence of  $L^{(pinch)}(t)$  (Fig. 41c,e,g). We found that the greater the bottleneck strength, i.e., the smaller  $q^{(cong)}$ , the lower the speed and the greater the density within the pinch region. However, the lower the speed and the greater the density in synchro-

nized flow, the smaller the nucleus required for the emergence of a flow interruption interval, i.e., the smaller the mean random time delay  $T_{SJ}^{(mean)}$  ([38], ► [Traffic Congestion, Modeling Approaches to](#)).

Thus the greater the bottleneck strength, i.e., the smaller  $q^{(cong)}$ , the greater the probability for the occurrence of a negligibly small  $T_{SJ}$  and the emergence of wide moving jams directly upstream of the bottleneck; in the latter case  $L^{(pinch)} = 0$ . This explains why under condition (11) the regular structure of the GP breaks due to the random disappearance of the pinch region during some time intervals. This explains also why  $L_{mean}^{(pinch)}$  decreases up to zero, when a strong increase in the bottleneck strength resulting in the subsequent strong decrease of  $q^{(cong)}$  causes a decrease in  $T_{SJ}^{(mean)}$  up to zero, i.e., when under condition (12) the pinch region does not exist.

As within the pinch region of synchronized flow of a GP, at greater bottleneck strengths resulting in smaller flow rates  $q^{(cong)}$  within congested patterns, the synchronized flow speed between wide moving jams is also low (compare synchronized flow speeds between the jams in Fig. 43 for different  $T_B$ ) and the density is great. Therefore, the probability of the emergence of flow interruption intervals between the jams, i.e., the emergence of new wide moving jams, increases rapidly with the bottleneck strength. As a result, many short flow interruption intervals appear at greater bottleneck strengths. On the other hand, short flow interruption intervals can dissolve easily, when the density decreases randomly upstream of these intervals. A spatiotemporal competition between the random emergence and dissolution of flow interruption inter-





**Traffic Congestion, Spatiotemporal Features of, Figure 48**

Simulated single-vehicle speed data within mega-jam presented in space and time by regions with variable darkness (the lower the speed, the darker the region) in the left (*figures left*) and right lanes (*right*). *White regions* (single-vehicle speeds higher than 1.8 km/h) are related to moving blanks. *Black regions* (single vehicle speeds are equal to zero) are related to flow interruption intervals.  $T_B = 60$  sec. The upstream boundary of the bottleneck is at location  $x = 16$  km

vals can explain the non-regular spatiotemporal jam dynamics of Subject. “[Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics](#)”.

There are several sources for random speed disturbances in the traffic flow model used for simulations considered in Subject. “[Evolution of Traffic Phases at Heavy Bottlenecks](#)”–“[Microscopic Spatiotemporal Structure of Mega-Jam](#)”, whose growth leads to the emergence of flow interruption intervals in synchronized flows: lane changing, the variety of the random delays in vehicle acceleration and decelerations ([38], ► [Traffic Conges-](#)

[tion, Modeling Approaches to](#)). In addition, within the bottleneck vehicles are forced to move at much longer safe time headways  $\tau^{(\text{safe})} = T_B$  than away of the bottleneck  $\tau^{(\text{safe})} = 1$  sec. For this reason, the mean amplitude of speed disturbances in the pinch region is considerably greater than the one in synchronized flows between the jams on a homogeneous road. This explains why the non-regular jam dynamics is considerably visible only at greater bottleneck strengths than the critical bottleneck strength for the occurrence of GPs with a non-regular pinch region.

When the bottleneck strength increases strongly and condition (10), i.e., (13) is satisfied, wide moving jams merge into a mega-jam. The whole flow rate within any mega-jam is supplied by moving blanks *only*. As a result, under condition (10) we get

$$q^{(\text{cong})} = q_{\text{mega}}^{(\text{blanks})}, \quad (18)$$

where  $q_{\text{mega}}^{(\text{blanks})}$  denotes the average flow rate associated with moving blanks within the mega-jam, which satisfies the condition

$$q_{\text{mega}}^{(\text{blanks})} \leq q^{(\text{blanks})}, \quad (19)$$

where, as defined above,  $q^{(\text{blanks})}$  is the average flow rate associated with moving blanks within wide moving jams separated by non-interrupted flows. The equality in (19) is related to (9). Thus if under condition (10) the bottleneck strength increases and therefore  $q^{(\text{cong})}$  decreases, then in accordance with (18), the flow rate  $q_{\text{mega}}^{(\text{blanks})}$  decreases.

As mentioned in Subsect. “Evolution of Traffic Phases at Heavy Bottlenecks” (see(15)), the threshold bottleneck strength for the pinch region existence is smaller than the critical bottleneck strength for the mega-jam formation. To explain that this result has a general character, let us assume that the bottleneck strength is initially greater than the critical one for the mega-jam formation. In this case,  $q^{(\text{cong})} = q_{\text{mega}}^{(\text{blanks})} < q^{(\text{blanks})}$ . If now the bottleneck strength decreases gradually, then the flow rate  $q^{(\text{cong})}$  increases. The flow rate  $q^{(\text{cong})}$  can be supplied by moving blanks only up to the critical bottleneck strength satisfying condition (9), i.e.,  $q^{(\text{cong})} = q^{(\text{blanks})}$ . When the bottleneck strength decreases further and becomes smaller than the critical one for the mega-jam formation, then  $q^{(\text{cong})} > q^{(\text{blanks})}$ . The difference  $\Delta q_{\text{blanks}} = q^{(\text{cong})} - q^{(\text{blanks})}$  should be supplied by vehicles accelerating from wide moving jams; this means that wide moving jams separated by synchronized flows must appear in congested traffic. In particular, such synchronized flows appear just upstream of the bottleneck as a result of the upstream propagation of wide moving jams that emerge directly upstream of the bottleneck. Nevertheless, a pinch region of the synchronized flow at the bottleneck does not still appear as long as the difference  $\Delta q_{\text{blanks}}$  is small enough. Indeed, the pinch region appears only, when at least one of the wide moving jams occurs at a *finite distance* upstream of the bottleneck in the synchronized flow due to a growing narrow moving jam that emerges initially in this flow. This is possible only, if the former wide moving jam due to its upstream propagation is at a great enough distance from the bottleneck. The latter can be realized, when the

difference  $\Delta q_{\text{blanks}}$  exceeds also some *finite value*, i.e., the threshold bottleneck strength for the pinch region existence should be always smaller than the critical one for the mega-jam formation. Thus condition (15) is a general result for any heavy bottleneck.

### Comparison with Empirical Results

To compare the above theory of traffic congestion at heavy bottlenecks [39,40] with empirical congested patterns, one should have measured data for traffic congestion at a bottleneck, whose strength should be manually continuously changeable from the one associated with usual bottlenecks like on- and off-ramps to great bottleneck strengths associated with heavy bottlenecks caused by bad weather conditions or accidents. Unfortunately, such measured data is not available. However, we can compare the theory with measured data related to two limiting cases:

- (i) Traffic congestion at an usual on-ramp bottleneck (Fig. 16).
- (ii) Traffic congestion at a heavy bottleneck caused by bad weather conditions – snow and ice on a road (Fig. 49).

The speed distributions with congested patterns found in simulations (Fig. 38a,b) for the range of  $T_B = 1.6\text{--}2.4$  sec associated with the flow rate range

$$q^{(\text{cong})} = q^{(\text{pinch})} = 1120\text{--}1800 \text{ vehicles/h/lane} \quad (20)$$

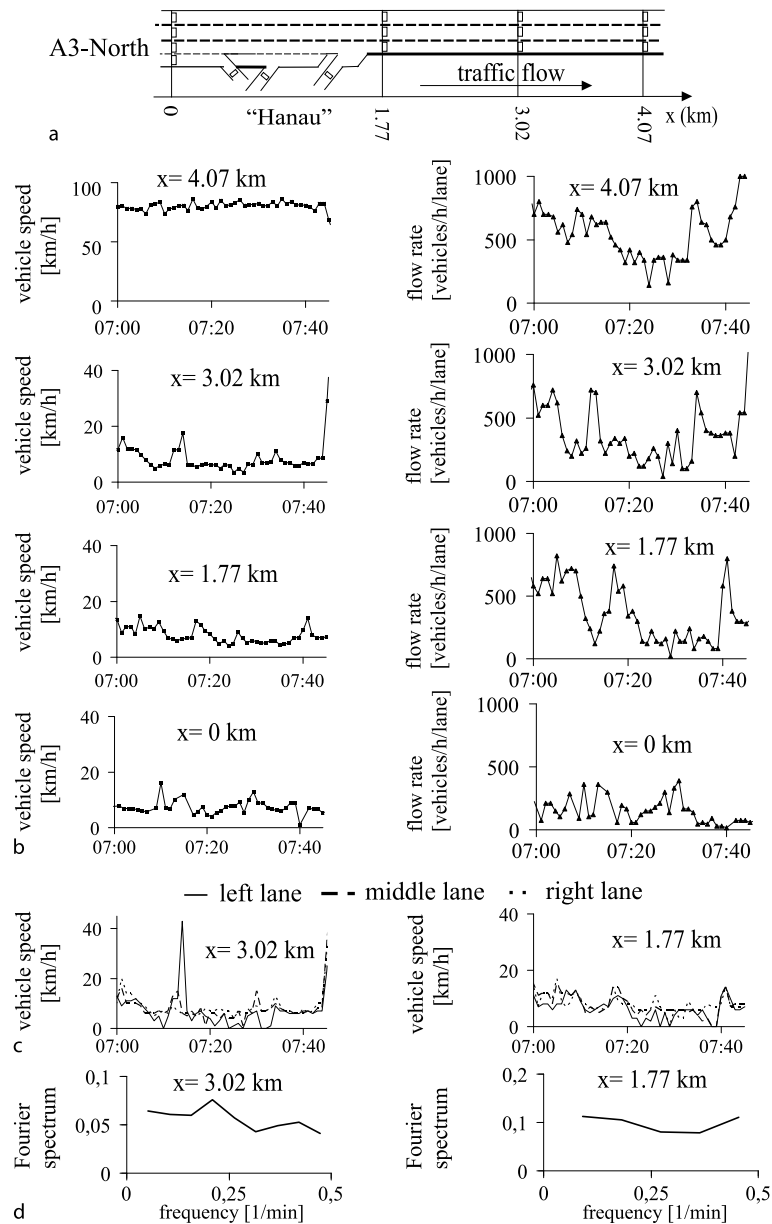
are qualitatively the same as those in an empirical GP shown in Fig. 16 and in all other known empirical GPs [38]. Moreover, quantitative values of empirical flow rates  $q^{(\text{pinch})}$  (7) are approximately associated with the theoretical result (20). In particular, within the pinch region of the GP shown in Fig. 16a, the flow rate averaged between 7:00 and 8:00 and across the road is  $q^{(\text{pinch})} = 1200$  vehicles/h/lane.

In contrast, for  $T_B > 6$  sec associated with the average flow rate  $q^{(\text{cong})} < 625$  vehicles/h/lane simulated congested patterns at heavy bottlenecks (Fig. 38c,d) are qualitatively the same as those we found in empirical data measured on many various days (and years) on different free-ways, when very heavy bottlenecks caused by bad weather conditions or accidents occur for which empirical flow rates

$$q^{(\text{cong})} \lesssim 600 \text{ vehicles/h/lane}. \quad (21)$$

In this case, rather than regular structure of traffic congestion of GPs (Figs. 38a,b and 16), both theoretical (Figs. 38c,d, 40i) and empirical traffic congested patterns (Fig. 49) exhibit non-regular spatiotemporal structure of





**Traffic Congestion, Spatiotemporal Features of, Figure 49**

Empirical structure of congestion caused by snow and ice: **a** Scheme of road detector arrangement on a section of the freeway A3-North in Germany near the intersection "Hanau". **b** Average speed and flow rate across the freeway at different locations. **c** Speed in different lanes at two locations. **d** Fourier spectra of time-dependences of the speed in the left lane at two locations. 1 min average data

congestion in which *no* sequences of wide moving jams can be distinguished in 1 min average data.

An example of such an empirical congested traffic pattern is shown in Fig. 49. Rather than the regular structure of congestion within the GP (Fig. 16), in measured data associated with bad weather conditions a non-regular spa-

tiotemporal structure of congestion is observed (Fig. 49). A heavy bottleneck appears on February 02, 2006 between locations 4.07 and 3.02 km due to snow and ice. Upstream of the bottleneck, very low speed and flow rate patterns ( $x \leq 3.02$  km in Fig. 49b) are observed. Downstream of the bottleneck ( $x = 4.07$  km) vehicles have escaped from

the congestion (speed is high), however, the flow rate is very small because the bottleneck reduced the average flow rate within the congestion strongly. For example, at  $x = 3.02$  km, the flow rate  $q^{(\text{cong})}$  averaged between 7:00 and 7:40 and across the road is 513 vehicles/h/lane. In contrast with the GP shown in Fig. 16, within traffic congestion shown in Fig. 49 non-regular low speed patterns are observed in which *no* sequence of wide moving jams can be distinguished ( $x \leq 3.02$  km). This conclusion is regardless of the flow rates to on- and off-ramps, which in the data set lead to an increase in  $q^{(\text{cong})}$  at  $x = 1.77$  km in comparison with  $q^{(\text{cong})}$  at  $x = 0$  km (Fig. 49a,b). If speeds in different lanes are compared (Fig. 49c), we find even more non-regular speed time-dependences: whereas no vehicles pass a detector in one of the lanes, i. e., the speed is zero, during the same time interval the average speed in other lanes can be higher than zero. This explains why in time-dependences of speeds averaged across the road (Fig. 49b), this very low speed is seldom equal to exactly zero.

As congested traffic in measured data (Fig. 49b,c), simulated congested traffic at a heavy bottleneck is also non-homogeneous in space and time (Fig. 40i,j). In the measured and simulated data, as follows from Fourier spectra of speed time-dependences, these non-homogeneous congested patterns exhibit non-regular spatiotemporal behavior (Figs. 49d and 40l). This is in contrast with the regular structure of traffic congestion within the GP (Fig. 16).

As in measured data (Fig. 49b,c), in 1 min average data related to traffic congestion in simulations very low speed and flow rate patterns are realized upstream of the bottleneck in which *no* sequence of wide moving jams can be distinguished; downstream of the bottleneck ( $x > 16.3$  km) vehicles have escaped from the congestion (speed is high) (Fig. 38c), however, the flow rate is very small because the bottleneck reduced the average flow rate within the congestion strongly. A summary of the above comparison of theory of traffic congestion with measured data is as follows:

- (i) Spatiotemporal distributions of the average speed and flow rate within GPs occurring at usual bottlenecks are qualitatively the same in the above theory and measured data: in both simulated and measured data, GPs consist of the pinch region and wide moving jams upstream. Measured and simulated speed time-dependences show *regular* character of wide moving jam propagation.
- (ii) In contrast with item (i), in 1 min average data related to simulated congested patterns and to measured data for traffic congestion at heavy bottlenecks *no* sequence

of wide moving jams can be distinguished. In simulations and measured data, spatiotemporal distributions of the average speed and flow rate in these congested patterns are non-homogeneous in space and time. Fourier spectra of the associated non-homogeneous time-dependences of the speed show non-regular character of traffic congestion at the heavy bottlenecks both in simulations (Fig. 40l) and measured data (Fig. 49d).

As follows from Subsects. “[Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics](#)” and “[Microscopic Spatiotemporal Structure of Mega-Jam](#)”, the non-regular pattern behavior of item (ii) can be explained by two different reasons:

- (1) The occurrence of a GP with the non-regular dynamics of wide moving jams.
- (2) The occurrence of a mega-jam.

However, based on 1 min average data these two reasons of non-regular traffic congestion cannot be distinguished from each other. Indeed, theoretical results of Subsects. “[Microscopic Spatiotemporal Features of Non-regular Moving Jam Dynamics](#)” and “[Microscopic Spatiotemporal Structure of Mega-Jam](#)” allow us to suggest that at a very heavy bottleneck caused for example by bad weather conditions or accidents there should be a very fine microscopic spatiotemporal structure of traffic congestion associated with non-regular spatiotemporal dynamics of wide moving jams or/and flow interruption intervals as well as moving blanks. However, this theoretical fine structure of traffic congestion, even if it exists in real traffic, would be averaged in 1 min average measured data, i. e., they could not be found in this data. Thus to make a more detailed comparison of the theory of traffic congestion at heavy bottlenecks with measured data, single vehicle data measured within traffic congestion at heavy bottlenecks is required. Unfortunately, currently this measured single vehicle data is not available. Such a comparison of the theory of traffic congestion at heavy bottlenecks presented in Sect. “[Congested Patterns at Heavy Bottlenecks](#)” with measured single vehicle data is a separate and important task of further investigations.

## Conclusions.

### Fundamental Empirical Features of Spatiotemporal Congested Freeway Traffic Patterns

There are empirical features of phase transitions and spatiotemporal congested patterns at freeway bottlenecks that are reproducible in traffic observations on numerous days and years on different freeways in various countries. Thus,

these qualitative empirical features remain in traffic flows with very different driver behavioral characteristics and vehicle parameters. These fundamental empirical macroscopic and microscopic features of the phase transitions and congested traffic are explained within the framework of three-phase traffic theory.

Fundamental empirical features of phase transitions and Congested patterns at freeway bottlenecks are as follows:

- (i) Traffic can be either “free” or “congested.” There are two types of the onset of congestion (traffic breakdown) in free flow at an effectual bottleneck, *spontaneous* and *induced*. The traffic breakdown and freeway capacity at the bottleneck possess a probabilistic nature.
- (ii) An induced speed breakdown in free flow at the bottleneck can be caused either by moving jam propagation through the bottleneck or the upstream propagation of a synchronized flow pattern (SP) that has initially occurred downstream of the bottleneck.
- (iii) In congested traffic, two different traffic phases can be distinguished: the synchronized flow phase and the wide moving jam phase. These traffic phases are defined through the use of the spatiotemporal empirical (objective) criteria [J] and [S]. Thus, there are three traffic phases in freeway traffic: (1) free flow; (2) synchronized flow; (3) wide moving jam.
- (iv) Traffic breakdown in free flow at the freeway bottleneck is associated with a phase transition from free flow to synchronized flow ( $F \rightarrow S$  transition).
- (v) Wide moving jams do not emerge spontaneously in free flow. Wide moving jams can emerge spontaneously only in synchronized flow; wide moving jams emerge due to the sequence of  $F \rightarrow S \rightarrow J$  transitions.
- (vi) The higher the density in synchronized flow, the higher the frequency of spontaneous moving jam emergence in that synchronized flow. In synchronized flow of higher speeds and lower densities, wide moving jams should not necessarily emerge spontaneously.
- (vii) During the dynamics of transformation of a narrow moving jam into a wide moving jam the mean velocity of the downstream jam front tends to the characteristic velocity. Wide moving jams possess *characteristic* parameters and features that do not depend on initial conditions and perturbations in traffic. The characteristic parameters are the same for different wide moving jams. The characteristic parameters can depend on control parameters of traffic (weather, road conditions, etc.). These characteristic parameters and features are as follows:
  - (a) The mean velocity of the downstream front of a wide moving jam. The wide moving jam propagates upstream through any traffic state and through any bottleneck while maintaining the mean velocity of the downstream jam front.
  - (b) When free flow occurs in the wide moving jam outflow, the flow rate and density in the jam outflow are also characteristic parameters.
  - (c) The flow rate in the jam outflow is lower than the maximum possible flow rate in free flow.
- (viii) There are two main types of congested patterns at an isolated effectual freeway bottleneck: synchronized flow (SP) and general patterns (GP).
- (ix) In contrast to wide moving jam propagation, an SP that propagates upstream is caught at the bottleneck at which free flow has been before (catch effect) when the SP reaches the bottleneck. Due to the catch effect, an induced  $F \rightarrow S$  transition can occur.
- (x) There are three types of SPs: a moving SP, a widening SP, and a localized SP.
- (xi) There can be complex spontaneous transformations between various congested patterns at the bottleneck over time.
- (xii) At the same traffic demand various congested patterns can be formed at the bottleneck (the probabilistic nature of spatiotemporal congested patterns at the bottleneck).
- (xiii) If two or more adjacent effectual bottlenecks are close to one another, expanded congested patterns (EP) often emerge. In an EP, synchronized flow affects at least two adjacent effectual bottlenecks.
- (xiv) When a wide moving jam that has initially emerged downstream of an effectual bottleneck propagates through synchronized flow upstream of the bottleneck, this foreign wide moving jam can considerably influence other moving jams, which are just emerging in that synchronized flow. In particular, the foreign wide moving jam can suppress the growth of a narrow moving jam that is close enough to the downstream front of the foreign wide moving jam.
- (xv) For each effectual freeway bottleneck, or each set of several adjacent effectual bottlenecks where congested patterns occur, the spatiotemporal structure of a congested pattern possesses some predictability, i.e., characteristic, unique, and reproducible features.
- (xvi) Rather than regular wide moving jam propagation observed upstream of usual bottlenecks like on- and

off-ramp bottlenecks (item iii-xv), no sequence of wide moving jams can be distinguished in 1 min average measured data for non-regular and non-homogeneous congested traffic, which is usually observed upstream of a heavy bottleneck caused for example by bad weather conditions or accidents. This empirical phenomenon can be explained by results of a theory of traffic congestion at heavy bottlenecks of Sect. “Congested Patterns at Heavy Bottlenecks”, in particular by random disappearance and appearance of the pinch region of synchronized flow of a GP over time and a non-regular dynamics of wide moving jams within the GP, as well as by the merger of wide moving jams into a mega-wide moving jam (mega-jam) that should occur upstream of a very heavy bottleneck.

### Future Directions

Although many empirical macroscopic spatiotemporal congested pattern features have already been understood, there are still many problems in understanding empirical microscopic features of these patterns as well as phase transitions within congested patterns, in particular occurring on freeway sections with many adjacent bottlenecks.

Microscopic features of phase transitions between the synchronized flow and wide moving jam phases have not still sufficiently been understood.

Spatiotemporal features of congested patterns associated with freeway networks, e. g., possible induced phase transitions and possible complex pattern transformations caused by upstream propagation of congested patterns onto neighborhood freeway sections of the network have almost not been found.

Traffic flow models in the framework of three-phase traffic theory are only at the beginning of their development. There are almost no analytical studies of the spatiotemporal characteristics of congested patterns within this theory.

An empirical microscopic spatiotemporal dynamics of traffic congestion at heavy bottlenecks has not still been studied.

These and many other unsolved problems are interesting and important fields of the future empirical and theoretical investigations of spatiotemporal congested traffic patterns.

We can also expect that features of spatiotemporal dynamics of traffic congestion reviewed in this article can be applied for a development of new effective methods of traffic control and management as well as driver assistant systems that increase vehicle safety in traffic.

### Acknowledgments

I would like to thank Sergey Klenov, Andreas Hiller, Hubert Rehborn, Mario Aleksić, Ines Maiwald-Hiller and Olivia Brickley for help and useful suggestions.

### Bibliography

1. Bellomo N, Coscia V, Delitala M (2002) On the Mathematical Theory of Vehicular Traffic Flow I. Fluid Dynamic and Kinetic Modelling. *Math Mod Meth Appl Sci* 12:1801–1843
2. Chowdhury D, Santen L, Schadschneider A (2000) Statistical Physics of Vehicular Traffic and Some Related Systems. *Phys Rep* 329:199
3. Cremer M (1979) *Der Verkehrsfluss auf Schnellstrassen*. Springer, Berlin
4. Daganzo CF (1997) *Fundamentals of Transportation and Traffic Operations*. Elsevier, New York
5. Davis LC (2004) Multilane simulations of traffic phases. *Phys Rev E* 69:016108
6. Davis LC (2006) Controlling traffic flow near the transition to the synchronous flow phase. *Physica A* 368:541–550
7. Davis LC (2006) Effect of cooperative merging on the synchronous flow phase of traffic. *Physica A* 361:606–618
8. Davis LC (2007) Effect of adaptive cruise control systems on mixed traffic flow near an on-ramp. *Physica A* 379:274–290
9. Edie LC (1961) Car-Following and Steady State Theory for Non-Congested Traffic. *Oper Res* 9:66–77
10. Edie LC, Foote RS (1958) Traffic Flow in Tunnels. *Highw Res Board Proc Ann Meet* 37:334–344
11. Edie LC, Foote RS (1960) Effect of Shock Waves on Tunnel Traffic Flow. In: *Highway Research Board Proceedings*, vol 39. HRB, National Research Council, Washington DC, pp 492–505
12. Edie LC, Herman R, Lam TN (1980) Observed Multilane Speed Distribution and the Kinetic Theory of Vehicular Traffic. *Transp Sci* 14:55–76
13. Elefteriadou L, Roess RP, McShane WR (1995) Probabilistic Nature of Breakdown at Freeway Merge Junctions. *Transp Res Rec* 1484:80–89
14. Fukui M, Sugiyama Y, Schreckenberg M, Wolf DE (eds) (2003) *Traffic and Granular Flow' 01*. Springer, Berlin
15. Gao K, Jiang R, Hu SX, Wang BH, Wu QS (2007) Cellular-automaton model with velocity adaptation in the framework of Kerner's three-phase traffic theory. *Phys Rev E* 76:026105
16. Gartner NH, Messer CJ, Rathi A (eds) (1997) *Special Report 165: Revised Monograph on Traffic Flow Theory*. Transportation Research Board, Washington DC
17. Haight FA (1963) *Mathematical Theories of Traffic Flow*. Academic Press, New York
18. Hall FL, Agyemang-Duah K (1991) Freeway capacity drop and the definition of capacity. *Trans Res Rec* 1320:91–98
19. Hall FL, Hurdle VF, Banks JH (1992) Synthesis of recent work on the nature of speed-flow and flow-occupancy (or density) relationships on freeways. *Transp Res Rec* 1365:12–18
20. Helbing D (2001) Traffic and related self-driven many-particle systems. *Rev Mod Phys* 73:1067–1141
21. Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) (2000) *Traffic and Granular Flow' 99*. Springer, Berlin
22. Hoogendoorn SP, Luding S, Bovy PHL, Schreckenberg M, Wolf DE (eds) (2005) *Traffic and Granular Flow' 03*. Springer, Berlin

23. Jiang R, Wu QS (2004) Spatial-temporal patterns at an isolated on-ramp in a new cellular automata model based on three-phase traffic theory. *J Phys A: Math Gen* 37:8197–8213
24. Jiang R, Wu QS (2005) First order phase transition from free flow to synchronized flow in a cellular automata model. *Eur Phys J B* 46:581–584
25. Jiang R, Wu QS (2005) Toward an improvement over Kerner-Klenov-Wolf three-phase cellular automaton model. *Phys Rev E* 72:067103
26. Jiang R, Wu QS (2007) Dangerous situations in a synchronized flow model. *Physica A* 377:633–640
27. Jiang R, Hua MB, Wang R, Wu QS (2007) Spatiotemporal congested traffic patterns in macroscopic version of the Kerner-Klenov speed adaptation model. *Phys Lett A* 365:6–9
28. Kerner BS (1998) Theory of Congested Traffic Flow. In: Rysgaard R (ed) *Proceedings of the 3rd Symposium on Highway Capacity and Level of Service*, vol 2. Road Directorate, Ministry of Transport – Denmark, pp 621–642
29. Kerner BS (1998) Traffic Flow: Experiment and Theory. In: *Traffic and Granular Flow' 97. Proceedings of the International Workshop on Traffic and Granular Flow*. Springer, Singapore, pp 239–267
30. Kerner BS (1998) Empirical features of self-organization in traffic flow. *Phys Rev Lett* 81:3797–3400
31. Kerner BS (1999) Congested Traffic Flow: Observations and Theory. *Trans Res Rec* 1678:160–167
32. Kerner BS (1999) Theory of Congested Traffic Flow: Self-Organization without Bottlenecks. In: Ceder A (ed) *Transportation and Traffic Theory*. Elsevier, Amsterdam, pp 147–171
33. Kerner BS (1999) The Physics of Traffic. *Phys World* 12:25–30
34. Kerner BS (2000) Phase Transitions in Traffic Flow. In: *Traffic and Granular Flow' 99*. Springer, Berlin, pp 253–284
35. Kerner BS (2000) Experimental features of the emergence of moving jams in free traffic flow. *J Phys A: Math Gen* 33:L221–L228
36. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
37. Kerner BS (2004) Three-phase traffic theory and highway capacity. *Physica A* 333:379–440
38. Kerner BS (2004) *The Physics of Traffic*. Springer, Berlin
39. Kerner BS (2007) Features of Traffic Congestion caused by bad Weather Conditions and Accidents. E-print [arXiv:0712.1728](https://arxiv.org/abs/0712.1728)
40. Kerner BS (2008) A Theory of Traffic Congestion at Heavy Bottlenecks. *J Phys A Math Theor* 41:215101
41. Kerner BS, Klenov SL (2002) A microscopic model for phase transitions in traffic flow. *J Phys A: Math Gen* 35:L31–L43
42. Kerner BS, Klenov SL (2003) Microscopic theory of spatiotemporal congested traffic patterns at highway bottlenecks. *Phys Rev E* 68:036130
43. Kerner BS, Klenov SL (2004) Spatiotemporal patterns in heterogeneous traffic flow with a variety of driver behavioural characteristics and vehicle parameters. *J Phys A: Math Gen* 37:8753–8788
44. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A: Math Gen* 39:1775–1809
45. Kerner BS, Klenov SL, Wolf DE (2002) Cellular automata approach to three-phase traffic theory. *J Phys A: Math Gen* 35:9971–10013
46. Kerner BS, Klenov SL, Hiller A (2006) Criterion for traffic phases in single vehicle data and empirical test of a microscopic three-phase traffic theory. *J Phys A: Math Gen* 39:2001–2020
47. Kerner BS, Klenov SL, Hiller A, Rehborn H (2006) Microscopic features of moving traffic jams. *Phys Rev E* 73:046107
48. Kerner BS, Klenov SL, Hiller A (2007) Empirical test of a microscopic three-phase traffic theory. *Non Dyn* 49:525–553
49. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2002) Single-vehicle data of highway traffic: Microscopic description of traffic phases. *Phys Rev E* 65:056133
50. Knospe W, Santen L, Schadschneider A, Schreckenberg M (2004) Empirical test for cellular automaton models of traffic flow. *Phys Rev E* 70:016115
51. Koshi M, Iwasaki M, Ohkura I (1983) Some Findings and an Overview on Vehicular Flow Characteristics. In: Hurdle VF (ed) *Proc. 8th International Symposium on Transportation and Traffic Theory*. University of Toronto Press, Toronto, pp 403
52. Laval JA (2007) Linking Synchronized Flow and Kinematic Waves. In: *Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow*. Springer, Berlin, pp 521–526
53. Lee HK, Barlović R, Schreckenberg M, Kim D (2004) Mechanical Restriction versus Human Overreaction Triggering Congested Traffic States. *Phys Rev Lett* 92:238702
54. Leutzbach W (1988) *Introduction to the Theory of Traffic Flow*. Springer, Berlin
55. Mahmassani HS (ed) (2005) *Transportation and Traffic Theory. Proceedings of the 16th Inter. Sym. on Transportation and Traffic Theory*. Elsevier, Amsterdam
56. Mahne R, Kaupužs J, Lubashevsky I (2005) Probabilistic description of traffic flow. *Phys Rep* 408:1–130
57. May AD (1990) *Traffic Flow Fundamentals*. Prentice-Hall, New Jersey
58. Nagatani T (2002) The physics of traffic jams. *Rep Prog Phys* 65:1331–1386
59. Nagel K, Wagner P, Woessler R (2003) Still Flowing: Approaches to Traffic Flow and Traffic Jam Modeling. *Oper Res* 51:681–716
60. Neubert L, Santen L, Schadschneider A, Schreckenberg M (1999) Single-vehicle data of highway traffic: A statistical analysis. *Phys Rev E* 60:6480–6490
61. Persaud BN, Yagar S, Brownlee R (1998) Exploration of the Breakdown Phenomenon in Freeway Traffic. *Trans Res Rec* 1634:64–69
62. Pottmeier A, Thiemann C, Schadschneider A, Schreckenberg M (2007) Mechanical Restriction versus Human Overreaction: Accident Avoidance and Two-Lane Simulations. In: *Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow*. Springer, Berlin, pp 503–508
63. Prigogine I, Herman R (1971) *Kinetic Theory of Vehicular Traffic*. Elsevier, New York
64. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and Granular Flow' 05. Proceedings of the International Workshop on Traffic and Granular Flow*. Springer, Berlin
65. Schreckenberg M, Wolf DE (eds) (1998) *Traffic and Granular Flow' 97. Proceedings of the International Workshop on Traffic and Granular Flow*. Springer, Singapore
66. Sugiyama Y, Fukui M, Kikuchi M, Hasebe K, Nakayama A, Nishinari K, Tadaki S, Yukawa S (2008) Traffic jam without bottleneck – Experimental evidence for the physical mechanism of forming a jam. *New J Phys* 10:033001



67. Tilch B, Helbing D (2000) Evaluation of Single Vehicle Data in Dependence of the Vehicle-Type, Lane, and Site. In: *Traffic and Granular Flow '99*. Springer, Berlin, pp 333–338
68. Treiterer J (1975) Investigation of Traffic Dynamics by Aerial Photogrammetry Techniques. Technical Report PB 246 094. Ohio State University, Columbus
69. Treiterer J, Myers JA (1974) The Hysteresis Phenomenon in Traffic Flow. In: Buckley DJ (ed) *Procs. 6th International Symposium on Transportation and Traffic Theory*. Reed, London, pp 13–38
70. Treiterer J, Taylor JI (1966) Traffic Flow Investigations by Photogrammetric Techniques. *Highway Res Rec* 142:1–12
71. Wang R, Jiang R, Wu QS, Liu M (2007) Synchronized flow and phase separations in single-lane mixed traffic flow. *Physica A* 378:475–484
72. Whitham GB (1974) *Linear and Nonlinear Waves*. Wiley, New York
73. Wiedemann R (1974) *Simulation des Verkehrsflusses*. University of Karlsruhe, Karlsruhe
74. Wolf DE (1999) Cellular automata for traffic simulations. *Phys A* 263:438–451

## Traffic and Crowd Dynamics: The Physics of the City

ARMANDO BAZZANI, BRUNO GIORGINI,  
SANDRO RAMBALDI  
“Fisica della Città” Laboratory, Center L. Galvani  
for Biocomplexity, Physics Department and INFN  
Sezione di Bologna, Bologna, Italy

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Automata Gas and Crowd Behavior  
Automata Gas and Traffic Behavior  
Simulation Results and Empirical Observations  
Future Directions  
Bibliography

### Glossary

**Automaton** A physical particle with internal cognitive dynamics able to process information and to make decisions in order to achieve an aim.

**Cognitive dynamics** A dynamical system that mimics the individual decision mechanisms pointed out by neuroscience and behavioral studies.

**Emergent properties** Macroscopic properties of a statistical systems that are not the result of the superposition of the microscopic states of individual particles.

**Global positioning system** A system to detect the geographical coordinates of a receiver on the earth using a satellite network.

**Mean field theory** a technique introduced in statistical physics that studies the macroscopic dynamics using the dynamics of averaged quantities.

**Self-organized dynamics** Stationary macroscopic states that are the result of collective interactions among the microscopic components.

**Utility function** A function that measures the expected advantages or the disadvantages of future decisions.

### Definition of the Subject

The physics of the city has been proposed as an interdisciplinary research field that applies the techniques of exact sciences to describe and understand the dynamical states of the new metropolis. The continuous growth and the social changes outlined by urban planners and sociologists, give to the actual cities a complex structure whose study requires a new generation of dynamical models. From this point of view, the traffic and crowd dynamics have a fundamental importance due to their impact on individual freedom and life quality. The study of the observed macroscopic phenomena using a complex systems approach, the search for the control parameters, and the understanding of the individual mobility demand are key points for future contributions of the complexity science to real human problems.

### Introduction

Crowd and traffic dynamics are a crucial issue for modeling the complex urban mobility [5,8,11]. The idea that the city itself can be considered a whole laboratory to understand the mobility is at the base of a new research field: the physics of the city [13]. “Aggregate of beings that hold their biological history into its borders and model it within all their intentions proper to thinking creatures, the city results at the same time by the biological generation, the organic evolution, and the aesthetics creativity. The city is contemporary a natural object and a subject of culture”. From these words of Claude Levi Strauss clearly emerges the complex nature of every urban system that today it is generally accepted, and considered in order to study the city, its development and evolution [8]. The city shows itself to be polymorphous, polysemic and polyglot [22], stratified in time and crossed by actors and objects whose dynamics are extremely different and can be conflictive to the point to engender a feeling of chaos, i. e., one can say that cities live at the edge of chaos and the problem is exactly to govern the system emerging complexity, and not

real chaos situations. To try to express the quality of this complexity in quantitative terms, using the instruments of the exact sciences, without losing its texture, one has to reduce the semantic, logical, syntactic and phenomenological field in which to articulate the possibilities of constructing models able to be descriptive, explanatory and, at least to some extent, predictive [78]. These models will never be completely isomorphic to reality, but can simulate some of its aspects and characteristics considered salient by the observer, thus playing a fundamental role in the choice of the significant observables and control parameters, which obviously depend from what one wants to control and govern. A simple consideration can help us in the reduction process: regardless of the variety and complexity of flows, forms and information, an urban system exists insofar as it is inhabited. Therefore our physics of the city will be essentially physics of an inhabited city, and given the large number of elementary components, this means non-equilibrium statistical physics [61,77], because a town is an open system. Moreover, for the elementary components moving in urban space-time, this also means the physics of dynamical systems, and since the individuals in the system have free will, probabilistic physics has to be considered (Pascal firstly modeled the human free will by the probability function, using the game of chance as paradigm). To finish, since the elementary components have memory and are capable of drawing information from the environment, processing it according to intentions, choices and decisions, the physics of the city must also be intentional, cognitive and decisional. As everyone knows, another basic ingredient of physics, besides the elementary components, is a space-time where the dynamics can develop (*ubi materia, ibi geometria* – Kepler). So one can ask if it is possible to identify a space-time structure proper to a generic urban system. Obviously we can describe the street network, and the different morphologies with a spatial metric that usually is not Euclidean, but is not sufficient to develop an urban dynamics. One needs also a clock, a time structure scale invariant. This structure can be modeled by the *chronotopoi* (literally, places of time), the primal agents of urban temporal dynamics able to generate time correlations that would not exist without them [10]. In the city planning language, they are defined as areas where are implanted, temporal scheduled activities, for example a hospital, the university, a shopping center and so on, that generates/attract mobility. Urban topology thus becomes chronotopic, and the interaction of the individual's agenda with the pulsation of the *chronotopos* produces complex urban mobility [36].

Assuming the mobility as a property of the individuals able to process information, the microscopic dynam-

ics one generates both by properly physical interactions and by some decisional/cognitive mechanism [1]. Moreover a mesoscopic dynamics for individuals aggregate has been developed to study self-organization behaviors in the context of a non-equilibrium and non-linear evolution of the system (being the urban mobility a complex phenomenon that develops itself in an open complex system, the city) [39,45]. Using a computer different mobility situations are simulated and analyzed, in order to study the control parameters, phase transitions and shift from order to chaos, criticalities and the whole range of possible orbits, also those that are very difficult to reach in material experiments [24,25]. Obviously there is a problem of realism, because the simulations can be consistent and reasonable but far, sometime very far, from the effective mobility.

In general social systems, like urban traffic and crowd dynamics, consist of many autonomous intelligent entities, which are distributed over a structured space and interact each other to achieve certain goals [33,68]. A powerful method to study such systems are multi-agent models, which allow one introduce a cognitive behavior in the dynamics of individual entities. The concept of "agent" has been proposed in many research areas [17,67,87], from computer science and artificial life research to economy and sociology, so that its definition is quite generic: agent is anything able to act in a given environment. To suggest a more specific agent definition in the simulation of traffic and crowd dynamics [6,7,94], the crucial issue concerning the meaning and the utility of a social system model has to be firstly discussed [19,20,33]. Complexity science faces the problem of studying systems whose dynamics cannot be reduced to physical fundamental principles. The different scales of description turn out to be interconnected, not only by a bottom-up interaction structure typical of physical systems (the dynamics at microscopic levels determines the evolution of macroscopic variables), but also by top-down interactions, so that the microscopic behavior is influenced by the macroscopic states of the system [30]. A classical statistical physics approach is questionable in such a case. A further difficulty is that even the interactions among the elementary components of a complex system are often unknown, and they can only be described in a qualitative way according to some hypotheses, consequences of real phenomena direct observations [78]. Usually a model is a simplified representation of the reality, but not too simple in order to describe the interested phenomenon; moreover any model should have sufficiently generic features to be able to describe analogous phenomena observed in different systems from social to physical or biological ones [76,90]. In complexity science the role of the models is modified becoming not a specific

realization of a general theory (think for example to different cosmological models which correspond to different boundary conditions and solutions for the Einstein equations), but rather an instrument of knowledge to perform *in silico* experiments that allow to verify the consistency of the assumptions with the considered phenomena. In other words the models realize a “virtual reality” where the simulations point out the relevance of different hypotheses on the appearance and evolution of emergent properties identifying the control parameters. This modeling cannot be validated in the usual Galilean sense, since the models are themselves experiments. Moreover there is an intrinsic difficulty in defining Galilean experiments on complex systems, due to the great information needed for the reproducibility [9,34]. From a physical point of view, one has to look for statistical laws, that consider the minimal conditions for the existence of phase transitions toward particular self-organized states.

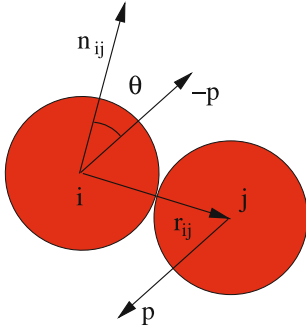
In our opinion, a paradigmatic physical model for complex systems can be the “automata gas” [95]: a statistical system whose elementary components perform both a physical dynamics and a cognitive behavior based on information exchanges [35,92]. In the sequel an automata gas model to describe the crowd and traffic dynamics is described. As a general convention the capital letters will denote scalar variables, whereas the lower letters will indicate vectors.

### Automata Gas and Crowd Behavior

The pedestrian mobility is a basic human activity strictly linked with the detailed structure of urban environment and the social life [14,32,81]. Indeed due to the human being social nature, the pedestrian mobility will continue to play a fundamental role for the life quality of metropolis. The development of fast and comfortable transportation has reduced the spatial scale of walking activity to an average dimension of  $\simeq 500$  m. The direct observations on video films turns out to be the most useful tool to extract detailed information on pedestrian dynamics [50,99]. Various research group have pointed out the following items [39]:

- I Pedestrians tend to walk at a desired velocity toward a local destination. The desired walking speed depends both on social characters (for example age, sex, job, ...) and on the cognitive internal state. Moreover individuals develop strategies in order to walk in the more comfortable way (for example avoiding collisions and least energy consuming), and following a best path according to their propensities and knowledge of the urban space.
- II Pedestrians like to keep a certain distance from other pedestrians and borders as walls or obstacles. This distance can be interpreted as a social space preserving individuality, and it depends on various factors like crowding conditions, individual hurry and space geometry. Usually individuals don't want their social space violated by extraneous people. Moreover individuals who know each other, tend to form clusters that move as single entities performing a flocking dynamics [24,25].
- III To avoid physical contact, individuals reduce the walking velocity and take local detours, but they feel aversion to move in opposite to their desired direction. In case of physical collisions, the incompressible nature of bodies play an important role and pedestrians move their body, so that they can slide on each other (if it is possible).
- IV The best pedestrian paths usually are not the minimal time or space geodesic, because many other factors contribute to the choice of the trajectory, so a minimal action principle cannot model dynamics [15,54].
- V The pedestrian paths seem to optimize some utility functions time or energy consuming [64], but depend on the information level of the individual on the urban space and single's propensities. They may be dynamically modified by decision mechanisms [40,46] in the presence of multiple choices, particular chronotopi and geometrical or crowding conditions [11]. In normal crowding conditions, there are strategies automatically applied at particular points, like at a bottleneck, where the priority convention avoid criticalities. The decision mechanism reflects the individual free will, which is an unpredictable relevant character of pedestrian dynamics, but there are also some average behaviors, such as the herding effect or habit [40,49].
- VI When panic phenomena occur, pedestrians lose their rationality and the dynamics are dominated by hard physical interactions [39]. People try to move considerably faster than normal and start pushing. As a consequence the pressure become extremely dangerous in jammed crowds, arching and clogging are observed at the exits so that even escape become very difficult or impossible [37].

In order to define an automata gas model for pedestrian dynamics that takes into account the previous observations, physical and information-based interactions are treated separately. Let us consider the following axioms, that define a common framework for several pedestrian dynamical models:



**Traffic and Crowd Dynamics: The Physics of the City, Figure 1**  
Sketch of automata collision in the center of a mass system:  $p = \mu(v_j - v_i)$  is the center of mass momentum and  $\theta$  the collision angle with respect to the normal to the collision direction  $r_j - r_i$

- A1) each automaton  $i$  has an inertial mass  $m_i$ , an incompressible body-space of fixed radius  $R_b$  and a social space of radius  $R_s$  around its center  $r_i$ .  
A2) Each automaton  $i$  has a desired velocity  $\bar{v}_i$ , and it continuously modifies its actual velocity  $v_i$  according to

$$\dot{v}_i = \gamma^{-1}(v_i - \bar{v}_i) \quad (1)$$

- A3) When two automata  $i, j$  have distance  $D_{ij} = \|r_j - r_i\| = 2R$  and their velocities satisfy the condition  $(v_j - v_i) \cdot (r_j - r_i) < 0$ , then a physical collision and the change of momenta  $p_i = m_i v_i$  are computed according to

$$\begin{aligned} p'_i &= \alpha (p_i - (\mathcal{R}(\phi) - \mathcal{I})\mu(v_j - v_i)) \\ p'_j &= \alpha (p_j + (\mathcal{R}(\phi) - \mathcal{I})\mu(v_j - v_i)) \end{aligned} \quad (2)$$

where  $\mu = m_i m_j / (m_i + m_j)$  is the reduced mass,  $\alpha \leq 1$  the dissipation coefficient, and  $\mathcal{R}(\phi)$  is a counterclockwise rotation of an angle  $\phi > \theta$  (see Fig. 1), whereas  $\mathcal{I}$  is the identity transformation.

- A4) Each automaton has a local visual space given by a semicircle of radius  $R_v$ , symmetric with respect to the desired velocity direction. At a given time interval  $\Delta t$ , the automaton obtains information on positions and velocities of any other automaton or obstacle in its visual space.  
A5) Each automaton  $i$  rotates and reduces its velocity

$$\dot{v}_i = \pm \omega \times v_i - \beta v_i \quad \beta > 0 \quad (3)$$

when it realizes that another automaton or an obstacle will enter into its social space, within a reaction time interval  $\tau$ .

The axiom A1 concerns the physical properties of the automata gas: the incompressible body is usually considered by models for simulation of crowd dynamics and pedestrian panic [44]. In principle one has to consider different body radius for different automata to take into account the diversity among individuals in a crowd; this can be relevant studying the dynamics at bottlenecks [43,52]. The “inertial mass”  $m$  is also introduced to differentiate the automata behavior during collisions or interactions: an automaton with a great mass behaves more aggressively by forcing other automata to deviate from its trajectory. The existence of a social space has been proposed by sociologists who related this space with the distance between two individuals shaking hands with each other (in the West civilization). However this is not a fixed space, and it depends on the single characteristics (for example philanthropic or misanthropic one), and on crowding conditions and/or presence of attractive points in the space where people accept to be closer. One should also consider the possibility that individuals form compact groups when there is empathy among them.

The existence of a desired velocity (axiom A2) is almost obvious, but of course its determination is the result of a cognitive process where both determinism and individual free will have an important role. Eq. (1) is the most simple dynamics to simulate the automaton tendency to move at the desired velocity, and the same choice has been proposed by other models, where the only relevant parameter is  $\gamma$ : i. e., the “relaxation time scale” [45].

The collision dynamics defined by axiom A3 is inspired by granular flows interactions [18,38], and it allows one to simulate both the body repulsion, and a sliding effect that mimics the real effect due to the possibility of wheeling the body when two pedestrians collide. The relevant parameters are the rotation angle  $\phi$  and the dissipation parameter  $\alpha \in [0, 1]$ : when  $\phi = 2\theta$  (see Fig. 1) and  $\alpha = 1$  one has a perfect elastic collision, whereas  $\phi = \theta$  and  $\alpha = \cos \theta$  gives an anelastic collision between two spheres. The direct observation of video films suggests the existence of a  $\alpha$  dependence on  $\theta$ , which has a maximal value at  $\theta = \pi/2$  (i. e., two pedestrians tend to reduce suddenly their velocity in a head on collision). Conversely the choice of the  $\phi$  within the interval  $[\theta, 2\theta]$  seems to be suitable for the simulations.

The physical automata dynamics, which is the combination of Eqs. (1) and (2), can be defined in an equivalent way by using the social force model, where a second-order dynamics is used to simulate pedestrian movements with interaction “social forces”, which depend both on the distance and on the reciprocal orientation of pedestrians (non-Newtonian character) [41]. The collisional approach

used in the automata gas model gives an advantage using essentially a first-order dynamics in the simulations, so that the numerical instability due to high force gradients are automatically solved and a longer integration time step can be used. Conversely the cellular automata models [3,48,70] solve the collisions problem by discretizing the space and are extremely convenient from a computational point of view, but they can introduce some bias in the microscopic dynamics whose macroscopic effect should be carefully considered [60].

The axioms A4 and A5 introduce a local vision mechanism: each automaton sees the others in front of it whose actual position is inside its visual space. For each of them, it computes the forecast position (according to the actual velocity) after a time  $\tau$ . Then it selects the automata whose forecast positions are inside its social space, therefore it rotates and reduces its velocity, in order to avoid future collisions choosing the suitable sign in Eq. (3). The vision has some typical features of an information based interactions [84]:

- a) It depends on the forecast positions of other automata, according to the available information;
- b) It has a dichotomic nature, since only the automata that will eventually enter the social space are considered;
- c) The sign choice of the rotation velocity is the result of an utility function evaluation.

Finally the interval  $\Delta t$  defines the decision timescale, when new information is achieved; by assumption, the following inequality holds

$$\Delta t \bar{v}_i \ll R_b \ll \tau \bar{v}_i \simeq R_v. \quad (4)$$

Therefore the automaton dynamics is the result of many decision processes, and the vision mechanism determines a collective interaction dynamics in the automata model, which is an essential feature for the appearance of emergent properties. According to the inequality (4), the average effect of local vision is to rotate the velocity in the direction where the “density” of counteracting automata is minimal, and to reduce the velocity proportionally to the surrounding automata density. From this point of view, the local vision is equivalent to a non-Newtonian repulsive long-range force with a spatial decaying scale  $\simeq R_v$ , as proposed in the social force model. Indeed both the approaches give rise to the same macroscopical states in many cases of study. But the individual character of information-based interactions allows one to introduce the free will in the model, using the concept of cognitive state, without the artifact of external random perturbations in

the dynamics. Using the idea of cognitive state to define the automata desired velocity, and the choice among multiple possibilities permits one to state two other axioms:

- A6) Each automaton has a cognitive network representation of the physical surrounding space, dependent on its knowledge level. Each network node is related to a possible decision, whereas the arcs introduce connections between successive decisions (see for example Fig. 2a).
- A7) The decision mechanism is defined by a subjective probability evaluation [27] based on the existence of an utility function and an internal cognitive dynamics [16,28]; the probability are computed taking into account the available information and the previous history.

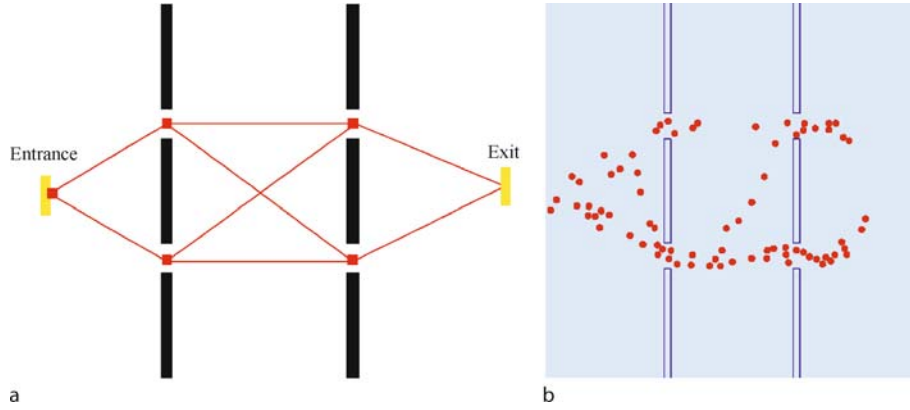
The axiom A6 implies an internal dynamical structure for the particle automaton, which acquires a knowledge capability of urban space. The nodes are identified with particular regions of interest in the urban space (for example in Fig. 2 the nodes correspond to doors neighborhoods). The decision process is discontinuous, since the automaton gets new information at the node points. One can compare the cognitive network structure with a best paths origin-destination algorithm, based on the deterministic minimal action principle, which has been proposed to study the accessibility of urban spaces [15,54]. However, this approach gives a global and rigid description of the space that turns out to be unrealistic and, sometimes, useless, when one studies crowd dynamics where the automata paths are the results of a continuously evolving process, dominated by collective interactions and individual free-will. On the contrary, the space representation using a cognitive network that focuses the attention only on some particular points, it is both robust and flexible to allow a single automaton to create its own individual path with optimization criteria based on local information.

The axiom A7 establishes a relation between an internal cognitive dynamics [93] and a probabilistic choice among different possibilities [62]. Let us consider the elementary dichotomic decision between the choices  $A$  and  $B$ , by associating  $A$  or  $B$  to the values of the variable  $x_i$  (for example  $A$  when  $x_i > 0$ ) that defines the “cognitive state” of the  $i$  automaton. One assumes the existence of a stochastic dynamical system that represents the neuronal activities [83]

$$\dot{x}_i = \frac{\partial U}{\partial x}(x_i; I) + \sqrt{2T_i} \xi_t \quad (5)$$

where  $U(x; I)$  is an utility function with two local maxima  $U_A$  and  $U_B$ . The “utility concept”, introduced in econ-





**Traffic and Crowd Dynamics: The Physics of the City, Figure 2**

**a** Cognitive network representation of a space with two parallel walls and four connecting doors; the automata enter from the left side and have to reach the exit on the right. The squares are the decision node where automata obtain new information and change the desired velocity. **b** Snapshot of the automata dynamics in the space shown in left picture. The cognitive network causes a flux organization along the shortest paths, but the subjective probabilistic decision implies the existence of “irrational” automata which follow longer paths

omy [31,96], is a measure of the perceived advantages of a certain decision with respect to the automaton final purpose, whereas the information  $I$  is any external variable that may influence the decision. Under the very general regularity hypotheses on  $U$ , the dynamics (5) evolves towards one of the local maxima in a stochastic way, due to the presence of the white noise  $\xi_t$  (unpredictable factors in the decision mechanism) and the “social temperature”  $T_i$ , which is a measure of the “irrational behavior” of the automaton. As a consequence even if all the automata have the same decision mechanism, the effective decision is subjective due to the different information  $I$  and the different temperature  $T_i$ . According to the Arrhenius’ law, the transition probabilities from the choice  $A$  to  $B$  and vice versa read

$$P_{AB} \propto \exp\left(-\frac{U_A(I)}{T_i}\right) \quad P_{BA} \propto \exp\left(-\frac{U_B(I)}{T_i}\right) \quad (6)$$

By introducing an average balance equation for the stationary probabilities to choose  $A$  or  $B$

$$P_A P_{AB} - P_B P_{BA} = 0$$

one obtains [16,46]

$$\begin{aligned} P_A &= \frac{\exp\left(\frac{U_A}{T_i}\right)}{\exp\left(\frac{U_A}{T_i}\right) + \exp\left(\frac{U_B}{T_i}\right)} \\ P_B &= \frac{\exp\left(\frac{U_B}{T_i}\right)}{\exp\left(\frac{U_A}{T_i}\right) + \exp\left(\frac{U_B}{T_i}\right)} \end{aligned} \quad (7)$$

The non-linear relation between utility and probability has two limit behaviors: for low temperature values, Eq. (7) defines a sigmoidal like function with a sharp transition between the two choices, whereas for high temperature values the relation is almost linear. These two regimes correspond to two different social behaviors of the automaton: in the first case there is a sudden change in the decision as utility overcomes a certain threshold (rational automaton), in the second case, the automaton choice becomes insensitive to utility changes (irrational behavior).

### Automata Gas and Traffic Behavior

The traffic behavior in a urban network is the consequence of citizens mobility demand driven by the chronotopic activity of modern cities [21]. The application of physics to traffic modeling dates back more than 40 years [80]. From one hand the microscopic traffic dynamics is forced on a one-dimensional space (the road), from the other hand its spatial scale is extended to the whole city. For this reason various researches have proposed to simplify the microscopic physical dynamics, and to focus the attention on the cognitive behavior at the base of the individual mobility agenda and on the fluxes optimization in the road network [2,42,86,94]. Nevertheless the microscopic drivers behavior is responsible of local congestions formation that may propagate in the road network and cause a macroscopic phase transition of the whole system. The proposed models are divided in two main classes: the cellular automata models which discretize the road dynamics using discrete stochastic maps [63,75], and the continu-

ous models based on stochastic differential equations [39]. The auto-mobilis model [12] is based on a simple microscopic continuous dynamics, but able to describe different driving styles and behaviors at the crossing points (traffic lights, roundabouts and stop signals at insertions). The automata in the model move on a road using alternatively the two main behaviors that has been proposed by various researchers [51]: the so called “car-following” [4] and the “optimal velocity” behavior [74]. In the first case each driver adjusts his velocity to the velocity of the successive vehicle; in the second case each driver chooses his velocity according to the safety distance with the successive vehicle. Both models can be described by the differential equation

$$\begin{aligned}\dot{s}_j &= v_j \\ \dot{v}_j &= a(s_{j-1} - s_j, v_j, v_{j-1}(t - \tau); \Theta)\end{aligned}\quad (8)$$

where  $(s_j, v_j)$  are the position and the velocity of the  $j$ -automaton on a road (by convention the  $j$ -driver follows the  $j - 1$ -driver),  $\Theta$  is the parameter set that characterizes the individual behavior,  $\tau$  is a reaction time to the velocity changes of the successive vehicle. The effect of line changes can be neglected since they are not frequent along urban street (of course this is not true at the intersections points that will be consider separately). To specify the acceleration in Eq. (8), the main ideas at the base of various models are summarized, by means of the following assumptions:

- B1) An automaton-driver has a finite dimension  $d_0$ , it keeps a minimal distance  $d_{\min}$  from the preceding automaton, it has a reaction time  $\tau$  to the velocity changes of the other vehicles, and it has a desired velocity  $v_{\max}$  (usually the maximal allowed velocity along the considered road, assuming that people respect the traffic rules)

$$\dot{v}_j = \alpha(v_j - v_{\max}) \quad (9)$$

where  $\alpha$  is the acceleration capacity in a comfortable driving.

By measuring  $s_j$  at the center position, the distance between two successive drivers is  $\Delta s_j = s_{j-1} - s_j - d_0$ .

- B2) Each automaton has a safety distance  $D_s$  and an attention distance  $D_{\text{att}}$  (i. e., a distance at which the driver considers the presence of the other drivers)

$$D_s = d_{\min} + 2\tau v_j \quad D_{\text{att}} = d_{\min} + 2.5\tau v_{\max} \quad (10)$$

Each time  $\Delta s_j \leq D_s$  it reduces the velocity according to

$$\dot{v}_j = -\beta(v_j - v_s) \quad (11)$$

where  $\beta$  is the breaking capacity.

Eq. (10) establishes a linear relation between reaction time  $\tau$  and the safety distance  $D_s$ ; this is consistent with direct observations for the typical urban velocity ranges. But to avoid accidents one requires a condition on the breaking capacity  $\beta\tau \geq 1/2$  (cf. Eq. (11)).

- B3) When the distance  $\Delta s_j$  is less than the attention distance  $D_{\text{att}}$ , the automaton adjusts its velocity to the velocity of the preceding vehicle

$$\dot{v}_j = -\alpha(v_j - v_{j-1}(t - \tau) - \gamma) \quad (12)$$

where the parameter  $\gamma \geq 0$  simulates the behavior of impatient drivers.

To be more realistic, one takes also into account that the automata cannot correctly evaluate small velocities so that when  $v_{j-1} \leq v_{\min}$  one set  $v_{j-1} = 0$  in Eq. (12).

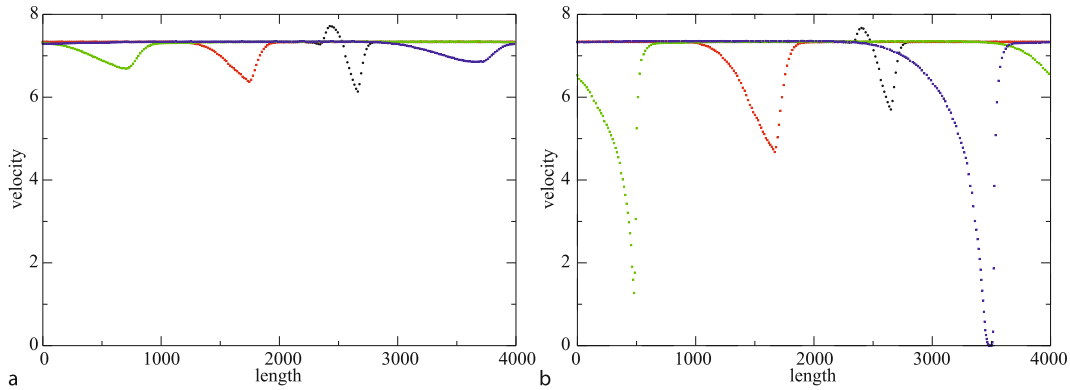
The Eqs. (9), (11), (12) define the microscopic dynamics along a road. All the parameters are individual characters, and have a random distribution among the automata populations. Indeed the driver behavior can be correlated to citizens social characters (sex, age, birth place etc.), so that several classes of drivers can be identified. In some realistic situations the presence of different driver classes is crucial to understand urban traffic. The model can simulate different traffic regimes: at low density the free flow (Eq. (9)) is described, when the density increases the automata perform a synchronized flow, which becomes unstable when the local density overcome  $1/D_s$  and gives rise to congestion effects. Indeed the model has infinite equilibria with automata moving at the same velocity  $v \leq v_s$ . However a linear stability analysis using the Fourier transform shows that the dynamics (12) is always stable with a relaxation time

$$T \simeq \frac{2(1 + \alpha\tau)^3}{\alpha(2\pi/\rho\lambda)^2} \quad (13)$$

in the limit  $\rho\lambda \ll 1$ , where  $\rho$  is the density and  $\lambda$  the wavelength of the perturbation, whereas the dynamics (11) become unstable when  $\beta\tau < 1$  [37,69]. Therefore when the local density satisfies  $\rho \leq 1/D_s$ , a congestion regime may rise; the instability can be driven by small percentage of “bad drivers” in the automata population, as shown in Fig. 3. This reflects the possibility that the different driving styles may cause local instability due to microscopic interactions, as suggested by direct observations [91].

To complete the dynamical model for urban traffic, one has to set the behavior at the crossing points in different situations assuming:

- B4) At a traffic light in case of yellow colors, the automata compute the time  $t_c$  required to pass through the



**Traffic and Crowd Dynamics: The Physics of the City, Figure 3**

**a** Stability study of the urban traffic micro-dynamics for an homogeneous automata population in a periodic road: the four curves (in the order: black, red, green, blue) give the vehicle velocity (m/s) as a function of the position (m) on the road at different time separated by 5 minutes. The model uses “realistic” parameters:  $d_0 = 4$  m,  $d_{\min} = 1$  m,  $v_{\max} = 16$  m/s,  $v_{\min} = 1$  m/s,  $\beta = 1.1$  s $^{-1}$ ,  $\tau = .75$  s and  $\alpha .5$  s $^{-1}$ . The initial equilibrium density  $1/16$  m $^{-1}$  is locally increased by a 10%. **b** The same as picture **a**, but the simulation considers a 20% of “bad drivers” randomly distributed on the road with reduced breaking capability  $\beta = .9$  s $^{-1}$ ; the blue curve shows as a stop-and-go instability has been developed by the microscopic dynamics

cross at the actual velocity: if  $t_c$  is longer than the yellow time-interval, they stop.

- B5) At a connection of two streets with priority rules, the vehicles in the secondary street are forced to reduce their velocity at a given value  $v_c$  ( $v_c = 0$  in case of a stop signal); they compute both the distance  $D$  from the crossing point and the velocity  $v$  of the incoming vehicle in the main street, if the distance  $D$  is greater than  $(v - v_c)/\beta$ , the other vehicles can enter in the main street; otherwise it has to stop. In Fig. 4 a snapshot of a simulation in a crossing point with priority rules is shown. As a matter of fact, the emergent global properties of urban traffic cannot be understood, described, and eventually predicted, only using the automata micro-dynamics on the road network [86,94]. A global cognitive time-dependent field interacting with individuals is needed.

- B6) In the urban space-time there exist the “chronotopoi”, which are adaptive agents generating the individual mobility demand.

Examples of chronotopoi are historical centers, shopping centers, university campus, great social events as concerts or football matches etc.; in our sense a chronotopos is not a geometrical or morphological property, but an entity that introduces timing correlations between the city macroscopic level (social time) and the automata microscopic dynamics (individual time). The chronotopoi provide a cognitive map [79] to the automata by means of cognitive fields. A sketch of chronotopic cognitive fields overlapped to a Manhattan-like road network is plotted in Fig. 5, where



**Traffic and Crowd Dynamics: The Physics of the City, Figure 4**

**Snapshot of vehicles dynamics simulation in a crossing point; the road in a lighter color at the cross has the priority. The different colors are related to the different vehicle directions**

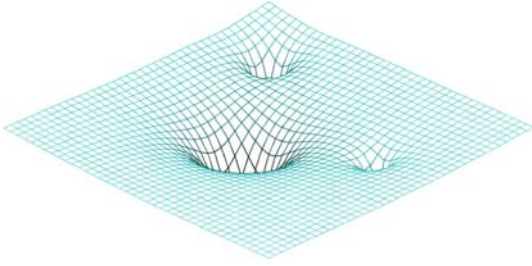
the intensity of the attraction is measured by the well amplitudes.

- B7) Every automaton constructs a daily mobility agenda based on its propensities and the available chronotopic information.

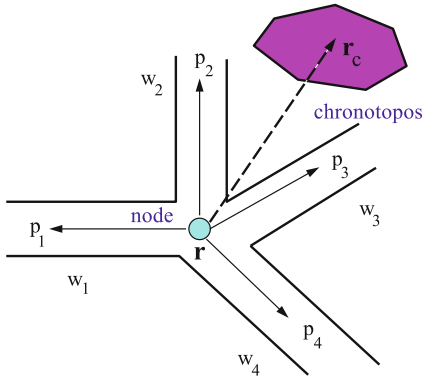
The decision mechanism is modeled by de Finetti subjective probability [27], taking free will into account.

- B8) The automaton trajectory is a path realization on a network, where the choice probabilities at the nodes depend on the chronotopic attraction.

The decision mechanism at a crossing point is outlined in Fig. 6, where the probabilities  $p_j$  depend on



**Traffic and Crowd Dynamics: The Physics of the City, Figure 5**  
Landscape of an abstract cognitive field superimposed to a road network



**Traffic and Crowd Dynamics: The Physics of the City, Figure 6**  
Scheme of the automata decision mechanism at a crossing point: the choice probabilities  $P_j$  are modulated according to the chronotopic attraction

a best path algorithms to reach the chronotopos, and the weights  $w_j$  measure the street quality (typically accessibility, security and aesthetics). The final probability choice is defined by

$$P_k = \frac{p_k w_k}{\sum_j p_j w_j} . \quad (14)$$

### Simulation Results and Empirical Observations

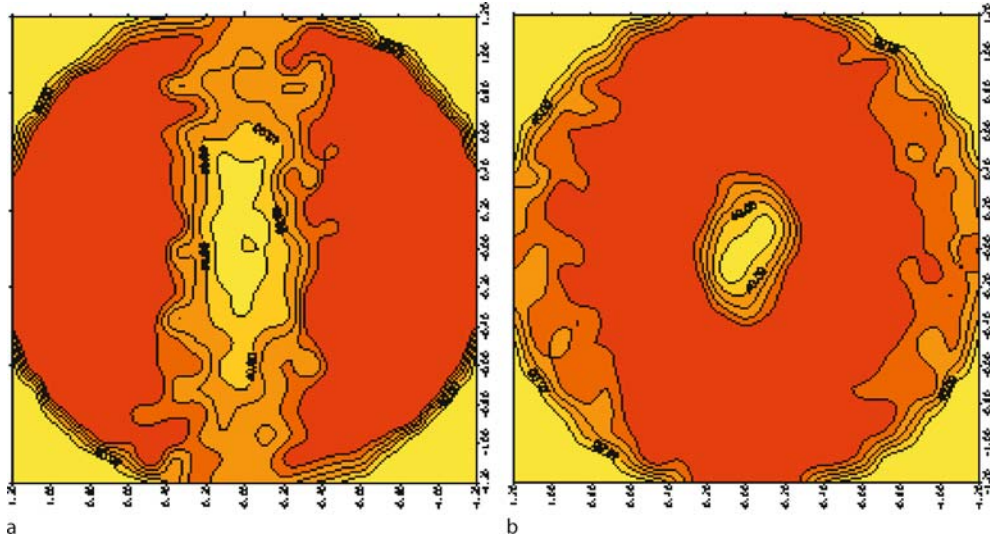
In complexity sciences, the simulations and the empirical observations play the role respectively of theory and laboratory experiments in the traditional physics framework. The main focus is to observe the emergent properties of the system, and to understand the class of microscopic interactions that can explain such properties in order to detect the control parameters. Simulations and empirical observations have to be compared in this optic [9]. In the physics of the city, the real urban systems are the natural laboratories to observe and study crowd and traffic flow.

### Crowd Behavior in Urban Spaces

An occasion to illustrate the crowd behavior in a whole pedestrian city, is Venice's Carnival, since about 100,000 tourists visit Venice each day during the last week. The observations are mainly based on video film analysis of particular regions of interest, like San Marco square and Riva degli Schiavoni. In principle the videos allow one to follow the single individual dynamics; however, the dimension of the considered regions is limited and a computer assisted analysis is difficult in crowded situations. Some quantitative information can be obtained by using software following the individual trajectories. The trajectories permit one to extract information both on the microscopic pedestrian dynamics, and on the individual cognitive behavior even if the noise level is still a serious problem [50]. Using a sample of  $\simeq 600$  trajectories in a moderately crowded area, we have computed for each individual the density of surroundings coming from the opposite direction and moving in a transverse direction within a distance of  $\simeq 2$  m. Then, after averaging all the local densities, the results are plotted in the Fig. 7. The observations indicate the existence of a low density area along the desired velocity direction, only if one focuses the attention on individuals moving in a counteracting way; conversely the individuals that move in other directions seems not to be considered. This is the sign for the existence of strategies to avoid collisions and to keep a certain area free along the direction of motion, based on a local vision mechanism. The automata gas dynamics can be illustrated performing some virtual experiments, where the emergent properties can be analyzed. The model parameters (cf. Eqs. (1), (2), (3)), have been fixed using  $R_b$  as space unit and measuring time in seconds: the desired velocity is set  $\|\bar{v}\| = 2R_b$  with a spread of 10% among the automata and the relaxation time  $\gamma = 1$ , whereas the collision (2) is defined by  $\phi = \theta + \pi/10$  (see Fig. 1) and  $\alpha = .9$ ; finally the local vision has a radius  $R_v = 2\bar{v}$ , the social space is  $2R_b$  and Eq. (3) uses  $\omega = 3\pi$  and  $\beta = .5$ . Moreover all the automata have the same mass.

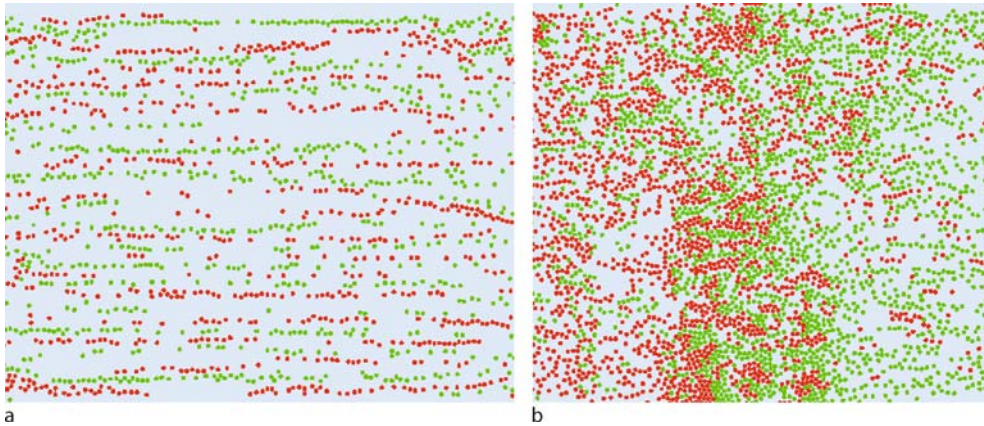
Let us consider two automata populations moving along opposite directions in an open space (for example, a square). The automata enter in the space from two sources located at the borders: the red ones move from the left to the right, whereas the yellow ones move from right to left. This system has two equilibrium regimes: the entropic regime where both the populations are uniformly spread in the space, and the organized regime where automata move along separated lines [47,89,98]. In the simulations, at low densities the automata relax to the entropic regime, whereas when the density increases one observes





**Traffic and Crowd Dynamics: The Physics of the City, Figure 7**

The scale color from red to yellow defines the average local density around a pedestrian that move in the vertical up direction as measured from the film video data during the Venice's Carnival. The pedestrians were moving in San Marco's Square in a moderately crowded condition. The considered radius around the pedestrian is 2 m. The picture a refers to the density of people moving into the opposite direction with respect to the considered individual. The picture b refers to the density of people moving into a transverse direction



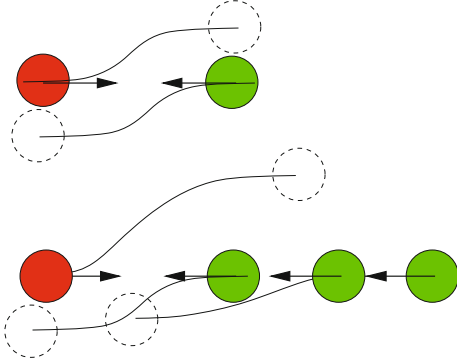
**Traffic and Crowd Dynamics: The Physics of the City, Figure 8**

a Simulation of two automata populations moving into opposite direction in a square. The parameters used in the model are (cf. previous section):  $R_b = 1$  (the space unit),  $\|\vec{v}\| = 2$ ,  $\gamma = 1$ ,  $\phi = \theta + \pi/10$ ,  $R_v = 4$ ,  $\omega = 3\pi$  and  $\beta = .5$ . The square dimensions are  $120 \times 120$  and the automata enter from two sources located at the left border (red automata) and the right border (yellow automata). The appearance of an organized regime with automata moving along lines is clear. b The same as the picture a, but the incoming flux from the sources has been increased by a factor 2. The lines of people start to interact and a turbulent regime appears

the lines formation in the space as shown in Fig. 8a. The automata form narrow straight non-interacting lines, separated by empty spaces: sometimes the lines are continuous, in other cases they interrupt or join together. Moreover there are always empty regions, so that the space distribution is not uniform. This phenomenon can be inter-

preted by a mean field argument: every time two automata meet, they deviate to avoid a collision and then they recover the desired velocity according to Eq. (1). The final result is a transverse displacement with respect to the desired velocity for both the automata. As far as binary interactions are considered, the transverse displacements





**Traffic and Crowd Dynamics: The Physics of the City, Figure 9**  
Sketch of automata interaction: the difference between the binary interaction dynamics and a collective interaction dynamics is outlined



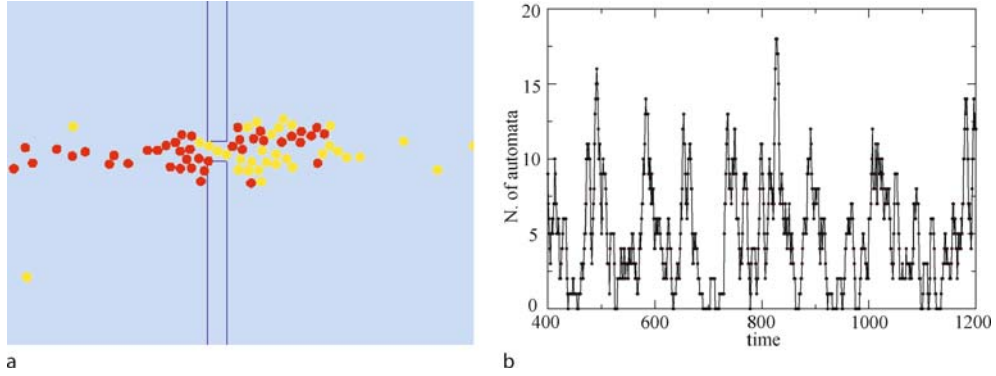
**Traffic and Crowd Dynamics: The Physics of the City, Figure 10**  
Snapshot of a video film of San Marco's square during the Venetian Carnival 2007: the video analysis shows the existence of lines of people moving in opposite directions

are symmetric and only the entropic regime is stable. But the local vision introduces long-range collective interactions, and aligned automata moving in the same direction behave as a simple entity [57]: in other words when an automaton meets a line of people moving in the opposite direction, it performs a larger transverse displacement than the automata in the line as sketched in Fig. 9. Therefore the line structure becomes stable and the system relaxes towards an organized regime. This behavior has been observed by various research groups in real situations [47,55,65]: an accelerated video film in the crowding conditions shown in Fig. 10, points out clearly the existence of stable lines of people moving in opposite directions, and a non-homogenous distribution of pedestrians

in the space. The lines survive also when the density increases, due probably to the slow down of the pedestrian desired velocity and the reduction of the social space. The simulations confirm this parametric dependence, and they also point out as to maintain the organized regime stability (for a fixed desired velocity and social space) it is essential to have some space among the automata lines, otherwise the interaction among the lines themselves will destroy the laminar character of the dynamics, and a new turbulent regime appears [45,71,72]. This implies the existence of a critical density value as it is shown in Fig. 8, where the simulation are repeated with a twice flux from the external sources. This turbulent regime has been observed in critical situations (like crowd stampede) when, under effect of panic, individuals behave irrationally, the velocity is increasing and the dynamics is mainly physical [50,59]. This was not the case during the Venetian Carnival and in the absence of panic, pedestrians take strategies that prevent critical situations. The automata gas model allows the study of the dynamics at a bottleneck, reducing the desired velocity when the density overcomes  $1/R_b^2$  (i. e., when automata have to share the social space). A bottleneck is a narrow passage at which pedestrian interactions do not allow the laminar flow formation. The dynamics at bottlenecks is a crucial task for understanding the crowd behavior in urban spaces, and it has been studied by various researchers [43,52,88]. The simulation of the passage of two counteracting flows at a bottleneck is shown in Fig. 11a: the automata reduce the desired velocity when they see a counteracting flux inside the bottleneck area. As a consequence there appears a self-organized oscillatory regime for the flux across the bottleneck (see Fig. 11b), as it has been observed in other models and in laboratory experiments [66]. A mean field theory approach suggests a possible interpretation of the oscillating fluxes regime. Let  $n^r$  and  $n^g$  the density of automata moving in opposite directions in the bottleneck area; each automaton can be in two states: waiting state  $P_w$  and crossing state  $P_c$ . The balance equations according to the following transition probabilities read

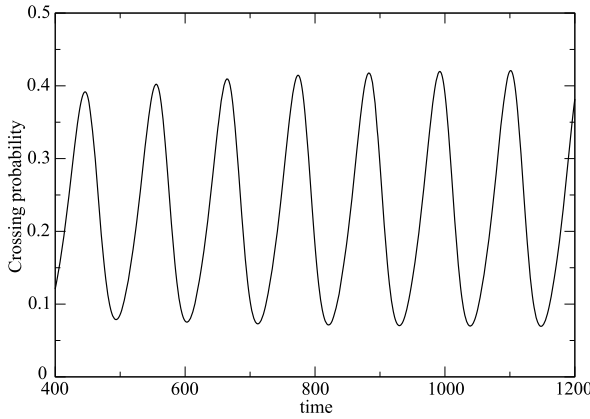
$$\begin{aligned} P_{c,w}^r &\propto n^g P_c^g (1 + c(n^r P_c^r - n^g P_c^g)) \\ P_{w,c}^r &\propto (1 - c(n^r P_c^r - n^g P_c^g)) \end{aligned} \quad (15)$$

based on the hypothesis that the probability to cross the bottleneck in the positive direction (red automata) increases if the flux  $n^r P_c^r - n^g P_c^g$  in the bottleneck is positive, and, on the contrary, the probability of returning in the waiting state increases when the flux is negative. Moreover if the flux is positive, the density  $n^r$  is decreased by a quantity  $\Delta n(n^r P_c^r - n^g P_c^g)$ , due to people passing



**Traffic and Crowd Dynamics: The Physics of the City, Figure 11**

**a** Dynamics of two counteracting flows at a bottleneck in automata gas simulation; the transverse bottleneck dimension is 2 time  $R_b$  with a constant incoming flux of one automaton each three time units. **b** Number of “red” automata that reach the right destination after crossing the bottleneck versus time: it is clear the appearance of an oscillating regime at the bottleneck



**Traffic and Crowd Dynamics: The Physics of the City, Figure 12**

**Solution of the balance Eqs. (17) showing an oscillating behavior for the crossing probabilities at a bottleneck under the hypotheses (15); the parameter  $c$  is the control parameter and the oscillations disappear below a critical value**



**Traffic and Crowd Dynamics: The Physics of the City, Figure 13**

**Snapshot of a video film of high density crowd crossing a narrow bridge during the Carnival; there is a long queue in front of the bridge whereas the flux across the bridge is constant**

trough the bottleneck, where  $\Delta n$  is a proportionality constant depending on the bottleneck geometry

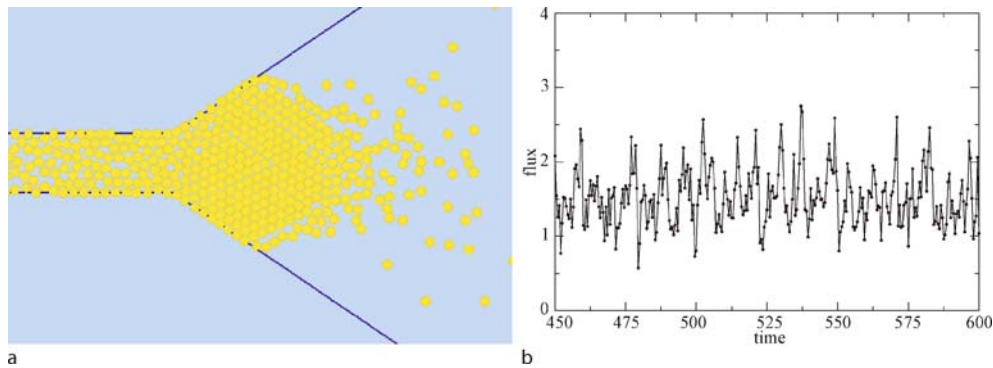
$$\dot{n}^r = \Delta n (n^r P_c^r - n^g P_c^g) \quad \text{if} \quad (n^r P_c^r - n^g P_c^g) > 0 \quad (16)$$

The balance equations (analogous equations hold for the yellow automata) read

$$\begin{aligned} \dot{P}_w^r &= -P_{w,c}^r P_w^r + P_{c,w}^r P_c^r \\ \dot{P}_c^r &= P_{w,c}^r P_w^r - P_{c,w}^r P_c^r \end{aligned} \quad (17)$$

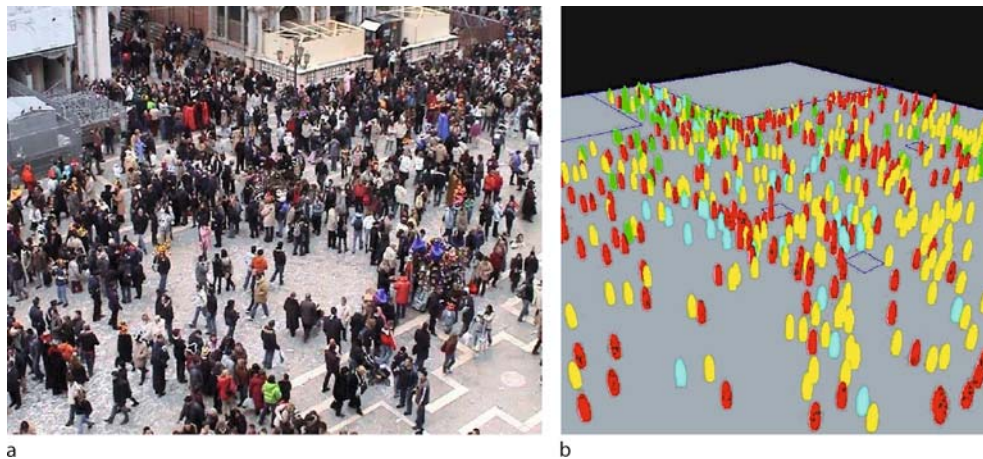
and the solution is plotted in Fig. 12 for a suitable set of parameters (the time unit is arbitrary), where the existence of a stationary oscillation regime in the crossing probability is evident. The parameter  $c$  (cf. Eq. (15)), which weights

the flux effect on the transition probabilities, is a control parameter for the dynamics: if  $c$  is below a critical value, no oscillation is observed and the stationary solution is constant. Finally another interesting situation is illustrated by Fig. 13, where people had to cross a narrow bridge during the Carnival: in the movie the people stop nearby the bridge forming a funnel-like queue, whereas the flux across the bridge is constant. As a result, in the queue the people move in a discontinuous way (stop and go) using the empty spaces left by individuals that enter the bridge; in such a situation there are waves back-propagating along the queue [50]. The measure of the flux across a given area can be performed by simulating the situation plotted in



**Traffic and Crowd Dynamics: The Physics of the City, Figure 14**

**a** Scheme of the bridge crossing geometry used in the gas automata simulation. **b** Automata flux in across small area inside the queue ( $\approx 10$  automata are considered simultaneously): fast periodic oscillations are detected due to short length back-propagating velocity waves



**Traffic and Crowd Dynamics: The Physics of the City, Figure 15**

**a** Mobility at the entrance of San Marco's square (snapshot from video movie). **b** Automata virtual experiment to study the mobility in a virtual space reproducing the urban space of the left picture

Fig. 14a; the results (see Fig. 14b) show a periodic structure that can be interpreted as short length back-propagating velocity waves. The finite automaton dimension has an essential role in the queue dynamics. As discussed previously, a quantitative comparison between empirical observations and virtual simulations have to face the intrinsic difficulty of reproducing the complexity of crowd behavior. However current computers can be used as virtual simulators that allow to study the crowd behavior in various realistic situations, becoming a helpful instrument for mobility planning [56]. Fig. 15 shows an example of crowd behavior in a real environment and in a virtual experiment by using the “Campus software” developed at the Physics of the City Laboratory of Bologna [99].

### Cognitive Behavior and Experimental Laws in Traffic Dynamics

As discussed in the previous sections, the cognitive behavior of individuals due to information-based interactions, plays a fundamental role to determine the emergent properties of urban traffic and crowd dynamics [100]. A crucial theoretical issue is to understand when it is possible to apply a statistical physics approach to derive the evolution of macroscopic variables from the single automata behavior [23]. The main difficulty is that from a cognitive point of view, the automata are all “different particles”, and the information-based interaction introduces long-range correlation among the automata dynamics, so that it is



very hard to justify a mean field approach. To illustrate a heuristic mean field approach, one can use the simple example of a dichotomic choice  $A$  and  $B$  for an automata gas, where two counteracting populations (red and yellow) share the same spatial environment [40,62]. An automaton  $i$  has an utility function  $U(N^r, N^y)$  in the decision mechanisms, depending on the number of automata of the two populations that have made the same choice. According to Eq. (7), the probability of the choice  $A$  is given by

$$P_A^{i,r} = \frac{1}{1 + \exp \left[ \left( U(N_B^r, N_B^y) - U(N_A^r, N_A^y) \right) / T_i \right]} \quad \text{if } i \text{ is red} \quad (18)$$

$$P_A^{i,y} = \frac{1}{1 + \exp \left[ \left( U(N_B^y, N_B^r) - U(N_A^y, N_A^r) \right) / T_i \right]} \quad \text{if } i \text{ is yellow.}$$

By using a statistical approach, one set  $N_A^r = N^r \langle P_A^i \rangle_r$  and  $N_A^y = N^y \langle P_A^i \rangle_y$  where  $\langle \rangle_{r,y}$  indicates the the average value on the population. Then the Eqs. (18) give a self-consistent system

$$N_A^r = \sum_i^r \frac{1}{1 + \exp \left[ \left( U(N^r - N_A^r, N^y - N_A^y) - U(N_A^r, N_A^y) \right) / T_i \right]}$$

$$N_A^y = \sum_i^y \frac{1}{1 + \exp \left[ \left( U(N^y - N_A^y, N^r - N_A^r) - U(N_A^y, N_A^r) \right) / T_i \right]} \quad (19)$$

where  $\sum^{r,y}$  indicates that the sum consider only the red or yellow population. In a symmetric condition between the two populations, the solutions of (19) have the property  $N_A^r = N_B^y$ ; therefore one reduces to the single equation

$$N_A = \sum_i \frac{1}{1 + \exp \left[ \left( U(N - N_A, N_A) - U(N_A, N - N_A) \right) / T_i \right]} \quad (20)$$

The obvious solution  $N_A = N_B = N/2$  exists, but if the following condition holds

$$\frac{1}{4T} \left. \frac{\partial V}{\partial N_A} \right|_{N_A=N/2} (P) > 1 \quad (21)$$

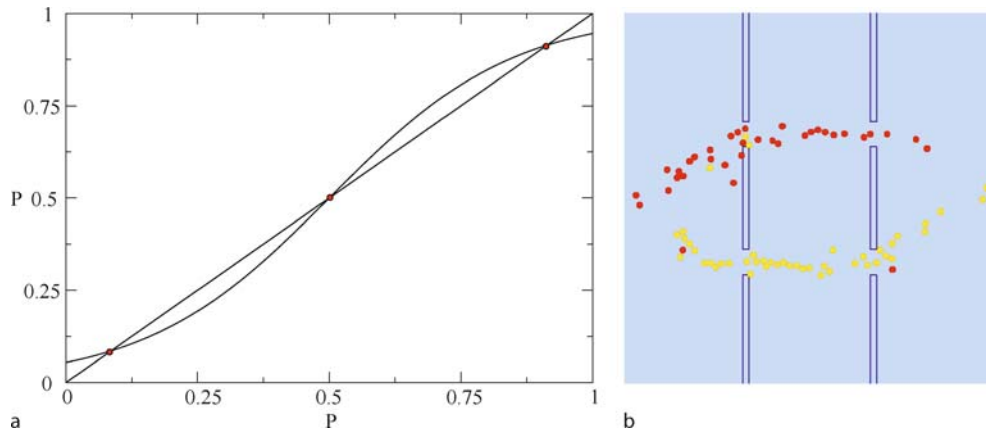
where  $V(N_A) = U(N - N_A, N_A) - U(N_A, N - N_A)$  and  $\bar{T}$  is the harmonic mean social temperature of the automata population, the symmetry is broken and two new

solutions appear where the two populations make opposite choices. Eq. (21) defines a critical temperature  $\bar{T}_*$  under which the phase transition occurs. This phenomenon allows one to simulate the herding cooperative effect, observed in crowd dynamics. In the Fig. 16, Eq. (20) is solved graphically for an utility function  $U(P) = P$ . The simulation results show that the statistical self-consistent approach (20), allowing to compute the correct value of critical temperature  $\bar{T}_*$ . In the simulation the automata use a utility function

$$U = \left( M + \frac{\phi \cdot \hat{v}}{N + 1} \right), \quad (22)$$

where  $M$  is a variable depending on the previous choice ( $M = 0$  if there is not a previous choice),  $\phi \cdot \hat{v}$  the flux at the door in the direction  $\hat{v}$  of the desired velocity, and  $N$  is the number of automata in a neighborhood of the door, and the social temperature is uniformly spread in the same interval for both the population. The memory  $M$  is introduced to enforce the choice of two successive doors along the same horizontal line. It is remarkable that there are always automata that move according to a minority choice, simulating the existence of free-will. The effect of a finite population size gives a critical value for  $N$ , that has to be overcome for the formation of a self-organized dynamics.

An experimental observation of statistical laws in crowd dynamics is very difficult due to the limited space extension of the video movie, whereas most of the data for traffic dynamics are on the road fluxes and not on the individual vehicles trajectories [53,58]. Recently a GPS system has been set up on  $\simeq 1\%$  of the vehicle population in Italy for insurance reasons, measuring position, velocity and signal quality. The data registration starts every time the engine is switched on, and ends when the engine is switched off. The GPS system repetition rate is one second, but the data are recorded each time a vehicle has covered  $\simeq 2$  km. The accuracy of GPS data has a standard uncertainty in the position of  $\simeq 10$  m. The GPS data allow a reconstruction of the individual dynamics so that it is possible to look for the existence of statistical macroscopic laws [82]. Different size cities have been considered: a small town, (Senigallia,  $\simeq 5 \times 10^4$  inhabitants, near the Adriatic see), a medium city (Bologna,  $\simeq 4 \times 10^5$  inhabitants) and a metropolis (Rome,  $\simeq 3 \times 10^6$  inhabitants). The three cities have a very different topology: the Senigallia topology is substantially a strip of  $2 \times 12 \text{ km}^2$  bounded by the sea, with few accesses, whereas Bologna has a radial configuration around an historical center with a radius of  $\simeq 6$  km. The urban area of Rome is one order of magnitude larger. To understand the mobility demand [16,29], the distribution of inner vehicle trajectories has been stud-



**Traffic and Crowd Dynamics: The Physics of the City, Figure 16**

**a** Graphical solution of the self-consistent Eq. (20) using an utility function  $U = N_A/N$  for both choices and a temperature below the critical threshold 0.5. The solutions are indicated by red dots; beyond the trivial solution  $P = N_A/N = 1/2$ , two further solutions exist that breaks the system symmetry. **b** Self-organized dynamics by simulating two populations moving into opposite direction in the space sketched in Fig. 2



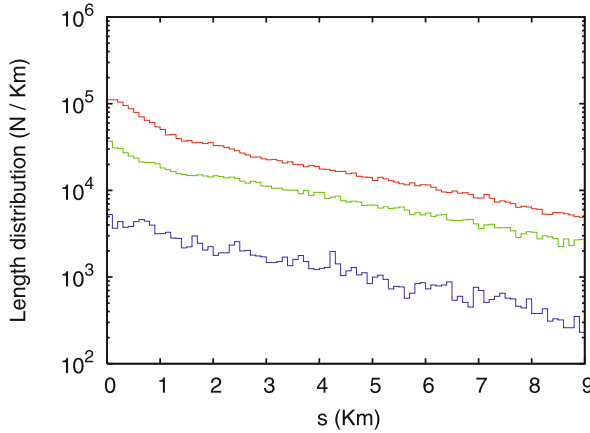
**Traffic and Crowd Dynamics: The Physics of the City, Figure 17**

The dots are the GPS recorded data in the Bologna urban area during the month of June 2006; the dots correspond to more than  $1.5 \times 10^6$  data. The colors label the recorded velocity: red means  $v < 30$  km/h, yellow  $30 < v < 60$  km/h, green  $60 < v < 90$  and blue  $90 < v$  km/h

ied: i. e., the trajectories fully contained in the considered urban area. The area of the considered regions are  $130 \text{ km}^2$  for Senigallia,  $160 \text{ km}^2$  for Bologna and  $400 \text{ km}^2$  (the area containing the “Grande Raccordo Anulare”) for Rome. In Fig. 17 the plot the recorded data in the Bologna urban

area for the whole month June 2006 is shown. In Fig. 18 the trajectories length distribution  $\rho(s)$  is plotted for one month of data recording: the lengths are computed using the polygonal metric in the three considered cases: Rome (February 2007), Bologna (June 2006) and Senigallia (June





**Traffic and Crowd Dynamics: The Physics of the City, Figure 18**  
Monthly length distribution of inner trajectories for the city of Rome (top) during February 2007, and for the cities of Bologna (center) and Senigallia (bottom) during June 2006. The vertical scale is logarithmic, and a polygonal metric for the lengths is used. The total number of trajectories is 1,176,763 for Rome, 79,154 for Bologna and 13,838 for Senigallia

2006). Even if there are almost two orders of magnitude between Senigallia and Rome's statistical samples, the path lengths distribution has the same behavior in the three cases. Moreover, it is really remarkable that the same exponential decay for the "long distance" trajectories is detected in all the considered cases, indicating a possible universal character at least in the class of Italian cities [82]. The exponential distribution law for individual trajectories has a statistical mechanics interpretation as the most probable distribution if each citizen has an independent mobility demand and there exists an average length for the urban trajectories. This means that the system behaves statistically as the citizens choose among randomly distributed destinations in the urban space, according to an "utility function" (proximity function) that increases linearly as a function of the trip length [16,29]. Let  $\{l_k\}$  be the set of all possible lengths of origin-destination trajectories in the city, and let us suppose that  $N$  citizens choose randomly the trajectories, with the constraint that the total covered length is finite

$$L = \sum_k n_k l_k, \quad \sum_k n_k = N, \quad (23)$$

where  $n_k$  is the number of citizens choosing a trajectory of length  $l_k$ . A simple combinatorial argument shows that the statistical weight of the distribution  $\{n_k\}$  is

$$P(\{n_k\}) \propto \frac{1}{n_1! \dots n_k! \dots}. \quad (24)$$

In the limit  $n_k \gg 1$  using the Stirling's approximation, one obtains

$$\log P(\{n_k\}) = - \sum_k (n_k \log n_k - n_k) + \text{const}. \quad (25)$$

The maximum of the logarithm (25) is computed using the Lagrangian multipliers method to take into account the constraints:

$$\frac{\partial}{\partial n_j} \left[ \sum_k (n_k \log n_k - n_k) + \alpha n_k + \beta n_k l_k \right] = 0 \quad (26)$$

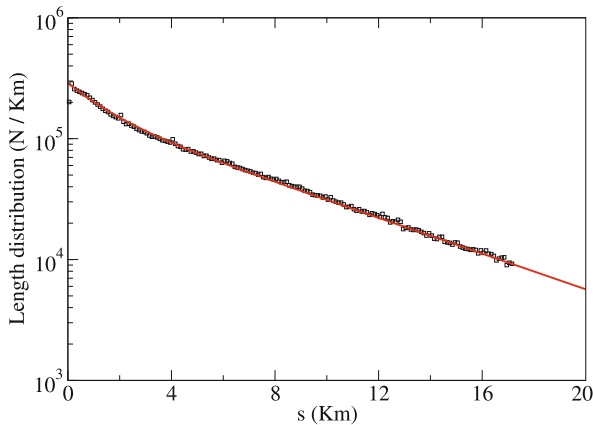
which implies

$$n_j = \exp(-\alpha - \beta l_j) \quad (27)$$

where the parameters  $\alpha$  and  $\beta$  are determined by the conditions (23). In the continuous limit, a statistical distribution  $\rho(s) \propto \exp(-\beta s)$  is achieved, consistently with experimental data. It is really remarkable that  $\beta^{-1}$  (i.e. the average length of urban trajectories) is the same for very different cities (see Fig. 18) [73]; this can be an indication for the existence of an individual mobility demand that can be related to a universal energy law in human travel behavior, according to the law proposed by D. Helbing and R. Kölbl in [64], based on the data of the *UK National Travel Surveys* during the years 1972–1998. Moreover, it allows the conjecture that the universal behavior of the length distribution (Fig. 18) can be explained as the result of a specific cognitive behavior of the populations. In other words, the spatial distance is the norm for the individual driving car decision and the "activities sprawling" in modern cities observed by urban planners [7], introducing a decorrelation in the system, which allows to apply the statistical physics techniques.

In the case of Rome, the lengths distribution has been also computed using the arc-length metric. The experimental results are plotted in Fig. 19, confirming the exponential behavior shown in Fig. 18, even if the decay slope is slightly different. This is a remarkable fact since one expects a decreasing of the path-lengths distribution when the length approaches to zero: it should be not convenient to use the car for short travels ( $<1$  km). A possible explanation is the presence of a relevant stochastic components in urban mobility (the so-called "asystematic mobility"), so that the individual mobility becomes the realization of a sequence of local destinations spread in the urban space. The experimental data in Fig. 19 suggest that the citizens use apparently two linear proximity functions, for the short and for the long length paths. Then the length distribution can be interpreted as an overlap of two exponential distributions

$$\rho(s) = A (\exp(-\beta_1 s) + c \exp(-\beta_2 s)). \quad (28)$$



**Traffic and Crowd Dynamics: The Physics of the City, Figure 19**  
**Squares:** length distribution of the inner trajectories in Rome during February 2007, computed using the arc-length metric. **Continuous curve:** interpolation of the experimental data with the analytical distribution (28). The interpolated formula has the following parameters:  $\beta_1 = .15 \text{ km}^{-1}$ ,  $\beta_s = .6 \text{ km}^{-1}$ , and  $c = .7$ . The vertical scale is logarithmic

The continuous curve in Fig. 19 shows the interpolation results using the parameters:  $\beta_1 = .15 \text{ km}^{-1}$ ,  $\beta_s = .6 \text{ km}^{-1}$  and  $c = .7$ .

However the existence of two exponentials curves for the length distribution should be studied more carefully, since the very short trajectories can be interpreted in various ways and the GPS signals have a low quality in such a case.

### Future Directions

From a theoretical point of view, we think that the gas of automata can be a good paradigmatic framework to investigate complexity in a general sense. For crowd and traffic flow, the previous model simulation and empirical results permit to tackle the problem of predictability and governance of the mobility system [26]. Last but not least, the prediction, even if partial or probabilistic, is relevant not only for the pleasure knowledge, but also for the application in order to minimize the impact of provisional city users on the town, and in particular on the mobility of local populations to prevent critical and potentially dangerous crowding phenomena, to prefigure and plan safety and security saving trajectories.

### Bibliography

- Ball P (2003) The Physical Modelling of Human Social Systems. *Complexus* 1:190
- Balmer M, Nagel K, Raney B (2004) Large-scale multi-agent simulations for transportation applications. *J Intell Transp Syst* 8(4):205
- Bandini S, Manzoni S, Vizzari G (2006) Crowd Modeling and Simulation. The role of multi-agent simulation in design support systems. In: Van Leeuwen JP, Timmermans HJP (eds) *Innovations in Design & Decision Support Systems in Architecture and Urban Planning*. Springer, Netherlands, pp 161
- Bando M et al (1995) Dynamical model of traffic congestion and numerical simulation. *Phys Rev E* 51(2):1035
- Batty M (1971) Modelling Cities as Dynamic Systems. *Nature* 231:425–428
- Batty M (2003) Agent-Based Pedestrian Modeling. In: Longley PA, Batty M (eds) *Advanced Spatial Analysis*. ESRI Press, Redlands, pp 81–105
- Batty M (2005) Agents, Cells and Cities: New Representational Models for Simulating Multi-Scale Urban Dynamics. *Environ Plan A* 37(8):1373–1394
- Batty M (2005) Cities and Complexity. Understanding Cities with Cellular Automata, Agent-Based Models, and Fractals. MIT Press, Cambridge
- Batty M, Torrens P (2001) Modeling Complexity: The Limits to Prediction CASA. Working Paper Series, vol 36. Available online at [www.casa.ucl.ac.uk](http://www.casa.ucl.ac.uk)
- Bazzani A et al (2003) A Chronotopic Model of Mobility in Urban Spaces. *Physica A* 325:517–530
- Bazzani A et al (2007) A Model for Asystematic Mobility in Urban Space. In: Albeverio S, Andrey D, Giordano P, Vancheri A (eds) *The Dynamics of Complex Urban Systems: An Interdisciplinary Approach*, vol 59. Physica, Heidelberg
- Bazzani A, Rambaldi S, Giorgini B, Turchetti G (2007) Complexity: modeling urban mobility. *Adv Complex Syst* 10(2):255
- Bazzani A, Giorgini B, Rambaldi S (eds) (2008) Physics and the City. *Adv Complex Syst* 10(2):215–377
- Bechtel R (1970) Human movement in architecture. In: Proshansky HM et al (eds) *Environmental Psychology*. Rinehart and Winston, New York
- Bellman R (1957) Dynamics Programming. Princeton University Press, Princeton
- Ben-Akiva M, Lerman SR (1985) Discrete choice analysis theory and application to travel demand. The MIT University Press, Cambridge
- Benenson I, Torrens PM (2004) Geosimulation: Automata-based modeling of urban phenomena. Wiley, London
- Biroli G (2007) Jamming: A new kind of phase transition? *Nat Phys* 3:222
- Brian WA, Durlauf NS, Lane AD (eds) (1997) The Economy as an Evolving Complex System II, Proceedings, vol XXVII. Addison-Wesley, Reading
- Byrne DS (1998) Complexity Theory and the Social Sciences: An Introduction. Routledge, London
- Cascetta E (2001) Transportation Systems Engineering: Theory and Methods. Kluwer, Boston
- Chomsky N (2000) New Horizons in the Study of Language and Mind. Cambridge University Press, Cambridge
- Chowdhury D, Santen L, Schadschneider A (2000) Statistical physics of vehicular traffic and some related systems. *Phys Rep* 329:199
- Chowdhury D, Nishinari K, Schadschneider K (2004) Self-organized patterns and traffic flow in colonies of organisms. *Phase Transit* 77:601624
- Cucker F, Smale S (2005) Emergent Behavior in Flocks <http://ttic.uchicago.edu/smale/papers.html>

26. Daganzo FC (2007) Urban gridlock: Macroscopic modeling and mitigation approaches. *Transp Res B* 41:49
27. de Finetti B (1972) *Probability, Induction and Statistics*. Wiley, New York
28. De Martino B et al (2006) Frames, biases and rational decision making in the human being. *Science* 313:684–687
29. Domencich TA, Domencich VD (1975) *Urban Travel Demand. A Behavioral Analysis*. North Holland, Amsterdam
30. Ellis GRF (2006) Physics and the real world. *Found Phys* 36:2:227
31. Fishburn PC (1983) Transitive Measurable Utility. *J Econ Theory* 31:293
32. Fruin JJ (1971) *Pedestrian planning and design*. Metropolitan Association of Urban Designers and Environmental Planners, New York
33. Garnsey E, McGlade J (eds) (2006) *Complexity and Co-evolution Continuity and Change in Socio-Economic Systems*. Edward Elgard, Chetentham
34. Giorgini B (2007) Philosophie naturelle de la causalité et du hasard dans un modèle de mobilité urbaine. In: Franceschelli S, Paty M, Roque T (eds) *Chaos et systèmes dynamiques. Collection Visions des Sciences*, vol 259. Editions Hermann, Paris
35. Giorgini B, Turchetti G (2005) From Newton-Boltzmann paradigms to complexity: a bridge to biosystems. In: *The Science of Complexity: chimera or reality?* Milan Research Centre for Industrial and Applied Mathematics, Esculapio Editore, Bologna, p 18
36. Giorgini B et al (2008) The Physics of the City: Modeling Complex Mobility. In: Diamantini D, Martinotti G (eds) *Urban science forward look*. Scripta Web, Napoli
37. Hayakawa H, Nakanishi K (1998) Universal behaviors in granular flows and traffic flows. *Prog Theor Phys Suppl* 130:57
38. Heinrich MJ, Nagel RS, Behringer RP (1996) Granular solids, liquids, and gases. *Rev Mod Phys* 68:1259
39. Helbing D (2001) Traffic and Related Self-Driven Many-Particle Systems. *Rev Mod Phys* 73:1067
40. Helbing D (2004) Dynamic decision behavior and optimal guidance through information services. In: Schreckenberg M, Selten R (eds) *Models and experiments in Human Behaviour and Traffic Networks*. Springer, Berlin, p 47
41. Helbing D, Molnár P (1995) Social force model for pedestrian dynamics. *Phys Rev E* 51:4282
42. Helbing D, Nagel K (2004) The physics of traffic and regional development. *Contemp Phys* 45(5):405–426
43. Helbing D et al (2006) Analytical approach to continuous and intermittent bottleneck flows. *Phys Rev Lett* 97:168001
44. Helbing D, Farkas I, Vicsek T (2000) Simulating dynamical features of escape panic. *Nature* 407:487
45. Helbing D, Farkas IJ, Vicsek T (2000) Freezing by Heating in a Driven Mesoscopic System. *Phys Rev Lett* 84:1240
46. Helbing D, Schönhof M, Kern D (2002) Volatile decision dynamics: experiments, stochastic description, intermittency control and traffic optimization. *New J Phys* 4:33
47. Helbing D, Molnár P, Farkas I, Bolay K (2003) Self-organizing pedestrian movement. *Environ Plan B* 28:361
48. Helbing D, Isobe M, Nagatani T, Takimoto K (2003) Lattice gas simulation of experimentally studied evacuation dynamics. *Phys Rev E* 67:067101
49. Helbing D, Buzna L, Johansson A, Werner T (2005) Self-organized pedestrian crowd dynamics: Experiments, simulations, and design solutions. *Transp Sci* 39–1:1
50. Helbing D, Johansson A, Al-Abideen HZ (2007) Dynamics of crowd disasters: an empirical study. *Phys Rev E* 75:046109
51. Hoogendoorn SP, Bovy PHL (2001) State-of-the-art of vehicular traffic flow modeling. In: *Proceedings of the: Institution of Mechanical Engineers. Part I. J Syst Control Eng* 215–4:283
52. Hoogendoorn SP, Daamen W (2005) Pedestrian Behavior at Bottlenecks. *Transp Sci* 39(2):147
53. Hoogendoorn SP et al (2003) Microscopic Traffic Data Collection by Remote Sensing. In: *Transportation Research Board (TRB) 82nd Annual Meeting*. Mira Digital Publishing, St. Louis, p 11
54. Hoogendoorn SP, Bovy PHL, Daamen W (2002) Microscopic Pedestrian Wayfinding and Dynamics Modelling. In: Schreckenberg M, Sharma SD (eds) *Pedestrian and Evacuation Dynamics*. Springer, Berlin, p 123
55. Isobe M, Adachi T, Nagatani T (2004) Experiment and simulation of pedestrian counter flow. *Physica A* 336:638
56. Jiang B (1998) SimPed: Simulating Pedestrian Flows in a Virtual Urban Environment. *J Geogr Inf Decis Anal* 3–1:21
57. John A, Schadschneider A, Chowdhury D, Nishinari K (2004) Collective effects in traffic on bi-directional ant trails. *J Theor Biol* 231:279
58. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A: Math Gen* 39:1775
59. Kirchner A, Nishinari K, Schadschneider A (2003) Friction Effects and Clogging in a Cellular Automaton Model for Pedestrian Dynamics. *Phys Rev E* 67:056122
60. Kirchner A, Klüpfel H, Nishinari K, Schadschneider A, Schreckenberg M (2004) Discretization Effects and the Influence of Walking Speed in Cellular Automata Models for Pedestrian Dynamics. *J Stat Mech* P10011
61. Klimontovich YL (1999) Entropy, information, and criteria of order in open systems. *Nonlinear Phenom Complex Syst* 2–4:1
62. Kluegl F, Bazzan ALC (2004) Route Decision Behaviour in a Commuting Scenario: Simple Heuristics Adaptation. In: Schreckenberg M, Selten R (eds) *Human Behaviour and Traffic Networks*. Springer, Berlin, p 285
63. Knospe W et al (2000) CA Models for traffic flow: Comparison with empirical single-vehicle data. In: Helbing D, Herrmann HJ, Schreckenberg M, Wolf DE (eds) *Traffic and Granular 1999: Social, Traffic, and Granular Dynamics*. Springer, Berlin
64. Kölbl RR, Helbing D (2003) Energy laws in human travel behavior. *New J Phys* 5:48.1
65. Kretz T et al (2006) Experimental study of pedestrian counterflow in a corridor. *J Stat Mech* P10001
66. Kretz T, Wölki M, Schreckenberg M (2006) Characterizing correlations of flow oscillations at bottlenecks. *J Stat Mech Theory Exp* P02005
67. Miller JH, Page SE (2007) *Complex Adaptive Systems*. Princeton Studies in Complexity
68. Millonas MM (1994) Swarms, phase transitions, and collective intelligence. In: Langton CG (ed) *Artificial Life III*. Addison Wesley, Reading, p 418
69. Mitarai N, Nakanishi H (1999) Stability Analysis of Optimal Velocity Model for Traffic and Granular Flow under Open Boundary Condition. *J Phys Soc Jpn* 68–8:2475
70. Mizar LF (2007) *Agent-Based Modelling and Micro-Simulations of Pedestrian Dynamics and Crowd Phenomena in Public Environments*. PHD Thesis, Università degli Studi, Milano Bicocca

71. Muramatsu M, Nagatani T (2000) Jamming transition in two-dimensional pedestrian traffic. *Phys A: Statist Mech Appl* 275:281
72. Muramatsu M, Irie T, Nagatani T (1999) Jamming transition in pedestrian counter flow. *Physica A* 267:487
73. Naess P (2006) Are short daily trips compensated by higher leisure mobility? *Environ Plan B: Plan Des* 33:197
74. Nagatani T (2002) The physics of traffic jams. *Report Prog Phys* 65:1331
75. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I* 2:2221
76. Nicolis G, Prigogine I (1989) *Exploring Complexity: An Introduction*. Freeman, New York
77. Parisi G (1993) *Statistical Physics and Biology*. Princeton University Press, p 284
78. Parisi G (1999) *Complex Systems: a Physicist's Viewpoint*. *Phys A* 263–1:557
79. Portugali J (2005) Cognitive Maps Are over 60. *COSIT* 251
80. Prigogine I, Andrews FC (1960) A Boltzmann like approach for traffic flow. *Oper Res* 8:789
81. Pushkarev B, Zupan JM (1975) *Urban space for pedestrians*. The MIT Press, Cambridge
82. Rambaldi S et al (2007) Mobility in modern cities: looking for physical laws. *Proceeding of the ECCS07 Conference, Dresden, 1–3 October 2007, paper n. 132*
83. Rizzolatti G (1997) *From Complexity to Creativity*. Plenum Press, New York
84. Roederer GJ (2003) On the concept of information and its role in Nature. *Entropy* 5:3
85. Schadschneider A, Pöschel T, Kühne R, Schreckenberg M, Wolf DE (eds) (2007) *Traffic and Granular Flow '05*. Springer, Heidelberg
86. Schreckenberg M, Neubert L, Wahle J (2001) Simulation of traffic in large road networks. *Futur Gener Comput Syst* 17–5:649
87. Schweitzer F (2003) *Brownian Agents and Active Particles*. Springer, Berlin
88. Tajima Y, Takimoto K, Nagatani T (2001) Scaling of pedestrian channel flow with a bottleneck. *Physica A* 294:257
89. Tajima Y, Takimoto K, Nagatani T (2002) Pattern formation and jamming transition in pedestrian counter flow. *Physica A* 313:709
90. Taylor M (2003) *The Moment of Complexity*. University of Chicago Press, Chicago
91. Treiber M, Hennecke A, Helbing D (2000) Congested Traffic States in Empirical Observations and Microscopic Simulations. *Phys Rev E* 62:1805
92. Turchetti G, Zanlungo F, Giorgini B (2007) Dynamics and thermodynamics of a gas of automata. *Europhys Lett* 78–5:58003
93. Van Gelder T (1998) The dynamical hypothesis in cognitive science. *Behav Brain Sci* 21:615
94. Vogel A, Nagel K (2005) Multi-agent based simulation of individual traffic in Berlin. Paper presented at CUPUM 2005 Conference, 29 June – 1 July 2005
95. Von Neumann J (1963) The general and logical theory of automata. In: *Collected works, vol V*. Pergamon Press, Oxford, p 288
96. Von Neumann J, Morgenstern O (1947) *Theory of Games and Economic Behavior*. Princeton University Press, Princeton
97. Wahle J et al (2000) Decision Dynamics in a Traffic Scenario. *Adv Complex Syst* 2:1
98. Weufeng F, Lizhong Y, Weicheng F (2003) Simulation of bi-directional pedestrian movement using a cellular automata model. *Physica A* 321:633
99. [www.physicsofthecitylab.unibo.it](http://www.physicsofthecitylab.unibo.it)
100. Yamori K (1998) Going with the flow: Micro-nacro dynamics in the macrobehavioral patterns of pedestrian crowds. *Psychol Rev* 105–3:530

---

## Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment

HESHAM RAKHA, ALY TAWFIK

Center for Sustainable Mobility, Virginia Tech  
Transportation Institute, Virginia Polytechnic Institute  
and State University, Blacksburg, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Driver Travel Decision Behavior Modeling](#)

[Static Traffic Routing and Assignment](#)

[Dynamic Traffic Routing](#)

[Traffic Modeling](#)

[Dynamic Travel Time Estimation](#)

[Dynamic or Time-Dependent Origin-Destination Estimation](#)

[Dynamic Estimation of Measures of Effectiveness](#)

[Use of Technology to Enhance System Performance](#)

[Related Transportation Areas](#)

[Future Directions](#)

[Appendix](#)

[Bibliography](#)

### Glossary

**Link or arc** A roadway segment with homogeneous traffic and roadway characteristics (e. g. same number of lanes, base lane capacity, free-flow speed, speed-at-capacity, and jam density). Typically networks are divided into links for traffic modeling purposes.

**Route or path** A sequence of roadway segments (links or arcs) used by a driver to travel from his/her point of origin to his/her destination.

**Traffic routing** The procedure that computes the sequence of roadways that minimize some utility objective function. This utility function could either be travel time or a generalized function that also includes road tolls.

**Traffic assignment** The procedure used to find the link flows from the Origin-Destination (O-D) demand. Traffic assignment involves two steps: (1) traffic routing and (2) traffic demand loading. Traffic assignment can be divided into static, time-dependent, and dynamic.

**User equilibrium traffic assignment** The assignment of traffic on a network such that it distributes itself in a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs.

**System optimum traffic assignment** The assignment of traffic such that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

**Static traffic assignment** Traffic assignment ignoring the temporal dimension of the problem.

**Time-dependent traffic assignment** An approximate approach to modeling the dynamic traffic assignment problem by dividing the time horizon into steady-state time intervals and applying a static assignment to each time interval.

**Dynamic traffic assignment** Traffic assignment considering the temporal dimension of the problem.

**Traffic loading** The procedure of assigning O-D demands to routes.

**Synthetic O-D estimation** The procedure that estimates O-D demands from measured link flow counts, which includes static, time-dependent, and dynamic.

**Traffic stream motion model** A mathematical representation (traffic flow model) for traffic stream motion behavior.

**Car-following model** A mathematical representation (traffic flow model) for driver longitudinal motion behavior.

**Marginal link travel time** The increase in a link's travel time resulting from an assignment of an additional vehicle to this link.

**Road pricing** Road pricing is an economic concept in which drivers are charged for the use of the road facility.

## Definition of the Subject

The dynamic nature of traffic networks is manifested in both temporal and spatial changes in traffic demand, roadway capacities, and traffic control settings. Typically, the underlying network traffic demand builds up over time at the onset of a peak period, varies stochastically during the peak period, and decays at the conclusion of the peak period. As traffic congestion builds up within a transporta-

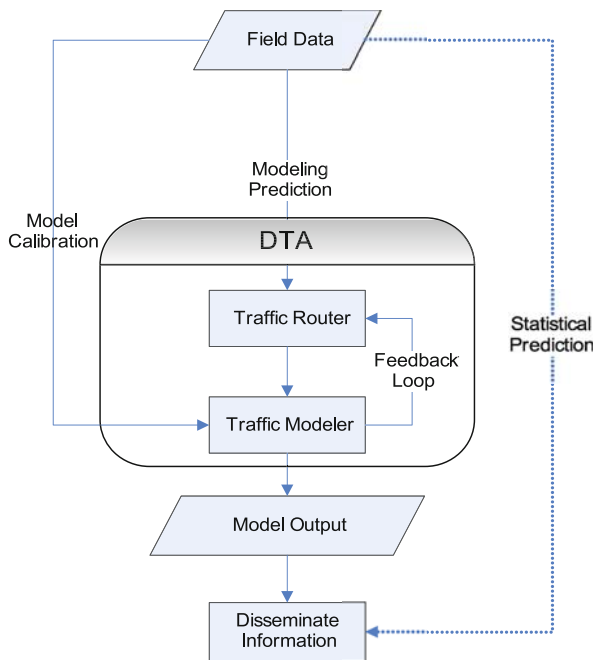
tion network, drivers may elect to either cancel their trip altogether, alter their travel departure time, change their mode of travel, or change their route of travel. Dynamic traffic routing is defined as the process of dynamically selecting the sequence of roadway segments from a trip origin to a trip destination. Dynamic routing entails using time-dependent roadway travel times to compute this sequence of roadway segments. Consequently, the modeling of driver routing behavior requires the estimation of roadway travel times into the near future, which may entail some form of traffic modeling.

In addition to dynamic changes in traffic demand, roadway capacities are both stochastic and vary dynamically as vehicles interact with one another along roadway segments. For example, the roadway capacity at a merge section varies dynamically as the composition of on-ramp and freeway demands vary [45,47,73,75,106,107,108,109,110,122,123,127,144,183]. To further complicate matters, traffic control settings (e.g. traffic signal timings) also vary both temporally and spatially, thus introducing another level of dynamics within transportation networks. All these factors make the dynamic assessment of traffic networks extremely complex, as shall be demonstrated in this article. The article is by no means comprehensive but does provide some insight into the various challenges and complexities that are associated with the assessment of dynamic networks.

## Introduction

Studies have shown that even drivers familiar with a trip typically choose sub-optimal routes thus incurring extra travel time in the range of seven percent on average [103]. Furthermore, the occurrence of incidents and special events introduces other forms of variability that drivers are unable to anticipate and thus result in additional errors in a driver's route selection. Consequently, advanced traveler information systems (ATISs), which are an integral component of intelligent transportation systems (ITSs), can assist the public in their travel decisions by providing real-time travel information via route guidance systems; variable message signs (VMSs); the radio, or the web. It is envisioned that better travel information can enhance the efficiency of a transportation system by allowing travelers to make better decisions regarding their time of departure, mode of travel, and/or route of travel. An integral component of an ATIS is a dynamic traffic assignment (DTA) system. A DTA system predicts the transportation network state over a short time horizon (typically 15- to 60-min time horizon) by modeling complex demand and supply relationships through the use of





**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 1**  
**Schematic of an ATIS Framework**

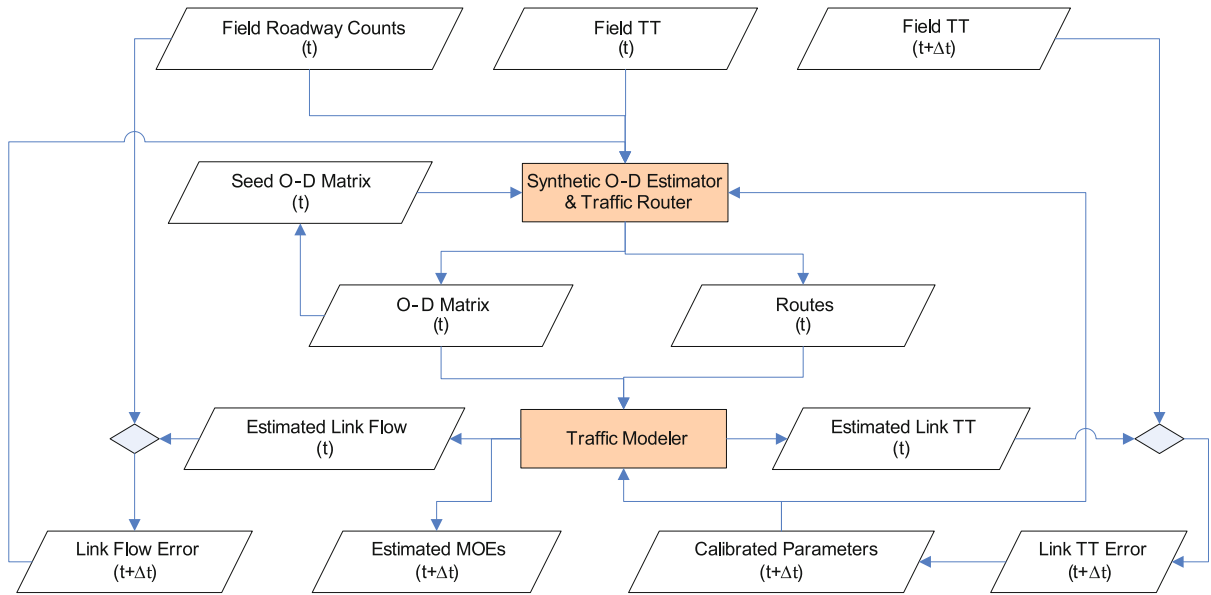
sophisticated models and algorithms. The DTA requires two sets of input, namely demand and supply data. Demand represents the demand for travel and is typically in the form of mode-specific time-dependent origin-destination (O-D) matrices. Alternatively, the supply component models the movement of individual vehicles along a roadway typically using roadway specific speed-flow-density relationships together with the explicit modeling of queue buildup and decay. Figure 1 illustrates schematically that an ATIS can utilize two approaches for the estimation of future traffic conditions, namely: statistical models or a DTA framework. This article focuses on the DTA approach and thus will be described in more detail. The DTA combines a traffic router and modeler, as illustrated in the figure. The traffic router estimates the optimum travel routes while the traffic modeler models traffic to evaluate the performance of traffic after assigning motorists to their routes. A feedback loop allows for the feedback of either travel times or marginal travel times, which in turn, are used by the traffic router to compute the optimum routes. This feedback continues until the travel times are consistent with the travel routes and there is no incentive for drivers to alter their routes.

A DTA can be applied off-line (in a laboratory) or on-line (in the field). An on-line application of a DTA en-

tails gathering traffic data in real-time at any instant  $t$  and feeding these data to the DTA to predict short-term traffic conditions  $\Delta t$  temporal units into the future (i.e. at time  $t + \Delta t$ ). As was mentioned earlier, the input to the DTA includes mode specific time dependent O-D matrices. Unfortunately, current surveillance equipment does not measure O-D matrices; instead they measure traffic volumes passing a specific point. Consequently, O-D estimation tools are required to estimate the O-D matrix from observed link counts, as illustrated in Fig. 2. However, the estimation of an O-D matrix requires identifying which O-D demands contribute to which roadway counts. The assigning of O-D demands to link counts involves what is commonly known in the field of traffic engineering as the traffic assignment problem. Traffic assignment in turn requires real-time O-D matrices and roadway travel times as input. Consequently, some form of feedback is required to solve this problem. A more detailed description of traffic assignment formulations and techniques is provided in Sect. “Static Traffic Routing and Assignment” and “Dynamic Traffic Routing”, while the estimation of route travel times is described in Sect. “Dynamic Travel Time Estimation” and the estimation of O-D matrices is described in Sect. “Dynamic or Time-Dependent Origin-Destination Estimation”.

The dynamic assessment of traffic networks using a DTA is both data driven (trapezoidal boxes) and model based (colored rectangular boxes), as illustrated in Fig. 2. This procedure involves: measuring raw field data, constructing model input data, executing a traffic model to predict future conditions, and advising a traveler in the case of control systems. The framework starts by measuring traffic states at instant “ $t$ ” (roadway travel times and link flows) and subsequently estimating these traffic states  $\Delta t$  in the future. Procedures for the estimation of dynamic roadway travel times are provided in Sect. “Dynamic Travel Time Estimation” of this article. Using the measured link flows and travel times, an O-D matrix is constructed using a synthetic O-D estimator. Section “Dynamic or Time-Dependent Origin-Destination Estimation” describes the various formulations for estimating a dynamic O-D matrix together with some heuristic practical approaches to estimate this O-D matrix.

Once the O-D demands are estimated the future states are predicted using a traffic modeler. Section “Traffic Modeling” provides a brief overview of the various state-of-the-practice modeling approaches. The model also computes various measures of effectiveness (MOEs) including delay, fuel consumption, and emissions, as will be described in Sect. “Dynamic Estimation of Measures of Effectiveness”. The traffic modeler can either combine traf-



**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 2**  
**Dynamic Traffic Assessment and Routing Framework**

fic modeling with traffic assignment or alternatively utilize the routes computed by the O-D estimator to route traffic. This closed loop optimal control framework can involve a single loop or in most cases may involve an iterative loop to attain equilibrium. The framework involves a feedback loop in which input model parameters are adjusted in real-time through the computation of an error between model predictions and actual measurements. This real-time calibration entails adjusting roadway parameters (e.g. capacity, free-flow speed, speed-at-capacity, and jam density) and traffic routes to reflect dynamic changes in traffic and network conditions. For example, the capacity of a roadway might vary because of changes in weather conditions and/or the occurrence of incidents. The system should be able to adapt itself dynamically without any user intervention.

This article attempts to synthesize the literature on the dynamic assessment and routing of traffic. The problem as will be demonstrated later in the paper is extremely complex because, after all, it deals with the human psychic, which not only varies from one person to another, but may also vary depending on the purpose of a trip, the level of urgency the driver has, and the psychic of the driver at the time the trip is made. This article is by no means comprehensive, given the massive literature on the topic, but does highlight some of the key aspects of the problem, how researchers have attempted to address this problem, and future research needs and directions.

The article discusses the various issues associated with the dynamic assessment of transportation systems. Initially, driver travel decision behavior modeling is presented and discussed. Subsequently, various traffic assignment formulations are presented together with the implementation issues associated with these formulations. Next, the mathematical formulations of these assignment techniques are discussed together with mathematical and numerical approaches to modeling dynamic traffic routing. Subsequently, the issues associated with the modeling of traffic stream behavior, the estimation of dynamic roadway travel times, and the estimation of dynamic O-D demands are discussed. Subsequently, the procedures for computation of various assessment measures are presented. Next, the use of technology to alter driver behavior is presented. Finally, directions for further research are presented.

### Driver Travel Decision Behavior Modeling

As with the general case of modeling human behavior, modeling driver travel behavior has always been complicated, never accurate enough, and in constant demand for further research. Among the early attempts to model human choice behavior is the economic theory of the “economic man”; who in the course of being economic is also “rational” [222]. According to Simon’s exact words, “*actual human rationality-striving can at best be an extremely*

*crude and simplified approximation to the kind of global rationality that is implied, for example, by game-theoretical models”.*

In general, traffic assignment (static or dynamic assignment) has undoubtedly been among the most researched transportation problems, if not the most, for more than the past half of a century. However, DTA in particular has had the bigger share for almost one third of a century now. Since the early work of Merchant and Nemhauser [142,143], researchers have attempted to improve available DTA models, hence, providing a very rich and vastly wide literature.

As a result of the rapid technological evolution over the last decade of the previous century (the 20th century); manifested in the communications, information and computational technological advances; a worldwide initiative to add information and communications technology to transport infrastructure and vehicles, termed as the intelligent transportation systems (ITS) program, was introduced to the transportation science. According to the Wikipedia Encyclopedia, among the main objectives of ITS is to “*manage factors that are typically at odds with each other such as vehicles, loads, and routes to improve safety and reduce vehicle wear, transportation times and fuel consumption*”. Needless to say, the ITS impact on route selection and roadway travel times has a direct effect on a DTA.

The main effect of ITS on DTA manifests itself within the area of advanced traveler information systems (ATIS). ATIS is primarily concerned with providing people, in general, and trip makers, in particular, with pre-trip and en-route trip-related information. According to the US Federal Highway Administration (FHWA), “*advanced traveler information includes static and real-time information on traffic conditions, and schedules, road and weather conditions, special events, and tourist information. ATIS is classified by how and when travelers receive their desired information (pre-trip or en-route) and is divided by user service categories. Operations essential to the success of these systems are the collection of traffic and traveler information, the processing and fusing of information - often at a central point, and the distribution of information to travelers. Important components of these systems include new technologies applied to the use and presentation of information and the communications used to effectively disseminate this information*” [157].

As will be discussed later, a significant amount of DTA research is directed towards developing data dissemination standards. These standards attempt to achieve the maximum possible benefits while complying with the ITS objectives. Although the provision of pre-trip informa-

tion may influence traveler departure time and route of travel (and in extreme cases, might result in a person canceling his/her trip all together), thus requiring further complicated DTA models that capture forgone and induced demand, as will be discussed later. Moreover, probably the greatest dimension for DTA model complexity was introduced to research when the disseminated ATIS information was to be designed as a control factor to change the manner by which trips are distributed over the network, for example from user equilibrium to system optimum.

Although ITS and ATIS were practically introduced a little more than a decade ago, and in spite of the significant research funds and efforts that have been devoted to the topic, current available DTA models are, at least, relatively undeveloped, which necessitates new approaches that can capture the challenges from the application domains as well as for the fundamental questions related to tractability and realism [180]. This will be discussed briefly in the following section.

Driver travel decision theory is a complicated research area. Research within this area encompasses a very wide range of research efforts. Before going over a brief list of these possible research areas, it should be noted that most of these research areas overlap with one another. Therefore, for a valid driver behavior model, all of the following aspects should be efficiently covered in a practical and realistic manner. This been said, the following is a brief list of some of the main research areas that are highly related to driver travel decision theory:

- Human decision theory, which can be reflected in the trip maker’s decision to make or cancel a scheduled trip, route and departure time selection, compliance with the pre-trip or en-route disseminated information, en-route path diversion and/or return, mode choice based on disseminated information, etc. Literature concerning human decision theory extends back to more than half a century ago and continues to be researched up to this date. Examples of the literature concerning the human decision theory include: administrative behavior [220,221], theory of choice [15], rational choice theory [222], game theory, and decision field theory [36]. Examples of the literature concerning driver decision theory include: decision field theory [229], approximate reasoning models [113], route choice utility models [89], inductive learning [151], effect of age on routing decisions [243], and rational learning [152].
- Design of disseminated information, which encompasses the criteria governing the dissemination of in-

formation, the structure and type of information to be disseminated, when data are disseminated, and identifying target drivers. This governs, to a large extent, the drivers' compliance rates in response to disseminated information. Hence, affecting the routes chosen by drivers, the traffic volumes on these routes and alternative routes, and different travel times, among others. Literature concerning the effect of ATIS and ATIS content on drivers behavior include: the required information that would reduce traffic congestion [14], the effect of ATIS on drivers route choice [1], commuters diversion propensity [214], the effect of traffic information disseminated through variable message signs on driver choices [172], drivers en-route routing decisions [111].

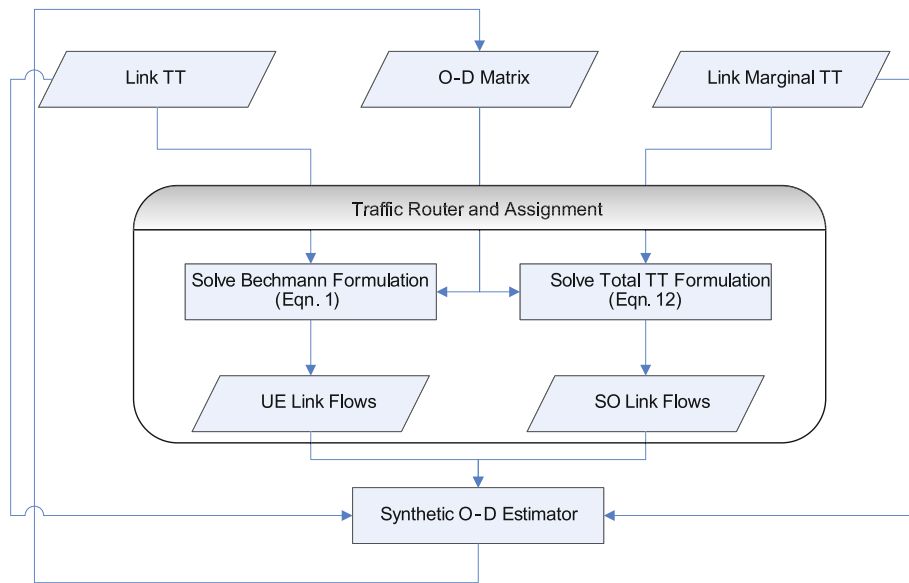
- Human perception based on experience and information provision, which is reflected in day-to-day variations in driver decisions. For example, given identical conditions on two separate days, the same person might select different routes and departure times; possibly due to different experiences on previous days. Examples of current literature include: models that include the incorporation of driver behavior dynamics under information provision [176], behavioral-based consistency seeking models [177], perception updating and day-to-day travel choice dynamics with information provision [104], the modeling of inertia and compliance mechanisms under real-time information [226], drivers psychological deliberation while making dynamic route choices [229], the effect of using in-vehicle navigational systems on driver behavior [11], the effect of network familiarity on routing decisions [128], and the effect of varying levels of cognitive loads on driver behavior [105].
- Among the challenges in modeling human decision theory are the possible data collection techniques. The current practice for data collection includes revealed and stated preference surveys. Research has demonstrated that surveyed stated preference results have significant biases; in comparison to real behavior. In addition to the research being performed to analyze, capture, and improve the reasons for such biases; other research directions are being performed to solve other survey problems. For example, the problems of low and slow survey participation rates, as well as under-represented groups in typical survey techniques. Examples of literature within this field include: stated preference for investigating commuters diversion propensity [214], using stated preference for studying the effect of advanced traffic information on drivers route choice [1], driver response to variable message sign-based traffic information according to stated pref-

erence data collected through three different survey administration methods, namely, an on-site survey, a mail-back survey and an internet-based survey [172], transferring insights into commuter behavior dynamics from laboratory experiments to field surveys [129] and [173], and the applicability of using driving simulators for data collection [113].

- Issues of uncertainty, which is a fundamental feature in most transportation phenomena. Research dealing with uncertainty has a wide application in DTA. It can be represented in the trip maker route travel time estimates, in the compliance rates of drivers to information, in the driver's trust in the disseminated information and its reliability, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the works of Birge and Ho [28], Peeta and Zhou [178,179], Cantarell and Cascetta [38], Ziliaskopoulos and Waller [263], Waller and Ziliaskopoulos [245], Waller [244], Peeta and Jeong [177], Jha et al. [104], Peeta and Paz [171], Koutsopoulos et al. [113], and Hawas [89].

### Static Traffic Routing and Assignment

Prior to describing the issues associated with dynamic routing, a description of static routing issues is first presented. This section describes two formulations for static traffic assignment, namely the User Equilibrium (UE) and System Optimum (SO) assignment. Traffic assignment is defined as the basic problem of finding the link flows given an origin-destination trip matrix and a set of link or marginal link travel times, as illustrated in Fig. 3. The solution of this problem can either be based on the assumption that each motorist travels on the path that minimizes his/her travel time – known as the UE assignment – or alternatively to minimize the system-wide travel time – known as the SO assignment. The traffic assignment initially computes the travel routes (paths) and then determines the unique link flows on the various network links. As will be discussed later, while the estimated link flows are unique the path flows that are derived from these link flows are not unique and thus require some computational tool to estimate the most-likely of these path flows (synthetic O-D estimator). If a time dimension is introduced to the assignment module the formulation is extended from a static to a dynamic context. However, as will be discussed later the addition of a time dimension deems the formulation non-convex and thus the mathematical program used to solve the problem becomes in-



**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 3**  
**Traffic Assignment Framework**

feasible and thus comes the need for a simulation-based solution approach.

Wardrop [246] was the first to explicitly differentiate between these two alternative traffic assignment methods or philosophies. Models based on Wardrop's first principle are referred to as UE, while those based on the second principle are deemed as SO. Wardrop's first principle states that "traffic on a network distributes itself in such a way that the travel costs on all routes used from any origin to any destination are equal, while all unused routes have equal or greater travel costs." Alternatively, Wardrop's second principle states that the average journey travel times of all motorists is a minimum, which implies that the aggregate vehicle-hours spent in travel is also minimum.

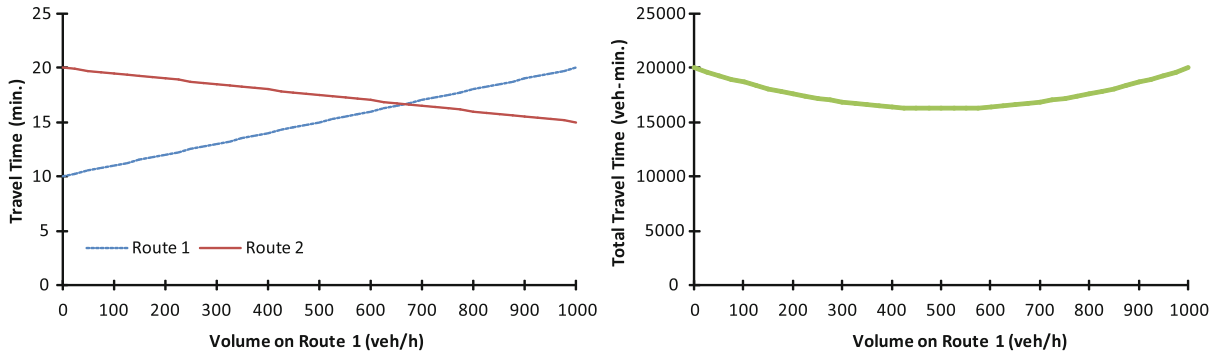
One of the most spectacular examples that illustrated that the UE flow in a network is in general different from the SO flow, is the Braess network [31]. In this network the system-optimal flow was obtained by completely suppressing the flow which would normally occur, on a certain link, at equilibrium. The Braess "paradox" was studied later in more detail [78,80,117,118,147,209,227,228]. For example, Stewart [228] illustrated three important facts using a very simple two-link network and the Braess paradox that included: (a) the equilibrium flow does not necessarily minimize the total cost; (b) adding a new link to a network may increase the total cost at equilibrium; (c) adding a new link to a network may increase the equilibrium travel cost for each individual motorist. Stewart also illustrated that a group of travelers having only one reasonable route may

be seriously inconvenienced by another group of travelers who choose the same route in order to obtain a slight improvement in their personal cost of travel.

### User Equilibrium vs. System Optimum Traffic Assignment

The differences between user and system optimum traffic assignment are best illustrated using an example illustration. The sample test network for this study is derived from an earlier study by Rakha [181]. The network consists of two one-way routes, numbered 1 and 2, from origin A to destination B. The travel time relationship for route 1 is characterized by the relationship  $10 + 0.010 v_1$  where  $v_1$  is the traffic volume on route 1 (veh). Alternatively, the travel time along route 2 is characterized by the relationship  $15 + 0.005 v_2$  where  $v_2$  is the traffic volume traveling along route 2 (veh). Considering at total demand of 1000 veh traveling between zones A and B, the travel time along routes 1 and 2 vary as a function of the volume on each of the routes, as illustrated in Fig. 4. The figure demonstrates that the travel times along routes 1 and 2 are equal at 16.5 min when 667 veh travel along route 1 and 333 veh travel along route 2. Alternatively, the system-optimum traffic assignment is achieved at a volume distribution of 500 veh on routes 1 and 2, respectively. From a traffic engineering point of view, the difference in total travel time between the system and user-optimum traffic assignment (16,250 versus 16,667 veh-min) is of interest. This difference rep-





**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 4**  
**Variation in Route and System Travel Time for Test Network**

represents the extent of possible benefits for a system versus user optimum routing for this particular network and traffic pattern. Figure 4 also illustrates how the average link travel times on routes 1 and 2 vary for the same range of possible routings of traffic between route 1 and 2. In this figure the difference between the travel times on route 1 and 2 (15.0 versus 17.5 min) represents the incentive that exists for vehicles on route 2 to change to route 1. When compared to the user equilibrium routing, the total difference in travel time is composed of two components, which represent the respective increases (route 1) and decreases (route 2) in average travel time that result from a shift from the system to user-optimum routing.

### Implementation Issues

While the simple example illustrated the potential benefits of system optimized routings and the incentive that exists for drivers to switch back to the original user equilibrium routings, it is clear that neither an exhaustive enumeration nor an analytical approach (solving the differential equations of the system travel time) are satisfactory for finding the system optimized routings when more than just a few possible routes are available.

Different static traffic assignment algorithms have been developed over the past half century. These methods are broadly divided into non-equilibrium and equilibrium methods. Non-equilibrium methods include all-or-nothing assignment, where all traffic is assigned to a single minimum path between two zones (path that incurs the minimum travel time). Example algorithms for computing minimum paths include models developed by Dantzig [67] and Dijkstra [69]. Other non-equilibrium methods include incremental, iterative, diversion models, multipath assignment [68], and combined models. According to Van Vliet [240] the incremental assignment method

(explained later) is capable of reaching an acceptable degree of convergence faster than an iterative method. With regards to diversion models, the most common diversion models include the California diversion curves [145] and the Detroit diversion curves [224]. Alternatively, multipath traffic assignment methods assign traffic stochastically. For example, the Dial method [68] stochastically diverts trips to alternate paths, but trips are not explicitly assigned to routes. Other multipath methods [34,35] assume that users do not know the actual travel times on each link, but a driver's estimate of link travel time is drawn randomly from a distribution of possible times. Finally, combined non-equilibrium models include combining capacity restraint models with probabilistic assignment [206], combining iterative with incremental assignment [253,254,255,256], or combining stochastic with equilibrium assignment [216].

Equilibrium assignment techniques are based on Wardrop's first principle [246]. These were classified by Matsoukis and Michalopoulos [138] into: assignments with fixed demand, assignments with elastic demands, and combined models. Only the first method will be discussed. The equilibrium assignment algorithm is a weighted combination of a sequence of all-or-nothing assignments. This produces a non-linear programming (NLP) problem which is subject to linear constraints. This NLP is very hard to solve and the approach seems to be of limited use for realistically sized equilibrium traffic assignment problems. The NLP problem can be replaced by a much simpler linear approximation and solved using the Frank-Wolfe algorithm [81]. This iterative linearization procedure still involves longer computational times than the iterative procedure. LeBlanc et al. [119] developed an iterative procedure solving one-dimensional searches and LP problems that minimize successively better linear approximations to the non-linear objective function. Nguyen [155] converted

the convex optimization problem into a set of simpler sub-problems that could be solved with the convex-simplex method.

One of the most common approaches to implement a user equilibrium traffic assignment involves the use of an incremental traffic assignment technique [121,137,231], [236,237,253,255]. Such a technique breaks down the total traffic demand that is to be loaded onto the network into a number of increments that are each loaded onto the network in turn. Each increment is loaded onto what appears to be the shortest route, after all the previous increments have been loaded. The link travel times are then recalculated, in order to re-compute the fastest route for the next increment to be loaded. When more than one route are to be used for travel between a given origin and destination, the increments are automatically assigned alternatively to each route, when each becomes faster again after previous increments head along the other route. In the end, the extent to which the overall assignment approaches an equilibrium state depends upon the number of increments utilized, with the average final error being roughly proportional to the final increment size.

Van Aerde and Rakha [181,195] demonstrated that the system-optimum traffic assignment can also be solved considering an incremental traffic assignment. Specifically, Van Aerde and Rakha [181,195] recognized the fact that the increase in system travel time caused by the addition of one vehicle is composed of the additional travel time incurred by the subject vehicle and the increase in travel time that is imparted on all other vehicles which are already on the link. While the former quantity is usually already available as a direct or indirect measurement on the link, the derivation of the latter quantity is more subtle. It is a function of the rate of change of the average travel time, per additional vehicle, and the number of vehicles already on the link. In mathematical terms, this is simply the product of the derivative of the travel time versus volume relationship, with respect to volume, multiplied by the volume already present on the link. Consequently, the standard objective function that is utilized in any minimum path algorithm, which searches for the user equilibrium routes, can be replaced by a new objective function that minimizes the total travel time. This routing can be achieved using an incremental assignment of vehicles based on their marginal travel time as opposed to their actual travel time, which results in a system optimum as opposed to a user equilibrium routing, as was demonstrated earlier in Fig. 3. Stated differently at dynamic system optimum, the time-dependent marginal cost on all the paths actually used are equal and less than the marginal cost on any unused paths. In the static case, the path marginal cost

(PMC) is the sum of the link marginal cost (LMC). However, in the dynamic case, the PMC evaluation is much more complicated since path flows are not assigned to links on the path simultaneously. However, within the dynamic context, most researchers [84,166] assume that the path flow perturbation travels along the path at the same speed as the additional flow unit. Shen et al. [217] demonstrated that this assumption is not necessarily correct. Furthermore, they presented a solution algorithm for path-based system optimum models based on a new PMC evaluation method. The approach was then tested and validated on a simple network.

### Dynamic Traffic Routing

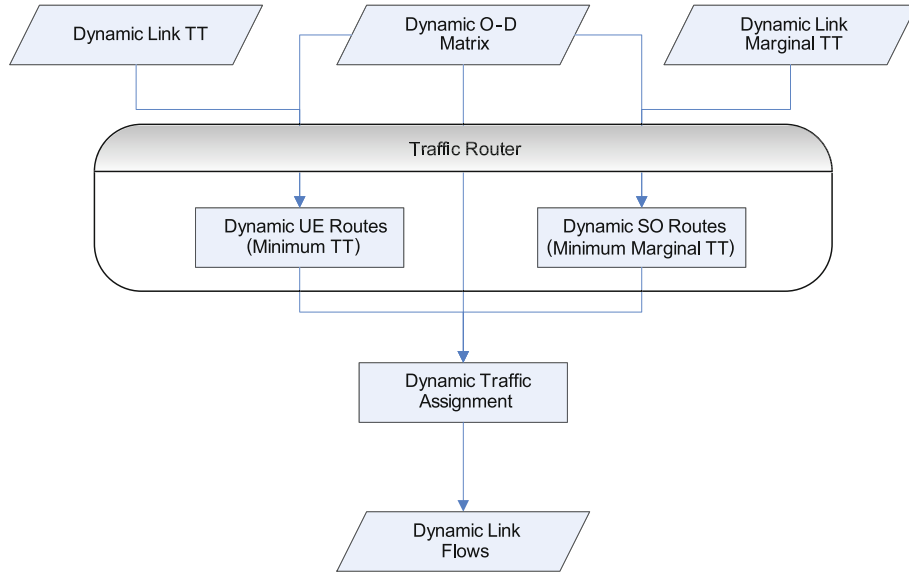
This section describes the mathematical formulations for the static routing problem together with some solution approaches to the problem. Subsequently, the extension of the problem for the dynamic context is presented together with state-of-the-art solution approaches.

The dynamic traffic assignment approach is summarized in Fig. 5 and involves three input variables, namely: dynamic link travel times (in the case of the UE assignment), dynamic marginal travel times (in the case of the SO assignment), and dynamic O-D matrices. In the case of the UE assignment the Bechmann formulation is solved (Eq. (1) if we use a time-dependent static (or quasi static) assignment as will be discussed in detail in the following sections, while in the case of the SO assignment Eq. (12) is solved. Within the static context these formulations are solved analytically using a mathematical program given that the objective function and feasible region are convex. Alternatively, in the dynamic context the objective function is non-convex and thus is more difficult to solve necessitating the use of a modeling approach to solve the problem.

After solving these two formulations the link flows are computed and input into an O-D estimator to provide an estimate of the O-D demand which is then compared to the initial solution. This feedback loop continues until the difference in either link flows or O-D flows is within a desired margin of error or the maximum number of iterations criteria is met.

### Mathematical Formulations

Following the notation presented by Sheffi [215] we present the network notations that are used in the mathematical formulation of a static traffic assignment problem. Initially, the variable definitions are presented followed by the vector definitions (bold variables).



**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 5**  
**Dynamic Traffic Assignment Framework**

- $N$  Set of network nodes  
 $A$  Set of network arcs (links)  
 $R$  Set of origin centroids  
 $S$  Set of destination centroids  
 $k_{rs}$  Set of paths connecting O-D pair  $(r-s)$ ;  $r \in R, s \in S$   
 $x_a$  Flow on arc  $(a)$   
 $t_a$  Travel time on arc  $(a)$   
 $f_k^{rs}$  Flow on path  $(k)$  connecting O-D pair  $(r-s)$   
 $c_k^{rs}$  Travel time on path  $(k)$  connecting O-D pair  $(r-s)$   
 $q_{rs}$  Trip rate between origin  $(r)$  and destination  $(s)$   
 $\delta_{a,k}^{rs}$  Indicator variable:

$$\delta_{a,k}^{rs} = \begin{cases} 1, & \text{if arc } (a) \text{ is on path } (k) \text{ between} \\ & \text{O-D pair } (r-s) \\ 0, & \text{otherwise} \end{cases}$$

Using vector notations (bold variables) the variables are defined as,

- $\mathbf{x}$  Vector of flows on all arcs,  $= (\dots, x_a, \dots)$   
 $\mathbf{t}$  Vector of travel times on all arcs,  $= (\dots, t_a, \dots)$   
 $\mathbf{f}^{rs}$  Vector of flows on all paths connecting O-D pair  $r-s$ ,  $= (\dots, f_k^{rs}, \dots)$   
 $\mathbf{f}$  Matrix of flows on all paths connecting all O-D pairs,  $= (\dots, \mathbf{f}^{rs}, \dots)$   
 $\mathbf{c}^{rs}$  Vector of travel times on all paths connecting O-D pair  $r-s$ ,  $= (\dots, c_k^{rs}, \dots)$   
 $\mathbf{c}$  Matrix of travel times on all paths connecting all O-D pairs,  $= (\dots, \mathbf{c}^{rs}, \dots)$

- $q$  Origin-destination matrix (with elements  $= q_{rs}$ )  
 $\Delta^{rs}$  Link-path incidence matrix (with  $\delta_{a,k}^{rs}$  elements) for O-D pair  $r-s$ , as discussed below  
 $\mathbf{\Delta}$  Matrix of link-path incidence matrices (for all O-D pairs),  $= (\dots, \Delta^{rs}, \dots)$

The link-path incident matrix is of size equal to the number of links or arcs in the network (number of rows) and number of paths between origin  $(r)$  and destination  $(s)$ . The element in the  $a$ th row, and  $k$ th column of  $\Delta^{rs}$  is  $\delta_{a,k}^{rs}$ . In other words,  $(\Delta^{rs})_{a,k} = \delta_{a,k}^{rs}$ .

The following basic relations are fundamental to the mathematical program formulation:

- A link performance function, which is also known as the volume-delay curve or the link congestion function, represents the relationship between flow and travel time on a link  $(a)$  ( $t_a = t_a(x_a)$ ).
- The mathematical program formulations assume that travel time on a given link is only dependent on the flow on the subject link (the model does not capture the effect of opposing flows on the delay of opposed flows), or mathematically

$$\frac{\partial t_a(x_a)}{\partial x_b} = 0 \quad \forall a \neq b \quad \text{and} \quad \frac{\partial t_a(x_a)}{\partial x_a} > 0$$

$\forall a$  where  $x_b$  is the flow on link  $(b)$ .

- The travel time on a particular path equals the sum of the travel times on the links comprising that path as

$c_k^{rs} = \sum_a t_a \delta_{a,k}^{rs} \quad k \in k_{rs}, \forall r \in R, \forall s \in S$  or  $\mathbf{c} = \mathbf{t} \cdot \Delta$  considering the vector notation.

- The flow on each link equals the sum of the flows on all paths traversing the subject link as  $x_a = \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) \quad \forall a \in A$  or  $\mathbf{x} = \mathbf{f} \cdot \Delta^T$ .
- The above formula uses the incidence relationships to express link flows in term of path flows, i. e.  $\mathbf{x} = \mathbf{x}(\mathbf{f})$ . The incidence relationships also mean that the partial derivative of the link flow can be defined with respect to a particular path flow as follows,

$$\frac{\partial x_a(f)}{\partial f_l^{mn}} = \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) = \delta_{a,l}^{mn},$$

where  $\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = 0$  if  $k \neq l$  or  $r - s \neq m - n$ .

Where,  $f_l^{mn}$  is the flow on path ( $l$ ) connecting O-D pair ( $m - n$ ). Since the function  $x_a(f)$  includes a flow summation using the subscripts  $r, s$ , and  $k$ , the variable with respect to which the derivative is being taken is sub-scripted by  $m, n$ , and  $l$ , to avoid the confusion in differentiation.

### User Equilibrium

As mentioned earlier, the UE model is based on the assumption that each traveler takes the path that minimizes his/her travel time from their origin to their destination, regardless of any effect this might have on the other network users. In other words, at equilibrium, none of the travelers will be able to reduce their travel times by unilaterally switching to another path. This implies that at equilibrium the link flow pattern is such that the travel times on all of the used paths connecting any given O-D pair will be equal. The travel time on all of these used paths will also be less than or equal to the travel time on any of the unused paths.

The mathematical program that represents this model can be cast using Beckmann's transformation as,

$$\begin{aligned} \min. \quad & z(\mathbf{x}) = \sum_a \int_0^{x_a} t_a(w) dw \\ \text{s.t.} \quad & \sum_k f_k^{rs} = q_{rs} \quad \forall r, s \\ & \text{(Flow conservation constraints)} \quad (1) \end{aligned}$$

$$f_k^{rs} \geq 0 \quad \forall k, r, s$$

(Non-negativity constraints)

$$x_a = \sum_r \sum_s \sum_k (f_k^{rs} \cdot \delta_{a,k}^{rs}) \quad \forall a.$$

It is worth mentioning that this formulation “has been evident in the transportation literature since the mid-1950's, but its usefulness became apparent only when solution algorithms for this program were developed in the late 1960's and early 1970's” [215].

In order to prove that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment, first the equivalence conditions will be discussed followed by the uniqueness conditions. In the equivalence conditions it will be shown that the first-order conditions for the minimization program are identical to the equilibrium conditions. Whereas, in the uniqueness conditions, it will be shown that the user-equilibrium equivalent minimization program has only one solution. Hence, proving that the solution of Beckmann's transformation program satisfies the user-equilibrium assignment problem.

**Equivalency Conditions** Beckmann's transformation program is a minimization program with linear equality and non-negativity constraints. In order to find the first-order conditions for such a program, the Lagrangian with respect to the equality constraints can be written as

$$L(f, u) = z[x(f)] + \sum_r \sum_s u_{rs} \left( q_{rs} - \sum_k f_k^{rs} \right), \quad (2)$$

where  $u_{rs}$  denotes the dual variable associated with the flow conservation constraint for O-D pair ( $r-s$ ). At the stationary point of the Lagrangian, the following first-order conditions have to hold with respect to the path-flow variables and the dual variables. First, with respect to the path-flow variables

$$\begin{aligned} f_k^{rs} \frac{\partial L(f, u)}{\partial f_k^{rs}} &= 0 \quad \forall k, r, s \\ \text{and} \quad \frac{\partial L(f, u)}{\partial f_k^{rs}} &\geq 0 \quad \forall k, r, s \end{aligned} \quad (3)$$

must hold. Alternatively, with respect to the dual variables

$$\frac{\partial L(f, u)}{\partial u_{rs}} = 0 \quad \forall r, s \quad (4)$$

must hold. In addition to the following non-negativity constraints,

$$f_k^{rs} \geq 0 \quad \forall k, r, s. \quad (5)$$

Note that the formulation of this Lagrangian is given in

terms of path flow by using the incidence relationships,  $x_a = x_a(f)$ .

The partial derivative of  $L(x, u)$  with respect to the flow variables  $f_l^{mn}$  can be given by

$$\begin{aligned} \frac{\partial L(f, y)}{\partial f_l^{mn}} &= \frac{\partial}{\partial f_l^{mn}} z[x(f)] \\ &+ \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left( q_{rs} - \sum_k f_k^{rs} \right). \end{aligned} \quad (6)$$

Using the chain rule the first term can be solved as

$$\begin{aligned} \frac{\partial}{\partial f_l^{mn}} z[x(f)] &= \sum_{b \in A} \frac{\partial z(x)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\ &= \sum_b \left[ \left( \frac{\partial}{\partial x_b} \sum_a \int_0^{x_a} t_a(w) dw \right) \cdot \left( \frac{\partial x_b}{\partial f_l^{mn}} \right) \right] \\ &= \sum_b t_b \cdot \delta_{b,l}^{mn} = c_l^{mn}. \end{aligned} \quad (7)$$

The second term can be solved as

$$\frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s u_{rs} \left( q_{rs} - \sum_k f_k^{rs} \right) = -u_{mn}, \quad (8)$$

because (a)  $u_{rs}$  is not a function of  $f_l^{mn}$ ; (b)  $q_{rs}$  is constant; and

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise} \end{cases}.$$

Consequently, Eq. (3) and (4) can be solved to derive

$$\frac{\partial L(f, u)}{\partial f_l^{mn}} = c_l^{mn} - u_{mn}.$$

Hence, we can derive the following first-order conditions,

$$\begin{aligned} f_k^{rs} (c_k^{rs} - u_{rs}) &= 0 \quad \forall k, r, s \\ c_k^{rs} - u_{rs} &\geq 0 \quad \forall k, r, s \\ \sum_k f_k^{rs} &= q_{rs} \quad \forall r, s \\ f_k^{rs} &\geq 0 \quad \forall k, r, s. \end{aligned} \quad (9)$$

We can imply the following from these conditions, (1) The first two conditions, for any path ( $k$ ) connecting any O-D

pair ( $r - s$ ), either (a) The flow on that path,  $f_k^{rs}$ , equals zero, in which case, the travel time on this path,  $c_k^{rs}$ , will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier,  $u_{rs}$ , or, (b) the flow on that path will have a value (greater than zero), in which case, the travel time on this path will have a value equal to the value of the O-D specific Lagrange multiplier,  $u_{rs}$ . In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the travel time on all other paths connecting the same O-D pair. Hence, this value of the Lagrange multiplier is the minimum path travel time between this O-D pair thus proving that the solution of Beckman's transformation program satisfies the user-equilibrium assignment.

The last two conditions satisfy the flow conversation and non-negativity constraints, respectively. The proof can further be explained as follows, paths connecting O-D pair ( $r - s$ ) can be divided into two groups, (1) Paths with zero flow, and are characterized by a travel time which is either greater than or equal to the minimum travel time; and (2) Paths with non-negative flows, and are characterized by minimum travel times. Thus, confirming the user-equilibrium notion which states that no user can improve his/her travel times by unilaterally changing their routes.

The above proves that user-equilibrium conditions are satisfied at any stationary point of Beckman's transformation program. The following section proves that there is only one solution for Beckman's transformation program. It proves that Beckman's transformation program has only one stationary point, and that this point is a minimum.

**Uniqueness Condition** In order to prove that Beckmann's transformation program has only one solution, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints ensure a convex feasible region, and that the addition of the non-negativity constraints does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven in two different ways. The first way can be achieved by the application of the properties of convex functions, on the link-performance functions. On the other hand, the second proof is achieved by proving that the Hessian matrix of the objective function is positive definite.

Link performance functions are known to be continuously increasing functions. Hence, link-performance functions are convex functions. The objective function equals the summation of the integral of the link-perfor-



mance functions of all links. Properties of convex functions state that integrals of convex functions are also convex functions, and that the summation of convex functions is also a convex function. Hence, proving that the objective function is convex everywhere. Subsequently proving that there is only one solution for Beckman's transformation program, with respect to link flows, and that solution is a minimum.

Recalling that

$$\frac{\partial^2 z(x)}{\partial x_m \partial x_n} = \frac{\partial t_m(x_m)}{\partial x_n} = \begin{cases} 1, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

the Hessian matrix for the objective function can be calculated to be as follows,

$$\begin{aligned} \nabla^2 z(x) &= \begin{bmatrix} \frac{\partial^2 z(x)}{\partial x_1^2} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_1} \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 z(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 z(x)}{\partial x_A \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 z(x)}{\partial x_1 \partial x_A} & \frac{\partial^2 z(x)}{\partial x_2 \partial x_A} & \cdots & \frac{\partial^2 z(x)}{\partial x_A^2} \end{bmatrix} \\ &= \begin{bmatrix} \frac{dt_1(x_1)}{dx_1} & 0 & \cdots & 0 \\ 0 & \frac{dt_2(x_2)}{dx_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{dt_A(x_A)}{dx_A} \end{bmatrix}. \quad (11) \end{aligned}$$

Obviously, the matrix is definite positive, proving that the objective function is strictly positive, and subsequently, has a unique minimum solution.

It is worth mentioning that the Beckman's transformation program is not convex with respect to path flows, and therefore, the equilibrium conditions themselves are not unique with respect to path flows. In other words, while there is actually only one unique solution for link flows, there are an infinite number of paths flows solutions that would produce this unique link flows solution, which raises the need to compute the most likely of these solutions using a synthetic O-D estimator as was described earlier and will be discussed later in more detail.

### System Optimum

As mentioned earlier, the SO model attempts to minimize the total travel time spent in the network. Hence, it might assign certain trips to a slightly longer path (in terms of travel time), in order to reduce the travel time of other user trips by a value which is greater than the value of the increased travel time, and thus achieving a reduced total network travel time. Opposite to user equilibrium, in the system optimum state, users can reduce their travel times by

unilaterally switching to alternative paths, which becomes a challenge to implement such a strategy. Therefore, the solution is not stable. SO network travel time mainly serves as a yardstick that measures the performance of a network.

The mathematical program that represents this model can be written as follows,

$$\begin{aligned} \min. \quad & \tilde{z}(x) = \sum_a x_a \cdot t_z(x_a) \\ \text{s.t.} \quad & \sum_k f_k^{rs} = q_{rs} \quad \forall r, s \\ & \text{(Flow conservation constraints)} \quad (12) \\ & f_k^{rs} \geq 0 \quad \forall k, r, s \\ & \text{(Non-negativity constraints)} \\ & x_a = \sum_r \sum_s \sum_k \left( f_k^{rs} \cdot \delta_{a,k}^{rs} \right) \quad \forall a. \end{aligned}$$

As can be seen, the only difference between user-equilibrium and system optimum programs is the objective function. The SO optimum objective function equals the summation of the products of the travel time on each link times the traffic volume assigned to this link, for all links. Hence, it works on minimizing the total travel time experienced by all vehicles traveling on all links of the networks. On the other hand, the UE objective function equalled the summation of only the travel times of all links.

It can also be seen that the constraints in the SO model are exactly the same as in the UE model. Consequently, similar to the case with the user-equilibrium equivalent program, the solution of this program can be found by solving for the first-order conditions for a stationary point of the following Lagrangian

$$\tilde{L}(f, u) = \tilde{z}[x(f)] + \sum_r \sum_s \tilde{u}_{rs} \left( q_{rs} - \sum_k f_k^{rs} \right), \quad (13)$$

where  $\tilde{u}_{rs}$  denotes the dual variable associated with the flow conservation constraint for O-D pair (r-s). At the stationary point of the Lagrangian, the following first-order conditions have to hold with respect to the path-flow variables

$$\begin{aligned} f_k^{rs} \frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_k^{rs}} &= 0 \quad \forall k, r, s \\ \text{and} \quad \frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_k^{rs}} &\geq 0 \quad \forall k, r, s. \end{aligned} \quad (14)$$

With respect to the dual variables

$$\frac{\partial \tilde{L}(f, u)}{\partial \tilde{u}_{rs}} = 0 \quad \forall r, s. \quad (15)$$

In addition to the non-negativity constraints

$$f_k^{rs} \geq 0 \quad \forall k, r, s. \quad (16)$$

Note that, the formulation of this Lagrangian is given in terms of path flow by using the incidence relationships,  $x_a = x_a(f)$ .

The partial derivative of  $\tilde{L}(x, \tilde{u})$  with respect to the flow variables  $f_l^{mn}$  can be given by

$$\begin{aligned} \frac{\partial \tilde{L}(f, \tilde{u})}{\partial f_l^{mn}} &= \frac{\partial}{\partial f_l^{mn}} \tilde{z}[x(f)] \\ &\quad + \frac{\partial}{\partial f_l^{mn}} \sum_r \sum_s \tilde{u}_{rs} \left( q_{rs} - \sum_k f_k^{rs} \right) \\ \frac{\partial}{\partial f_l^{mn}} \tilde{z}[x(f)] &= \sum_{b \in A} \frac{\partial \tilde{z}(x)}{\partial x_b} \cdot \frac{\partial x_b}{\partial f_l^{mn}} \\ &= \sum_b \left[ \left( \frac{\partial}{\partial x_b} \sum_a x_a \cdot t_a(x_a) \right) \frac{\partial x_b}{\partial f_l^{mn}} \right] \\ &= \sum_b \left[ t_b(x_b) + x_b \frac{dt_b(x_b)}{dx_b} \right] \delta_{b,l}^{mn} \\ &= \sum_b \tilde{t}_b \cdot \delta_{b,l}^{mn} = \tilde{c}_l^{mn}. \end{aligned}$$

Assuming  $\tilde{t}_b = t_b(x_b) + x_b \frac{dt_b(x_b)}{dx_b}$  and  $\partial/(\partial f_l^{mn}) \sum_r \sum_s \tilde{u}_{rs} (q_{rs} - \sum_k f_k^{rs}) = -\tilde{u}_{mn}$  because (a)  $u_{rs}$  is not a function of  $f_l^{mn}$ ; (b)  $q_{rs}$  is constant; and (c)

$$\frac{\partial f_k^{rs}}{\partial f_l^{mn}} = \begin{cases} 1, & \text{if } r = m, s = n, \text{ and } k = l \\ 0, & \text{otherwise.} \end{cases}$$

Therefore  $(\partial \tilde{L}(f, \tilde{u})/(\partial f_l^{mn})) = \tilde{c}_l^{mn} - \tilde{u}_{mn}$ .

Where,  $\tilde{t}_a$  is a summation of two terms, (1)  $t_a(x_a)$ , which is the travel time experienced by this additional driver when the total link flow is  $(x_a)$  and (2)  $\frac{dt_a(x_a)}{dx_a}$ , which is the additional travel time burden that this driver inflicts on each one of the other  $(x_a)$  travelers already using link  $a$ .

In summary, it can be interpreted as the marginal contribution of an additional traveler – or an infinitesimal flow unit – on the  $a$ th link to the total travel time on that link.

Substituting the above results into Eqs. (14) through (16), we get the following first-order conditions

$$\begin{aligned} f_k^{rs} (\tilde{c}_k^{rs} - \tilde{u}_{rs}) &= 0 & \forall k, r, s \\ \tilde{c}_k^{rs} - \tilde{u}_{rs} &\geq 0 & \forall k, r, s \end{aligned}$$

$$\begin{aligned} \sum_k f_k^{rs} &= q_{rs} & \forall r, s \\ f_k^{rs} &\geq 0 & \forall k, r, s \end{aligned}$$

Similar to the interpretation of the user equilibrium conditions, the following can be implied from the above, (1) the first two Conditions, for any path  $(k)$  connecting any O-D pair  $(r-s)$ , either (a) the flow on that path,  $f_k^{rs}$ , equals zero whenever the marginal total travel time on this path,  $\tilde{c}_k^{rs}$ , will have a value that is greater than or equal to the value of the O-D specific Lagrange multiplier,  $\tilde{u}_{rs}$ , or, (b) the flow on that path,  $f_k^{rs}$ , will have a value (greater than zero) whenever the marginal total travel time on this path,  $\tilde{c}_k^{rs}$ , will have a value equal to the value of the O-D specific Lagrange multiplier,  $\tilde{u}_{rs}$ . In both cases, the value of the O-D specific Lagrange multiplier is always less than or equal to the marginal total travel time on all other paths connecting the same O-D pair, i.e. the value of the Lagrange multiplier is the marginal travel time on the used paths between this O-D pair. (2) The last two conditions satisfy the flow conservation and non-negativity constraints, respectively.

The proof can further be explained as follows, paths connecting O-D pair  $(r-s)$  can be divided into two groups, (1) Paths with zero flow, and are characterized by a total marginal travel time which is either greater than or equal to the marginal travel time of the used networks (or the Lagrange multiplier). (2) Paths with non-negative flows, and are characterized by equal marginal travel times.

In order to prove that the SO program has only one solution, as was the case with the user equilibrium program, it is sufficient to prove that the objective function is strictly convex in the vicinity of the solution point, convex everywhere else (within the feasible solution region), and that the feasible region (defined by the constraints) is convex.

It is known that linear equality constraints assure a convex feasible region, and that the addition of the non-negativity constraints does not alter this fact. The convexity of the objective function, with respect to link flows, can be proven if the Hessian matrix of the objective function is positive definite.

Recalling that,

$$\begin{aligned} \frac{\partial^2 \tilde{z}(x)}{\partial x_m \partial x_n} &= \frac{\partial}{\partial x_n} \cdot \left[ \frac{\partial}{\partial x_m} \sum_a x_a \cdot t_a(x_a) \right] \\ &= \frac{\partial}{\partial x_n} \left[ t_m(x_m) + x_m \frac{dt_m(x_m)}{dx_m} \right] \\ &= \begin{cases} 2 \frac{dt_m(x_m)}{dx_m} + x_m \frac{d^2 t_m(x_m)}{dx_m^2}, & \text{for } m = n \\ 0, & \text{otherwise} \end{cases} \end{aligned}$$

As in the user equilibrium program, the Hessian matrix for the objective function can be calculated to be as

$$\nabla^2 z(x) = \begin{bmatrix} 2 \frac{dt_1(x_1)}{dx_1} + x_1 \frac{d^2 t_1(x_1)}{dx_1^2} & 0 & \cdots & 0 \\ 0 & 2 \frac{dt_2(x_2)}{dx_2} + x_n \frac{d^2 t_2(x_2)}{dx_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 2 \frac{dt_A(x_A)}{dx_A} + x_A \frac{d^2 t_A(x_A)}{dx_A^2} \end{bmatrix}.$$

This Hessian matrix is positive definite if all the diagonal terms are positive, which is manifested if the link performance functions are positive. Based on the earlier discussion in the user equilibrium section, it was demonstrated that link-performance functions are convex, and thus demonstrating that the objective function is strictly positive, and subsequently, has a unique minimum solution – with respect to link flows.

It is worth noting that user equilibrium and system optimum produce identical results in any of the following: (1) If congestion effects were ignored, i. e.  $t_a(x_a) = t'_a$  (a constant value per arc) or (2) In case of minimal traffic volumes, that would have negligible effects on the arc specific travel times,  $t_a(x_a)$ .

### Dynamic Traffic Assignment Solution Approach

The extension from a static to a dynamic formulation involves the introduction of two time indices into the formulation. The first time index identifies the time at which the path flow leaves its origin while the second time index identifies when the path flow is observed on a specific link. Unfortunately, the introduction of these time indices deems the objective function non-convex and thus two approaches are considered in solving this problem. The first approach is to divide the analysis period into time intervals while assuming that conditions are static within each time interval (time-dependent static or quasi static). The duration of these intervals are network dependent and should be sufficiently long enough to ensure that motorists can complete their trip within the time interval. The static UE and SO mathematical programs can then be solved for each time interval using the standard static formulations that were presented earlier. The mathematical solution approach requires a closed form solution using an analytical modeling approach. Analytical modeling of the network aims at finding the correct

mathematical presentation of DTA models that would realistically reflect the real world problem with minimum compromises in the modeling of traffic behavior. The solution of such models should guarantee theoretical existence, uniqueness, and stability. Analytical models are valuable because theoretical insights can be analytically derived. Different analytical network modeling may include mathematical programming formulations, optimal control formulations, and variational inequality formulations [180]. Literature within this area of research is extensive. In general, models within the group may be classified into [180]: i) mathematical programming formulations, as the works of Merchant and Nemhauser [142,143], Ho [94], Carey [39,40,41], Janson [98,99], Birge and Ho [28], Ziliaskopoulos [262], Carey and Subrahmanian [42]; ii) optimal control formulations, as in the works of Friesz et al. [83], Ran and Shimazaki [204,205], Wie [249], Ran et al. [202], Boyce et al. [30]; and iii) variational inequality formulations, as with the works of Dafermos [64], Friesz et al. [82], Wie et al. [37], Ran and Boyce [203], Ran et al. [201], Chen and Hsueh [96].

Alternatively, the second approach involves the use of a simulation solution approach. Simulation models on the other hand, in spite of solving the DTA problem within a simulation environment, still use some form of mathematical abstraction of the problem. According to Peeta [180], “the terminology simulation-based models may be a misnomer. This is because the mathematical abstraction of the problem is a typical analytical formulation, mostly of the mathematical programming variety in the current literature. However, the critical constraints that describe the traffic flow propagation, and the spatio-temporal interactions, such as the link-path incidence relationships, flow conservation, and vehicular movements are addressed through simulation instead of analytical evaluation while solving the problem. This is because analytical representations of traffic flows that adequately replicate traffic theoretic relationships and yield well-behaved mathematical formulations are currently unavailable. Hence, the term simulation-based primarily connotes the solution methodology rather than the problem formulation. A key issue with simulation-based models is that theoretical insights cannot be analytically derived as the complex traffic interactions are modeled using simulation. On the other hand, due to the inherently ill-behaved nature of the DTA problem, notions of convergence and uniqueness of the associated solution may not be particularly meaningful from a practical standpoint. In addition, due to their better fidelity vis-à-vis realistic traffic modeling, simulation-based models have gained greater acceptability in the context of real-world deployment”.

One of the early simulation DTA tools is the Simulation and Assignment in Urban Road Networks (SATURN) approach. The SATURN algorithm utilizes an equilibrium technique which optimally combines a succession of all-or-nothing assignments (i. e. it is an iterative equilibrium assignment based on iterative traffic loading) [29,87,241]. This model treats platoons of traffic rather than individual vehicles but delays vehicles but delays at intersections are treated in considerable detail. The model consists of two parts: a simulation component and a traffic assignment component. The traffic simulation component fits a delay-flow power curve to three points, namely: zero flow, current flow, and capacity. This delay-flow curve is used by the assignment model to route vehicles. For each traffic signal four cyclic flow profiles are considered: the IN pattern, the ARRIVE pattern, the ACCEPT pattern, and the OUT pattern. SATURN can account for delays caused by opposing flows, delays caused by vehicles on the same roadway, the shape of the arriving platoon, the effect of traffic signal phasing structure and offsets, and individual lane capacities. Arrival rates that exceed capacity are assumed to form queues that build up at constant rates. SATURN can model networks at two levels of detail, namely: inner and buffer. The model was used in studies in the U.K., Australia, and New Zealand. The limitations of the model include: (a) it assumes steady-state conditions for periods of 15–30 minutes and thus is a time-dependent dynamic assignment approach; (b) queues are modeled vertically and thus they cannot spillback to upstream intersections; (c) it is unsuitable for freeways; (d) it cannot model over-saturated conditions explicitly.

Another early simulation DTA that was developed in the late 1970s is the CONTRAM model (Continuous TRaffic Assignment Model). CONTRAM is similar to SATURN in that it combines traffic assignment with traffic simulation [121]. CONTRAM is a computer based time-varying assignment and queuing model. Unlike SATURN, vehicles are grouped within CONTRAM into packets where each packet is treated in the same way as a single vehicle when assigning it to its minimum path. Time varying flow conditions are modeled by dividing the simulation period into a number of consecutive time intervals, which need not be of the same length, and the packets leave each origin at a uniform rate through each such interval. The assignment is an incremental iterative technique where during the first iteration; packets are routed based on link-travel times of previous packets. However, in successive iterations, they are routed based on link travel times that reflect a weighting of travel times during previous iterations and previous packets. Prior to routing a packet, the packet volume is removed from its previ-

ously used links. An advantage of this assignment model is that it takes into account the effects of packets leaving later on the routing of packets which leave earlier. Thus, it decides upon the path based on a fully loaded network, rather than on one in which has only been loaded to the extent of any previous increments. This model is more dynamic than most models because vehicles are able to change their routing decisions while en-route, if traffic conditions alter. Satisfactory convergence is usually achieved in 5 to 10 iterations. The limitations of the model are: (a) introduction of signal optimization makes the model unable to converge; (b) vehicles queue vertically on a link; (c) no limitation of the storage capacity of a link is introduced; (d) it is unsuitable for freeway networks; and (e) it can only assign vehicles based on Wardrop's first principle.

A number of contemporary DTA models were developed using the basic CONTRAM concept, including the INTEGRATION [195,196,208,209,210,211,231,232,233,234,236,237], DYNASMART [2,3,4,5,53,100,101,102,170,225] and DYNAMIT [19,22,113,257] modeling approaches. In this section the INTEGRATION dynamic traffic assignment and modeling framework is briefly described as an example illustration of a microscopic traffic assignment and simulation approach. The INTEGRATION model is similar to the CONTRAM model in that it models individual vehicles (packets of unit size). Unlike, other traffic assignment models, the INTEGRATION traffic simulation logic is microscopic in that it models vehicles at a deci-second level of resolution. The software combines car-following, vehicle dynamics, lane-changing, energy, and emission models. Thus, mobile source emissions can be directly estimated from instantaneous speed and acceleration levels. Furthermore, the traffic and emission modeling modules have been tested and validated extensively. For example, the software, which was developed over the past two decades, has not only been validated against standard traffic flow theory [185,199], but has also been utilized for the evaluation of real-life applications [190,198]. Furthermore, the INTEGRATION software offers unique capability through the explicit modeling of vehicle dynamics by computing the tractive and resistance forces on the vehicle each deci-second [187,188,193].

The INTEGRATION software uses car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or acceler-

ating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics [185,192]. The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by [194]. Alternatively, lane-changing behavior describes the lateral behavior of vehicles along a roadway segment. Lane changing behavior affects the vehicle car-following behavior especially at high intensity lane changing locations such as merge, diverge, and weaving sections.

The INTEGRATION model provides for 7 basic traffic assignment/ routing options: (a) Time-Dependent Method of Successive Averages (MSA); (b) Time-Dependent Sub-Population Feedback Assignment (SFA); (c) Time-Dependent Individual Feedback Assignment (IFA); (d) Time-Dependent Dynamic Traffic Assignment (DTA); (e) Time-Dependent Frank-Wolf Algorithm (FWA); (f) Time-Dependent External; and (g) Distance Based Routing.

The derivation of a time series of MSA traffic assignments involves analyzing each time slice in isolation of either prior or subsequent time slices (time-dependent static or quasi static). The link travel times, upon which the route computations are based, are estimated based on the prevailing O-D pattern and an approximate macroscopic travel time relationship for each link. Multiple paths are computed in an iterative fashion, where the tree for each subsequent iteration is based on the travel times estimated during the previous iterations. The weight assigned to each new tree is  $1/N$  where  $N$  is the iteration number.

In the case of the feedback assignment vehicles base their routings on the experience of previous vehicle departures (incremental traffic assignment). In the case of the SFA assignment all drivers of a specific type are divided into 5 sub-populations each consisting of 20% of all drivers. The paths for each of these sub-populations are then updated every  $t$  seconds during the simulation based on real-time measurements of the link travel times for that specific vehicle class. The value of  $t$  is a user-specified value. Furthermore, the minimum path updates of each vehicle sub-population are staggered in time, in order to avoid having all vehicle sub populations update their paths at the same time. This results in 20% of the driver paths being updated every  $t/5$  seconds. In the case of the IFA assignment all paths are customized to each individual driver and may therefore be unique relative to any other drivers. This incremental traffic assignment accounts the effect of earlier vehicle departures on the travel time of later which is very similar to the CONTRAM approach.

However, unlike CONTRAM no iterations are made to re-assign all the vehicles.

The INTEGRATION DTA computes the minimum path for every scheduled vehicle departure, in view of the link travel times anticipated in the network at the time the vehicle will reach these specific links. The anticipated travel time for each link is estimated based on anticipated link traffic volumes and queue sizes. This routing involves the execution of a complete mesoscopic DTA model prior to the simulation of the traffic. During this DTA, the routes of all vehicles are computed using the above procedure. Upon completion of this DTA, the actual simulation simply implements the routings computed as per the DTA.

Clearly the validity of any of these modeling approaches hinges on the ability of the traffic simulation model to reflect real-life behavior and capture all the complexities of traffic modeling. Clearly, no modeling approach can claim that it is capable of capturing every aspect of empirical traffic flow behavior and thus the output of such models should be interpreted within the context of how they model the spatio-temporal behavior of drivers.

It should also be noted that the models that were described in this section are heuristic approaches attempting to solve the mathematical formulations that were presented earlier and thus there is no guarantee that they converge to a single (unique) solution for UE and/or SO assignment problems for a complex dynamic network. Furthermore, it is not clear if drivers actually attain such an equilibrium state in such networks. Consequently, research is needed to study and develop models on how drivers select routes, how they respond to the dissemination of traffic information, and how their routing decisions vary temporally in the short- and long-term.

## Traffic Modeling

A key component of a DTA is the modeling of traffic stream behavior in order to predict traffic states into the near future and compute link travel times and various measures of effectiveness, as was illustrated earlier in Fig. 2. This section briefly summarizes the various state-of-the-practice approaches to traffic modeling. Researchers have demonstrated that these approaches are unable to predict empirical spatio-temporal aspects [106] observed in the field. Conversely, others have argued that these approaches, while not perfect, capture the main aspects of empirical data. While our objective is not to argue either way, it is sufficient to note that these tools are being used by transportation professionals to assess dynamic networks and thus are presented in this section. These



approaches can be classified into three categories, which include: macroscopic, mesoscopic, and microscopic approaches. Each of these approaches is briefly described in this section. Again the description is by no means comprehensive but does provide a general overview of these approaches. The interested reader should consider reading the wealth of literature on this topic.

Prior to describing the specifics of the various modeling approaches it is important to note that with the exception to research conducted by Kerner [106,107], most existing approaches are based on the famous one-dimensional kinematic waves (KW) theory, which was proposed by Lighthill and Witham [126] and independently proposed by Richards [207]. The key postulate of the theory is that there exists a functional relationship between the traffic stream flow rate  $q$  and density  $k$  that might vary with location  $x$  but does not vary with time  $t$  (this contradicts the definition of dynamic given that within a dynamic process variables vary spatially and temporally). It should be noted that in microscopic approaches, as will be described later, the fundamental diagram varies temporally as a function of the traffic composition, thus overcoming some of the drawbacks of this approach. The fundamental hypothesis of all traffic flow theories is the existence of a site-specific unique relationship between traffic stream flow and traffic stream density, commonly known as the fundamental diagram, the traffic stream motion model, or the car-following model at the microscopic level. The assumption is that all steady-state model solutions lie on the fundamental diagram and thus are referred to as fundamental diagram approaches [106]. Given that traffic stream space-mean speed can be related to traffic stream flow and density, a unique speed-flow-density relationship (in the macroscopic approach) is derived from the fundamental diagram for each roadway segment. This relationship can also be cast at the micro-level by relating the vehicle speed to its spacing, given that vehicle spacing is the inverse of traffic stream density. Some researchers have argued that the fundamental diagram approach cannot capture the spontaneous traffic stream failure that is observed in the field and thus these researchers have proposed other theories.

One of these theories is the three-phase traffic flow theory proposed by Kerner [106,107], which attempts to explain empirical spatiotemporal features of congested patterns. The theory divides traffic into three phases: free-flow, synchronized flow, and wide moving jams. The free-flow phase is consistent with the uncongested regime on a fundamental diagram and thus is not discussed further. The *synchronized flow* phase involves continuous traffic flow with no significant stoppage. The word “flow” reflects

this feature. Within this phase there is a tendency towards synchronization of vehicle speeds and flows across the different lanes on a multilane roadway, and thus comes the name “synchronized.” This synchronization of speeds is a result of the relatively low probability of passing within this phase. The third phase, *wide moving jam*, is a phase that involves traffic jams that propagate through other states of traffic flow and through any bottleneck while maintaining the velocity of the downstream jam front. The phrase “moving jam” reflects the propagation as a whole localized structure on a road. To distinguish wide moving jams from other moving jams, which do not characteristically maintain the mean velocity of the downstream jam front, Kerner uses the term wide moving jam. Kerner indicates that if a moving jam has a width (in the longitudinal direction) considerably greater than the widths of the jam fronts, and if vehicle speeds inside the jam are zero, the jam always exhibits the characteristic feature of maintaining the velocity of the downstream jam front.

Kerner distinguishes his three-phase traffic flow theory from fundamental diagram approaches in a number of aspects. He demonstrates that the fundamental diagram approach cannot capture two key empirically observed phenomena in traffic, namely: (a) the probabilistic nature of free-flow to synchronized flow transition (flow breakdown), and (b) the spontaneous formation of general patterns (GP), which include moving and wide moving jams. Alternatively, it could be hypothesized that by modeling individual driver behavior (micro or nano modeling), capturing vehicle acceleration constraints, and introducing stochastic differences between drivers that this may be sufficient to model these two key phenomena.

### Macroscopic Modeling Approaches

In order to solve for the three traffic stream variables ( $q$ ,  $k$ , and  $u$ ) three equations are introduced. The first is the functional relationship between flow and density, or what is commonly known as the fundamental diagram. Typical functions include the Pipes triangular function (3 parameters), the Greenshields parabolic function [85] (2 parameters), or the Van Aerde [188] function (4 parameters). The second equation is the flow conservation equation (equation of continuity) that can be expressed as  $\partial k(x, t)/\partial t + \partial q(x, t)/\partial x = 0$ , considering no entering or exiting traffic. The third and final equation relates the traffic stream flow rate ( $q$ ) to the traffic stream density ( $k$ ) and space-mean speed ( $u$ ) as  $q = ku$ . The numerical solution of the KW problem involves partitioning the network into small cells of length  $\Delta x$  and discretizing time into steps of duration  $\Delta t$ . For numerical stability  $\Delta x = u\Delta t$ .

The problem is solved by stepping through time and solving for the variables in every cell using the incremental transfer (IT) principle (essentially explicit finite difference method). Extensions to the standard KW solution have introduced IT solutions for each lane along a freeway where the freeway is modeled as a set of interacting streams linked by lane changes. Lane-changing vehicles can be treated as a fluid that can accelerate instantaneously, however this approach does not capture the reduction in capacity that is associated with lane changes. Consequently, further improvements have been introduced through the use of a hybrid approach [115] that combines microscopic and macroscopic models. Specifically, slow vehicles are treated as moving bottlenecks in a single KW stream, while lane changing vehicles are modeled as discrete particles with constrained motion. The model requires identifying a lane-change intensity parameter in addition to the functional relationship parameters that were described earlier. It is not clear how such a parameter is derived. The major drawbacks of this modeling approach are that it does not account for the dynamic changes in roadway capacity (e.g. the capacity of a weaving section varies as a function of the traffic composition), it cannot capture the spontaneous traffic stream failure that is observed in the field, it cannot capture the impact of opposing flows on the traffic behavior of an opposed flow (e.g. how the capacity of an opposed left turn movement is affected by the opposing through movement), and it ignores the stochastic nature of traffic.

### Mesosopic Modeling Approaches

The mesoscopic analysis tracks individual vehicles as they travel through the network along a sequence of links that are determined by the traffic assignment. The level of tracking involves computing the vehicle's travel speed on each link based upon the density on the link together with a user specified speed/density relationship. The vehicle is then held on the link for the duration of its travel time. At the vehicle's scheduled departure time, the vehicle is allowed to exit the link if the link privileges permit it to leave; otherwise, the vehicle is held on the link until the link privileges so permit. Link exit privileges may be controlled by traffic signals at the downstream end of the link or by any queues that may be present on the lane. Queues are stored for each lane separately to account for any queue length differentials that may occur (e.g., longer queues on left turn opposed lanes). The mesoscopic analysis captures the operational level of detail (e.g., the reduction in lane capacity as a result of an opposed flow) without having to track each vehicle's instantaneous speed profile.

This means that the computational requirements for such a type of modeling are more than that required by a macroscopic analysis, but less than that required by a microscopic analysis. The INTEGRATION 1.50, DynaSMART, and DynaMIT models are examples of such modeling approaches. This approach suffers from similar drawbacks as identified in the macroscopic analysis procedures, namely an inability to capture correct spatiotemporal propagation of congestion, a failure to capture dynamic changes in capacity, a failure to capture for spontaneous breakdown in a traffic stream, and failure to capture the stochastic nature of traffic.

### Microscopic Modeling Approaches

The third approach to modeling traffic is the microscopic analysis, which tracks each vehicle as it travels through the network on a second-by-second or deci-second level of resolution using detailed car-following and lane-changing models. Microscopic simulation software use car-following models to capture the longitudinal interaction of a vehicle and its preceding vehicle in the same lane. The process of car-following is modeled as an equation of motion for steady-state conditions (also referred to as stationary conditions in some literature) plus a number of constraints that govern the behavior of vehicles while moving from one steady-state to another (decelerating and/or accelerating). The first constraint governs the vehicle acceleration behavior, which is typically a function of the vehicle dynamics. The second and final constraint ensures that vehicles maintain a safe position relative to the lead vehicle in order to ensure asymptotic stability within the traffic stream. A more detailed description of the longitudinal modeling of vehicle motion is provided by Rakha et al. [194]. While there are a number of commercially available software packages that simulate traffic microscopically (CORSIM, Paramics, FREEVU, VISSIM, AIMSUN2, and INTEGRATION), these approaches are computationally intensive and cannot run in real-time. The INTEGRATION software has been able to capture the stochastic nature of traffic stream capacity by randomly modeling vehicle-specific car-following models. Furthermore, the model captures the capacity loss associated with recovery from breakdown through the vehicle acceleration constraints. The stochastic nature of car-following and lane-changing behavior may allow the model to capture spontaneous breakdown in traffic stream flow.

The amount of computation and memory necessary for simulating a large transportation network at a level of detail down to an individual traveler and an individual vehicle may be extensive. Hence a microscopic massively

parallel simulation approach entitled “cellular automata” (CA) is sometimes proposed to simulate large networks. The cellular automata approach essentially divides every link on the network into a finite number of cells. At a one second time step, each of these “cells” is scanned for a vehicle presence. If a vehicle is present, the vehicle position is advanced, either within the cell or to another cell, using a simple rule set [148,149,150]. The rule set is made simple to increase the computational speed necessary for a large simulation. Vehicles are moved from one grid cell to another based on the available gaps ahead, with modifications to support lane changing and plan following, until they reach the end of the grid. There, they wait for an acceptable gap in the traffic or for protection at a signal before moving through the intersection onto the next grid. This continues until each vehicle reaches its destination, where it is removed from the grid. Reducing the size of the “cell”, expanding the rule set, and adding vehicle attributes increases the fidelity of the simulator, but also greatly affects the computational speed. The size of 7.5 meters in length and a traffic lane in width is often chosen as a default size for the “cell” as was applied with the TRANSIMS software [149]. This approach suffers from a number of drawbacks including the inability to capture the dynamic nature of roadway capacity, the inability to capture spontaneous breakdown of traffic stream, and the inability to capture opposing flow impacts on opposed flow saturation flow rates (e.g. the impact an opposing through movement flow has on the capacity of a permitted left turn movement that has to find a gap in this opposing flow).

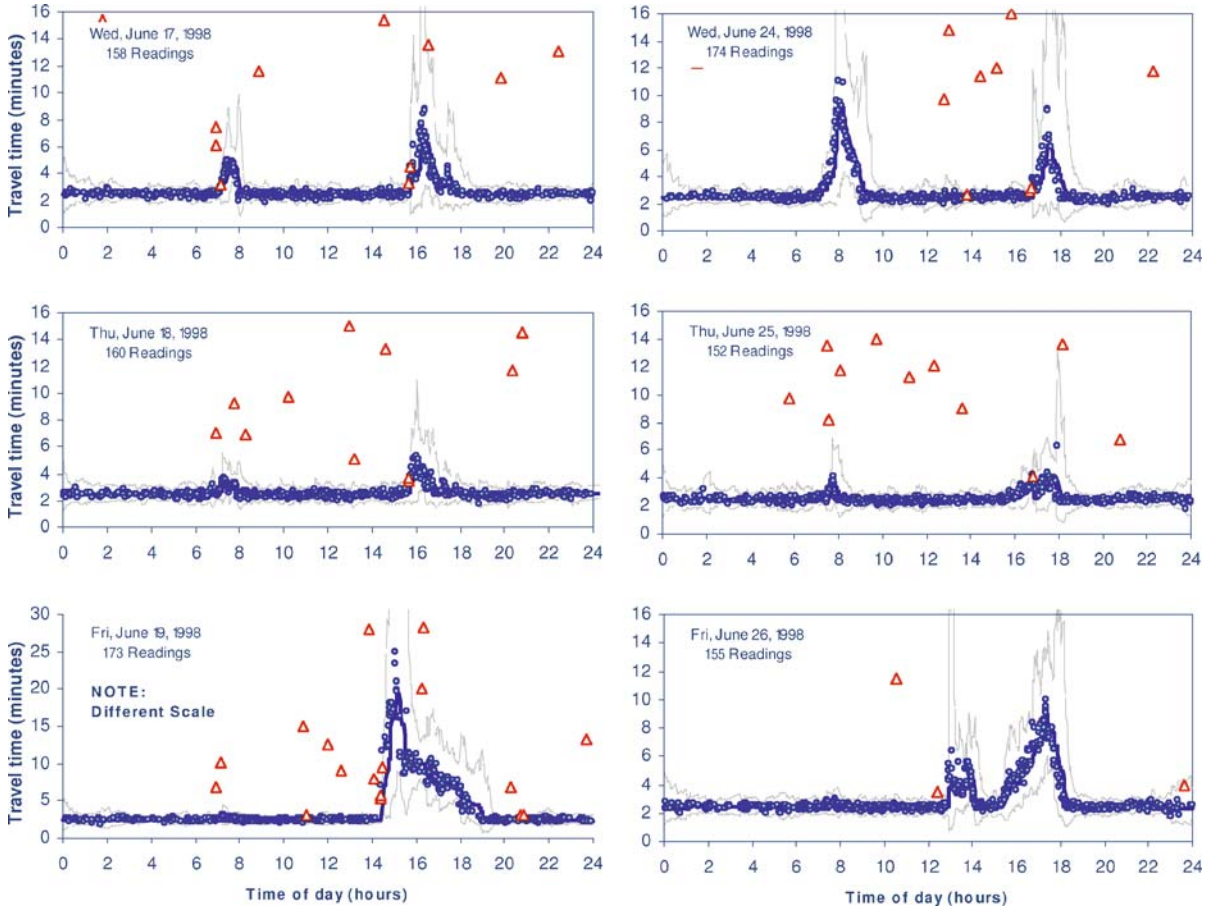
### Dynamic Travel Time Estimation

As was demonstrated earlier in the paper, the DTA requires arc (link) travel times in order to compute minimum paths. There are several systems commercially available that are capable of estimating real-time travel times. These can be broadly classified into spot speed measurement systems, spatial travel time systems, and probe vehicle technologies. Spot speed measurement systems, specifically inductance loop detectors, have been the main source of real-time traffic information for the past two decades. Other technologies for measuring spot speeds have also evolved, such as infrared and radar technologies. Regardless of the technology, the spot measurement approaches only measure traffic stream speeds over a short roadway segment at fixed locations along a roadway. These spot speed measurements are used to compute spatial travel times over an entire trip using space-mean-speed estimates. In addition, new approaches that match vehicles

based on their lengths have also been developed [54,55,56,57]. However, these approaches require raw loop detector data as opposed to typical 20- or 30-second aggregated data. Alternatively, spatial travel time measurement systems use fixed location equipment to identify and track a subset of vehicles in the traffic stream. By matching the unique vehicle identifications at different reader locations, spatial estimates of travel times can be computed. Typical technologies include AVI and license-plate video detection systems. Finally, probe vehicle technologies track a sample of probe vehicles on a second-by-second basis as they travel within a transportation network. These emerging technologies include cellular geo-location, Global Positioning Systems (GPS), and Automatic Vehicle Location (AVL) systems.

Traffic routing strategies under recurring and non-recurring strategies should be based on forecasting of future traffic conditions rather than historical and/or current conditions. In general the traffic prediction approaches can be categorized into three broad areas: (i) statistical models, (ii) macroscopic models, and (iii) route choice models based on dynamic traffic assignment [24,25,26,28,168]. Time series models have been used in traffic forecasting mainly because of their strong potential for online implementation. Early examples of such approaches include [7] and more recently [120] and [97]. In addition, researchers have applied Artificial Neural Network (ANN) techniques for the prediction of roadway travel times [6,159,160,161,162,165]. These studies demonstrated that prediction errors were affected by a number of variables pertinent to traffic flow prediction such as spatial coverage of surveillance instrumentation, the extent of the loop-back interval, data resolution, and data accuracy.

An earlier publication [70] developed a low-pass adaptive filtering algorithm for predicting average roadway travel times using Automatic Vehicle Identification (AVI) data. The algorithm is unique in three aspects. First, it is designed to handle both stable (constant mean) and unstable (varying mean) traffic conditions. Second, the algorithm can be successfully applied for low levels of market penetration (less than 1 %). Third, the algorithm works for both freeway and signalized arterial roadways. The proposed algorithm utilizes a robust data-filtering procedure that identifies valid data within a dynamically varying validity window. The size of the validity window varies as a function of the number of observations within the current sampling interval, the number of observations in the previous intervals, and the number of consecutive observations outside the validity window. Applications of the algorithm to two AVI datasets from San Antonio, one from



**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 6**  
**Sample Application of AVI Travel Time Estimation Algorithm [70]**

a freeway link and the other from an arterial link, demonstrated the ability of the proposed algorithm to efficiently track typical variations in average link travel times while suppressing high frequency noise signals.

Within the filtering algorithm, the expected average travel time and travel time variance for a given sampling interval are computed using a moving average (MA) technique. As shown in Eqs. (17) and (18), it estimates the expected average travel time and expected travel time variance within a given sampling interval based on the set of valid travel time observations in the previous sampling interval and the corresponding previously smoothed moving average value using an adaptive exponential smoothing technique. In both equations, calculations of the smoothed average travel time and travel time variance are made using a lognormal distribution to reflect the fact that the distribution is right skewed (skewed towards longer travel times). Field data from the San Antonio AVI system

demonstrated that this assumption is reasonable.

$$\tilde{t}_{i,k} = \begin{cases} e^{\alpha \cdot \ln t_{i,k-1} + (1-\alpha) \cdot \ln \tilde{t}_{i,k-1}}, & \text{if } n_{i,k} > 0, \\ \tilde{t}_{i,k-1}, & \text{if } n_{i,k} = 0, \end{cases} \quad (17)$$

$$\tilde{\sigma}_{i,k}^2 = \begin{cases} \alpha \cdot \sigma_{i,k-1}^2 + (1-\alpha) \cdot \tilde{\sigma}_{i,k-1}^2, & \text{if } n_{i,k} > 1, \\ \tilde{\sigma}_{i,k-1}^2, & \text{if } n_{i,k} \leq 1. \end{cases} \quad (18)$$

It should be noted that  $t_{i,k}$  is the observed average travel time along link  $i$  within the  $k$ th sampling interval ( $s$ ),  $\tilde{t}_{i,k}$  is the smoothed average travel time along link  $i$  in the  $k$ th sampling interval ( $s$ ),  $\sigma_{i,k}^2$  is the variance of the observed travel times relative to the observed average travel time in the  $k$ th sampling interval ( $s^2$ ),  $\tilde{\sigma}_{i,k}^2$  is the variance of the observed travel times relative to the smoothed travel time in the  $k$ th sampling interval ( $s^2$ ),  $n_{i,k}$  is the number of valid travel time readings on link  $i$  in the  $k$ th sampling interval, and  $\alpha = 1 - (1 - \beta)^{n_{i,k}}$  for all  $i$  and  $k$  is an exponential

smoothing factor that varies as a function of the number of observations  $n_{i,k}$  within the sampling interval, where  $\beta$  is a constant that varies between 0 and 1.

Figure 6 shows an example application of the algorithm using AVI data along I-35 in San Antonio, TX. The figure illustrates the average travel time estimate (thick line), the validity window bounds (thin lines), what are considered to be valid data (circular), and the observations that are considered to be outliers (triangles). The figure clearly illustrates the effectiveness of the algorithm in estimating roadway travel times for low levels of market penetration of AVI tags.

Once link travel times have been estimated, the expected trip or path travel times can be computed by summing the relevant smoothed link travel times. In addition, the trip travel time reliability, which is the probability that a trip can reach its destination within a given period at a given time of day, can be computed for use in traffic routing. Travel time reliability is a measure of the stability of travel time, and therefore is subject to fluctuations in flow [21]. Typically, when flow fluctuations are large, travel time is often longer than expected. As levels of congestion in transportation networks grow, generally the stability of travel time will have greater significance to transportation users. The trip travel time reliability can be computed as the probability  $P(T \leq t)$  that the trip travel time ( $T$ ) is less than some arbitrary travel time ( $t$ ), using the cumulative distribution function estimated from an analysis of AVI field data. The current state-of-the-art in predicting trip travel time variability is to assume that the travel times on all the links along a path are generated by statistically independent normal distributions. Consequently, the trip variance can be computed as the summation of the link travel time variances for all links along a path. As part of the proposed research effort, different statistical techniques (not assuming independent normal variates) will be devised to estimate the trip travel time variance, as discussed in the Proposed Research Tasks section. These techniques will be tested using data from the video detection system that is currently implemented in the Blacksburg Area.

In addition, research has been conducted to estimate the optimum locations of surveillance equipment for the estimation of travel times. Specifically, an earlier publication developed an algorithm for optimally locating Automatic Vehicle Identification tag readers by maximizing the benefit that would accrue from measuring travel times on a transportation network [219]. The problem is formulated as a quadratic 0–1 optimization problem where the objective function parameters represent benefit factors that capture the relevance of measuring travel times as re-

flected by the demand and travel time variability along specified trips. An optimization approach based on the Reformulation-Linearization Technique coupled with semi-definite programming concepts was designed to solve the formulated reader location problem. Alternatively, a Genetic Algorithm (GA) approach was developed to optimally locate the AVI readers [13].

### Dynamic or Time-Dependent Origin-Destination Estimation

As was demonstrated earlier the Bechmann UE and the SO formulations do not provide unique path flows and thus a synthetic O-D estimator is required to estimate the path flows from the unique link flows. The techniques used to estimate O-D demands can be categorized based on different factors, as will be discussed in detail. The first categorization, of the available O-D estimation techniques, relates to whether the O-D's to be estimated are static, and apply to only one observation time period, or whether estimates are required for a series of linked dynamic time periods. The next breakdown relates to whether the estimation is based on information about the magnitude of trip ends only, or whether information is available on additional links along the route of each trip. The former problem is commonly referred to as the trip distribution problem in demand forecasting, while the latter problem is commonly referred to as the synthetic O-D generation problem. Both problems are discussed, but this section will focus on the latter synthetic O-D generation problem. The former is viewed simply as a simpler subset of the latter.

Within the overall static synthetic O-D generation problem, there are two main flavors. The first exists when the routes that vehicles take through the network are known a priori. The second arises when these routes need to be estimated concurrently while the O-D is being estimated. A priori knowledge of routes can arise automatically when there is only one feasible route between each O-D pair, or when observed traffic volumes are only provided for the zone connectors at the origins and destinations in the network. The first condition is common when O-D's are estimated for a single intersection or arterial, or a single interchange or freeway. The second condition is the default for any trip distribution analysis. This section will initially focus on situations where the routes are known a priori. Subsequently, a solution to the more general problem which involves an iterative use of the solution approach when routes are not known a priori will be discussed.

Within the static/dynamic synthetic O-D generation problem, for scenarios where routes are known a priori



(or are assumed to be known a priori) there exist two sub-problems. The first of these problems relates to situations where flow continuity exists at each node in the network, and multiple O-D matrices can be shown to match these observed flows exactly. In this case, the most likely of these multiple O-D matrices needs to be identified. The second sub-problem relates to situations where flow continuity does not exist at either the node level or at the network level. In other words, the observed traffic flows are such that no matrix exists that will match the observed flows exactly. In this case, Van Aerde et al. [235] introduced a new set of complementary link flows that maintain flow continuity by introducing minimum alterations to the observed flows to solve the maximum likelihood problem.

The static synthetic O-D generation problem, for scenarios where flow continuity does exist, can be formulated in two different ways [242,250]. The first of these considers that the fundamental unit of measure is the individual trip, while the second considers that the fundamental unit of measure is the observation of a single vehicle on a particular link. The availability of a seed or target O-D matrix is implicit in the latter formulation, but can be dropped in the former formulation, as was demonstrated in an earlier publication [235]. However, only when a seed matrix is properly included in the former formulation is it guaranteed to yield consistent results with the latter formulation. In other words, the absence of a seed matrix in the trip based formulation can be shown to yield inconsistent results, at least for some networks in which the multiple solutions result in a different number of total trips.

An additional and related attribute, of the trip-based formulation of maximum likelihood, is the presence of a term in the objective function that is based on the total number of trips in the network. This term, referred to as  $T$ , is often dropped in some approximations. However, it can be shown that dropping this term can yield solutions that represent only a very poor approximation to the true solution [186]. In contrast, approximations involve the use of Stirling's approximation, for representing the logarithm of factorials, were shown to yield consistently very good approximations [186]. This finding is critical because use of Stirling's approximation is critical to being able to compute the derivatives that are needed to numerically solve the problem (it is difficult to take derivatives of terms that include factorials).

Other examples from literature include the works of Cremer and Keller [59], Cascetta et al. [43], Wu and Chang [252], Sherali et al. [218], Ashok and Ben-Akiva [18], Hu et al. [95], Chang and Tao [51], Pavlis and Papageorgiou [165], Peeta and Zhou [178,179], Peeta and Yang [174,175], Yang [258], Peeta and Bulusu [167].

### Comparison of Synthetic O-D and Trip Distribution Formulations

Within the four-step planning process O-D matrices are estimated in the trip distribution step. Several methods are used for trip distribution including the gravity, growth factor, and intervening opportunities models. The gravity model is most utilized because it uses the attributes of the transportation system and land-use characteristics and has been calibrated and applied extensively to the modeling of numerous urban areas. The model assumes that the number of trips between two zones  $i$  and  $j$  ( $T_{ij}$ ) is directly proportional to the number of trip productions from the origin zone ( $P_i$ ) and the number of attractions to the destination zone ( $A_j$ ) and inversely proportional to a function of travel time between the two zones ( $F_{ij}$ ) as

$$T_{ij} = P_i \left( \frac{A_j F_{ij} K_{ij}}{\sum_j A_j F_{ij} K_{ij}} \right). \quad (19)$$

Typically the values of trip productions and attractions are computed based on trip generation procedures. The values of  $F_{ij}$  are computed using a calibration procedure that involves matching modeled and field trip length distributions. The socio-economic adjustment factors ( $K_{ij}$ ) values are used when the estimated trip interchange must be adjusted to ensure that it agrees with observed trips by attempting to account for factors other than travel time. The values of  $K$  are determined in the calibration process, but considered judiciously when a zone is considered to possess unique characteristics.

Because the O-D problem is under-specified, multiple O-D demands can generate identical link flows. For example, if one attempts to estimate an O-D matrix for a 100 zone network with, say 1000 links, one has more unknowns to solve for than there are constraints. In the case of the trip distribution process, there are 100x100 O-D cells to estimate, and only 2x100 trip end constraints. In the case of the synthetic O-D generation process, there are again 100x100 O-D cells to estimate, and only 1000 link constraints. Given the possibility of multiple solutions, both the trip distribution process and the synthetic O-D generation process invoke additional considerations to select a preferred matrix from among the multiple solutions.

In the case of synthetic O-D generation, the desire is to select from among all of the possible solutions, the most likely. This approach requires one to define a measure of the likelihood of each matrix. In general, there are two approaches to establish the likelihood of a matrix. One of them treats the trip as the basic unit of observa-

tion, while the other considers a volume count as the basic unit of observation. The first approach will be discussed in detail, while the interested reader might refer to the literature [235] for a more detailed description of the various formulations. It suffices to indicate that for any matrix with cells  $T_{ij}$ , the likelihood of the matrix can be estimated using a function  $L = f(T_{ij}, t_{ij})$ , where  $t_{ij}$  represents prior information. The prior information is often referred to as a seed matrix, and can be derived from a previous study or survey. In the absence of such prior information, all of the cells in this prior matrix should be set to a uniform set of values.

In the case of the trip distribution process, the additional information that is added is some form of impedance. For example, the original gravity model considered that the likelihood of trips between two zones was proportional to the inverse of the square of the distance between the two zones. Since that time, many more sophisticated forms of impedance have been considered, but for the purposes of this discussion, all of these variations can be generalized as being of the form  $F_{ij}$ , where  $F_{ij} = f(c_{ij})$  or the generalized cost of inter-zonal travel. What is less obvious, however, is the fact that the use of this set of impedance factors  $F_{ij}$ , is essentially equivalent to the use of a seed matrix  $t_{ij}$ .

Van Aerde et al. [235] demonstrated that solving the trip distribution problem, using zonal trip productions and attractions as constraints, together with a trip impedance matrix, is essentially the same as solving the synthetic O-D problem using zone connector in and out flows as constraints, and utilizing a seed matrix based on Eq. (20).

$$t_{ij} = T \frac{F_{ij}}{\sum_{ij} F_{ij}} \quad (20)$$

### Static Formulations

Entropy maximization and information minimization techniques have been used to solve a number of transportation problems [251]. The application of the entropy maximization principles to the static O-D estimation problem was first introduced by Willumsen [242,250]. Willumsen demonstrated that by maximizing the entropy, the most likely trip matrix could be estimated subject to a set of constraints.

The trip-based approach to defining maximum likelihood considers that the overall trip matrix is made up of uniquely identifiable individual trip makers. The most likely matrix is one where the likelihood function is maxi-

mized as

$$\max. \quad Z_1(T_{ij}) = \frac{T!}{\prod_{ij} (T_{ij}!)} \quad (21)$$

The above formulation does not take into account any prior information, from for example a previous survey. While the seed matrix does not necessarily have to satisfy the observed link flows, the seed matrix can be utilized to expand the maximum likelihood function to

$$\max. \quad Z_2(T_{ij}, t_{ij}) = \frac{T!}{\prod_{ij} (T_{ij}!)} \prod_{ij} \left( \frac{t_{ij}}{\sum_{ij} t_{ij}} \right)^{T_{ij}} \quad (22)$$

It can be noted that the likelihood of an individual trip from  $i$  to  $j$  is  $t_{ij} / \sum_{ij} t_{ij}$ , based on the above seed matrix. Consequently, the probability of  $T_{ij}$  trips being drawn is  $(t_{ij} / \sum_{ij} t_{ij})^{T_{ij}}$ .

The above formulations of objective functions for expressing likelihood require additional constraints in order to be complete [242,250]. The simplest of these constraints indicate that the sum of all trips crossing a given link must be equal to the link flow on that link as

$$V_a = \sum_{ij} T_{ij} p_{ij}^a \quad \forall a. \quad (23)$$

As will be shown later, the simplest mechanism, for including the above constraints in the earlier objective functions, is to utilize Lagrange multipliers. These multipliers permit an objective function with equality constraints to be transformed into an equivalent unconstrained objective function.

This simple set of equality constraints, while making the formulation complete, may at times also render the problem infeasible. A more general formulation that was proposed in the literature [235] is to minimize the link flow error, rather than eliminate the error. In other words, rather than finding the most likely O-D that exactly replicates the observed link flows, the problem is re-formulated as finding the most likely O-D matrix from among all of those that come equally close to matching the link flows. One expression that is proposed to capture the error to be minimized is shown in Eq. (24), and is subject to the flow continuity constraints in Eq. (25). The constraints in Eq. (25) can be introduced in Eq. (25) to yield an unconstrained objective function, yielding a set of complementary link flows  $V'_a$ . These complementary flows are those which deviate the least from the observed link flows, while satisfying link flow continuity. Given that these complementary link flows do satisfy flow continuity, they can now

be added in as rigid equality constraints to the objective function (21) or (22), and be guaranteed to yield a feasible solution.

$$\min. \quad Z_3(T_{ij}) = \sum_a (V_a - V'_a)^2 \quad (24)$$

$$V'_a = \sum_{ij} T_{ij} p_{ij}^a \quad \forall a. \quad (25)$$

Alternatively, one can incorporate Eq. (25) into Eq. (24) to yield

$$\min. \quad Z_4(T_{ij}) = \sum_a \left( V_a - \sum_{ij} T_{ij} p_{ij}^a \right)^2. \quad (26)$$

This equation should be minimized concurrently to maximizing the objective function (21) or (22). Unfortunately, it is not easy to combine one expression that desires to maximize likelihood with another that desires to minimize link flow error, as a Lagrangian can only add equality constraints to a constrained objective function. Van Aerde et al. proposed a solution to this problem which involves taking the partial derivatives of Eq. (26) with respect to each of the trip cells that are to be estimated as

$$\frac{\partial}{\partial T_{ij}} Z_4(T_{ij}) = \frac{\partial}{\partial T_{ij}} \sum_a \left( V_a - \sum_{ij} T_{ij} p_{ij}^a \right)^2 = 0 \quad \forall i, j. \quad (27)$$

$$0 = 2 \left( \sum_a (V_a \cdot p_{ij}^a) - \left( \sum_a p_{ij}^a \left( \sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \quad \forall i, j \quad (28)$$

This yields as many equations as there are trip cells, as shown in Eq. (27). Furthermore, setting these derivatives equal to 0 is equivalent to minimizing Eq. (27). However, while Eq. (24) could not be added to the maximum likelihood objective function, the equalities in Eq. (28) can. This produces an unconstrained objective function that always yields a feasible solution computed as

$$\max. \quad \frac{T!}{\prod_{ij} T_{ij}} \prod_{ij} \left( \frac{t_{ij}}{t} \right)^{T_{ij}} - \sum_{ij} \left( \lambda_{ij} \cdot 2 \left( \sum_a \left( V_a \cdot p_{ij}^a \right) - \left( \sum_a p_{ij}^a \left( \sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right), \quad (29)$$

where:  $T = \sum_{ij} T_{ij}$  and  $t = \sum_{ij} t_{ij}$ .

The net result, of the above process, is to suggest that most synthetic O-D generation problems consist of two sub-problems. One of these involves finding a new set of complementary link flows that do produce link flow continuity, at which point the maximum likelihood problem can be solved as before. Alternatively, one can compute the partial derivatives, that will yield link flow continuity, while deviating by the least amount from the observed link flows, and then utilize them directly in the maximum likelihood formulation using Lagrange multipliers. Both solutions can be shown to yield identical results.

A first challenge with maximizing Eq. (29) is that it yields very large numbers that are difficult to work with. Further more, as it is common to maximize objective functions by taking their derivatives, and as it is more difficult to contemplate the derivative of a discontinuous expression, such as those including factorials, a simple approximation is made. This approximation involves taking the natural logarithm of either objective function Eq. (21) or (22). Taking the natural logarithm of the objective function both makes the output easier to handle and permits the use of Stirling's approximation as a convenient continuous equivalent to the term  $\ln(x!)$  as

$$\ln(T!) = T \ln T - T \quad (30)$$

The resulting converted objective function using the Stirling approximation on the original objective function of Eq. (22) is computed as

$$\begin{aligned} \max. \quad & T \ln \left( \frac{T}{t} \right) - T - \sum_{ij} \left( T_{ij} \ln \left( \frac{T_{ij}}{t_{ij}} \right) - T_{ij} \right) \\ & - \sum_{ij} \left( \lambda_{ij} \cdot 2 \left( \sum_a \left( V_a \cdot p_{ij}^a \right) - \left( \sum_a p_{ij}^a \left( \sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right). \quad (31) \end{aligned}$$

Expanding and simplifying the various terms we derive

$$\begin{aligned} & T \ln \left( \frac{T}{t} \right) - T - \sum_{ij} \left( T_{ij} \ln \left( \frac{T_{ij}}{t_{ij}} \right) - T_{ij} \right) \\ & = T \ln \left( \frac{T}{t} \right) - \sum_{ij} T_{ij} \ln \left( \frac{T_{ij}}{t_{ij}} \right). \quad (32) \end{aligned}$$

When Eq. (32) is augmented with the previously mentioned partial derivatives that minimize the link flow error

we derive

$$\begin{aligned} \max. \quad & T \ln \left( \frac{T}{t} \right) - \sum_{ij} T_{ij} \ln \left( \frac{T_{ij}}{t_{ij}} \right) \\ & - \sum_{ij} \left( \lambda_{ij} 2 \left( \sum_a \left( V_a \cdot p_{ij}^a \right) \right. \right. \\ & \left. \left. - \left( \sum_a p_{ij}^a \left( \sum_{xy} T_{xy} p_{xy}^a \right) \right) \right) \right). \quad (33) \end{aligned}$$

This equation, when solved, yields the most likely O-D matrix of all of those matrices that come equally close to matching the observed link flows.

It should be noted that the objective function of Eq. (33) is composed of two components. The first being the error between the field observed flows and the flows that satisfy flow continuity with minimum difference from observed flows. The second component represents the likelihood of an O-D matrix table. The objective is to find the O-D matrix with the maximum likelihood. In the case that the seed matrix is the optimum matrix the likelihood component resolves to zero.

### Dynamic Formulations

The above formulations assume that the vehicles are assigned to all links simultaneously (i. e. a vehicle is present on all links along its path simultaneously). In order to address the dynamic nature of traffic, the analysis period can be divided into equally spaced time slices. Origin-destination demands are then indexed by the time slice they depart and the time slice they are observed on a link, as

$$\begin{aligned} \max. \quad & T_r \ln \left( \frac{T_r}{t_r} \right) - \sum_{rij} T_{rij} \ln \left( \frac{T_{rij}}{t_{rij}} \right) \\ & - \sum_{rij} \left( \lambda_{rij} 2 \left( \sum_{sa} \left( V_{sa} \cdot p_{rij}^{sa} \right) \right. \right. \\ & \left. \left. - \left( \sum_{sa} p_{rij}^{sa} \left( \sum_{rxy} T_{rxy} p_{rxy}^{sa} \right) \right) \right) \right). \quad (34) \end{aligned}$$

Where  $T_r$  is the total demand departing during time-slice  $r$ ,  $t_r$  is the total seed matrix demand departing during time-slice  $r$ ,  $T_{rij}$  is the traffic demand departing during time-slice  $r$  traveling between origin  $i$  and destination  $j$ ,  $t_{rij}$  is the seed traffic demand departing during time-slice  $r$  traveling between origin  $i$  and destination  $j$ ,  $\lambda_{rij}$  is the Lagrange multiplier for departure time-slice, origin, and destination combination  $rij$ ,  $V_{sa}$  is the observed volume on link  $a$  during time slice  $s$ , and  $p_{rij}^{sa}$  is the probability of

a demand between origin  $i$  and destination  $j$  during time-slice  $r$  is observed on link  $a$  during time-slice  $s$ . The solution of Eq. (34) is computationally extensive and has been demonstrated to not produce significantly better results than generating time-dependent O-D demands, as will be discussed.

Alternatively, the more common approach is to generate time-dependent O-D demands by solving Eq. (33) for each time-slice independently assuming that O-D demands can travel from the origin to destination zone within a time-slice (i. e. the trip travel time is less than the time-slice duration) without considering the interaction between time slices. This approach is computationally simpler and easier to implement and thus will be discussed in more detail. The formulation can be written as

$$\begin{aligned} \max. \quad & T_r \ln \left( \frac{T_r}{t_r} \right) - \sum_{ij} T_{rij} \ln \left( \frac{T_{rij}}{t_{rij}} \right) \\ & - \sum_{ij} \left( \lambda_{rij} \cdot 2 \left( \sum_a \left( V_a \cdot p_{rij}^a \right) \right. \right. \\ & \left. \left. - \left( \sum_a p_{rij}^a \left( \sum_{xy} T_{rxy} p_{rxy}^a \right) \right) \right) \right) \quad \forall r. \quad (35) \end{aligned}$$

Here the  $T_{rij}$  are solved for independent of other time-slices. It should be noted, that the approach ignores the interaction of demands across various time-slices which is a valid assumption if the network is not over-saturated. However, if the network is oversaturated the assumption of time slice independence may not be valid. The duration of the time-slice should be selected such that steady-state conditions are achieved within a time-slice.

### Solution Algorithms

The solution of the set of equations presented in (35) is hard given that the objective function is nonconvex and that in many cases the  $p_{rij}^a$  are not available and thus the problem becomes to solve for  $T_{rij}$  and  $p_{rij}^a$  that maximize the objective function.

Here we present a numerical heuristic that solves the above formulation for large networks when the number of equations and unknowns becomes extremely computationally intensive. This special purpose equation solver has been developed and implemented in the QUEENSOD software. This solver fully optimizes the objective function of Eq. (35). The software has been shown to produce errors less than 1% for the range of values and derivatives being typically considered in the problem. A sample application of the QUEENSOD software is presented later in the paper, however, initially the heuristic approach is described.

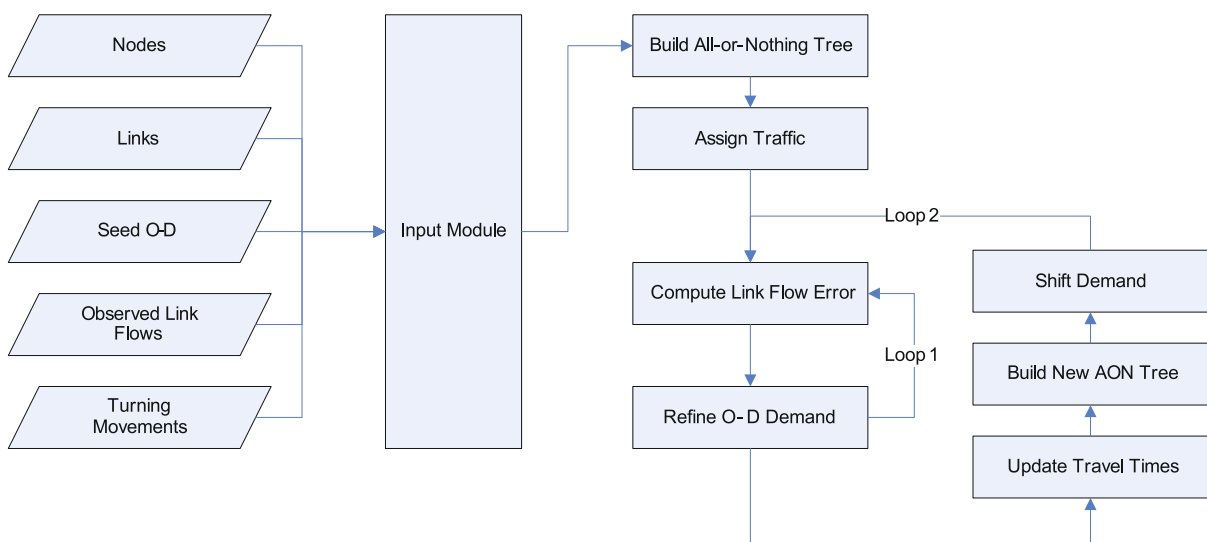
The numerical solution begins by building a minimum path tree and performing an all-or-nothing traffic assignment of the seed matrix, as illustrated in Fig. 7. A relative or absolute link flow error is computed depending on user input. Using the link flow errors O-D adjustment factors are computed and utilized to modify the seed O-D matrix. The adjustment of the O-D matrix continues until one of two criteria are met, namely the change in O-D error reaches a user-specified minimum or the number of iterations criterion is met. If additional trees are to be considered, the model builds a new set of minimum path trees (loop 2) and shifts traffic gradually to the second minimum path tree. The minimum objective function for two trees is computed in a similar fashion as described for the single tree scenario. The process of building trees and finding the optimum solution continues until all possible trees have been explored. The proposed numerical solution ensures that in the case that the seed matrix is optimum no changes are made to the matrix. In addition, the use of the seed matrix as a starting point for the search algorithm ensures that the optimum solution resembles the seed matrix as closely as possible while minimizing the link flow error. In other words, the seed matrix biases the solution towards the seed matrix.

In order to demonstrate the applicability of the QUEENSOD software, a sample application to a 3500-link network of the Bellevue area in Seattle is presented. Other applications of the QUEENSOD software are described in detail in the literature [71,196]. The O-D demand for the Bellevue network was calibrated to AM peak Single

Occupancy Vehicle (SOV) and High Occupancy Vehicle (HOV) flows. The seed matrix was created using the standard four-step transportation planning process by applying the EMME/2 model. The Seattle network was converted from EMME/2 format to INTEGRATION format.

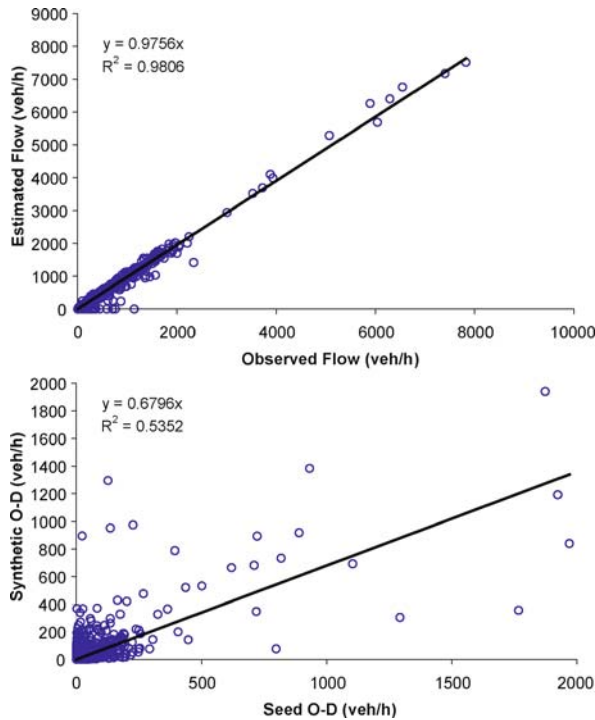
The calibration of the O-D demand to tube and turning movement counts was conducted using the QUEENSOD software using the planning trip distribution O-D matrix as the seed solution. The calibration resulted in a high level of consistency between estimated and field observed link flow counts (coefficient of determination of 0.98 between the estimated and observed flows), as illustrated in Fig. 8. Figure 8 demonstrates that in calibrating the O-D matrix to observed traffic counts, the trip distribution O-D matrix (seed matrix) was modified significantly (coefficient of determination of 0.56 between trip distribution and synthetic O-D matrix). Consequently, it is evident that a modification of the trip distribution matrix was required in order to better match observed link and turning movement counts. It should be noted however, that the total number of trips was increased by only 4 % as a result of the synthetic O-D calibration effort. Consequently, the illustrated calibration effort resulted in a significant modification of the trip table with minor modification to the total number of trips.

In addition to the above mentioned research, a significant number of problem formulations and applications have been documented in the literature. To name a few (in chronological order) Cascetta et al. [43] tested a method based on two generalized least squares estimators on the



**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 7**  
QueensOD Heuristic O-D Estimation Approach (Synthetic O-D Estimator)





**Traffic Networks: Dynamic Traffic Routing, Assignment, and Assessment, Figure 8**

**Example Application of QUEENSOD to the Bellevue Network in Seattle**

Brescia-Verona-Vicenza-Padua motorway in Italy. They found that the accuracy of their model depended heavily on the number of links with observed traffic counts. Van Aerde et al. [238] introduced the QUEENSOD method and demonstrated its applicability on a 35-km section of Highway 401 in Toronto, Canada. Ashok [16] evaluated the use of a Kalman filtering-based method, which was first presented by Okutani [158] and estimates unobserved link traffic counts from observed link traffic counts. The method used was formulated by Ashok and Ben Akiva [17] and Ashok [16] and was evaluated using actual data from the Massachusetts Turnpike, Massachusetts, a stretch of I-880 near Hayward, California and a freeway encircling the city of Amsterdam, Netherlands. Later, Hellinga and Van Aerde [93] compared a least square error model and a least relative error model on a 35-km section of Highway 401 in Toronto, Canada. Zhou and Sachse [260] compared the use of three different O-D estimators and on a motorway network in Europe. They concluded that the models, although characterized by different computational loads, produced satisfactory results. They also commented on the need to decide on locations of detectors and aggre-

gation time intervals. Van Der Zijpp and Romph [239] experimented their model on the Amsterdam Beltway. They tested their model using two different days worth of data and compared their model results with real and historical average data. While their model performed better in cases of accidents, the historical average data did, at least as good, in normal traffic. They stressed on the importance of correct modeling of the network and traffic flow characteristics for the production of good results. Kim et al. [112] introduced a genetic algorithm based method to overcome the shortcoming of the bi-level programming method when there is a significant difference between target and true O-D matrices. They tested their model on a small network of 9 nodes. Bierlaire and Crittin [27] formulated a least-square based method to overcome some of the shortcomings of the Kalman filter approach. They tested their method on a simple network as well as two real networks: a medium scale network, Central Artery Network, Boston, MA, and a large scale network, Irvine Network, Irvine, CA. Yun and Park developed a genetic algorithm based method with the purpose of solving dynamic O-D matrices for large networks. They compared their model's results with the results of QUEENSOD, and they tested their method on the City of Hampton network using the PARAMICS microscopic traffic simulation software. Nie et al. [156] developed a formulation that incorporates a decoupled path flow estimator in a generalized least squares framework with the objective of developing an efficient, simplified solution algorithm for realistic size networks. They tested their method on a small (9-node) and mid-size (100 nodes) network. Zhou and Mahmassani [259] developed a multi-objective optimization framework for the estimation of the O-D matrices using automatic vehicle identification data. They tested their method on a simplified Irvine testbed network (31 nodes). Finally, Castillo et al. [48] developed a method for the reconstruction and estimation of the trip matrix and path flows based on plate scanning and link observations. They tested their method on the Nguyen-Dupius Network, and concluded the superiority of plate scanning on link counts.

It should be noted at this point that the O-D estimation formulations and techniques that were presented and described in this section are heuristics and thus there is no mathematical proof that the algorithms converge to the unique optimum solution either in the static or dynamic context. While we have demonstrated that the solution matches the observed link flows for complex networks (Fig. 8, unfortunately the actual O-D demand is typically not available for real-life applications and thus it is not possible to measure how good the solution compares to the unique optimum O-D matrix.

## Dynamic Estimation of Measures of Effectiveness

Dynamic assessment of traffic network performance requires the estimation of various measures of effectiveness in a dynamic context. This section provides a brief overview of the procedures for estimating delay, vehicle stops, and vehicle energy consumption and emissions.

### Estimation of Delay

A key parameter in the dynamic assessment of traffic networks is the estimation of vehicle delay. The computation of delay requires the computation of travel times. Significant research has been conducted to develop analytical models for estimating delay especially at signalized intersections. Examples of such research efforts are provided for the interested reader [10,32,33,44,46,49,58,60,61,63,65,66,74,76,77,79,86,114,116,124,154,212,213,230].

Roadway travel times can be computed for any given vehicle by providing that vehicle with a *time card* upon its entry to any roadway or link. Subsequently, this *time card* is retrieved when the vehicle leaves the roadway. The difference between these entry and exit times provides a direct measure of the roadway travel time experienced by each vehicle. The delay can then be computed as the difference between the actual and free-flow travel time.

Alternatively, vehicle delay can be computed microscopically every deci-second as the difference in travel time between travel at the vehicle's instantaneous speed and travel at free-flow speed, as

$$d(t_i) = \Delta t \left( 1 - \frac{u(t_i)}{u_f} \right) \quad \forall i. \quad (36)$$

The summation of these instantaneous delay estimates over the entire trip provides an estimate of the total delay. This model has been validated against analytical time-dependent queuing models, shockwave analysis, the Canadian Capacity Guide, the Highway Capacity Manual (HCM), and the Australian Capacity Guide procedures [71]. The procedure has also been incorporated in the INTEGRATION traffic simulation software [233,234] and utilized with second-by-second Global Positioning System (GPS) data [72,190].

### Estimation of Vehicle Stops

Numerous researchers have dealt with the problem of estimating vehicle stops especially at signalized intersections. An important early contribution is attributed to Webster [247], who generated stop and delay relationships by simulating uniform traffic flows on a single-lane approach to an isolated intersection. In particular, the equa-

tions that Webster [247] derived have been fundamental to traffic signal setting procedures since their development. Later, Webster and Cobbe [248] developed a formula for estimating vehicle stops at under-saturated intersections assuming random vehicle arrivals. Other models were developed by Newell [153] and Catling [49]. Catling adapted equations of classical queuing theory to over-saturated traffic conditions and developed a comprehensive queue length estimation procedure that captured the time-dependent nature of queues to be applied to both under-saturated and over-saturated conditions. In addition, Cronje [60,61,62,63] developed stop and delay equations by treating traffic flow through a fixed-time signal as a Markov process. The approach assumed that the number of queued vehicles at the beginning of a cycle could be expressed by a geometric distribution. These models, however, were not designed to account for the partial stops that vehicles may incur. Furthermore, the models that account for partial stops do not estimate vehicle partial and full stops for over-saturated conditions. A study by Rakha et al. [187] developed a procedure for estimating vehicle stops while accounting for partial stops, as

$$S(t_i) = \frac{u(t_i) - u(t_{i-1})}{u_f} \quad \forall i \ni u(t_i) < u(t_{i-1}). \quad (37)$$

The sum of these partial stops is also recorded. This sum, in turn, provides a very accurate explicit estimate of the total number of stops that are encountered along a roadway. Again the model can be implemented within a microscopic traffic simulation software or applied to second-by-second speed measurements using a GPS system.

### Estimation of Vehicle Energy Consumption and Emissions

Estimating accurate mobile source emissions has gained interest among transportation professionals as a result of increasing environmental problems in large metropolitan urban areas. While current emission inventory models in the US, such as MOBILE and EMPAC, are capable of estimating large scale inventories, they are unable to estimate accurate vehicle emissions that result from operational-level projects. Alternatively, microscopic emission models are capable of assessing the impact of transportation projects on the environment and performing project-level analyzes. Consequently, the focus of this discussion will be on these microscopic and also mesoscopic models. Two models that are emerging include the Comprehensive Modal Emissions Model (CMEM) and the Virginia Tech Microscopic (VT-Micro) model. These models

are briefly described in terms of their structure, logic, and validity.

**Comprehensive Modal Emission Model** The Comprehensive Modal Emissions Model (CMEM), which is one of the newest power demand-based emission models, was developed by researchers at the University of California, Riverside [20]. The CMEM model estimates LDV and LDT emissions as a function of the vehicle's operating mode. The term "comprehensive" is utilized to reflect the ability of the model to predict emissions for a wide variety of LDVs and LDTs in various operating states (e. g., properly functioning, deteriorated, malfunctioning).

The development of the CMEM model involved extensive data collection for both engine-out and tailpipe emissions of over 300 vehicles, including more than 30 high emitters. These data were measured at a second-by-second level of resolution on three driving cycles, namely: the Federal Test Procedure (FTP), US06, and the Modal Emission Cycle (MEC). The MEC cycle was developed by the UC Riverside researchers in order to determine the load at which a specific vehicle enters into fuel enrichment mode. CMEM predicts second-by-second tailpipe emissions and fuel consumption rates for a wide range of vehicle/technology categories. The model is based on a simple parametrized physical approach that decomposes the entire emission process into components corresponding to the physical phenomena associated with vehicle operation and emission production. The model consists of six modules that predict engine power, engine speed, air-to-fuel ratio, fuel use, engine-out emissions, and catalyst pass fraction. Vehicle and operation variables (such as speed, acceleration, and road grade) and model calibrated parameters (such as cold start coefficients, engine friction factor) are utilized as input data to the model.

Vehicles were categorized in the CMEM model based on a vehicle's total emission contribution. Twenty-eight vehicle categories were constructed based on a number of vehicle variables. These vehicle variables included the vehicle's fuel and emission control technology (e. g. catalyst and fuel injection), accumulated mileage, power-to-weight ratio, emission certification level (tier0 and tier1), and emitter level category (high and normal emitter). In total 24 normal vehicle and 4 high emitter categories were considered [20].

**The Virginia Tech Microscopic Energy and Emission Model (VT-Micro Model)** The VT-Micro emission models were developed from experimentation with numerous polynomial combinations of speed and acceleration levels. Specifically, linear, quadratic, cubic, and fourth

degree combinations of speed and acceleration levels were tested using chassis dynamometer data collected at the Oak Ridge National Laboratory (ORNL). The final regression model included a combination of linear, quadratic, and cubic speed and acceleration terms because it provided the least number of terms with a relatively good fit to the original data ( $R^2$  in excess of 0.92 for all measures of effectiveness [MOE]). The ORNL data consisted of nine normal-emitting vehicles including six light-duty automobiles and three light-duty trucks. These vehicles were selected in order to produce an average vehicle that was consistent with average vehicle sales in terms of engine displacement, vehicle curb weight, and vehicle type. The data collected at ORNL contained between 1,300 to 1,600 individual measurements for each vehicle and MOE combination depending on the vehicle's envelope of operation [8].

This method has a significant advantage over emission data collected from a few driving cycles because it is difficult to cover the entire vehicle operational regime with only a few driving cycles. Typically, vehicle acceleration values ranged from  $-1.5$  to  $3.7 \text{ m/s}^2$  at increments of  $0.3 \text{ m/s}^2$  ( $-5$  to  $12 \text{ ft/s}^2$  at  $1\text{-ft/s}^2$  increments). Vehicle speeds varied from  $0$  to  $33.5 \text{ m/s}$  ( $0$  to  $121 \text{ km/h}$  or  $0$  to  $110 \text{ ft/s}$ ) at in increments of  $0.3 \text{ m/s}$  [8].

The model had the problem of overestimating HC and CO emissions especially for high acceleration levels. Since this problem arose from the fact that the sensitivity of the dependent variables to the positive acceleration levels is significantly different from that for the negative acceleration levels, a two-regime model for positive and negative acceleration regimes was developed as [8,183]

$$\ln(MOE_e) = \begin{cases} \sum_{i=0}^3 \sum_{j=0}^3 (L_{i,j}^e \times u^i \times a^j) & \text{for } a \geq 0 \\ \sum_{i=0}^3 \sum_{j=0}^3 (M_{i,j}^e \times u^i \times a^j) & \text{for } a < 0. \end{cases} \quad (38)$$

Where  $MOE_e$  is the instantaneous fuel consumption or emission rate (ml/s or mg/s);  $K_{i,j}^e$  is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j";  $L_{i,j}^e$  is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for positive accelerations;  $M_{i,j}^e$  is the model regression coefficient for MOE "e" at speed power "i" and acceleration power "j" for negative accelerations;  $u$  is the instantaneous speed (km/h); and  $a$  is the instantaneous acceleration rate (km/h/s).

Additionally, the VT-Micro model was expanded by including data from 60 light-duty vehicles (LDVs) and

trucks (LDTs). Statistical clustering techniques were applied to group vehicles into homogeneous categories using classification and regression tree (CART) algorithms. The 60 vehicles were classified into five LDV and two LDT categories [182]. In addition, HE vehicle emission models were constructed using second-by-second emission data. In constructing the models, HEVs are classified into four categories for modeling purposes. The employed HEV categorization was based on the comprehensive modal emission model (CMEM) categorization. The first type of HEVs has a chronically lean fuel-to-air ratio at moderate power or transient operation, which results in high emissions in NO. The second type has a chronically rich fuel-to-air ratio at moderate power, which results in high emissions in CO. The third type is high in HC and CO. The fourth type has a chronically or transiently poor catalyst performance, which results in high emissions in HC, CO, and NO. Each model for each category was constructed within the VT-Micro modeling framework. The HE vehicle model was found to estimate vehicle emissions with a margin of error of 10% when compared to in-laboratory bag measurements [9]. Furthermore, all the models were incorporated into the INTEGRATION software, and made it possible to evaluate the environmental impacts of operational level transportation projects [164].

### Use of Technology to Enhance System Performance

Due to the recent extensive developments within the fields of artificial intelligence, communications, and computation algorithms, transportation and traffic engineers' goals have evolved. As mentioned earlier in the paper, current spatio-temporal distribution of trips is far from being optimum, either with respect to driver satisfaction and/or network performance. A part of the contemporary DTA research is directed towards influencing, as opposed to modeling, dynamic spatio-temporal trip distributions. Advanced Traveler Information Systems (ATISs) are definitely the main tool for such influence, and understanding driver behavior is critical to the design and implementation of such systems. Research which is directly related to the possibility of enhancing system performance through the use of technology may be categorized in the following main areas of research:

- Validation of models, lab experiments and real world behavior, which is the area concerned with verifying the different theories and their implicit assumptions with regards to real-life situations. Due to the extreme complexity and questionable possibility of this task, several attempts have been made to verify the models with respect to lab experiments rather than the real world behavior. Moreover, comparison and verification of spatial and temporal transferability of the models might as well fall within this area. Examples of current literature include the works of Chang and Mahmassani [50] and Mahmassani and Jou [129].
- Calibration of algorithms and models, which as the name suggests, is the area related to the calibration of the algorithms and model parameters. This also entails spatial and temporal calibration, for certain models and/or parameters might only be valid for certain locations and time periods rather than others. Examples of current literature include the works of Chang and Mahmassani [50] and Rakha and Arafeh [184].
- Real time deployment, which focuses on the possibility of deploying DTA models into the real world. This area of research is concerned with developing deployable DTA algorithms. Current literature states that although "a mathematically tractable analytical model that is adequately sensitive to traffic realism vis-à-vis real-time operation is still elusive", yet even with currently available models there is a tradeoff between solution accuracy and computational efficiency. Other real-time deployment issues include computational tractability; consistency checking; model robustness, stability, and error and fault tolerance; and demand estimation and prediction [180]. Examples of current literature include the works of Mahmassani et al. [131,132,133,134], Ben-Akiva et al. [25,26], Mahmassani and Peeta [130,134,135], Peeta and Mahmassani [168], Hawas [88], Hawas et al. [91], Hawas and Mahmassani [90,92], Cantarell and Cascetta [38], Anastassopoulos [12], and Jha et al. [104].
- Issues of uncertainty, which is, as mentioned earlier, a fundamental feature in most transportation phenomena. Uncertainty can be represented in trip makers' knowledge of different route travel times, in the compliance rates of drivers to information, in the accuracy of the disseminated control information, in the driver's perception of disseminated information reliability, in the controller's predicted and/or refined dynamic travel times and/or O-D matrices, among others. Uncertainty-related research issues have been addressed through several approaches, like stochastic modeling, fuzzy control, and reliability indices. Examples of current literature include the works of: Birge and Ho [28], Peeta and Zhou [109,167], Cantarell and Cascetta [38], Ziliaskopoulos and Waller [262], Waller and Ziliaskopoulos [245], Waller [244], Peeta and Jeong [177], Jha et al. [104], Peeta and Paz [171].
- DTA control, which is the area of research concerned

with modifying how trips are distributed on the network. Research within this area focuses on capturing current network performance, and works on modifying the system elements, such as drivers route, and/or departure time selection, as well as mode choice (possibly through pricing and information dissemination); and traffic management (primarily through signal operation), in order to optimize system performance. Examples of current literature include the works of Peeta and Paz [171].

- Realism of other system characteristics, which is the research area concerned with capturing other system realities that are not considered in current available literature. Examples of such realities may include [180],
  - Person rather than driver assignment. It is an undeniable fact that many people tend to make their mode choices based on daily, real-time decisions, i.e. this is a dynamic and not a static process. It is further anticipated that with the current (and predicted) maturity of information technology within the transportation arena, would require explicit modeling within DTA models.
  - The effect of interaction between the different vehicle classes and road infrastructure. It is beyond doubt that certain vehicle classes (such as trucks and busses for example) will not be able to comply with certain diversion-requesting disseminated information, due to road infrastructure constraints. However, in other occasions, these vehicle classes might be able to divert routes, yet with travel time penalties (example if the turning radius was inadequate) that might not only affect these vehicle classes, but all other diverting vehicles as well.
  - Capturing latest traffic control technology and strategies. Traffic control technology and strategies have been rapidly developing during the past couple of decades. Examples of this include transit signal preemption, real-time adaptive signal traffic control, electronic toll collection, etc. For efficient DTA control, DTA algorithms should be able to sufficiently capture and consider them.

Examples of current literature include the works of Ran and Boyce [200], Peeta et al. [174], Ziliaskopoulos and Waller [262], Dion and Rakha [72], Sivanandan et al. [223], Rakha et al. [190,191], Rakha and Zhang [197].

### Road Pricing

A number of approaches have been and continue to be considered in an attempt to reduce urban congestion. These approaches include supply and demand manage-

ment strategies. Supply management strategies are addressed through the building of new infrastructure, expanding the existing infrastructure, or by increasing the utilization of existing infrastructure through the introduction of traffic management systems. Alternatively, demand strategies focus on either distributing the existing demand more evenly in space over the network or distributing the existing demand more evenly by spreading the peak period over a longer time. One of the emerging demand management strategies for enhancing traffic system performance is the use of electronic road pricing.

Road pricing entails the charging of the user of a road for the use of the facility. The road charges include fuel taxes, license fees, parking taxes, tolls, and congestion charges, including those which may vary by time-of-day, by the specific vehicle type being used. Road pricing has two distinct objectives: revenue generation, usually for road infrastructure financing, and congestion pricing for demand management purposes. Toll roads are the typical example of revenue generation. Changes for using high-occupancy toll (HOT) lanes or urban tolls for entering a restricted area of a city are typical examples of using road pricing for congestion management purposes.

In modeling route choice within the context of road pricing one has to consider at least two main types of route choices, one which is tolled and one which is not. An example of such a situation is a multi-lane freeway that splits at several locations to provide 3 lanes that are free of toll and 1 lane that involves a toll. While these are some unique traffic engineering issues associated with the weaving that can take place, when there are multiple breaks in the structure that divides the toll lanes from the free ones, many important equilibrium issues can already be analyzed by looking at single diverge sections. If traffic demand is low, and the distance and speed limits are similar on both types of lanes, there exists little incentive for drivers to utilize the toll lane alternative. However, as the traffic demand on the toll alternative increases, its speed will decrease. This will make its travel time become longer, making the free lanes less attractive. This sets up a condition in which it may become attractive to some, but not necessarily all drivers, to start utilizing the toll lane. The timing of when drivers start to use the toll lanes and the number of drivers who elect to use the toll lane depends primarily on two factors, namely the “value of the toll” and the “value of time” of the drivers involved.

The “value of the toll” and “value of time” are closely related quantities, as a higher toll value accompanied with a higher value of time will result in the same perceived disutility and therefore path impedance. Consequently, it is ratio of the “value of the toll” to the “value of time” that is



most important. In addition, one should also note that for the same toll value, those with the highest “value of time” will be least influenced by the toll. They will therefore also be the first ones to start using the toll lanes when the free lanes become busy and slower. In contrast, those with the lowest value of time perceive the toll as having disutility or travel time equivalency. They will therefore be the last ones to switch to the toll lane, if at all.

The decisions, of those drivers who have a lower value of time, are also influenced by the decisions of the drivers with a higher value of time. Specifically, any switch over of the latter drivers from the free lanes to the toll lanes will leave the free lanes less congested for those with a lower value of time who stay in the free lanes. This may leave it undesirable for those with a low value of time to also switch to the toll lane. However, if the percent of drivers with a high value of time is relatively small, or if the total flow A to B is very high, even some of those with a low value of time will perceive themselves as being better off if they use the toll lanes. The issue of toll lane usage therefore requires one to answer two interrelated questions, namely: (a) what group or sub-population of drivers will utilize the toll road, and (b) what fraction of this group or sub-population will utilize the toll road. Consequently, extensive research is currently underway to model driver route selection under various tolling conditions. Examples of some research efforts on this topic include [23, 139, 140, 141, 146]. This list is by no means comprehensive but does provide some applications on this emerging topic.

### Related Transportation Areas

Research within the following two transportation areas definitely precedes DTA research. However, their significance to the DTA field is based on the fact that DTA theories are mostly dependent on older theories stemming from these two areas. Hence, advances within these two areas could probably significantly affect the advances within the DTA arena.

- Traffic flow models encompass the mathematical representation, or perhaps simulation of the traffic flow characteristics, such as modeling traffic flow propagation, queue spillbacks, lane-changing, signal operation, travel time computation, etc. are crucial in determining driver expectations and behavior. In addition, these are also fundamental in the calculations of travel times, which are vital in the combined problem of departure time and route choice. The quantity of research available in this area is probably as big as the quantity of research done in the area of DTA all together, if not even more. However, as mentioned earlier, all of the research done within this area has direct influence on the realization of the traffic flow models, which are also used within the DTA models.
- Planning applications, which in spite of being a quite under-researched area at the moment, is a vitally important one. There is no doubt DTA models are superior to static models, hence, it is probably only a matter of time before the industry abandons static models for dynamic models. “Dynamic models are simply the natural evolution in the transportation field that like any other new effort suffers from early development shortcomings” [180]. Examples of current literature within this area of research include the works of Li [125], Friesz et al., Waller [244], Waller and Ziliaskopoulos [245], Ziliaskopoulos and Waller [263], Ziliaskopoulos and Wardell [261].

### Future Directions

Future research challenges and directions include:

- Enhance traffic flow modeling and driver behavior modeling. These include the modeling of person as opposed vehicle route choices, the separation of driver and vehicle within the traffic modeling framework, the explicit modeling of vehicle dynamics, enhancing car-following, lane-changing, and routing behavior.
- Develop more efficient algorithms that would be suitable for real-time deployment, without making any compromises in the computational accuracy, i. e. without trading-off the solution accuracy for the computational efficiency. In precise, without compromising any dimension of the traffic flow theory, nor driver behavior assumptions. As a matter of fact, further research should be done to capture more of the traffic, as well as the driver behavior theory. Hence, this should help in improving the realism of the available DTA models.
- Conduct more research on the driver behavior theory. Especially, since human factors cognitive research has significantly improved in the previous couple of decades, then modeling driver behavior from this perspective might lead to valuable outcomes.
- Critical examination of the validity of network equilibrium as a framework for network flow analysis [152]. Many of the current algorithms are based on the assumption that drivers become rational and homogeneous with learning. Hence, resulting in network equilibrium. A number of recent research efforts suggest that some drivers remain less rational, and heterogeneous drivers make up the system; drivers’ attitudes

toward uncertainty become bipolar; and some drivers are sometimes deluded. Further research is required to characterize and model such behavior.

- Validate current models by comparing current model outputs with real world experiments, and possibly with controlled lab experiments (as mid-way experiments before conducting real world evaluations).
- Enhance traffic modeling tools within DTA models to capture the effect of diversion compliance of different vehicle modes (especially heavy vehicles) to more geometrically restrictive highways.
- Possibly calibrating hybrid fuzzy-stochastic models and comparing results to traditional models. According to the work done by Chen [52], probabilistic methods are better than possibility-based methods if sufficient information is available, on the other hand, possibility-based methods can be better if little information is available. However, when there is little information available about uncertainties, a hybrid method may be optimum.
- Conduct further research on the dynamic synthetic O-D estimation from link flow measurements and vehicle probe data. Further research is required to quantify the impact of erroneous or missing data on the accuracy of O-D estimates.
- Conduct further research on the temporal distribution of demand, analyzing and modeling it. Then, including the estimation and forecast of time-dependent demand within the planning process, in addition to the dynamic traffic management and control processes. This should, hopefully, help to fill-in the gap between the three mentioned processes.
- Incorporating person assignment, rather than mode assignment in the DTA and planning models, for as mentioned earlier, mode split is currently more of a daily real-time dynamic, rather than a static decision.
- Research is needed to develop models for driver behavior to different ATIS systems: (types and/or scenarios). Current literature is mainly based on stated preference surveys, which are known for their lack of accuracy. Before the deployment of ATIS systems, stated preference surveys were the best approach for prediction and modeling drivers' reactions. However now, after the deployment of many ATIS systems, more research is needed to capture the actual (possibly revealed) drivers' behavior, rather than the stated behavior.
- Develop approaches that are capable of realistically capturing traffic flow, traffic control, and their interactions; and simultaneously optimizing traffic flow routing and control. In other words, developing algorithms that are actually capable of capturing real-time driver

behavior, and are able to control it, in order to improve network performance.

## Appendix

### Term Abbreviations

<b>ANN</b>	Artificial Neural Networks
<b>ATIS</b>	Advanced Traveler Information System
<b>AVI</b>	Automatic Vehicle Identification
<b>AVL</b>	Automatic Vehicle Location
<b>DTA</b>	Dynamic Traffic Assignment
<b>FHWA</b>	Federal Highway Administration
<b>GA</b>	Genetic Algorithm
<b>GPS</b>	Global Positioning System
<b>HCM</b>	Highway Capacity Manual
<b>HOV</b>	High Occupancy Vehicle
<b>ITS</b>	Intelligent Transportation Systems
<b>LDV</b>	Light Duty Vehicle
<b>LMC</b>	Link Marginal Cost
<b>LP</b>	Linear Programming
<b>MOE</b>	Measure of Effectiveness
<b>NLP</b>	Non-Linear Programming
<b>O-D</b>	Origin – Destination
<b>PMC</b>	Path Marginal Cost
<b>SO</b>	System Optimum
<b>SOV</b>	Single Occupancy Vehicle
<b>TT</b>	Travel Time
<b>UE</b>	User Equilibrium
<b>VMS</b>	Variable Message Sign

### Variable Definitions

$v_i$	Traffic volume on route $i$
$N$	Set of network nodes
$A$	Set of network arcs (links)
$R$	Set of origin centroids
$S$	Set of destination centroids
$k_{rs}$	Set of paths connecting O-D pair $(r-s)$ ; $r \in R$ , $s \in S$
$x_a$	Flow on arc $(a)$
$x_b$	Flow on arc $(b)$
$t_a$	Travel time on arc $(a)$
$t_b$	Travel time on arc $(b)$
$f_k^{rs}$	Flow on path $(k)$ connecting O-D pair $(r-s)$
$f_l^{mn}$	Flow on path $(l)$ connecting O-D pair $(m-n)$
$c_k^{rs}$	Travel time on path $(k)$ connecting O-D pair $(r-s)$
$q_{rs}$	Trip rate between origin $(r)$ and destination $(s)$
$\delta_{a,k}^{rs}$	Indicator variable, = 1 if arc $(a)$ is on path $(k)$ between O-D pair $(r-s)$ , and 0 otherwise
$x$	Vector of flows on all arcs, = $(\dots, x_a, \dots)$
$t$	Vector of travel times on all arcs, = $(\dots, t_a, \dots)$

$f^{rs}$	Vector of flows on all paths connecting O-D pair $r$ - $s$ , $= (\dots, f_k^{rs}, \dots)$	$t_r$	Total seed matrix demand departing during time-slice ( $r$ )
$f$	Matrix of flows on all paths connecting all O-D pairs, $= (\dots, f^{rs}, \dots)$	$T_{rij}$	Traffic demand departing during time-slice ( $r$ ) traveling between origin ( $i$ ) and destination ( $j$ )
$c^{rs}$	Vector of travel times on all paths connecting O-D pair $r$ - $s$ , $= (\dots, c_k^{rs}, \dots)$	$t_{rij}$	Seed traffic demand departing during time-slice ( $r$ ) traveling between origin ( $i$ ) and destination ( $j$ )
$c$	Matrix of travel times on all paths connecting all O-D pairs, $= (\dots, c^{rs}, \dots)$	$l_{rij}$	Lagrange multiplier for departure time-slice, origin, and destination combination ( $rij$ )
$q$	Origin-destination matrix (with elements $= q_{rs}$ )	$V_{sa}$	Observed volume on link ( $a$ ) during time-slice ( $s$ )
$\Delta^{rs}$	Link-path incidence matrix (with $\delta_{a,k}^{rs}$ elements) for O-D pair $r$ - $s$ , as discussed below	$p_{rij}^{sa}$	Probability of ( $a$ ) demand between origin ( $i$ ) and destination ( $j$ ) during time-slice ( $r$ ) is observed on link ( $a$ ) during time-slice ( $s$ )
$\Delta$	Matrix of link-path incidence matrices (for all O-D pairs), $= (\dots, \Delta^{rs}, \dots)$	$d(t_i)$	Vehicle delay at time ( $t_i$ )
$z$	Objective function	$u(t_i)$	Vehicle instantaneous speed at time ( $t_i$ )
$L$	Lagrange (transformation of the) objective function	$u_f$	Free-flow speed
$u_{rs}$	Dual variable associated with the flow conservation constraint for O-D pair ( $r$ - $s$ )	$S(t_i)$	Vehicle full and partial stops at time ( $t_i$ )
$t_{i,k}$	Observed average travel time along link $i$ within the $k$ th sampling interval	$MOE_e$	Instantaneous fuel consumption or emission rate
$\tilde{t}_{i,k}$	Smoothed average travel time along link $i$ in the $k$ th sampling interval	$K_{i,j}^e$	Model regression coefficient for MOE ( $e$ ) at speed power ( $i$ ) and acceleration power ( $j$ )
$s_{i,k}^2$	Variance of the observed travel times relative to the observed average travel time in the $k$ th sampling interval	$L_{i,j}^e$	Model regression coefficient for MOE ( $e$ ) at speed power ( $i$ ) and acceleration power ( $j$ ) for positive accelerations
$\tilde{s}_{i,k}^2$	Variance of the observed travel times relative to the smoothed travel time in the $k$ th sampling interval	$M_{i,j}^e$	Model regression coefficient for MOE ( $e$ ) at speed power ( $i$ ) and acceleration power ( $j$ ) for negative accelerations
$n_{i,k}$	Number of valid travel time readings on link $i$ in the $k$ th sampling interval	$u$	Vehicle instantaneous speed
$\alpha$	Exponential smoothing factor that varies as a function of the number of observations $n_{i,k}$ within the sampling interval	$a$	Vehicle instantaneous acceleration rate
$\beta$	Constant that varies between 0 and 1	$q$	Traffic stream flow (veh/h)
$T_{ij}$	Number of trips between production zone $i$ and attraction zone $j$	$k$	Traffic stream density (veh/km)
$P_i$	Number of trip productions from the origin zone	$u$	Traffic stream space-mean speed (km/h)
$A_j$	Number of trip attractions to the destination zone	$u_f$	Expected traffic stream free-flow speed (km/h)
$F_{ij}$	Impedance factor between production zone $i$ and attraction zone $j$	$u_c$	Expected traffic stream speed-at-capacity (km/h)
$K_{ij}$	Socio-economic adjustment factor for trips between production zone $i$ and attraction zone $j$	$k_j$	Expected traffic stream jam density (veh/km)
$c_{ij}$	Generalized cost of inter-zonal travel between production zone $i$ and attraction zone $j$	$q_c$	Expected traffic stream capacity (veh/km)
$t_{ij}$	Prior information on the number of trips between production zone $i$ and attraction zone $j$	$c_1$	Model coefficient (km/veh)
$V_a$	Traffic flow on link ( $a$ )	$c_3$	Model constant ( $\text{h}/\text{km}^2$ -veh)
$V'_a$	Complementary traffic flow on link ( $a$ )	$c_2$	Model constant ( $\text{h}^{-1}$ )
$p_{ij}^a$	Probability of traffic flow between origin ( $i$ ) and destination ( $j$ ) to use link ( $a$ )		
$T_r$	Total demand departing during time-slice ( $r$ )		

## Bibliography

1. Abdel-Aty MA, Kitamura R, Jovanis PP (1997) Using Stated Preference Data for Studying the Effect of Advanced Traffic Information on Drivers' Route Choice. *Transp Res Part C: (Emerg Technol)* 5(1):39–50
2. Abdelfatah AS, Mahmassani HS (2001) A simulation-based signal optimization algorithm within a dynamic traffic assignment framework. *IEEE Intelligent Transportation Systems Proceedings, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland*
3. Abdelghany AF, Abdelghany KF et al (2000) Dynamic traffic assignment in design and evaluation of high-occupancy toll lanes. *Transp Res Rec* 1733:39–48

4. Abdelghany KF, Mahmassani HS (2001) Dynamic trip assignment-simulation model for intermodal transportation networks. *Transp Res Rec* 1771:52–60
5. Abdelghany KF, Valdes DM et al (1999) Real-time dynamic traffic assignment and path-based signal coordination: Application to network traffic management. *Transp Res Rec* 1667:67–76
6. Abdulhai B, Porwal H, Recker W (2002) Short-term Freeway Traffic Flow Prediction Using Genetically Optimized Time-Delay-Based Neural Networks. *ITS J: Intell Transp Syst J* 7(1): 3–41
7. Ahmed M, Cook (1982) Analysis of Freeway Traffic Time Series Data by Using Box-Jenkins Techniques. *Transp Res Rec* 722:1–9
8. Ahn K, Rakha H et al (2002) Estimating vehicle fuel consumption and emissions based on instantaneous speed and acceleration levels. *J Transp Eng* 128(2):182–190
9. Ahn K, Rakha H et al (2004) Microframework for modeling of high-emitting vehicles. *Transp Res Rec* 1880:39–49
10. Akcelik R, Roupail NM (1994) Overflow queues and delays with random and platooned arrivals at signalized intersections. *J Adv Transp* 28(3):227–251
11. Allen RW, Stein AC, Rosenthal TJ, Ziedman D, Torres JF, Halati A (1991) Human factors simulation investigation of driver route diversion and alternate route selection using in-vehicle navigation systems. In: *Vehicle Navigation & Information Systems Conference*, Dearborn, 20–23 Oct 1991. *Proceedings Part 1 (of 2) Society of Automotive Engineers*. SAE, Warrendale, pp 9–26
12. Anastassopoulos I (2000) Fault-Tolerance and Incident Detection using Fourier Transforms. *Purdue University*, West-lafayette
13. Arafeh M, Rakha H (2005) Genetic Algorithm Approach for Locating Automatic Vehicle Identification Readers. In: *IEEE Intelligent Transportation System Conference*, Vienna, 2005. *Proceedings ITSV'05 IEEE Intelligent Conference on Transportation Systems*, p 1153–1158
14. Arnott R, de Palma A, Lindsey R (1991) Does providing information to drivers reduce traffic congestion? *Transp Res Part A (General)* 25A(5):309
15. Arrow KJ (1951) Alternative Approaches to the Theory of Choice in Risk-Taking Situations. *Econometrica* 19(4):404–437
16. Ashok K (1996) Estimation and Prediction of Time-Dependent Origin-Destination Flows. Boston, Massachusetts Institute of Technology. Ph D Thesis, Massachusetts Institute of Technology
17. Ashok K, Ben-Akiva ME (1993) Dynamic Origin-Destination Matrix Estimation and Prediction for Real-Time Traffic Management Systems. In: Daganzo CF (ed) *12th International Symposium on Transportation and Traffic Theory*. Elsevier, New York, pp 465–484
18. Ashok K, Ben-Akiva ME (2000) Alternative approaches for real-time estimation and prediction of time-dependent Origin-Destination flows. *Transp Sci* 34(1):21–36
19. Balakrishna R, Koutsopoulos HN et al (2005) Simulation-Based Evaluation of Advanced Traveler Information Systems. *Transp Res Rec* 1910:90–98
20. Barth M, An F et al (2000) Comprehensive Modal Emission Model (CMEM): Version 2.0 User's Guide. University of California, Riverside
21. Bell M, Iida Y (1997) *Transportation network analysis*. Iida Y translator. Wiley, Chichester / New York
22. Ben-Akiva M, Bierlaire M et al (1998) DynaMIT: a simulation-based system for traffic prediction. *DACCORD Short Term Forecasting Workshop*, Delft, February 1998
23. Ben-Akiva M, Bolduc D et al (1993) Estimation of travel choice models with randomly distributed values of time. *Transp Res Rec* 1413:88–97
24. Ben-Akiva M, Kroes E et al (1992) Real-Time Prediction of Traffic Congestion. *Vehicle Navigation and Information Systems*, IEEE, New York
25. Ben-Akiva M, Bierlaire M, Bottom J, Koutsopoulos H et al (1997) Development Of A Route Guidance Generation System For Real-Time Application. In: *8th International Federation of Automatic Control Symposium on Transportation Systems*, Chania, 16–18 June 1997
26. Ben-Akiva MMB, Koutsopoulos H, Mishalani R (1998) DynaMIT: a simulation-based system for traffic prediction. *DACCORD Short Term Forecasting Workshop*, Delft
27. Bierlaire M, Crittin F (2004) An efficient algorithm for real-time estimation and prediction of dynamic OD tables. *Oper Res* 52(1):116–27
28. Birge JR, Ho JK (1993) Optimal flows in stochastic dynamic networks with congestion. *Oper Res* 41(1):203–216
29. Bolland JD, Hall MD et al (1979) SATURN: Simulation and Assignment of Traffic in Urban Road Networks. In: *International Conference on Traffic Control Systems*, Berkeley
30. Boyce DE, Ran B, Leblanc LJ (1995) Solving an instantaneous dynamic user-optimal route choice model. *Transp Sci* 29(2):128–142
31. Braess D (1968) Über ein Paradoxon der Verkehrsplanung. *Unternehmensforschung* 12:258–268
32. Brilon W (1995) Delays at oversaturated unsignalized intersections based on reserve capacities. *Transp Res Rec* 1484:1–8
33. Brilon W, Wu N (1990) Delays at fixed-time traffic signals under time-dependent traffic conditions. *Traffic Eng & Control* 31(12):8
34. Burrell JE (1968) Multipath Route Assignment and its Application to Capacity-Restraint. *Fourth International Symposium on the Theory of Traffic Flow*, Karlsruhe
35. Burrell JE (1976) Multipath Route Assignment: A Comparison of Two Methods. In: Florian M (ed) *Traffic Equilibrium Methods*. Lecture Notes in Economics and Mathematical Systems, vol 118. Springer, New York, pp 210–239
36. Busemeyer JR, Townsend JT (1993) Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychol Rev* 100(3):432
37. Byung-Wook Wie TRL, Friesz TL, Bernstein D (1995) A discrete time, nested cost operator approach to the dynamic network user equilibrium problem. *Transp sci* 29(1):79–92
38. Cantarella GE, Cascetta ES (1995) Dynamic processes and equilibrium in transportation networks: towards a unifying theory. *Transp Sci* 29(4):305–329
39. Carey M (1986) Constraint qualification for a dynamic traffic assignment model. *Transp Sci* 20(1):55–58
40. Carey M (1987) Optimal time-varying flows on congested networks. *Oper Res* 35(1):58–69
41. Carey M (1992) Nonconvexity of the dynamic traffic assignment problem. *Transp Res Methodol* 26B(2):127
42. Carey M, Subrahmanian E (2000) An approach to mod-

- elling time-varying flows on congested networks. *Transp Res Methodol* 34B(3)
43. Cascetta E, Marquis G (1993) Dynamic estimators of origin-destination matrices using traffic counts. *Transp Sci* 27(4):363–373
  44. Cassidy MJ, Han LD (1993) Proposed model for predicting motorist delays at two-lane highway work zones. *J Transp Eng* 119(1):27–42
  45. Cassidy MJ, Rudjanakanoknad J (2005) Increasing the capacity of an isolated merge by metering its on-ramp. *Transp Res Part B Methodol* 39(10):896–913
  46. Cassidy MJ, Son Y et al (1994) Estimating motorist delay at two-lane highway work zones. *Transp Res Part A, Policy Pract* 28(5):433–444
  47. Cassidy MJ, Windover JR (1995) Methodology for assessing dynamics of freeway traffic flow. *Transp Res Rec* 1484:73–79
  48. Castillo E, Menendez JM, Jimenez P (2008) Trip matrix and path flow reconstruction and estimation based on plate scanning and link observations. *Transp Res Part B: Methodol* 42(5):455–481
  49. Catling I (1977) A Time-Dependent Approach to Junction Delays. *Traffic Eng Control* 18(11):520–523:526
  50. Chang GL, Mahmassani HS (1988) Travel Time Prediction And Departure Time Adjustment Behavior Dynamics In A Congested Traffic System. *Transp Res, Part B: Methodol* 22B(3):217–232
  51. Chang GL, Tao X (1999) Integrated model for estimating time-varying network origin-destination distributions. *Transp Res, Part A: Policy Pract* 33(5):381–399
  52. Chen SQ (2000) Comparing Probabilistic and Fuzzy Set Approaches for Design in the Presence of Uncertainty. In: *Aerospace and Ocean Engineering*. Ph D, Polytechnic Institute and State University, Blacksburg
  53. Chiu YC, Mahmassani HS (2001) Toward hybrid dynamic traffic assignment-models and solution procedures. In: *IEEE Intelligent Transportation Systems Proceedings, IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland*
  54. Coifman B (1998) New algorithm for vehicle reidentification and travel time measurement on freeways. In: *Proceedings of the 1998 5th International Conference on Applications of Advanced Technologies in Transportation, Newport Beach, Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering, ASCE, Reston*
  55. Coifman B, Banerjee B (2002) Vehicle reidentification and travel time measurement on freeways using single loop detectors-from free flow through the onset of congestion. In: *Proceedings of the seventh International Conference on: Applications of Advanced Technology in Transportation, Cambridge, 5-7 Aug 2002. Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering, American Civil Engineers*
  56. Coifman B, Cassidy M (2001) Vehicle reidentification and travel time measurement, Part I: Congested freeways. In: *IEEE Intelligent Transportation Systems Proceedings, Conference IEEE on Intelligent Transportation Systems, Proceedings, ITSC 2001, Oakland*
  57. Coifman B, Ergueta E (2003) Improved vehicle reidentification and travel time measurement on congested freeways. *J Transp Eng* 129(5):475–483
  58. Colyar JD, Roupail NM (2003) Measured Distributions of Control Delay on Signalized Arterials. *Transp Res Rec* 1852:1–9
  59. Cremer M, Keller H (1987) New Class Of Dynamic Methods For The Identification Of Origin-Destination Flows. *Transp Res, Part B: Methodol* 21(2):117–132
  60. Cronje WB (1983) Analysis of Existing Formulas for Delay, Overflow, and Stops. *Transp Res Rec* 905:89–93
  61. Cronje WB (1983) Derivation of Equations for Queue Length, Stops, and Delay for Fixed-Time Traffic Signals. *Transp Res Rec* 905:93–95
  62. Cronje WB (1983) Optimization Model for Isolated Signalized Traffic Intersections. *Transp Res Rec* 905:80–83
  63. Cronje WB (1986) Comparative Analysis of Models for Estimating Delay for Oversaturated Conditions at Fixed-Time Traffic Signals. *Transp Res Record* 1091:48–59
  64. Dafermos S (1980) Traffic equilibrium and variational inequalities. *Transp Sci* 14(1):42–54
  65. Daganzo CF, Laval JA (2005) On the numerical treatment of moving bottlenecks. *Transp Res Part B Methodol* 39(1):31–46
  66. Daniel J, Fambro DB et al (1996) Accounting for nonrandom arrivals in estimate of delay at signalized intersections. *Transp Res Rec* 1555:9–16
  67. Dantzig GB (1957) The Shortest Route Problem. *Oper Res* 5:270–273
  68. Dial R (1971) A Probabilistic Multipath Traffic Assignment Model which Obviates Path Enumeration. *Transp Res* 5: 83–111
  69. Dijkstra EW (1959) A Note on Two Problems in Connection with Graphics. *Numeriche Math* 1:209–271
  70. Dion F, Rakha H (2006) Estimating dynamic roadway travel times using automatic vehicle identification data for low sampling rates. *Transp Res Part B* 40:745–766
  71. Dion F, Rakha H et al (2004) Comparison of delay estimates at under-saturated and over-saturated pre-timed signalized intersections. *Transp Res Part B Methodol* 38(2):99–122
  72. Dion F, Rakha H et al (2004) Evaluation of potential transit signal priority benefits along a fixed-time signalized arterial. *J Transp Eng* 130(3):294–303
  73. Elefteriadou L, Fang C et al (2005) Methodology for evaluating the operational performance of interchange ramp terminals. *Transp Res Rec* 1920:13–24
  74. Engelbrecht RJ, Fambro DB et al (1996) Validation of generalized delay model for oversaturated conditions. *Transp Res Rec* 1572:122–130
  75. Evans JL, Elefteriadou L et al (2001) Probability of breakdown at freeway merges using Markov chains. *Transp Res Part B Methodol* 35(3):237–254
  76. Fambro DB, Roupail NM (1996) Generalized delay model for signalized intersections and arterial streets. *Transp Res Rec* 1572:112–121
  77. Fang FC, Elefteriadou L et al (2003) Using fuzzy clustering of user perception to define levels of service at signalized intersections. *J Transp Eng* 129(6):657–663
  78. Fisk C (1979) More Paradoxes in the Equilibrium Assignment Problem. *Transp Res* 13B:305–309
  79. Flannery A, Kharoufeh JP et al (2005) Queuing delay models for single-lane roundabouts. *Civ Eng & Environ Syst* 22(3):133–150
  80. Frank M (1981) The Braess Paradox. *Math Program* 20: 283–302



81. Frank M, Wolfe P (1956) An Algorithm of Quadratic Programming. *Nav Res Logist* 3:95–110
82. Friesz TL, DB, Smith TE, Tobin RL, Wie BW (1993) A Variational Inequality Formulation of the Dynamic Network User Equilibrium Problem. *Oper Res* 41(1):179–191
83. Friesz TL, JL, Tobin RL, Wie B-W (1989) Dynamic network traffic assignment considered as a continuous time optimal control problem. *Oper Res* 37(6):893–901
84. Ghali MO, Smith MJ (1995) A model for the Dynamic System Optimum Traffic Assignment Problem. *Transp Res* 29B(3):155–170
85. Greenshields BD (1934) A study of traffic capacity. In: *Proc Highway Research Board* 14:448–477
86. Hagring O, Rouphail NM et al (2003) Comparison of Capacity Models for Two-Lane Roundabouts. *Transp Res Rec* 1852:114–123
87. Hall MD, Van Vliet D et al (1980) SATURN - A Simulation-Assignment Model for the Evaluation of Traffic Management Schemes. *Traffic Eng Control* 4:167–176
88. Hawas YE (1995) A Decentralized Architecture And Local Search Procedures For Real-Time Route Guidance In Congested Vehicular Traffic Networks. University of Texas, Austin
89. Hawas YE (2004) Development and calibration of route choice utility models: Neuro-fuzzy approach. *J transp eng* 130(2):171–182
90. Hawas YE, Mahmassani HS (1995) A Decentralized Scheme For Real-Time Route Guidance In Vehicular Traffic Networks. In: *Second World Congress on Intelligent Transport Systems*, Yokohama, 1995, p 1965–1963
91. Hawas YE, Mahmassani HS (1997) Comparative analysis of robustness of centralized and distributed network route control systems in incident situations. *Transp Res* 1537:83–90
92. Hawas YE, Mahmassani HS, Chang GL, Taylor R, Peeta S, Ziliaskopoulos A (1997) Development of Dynasmart-X Software for Real-Time Dynamic Traffic Assignment. Center for Transportation Research, The University of Texas, Austin
93. Hellinga BR, Van Aerde M (1998) Estimating dynamic O-D demands for a freeway corridor using loop detector data. *Canadian Society for Civil Engineering*, Halifax, Montreal
94. Ho JK (1980) A successive linear optimization approach to the dynamic traffic assignment problem. *Transp Sci* 14(4):295–305
95. Hu SR, Madanat SM, Krogmeier JV, Peeta S (2001) Estimation of dynamic assignment matrices and OD demands using adaptive Kalman Filtering. *Intell Transp Syst J* 6:281–300
96. Huey-Kuo C, Che-Fu H (1998) A model and an algorithm for the dynamic user-optimal route choice problem. *Transp Res, Part B Methodol* 32B(3):219–234
97. Ishak S, Al-Deek H (2003) Performance Evaluation of a Short-Term Freeway Traffic Prediction Model. *Transportation Research Board 82nd Annual Meeting*, Washington DC
98. Janson BN (1991) Convergent Algorithm for Dynamic Traffic Assignment. *Transp Res Rec* 1328:69–80
99. Janson BN (1991) Dynamic Traffic Assignment For Urban Road Networks. *Transp Res, Part B Methodol* 25B:2–3
100. Jayakrishnan R, Mahmassani HS (1990) Dynamic simulation-assignment methodology to evaluate in-vehicle information strategies in urban traffic networks. *Winter Simulation Conference Proceedings*, New Orleans 1990. 90 Winter Simulation Conf Winter Simulation Conference Proceedings. IEEE, Piscataway (IEEE cat n 90CH2926–4)
101. Jayakrishnan R, Mahmassani HS (1991) Dynamic modelling framework of real-time guidance systems in general urban traffic networks. In: *Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering*, Minneapolis. ASCE, New York
102. Jayakrishnan R, Mahmassani HS et al (1993) User-friendly simulation model for traffic networks with ATIS/ATMS. *Proceedings of the 5th International Conference on Computing in Civil and Building Engineering - V ICCCBE*, Anaheim 1993. ASCE, New York
103. Jeffery DJ (1981) The Potential Benefits of Route Guidance. TRRL, Department of Transportation, Crowthorne
104. Jha M, Madanat S, Peeta S (1998) Perception updating and day-to-day travel choice dynamics in traffic networks with information provision. *Transp Res Part C Emerg Technol* 6C(3):189–212
105. Katsikopoulos KV, Duse-Anthony Y et al (2000) The Framing of Drivers' Route Choices when Travel Time Information Is Provided under Varying Degrees of Cognitive Load. *J Hum Factors Ergonomics Soc* 42(3):470–481
106. Kerner BS (2004) *The physics of traffic*. Springer, Berlin
107. Kerner BS (2004) Three-phase traffic theory and highway capacity. *Physica A* 333(1-4):379–440
108. Kerner BS (2005) Control of spatiotemporal congested traffic patterns at highway bottlenecks. *Physica A* 355(2-4):565–601
109. Kerner BS, Klenov SL (2006) Probabilistic breakdown phenomenon at on-ramp bottlenecks in three-phase traffic theory: Congestion nucleation in spatially non-homogeneous traffic. *Physica A* 364:473–492
110. Kerner BS, Rehborn H et al (2004) Recognition and tracking of spatial-temporal congested traffic patterns on freeways. *Transp Res Part C Emerging Technologies* 12(5):369–400
111. Khattak AJ, Schofer JL, Koppelman FS (1993) Commuters' enroute diversion and return decisions: analysis and implications for advanced traveller information systems. *Transp Res (Policy and Practice)* 27A(2):101
112. Kim H, Baek S et al (2001) Origin-destination matrices estimated with a genetic algorithm from link traffic counts. *Transp Res Rec* 1771:156–163
113. Koutsopoulos HN, Polydoropoulou A et al (1995) Travel simulators for data collection on driver behavior in the presence of information. *Transp Res Part C Emerging Technologies* 3(3):143
114. Krishnamurthy S, Coifman B (2004) Measuring freeway travel times using existing detector infrastructure. In: *Proceedings - 7th International IEEE Conference on Intelligent Transportation Systems, ITSC*, Washington, 2004
115. Laval JA, Daganzo CF (2006) Lane-changing in traffic streams. *Transp Res Part B Methodol* 40(3):251–264
116. Lawson TW, Lovell DJ et al (1996) Using input-output diagram to determine spatial and temporal extents of a queue upstream of a bottleneck. *Transp Res Rec* 1572:140–147
117. LeBlanc LJ (1975) An Algorithm for Discrete Network Design Problem. *Transp Sci* 9:183–199
118. LeBlanc LJ, Abdulaal M (1970) A Comparison of User-Optimum versus System-Optimum Traffic Assignment in Transportation Network Design. *Transp Res* 18B:115–121
119. LeBlanc LJ, Morlok EK et al (1974) An Accurate and Efficient Approach to Equilibrium Traffic Assignment on Congested Networks. *Transp Res Rec* 491:12–23
120. Lee S, Fambro D (1999) Application of the Subset ARIMA

- Model for Short-Term Freeway Traffic vol Forecasting. *Transp Res Rec* 1678:179–188
121. Leonard DR, Tough JB et al (1978) CONTRAM - A Traffic Assignment Model for Predicting Flows and Queues During Peak Periods. TRRL SR 568. Transport Research Laboratory, Crowthorne
  122. Lertworawanich P, Elefteriadou L (2001) Capacity estimations for type B weaving areas based on gap acceptance. *Transp Res Rec* 1776:24–34
  123. Lertworawanich P, Elefteriadou L (2003) A methodology for estimating capacity at ramp weaves based on gap acceptance and linear optimization. *Transp Res Part B Methodol* 37(5):459–483
  124. Li J, Routhail NM et al (1994) Overflow delay estimation for a simple intersection with fully actuated signal control. *Transp Res Rec* 1457:73–81
  125. Li Y (2001) Development of Dynamic Traffic Assignment Models for Planning Applications. Northwestern University, Evanston
  126. Lighthill MJ, Witham GB (1955) On Kinematic Waves. I: Flood Movement in Long Rivers, II. A Theory of Traffic Flow on Long Crowded Roads. In: *Proceedings of the Royal Society of London A* 229, pp 281–345
  127. Lorenz MR, Elefteriadou L (2001) Defining freeway capacity as function of breakdown probability. *Transp Res Rec* 1776: 43–51
  128. Lotan T (1997) Effects of familiarity on route choice behavior in the presence of information. *Transp Res Part C Emerg Technol* 5(3-4):225–243
  129. Mahmassani H, Jou R-C (2000) Transferring insights into commuter behavior dynamics from laboratory experiments to field surveys. *Transp Res Part A Policy Pract* 34A(4):243–260
  130. Mahmassani H, Peeta S (1992) System optimal dynamic assignment for electronic route guidance in a congested traffic network. In: Gartner NH, Improta G (eds) *Urban Traffic Networks. Dynamic Flow Modelling and Control*. Springer, Berlin, pp 3–37
  131. Mahmassani HS, Chiu Y-C, Chang GL, Peeta S, Ziliaskopoulos A (1998) Off-line Laboratory Test Results for the DYNASMART-X Real-Time Dynamic Traffic Assignment System. Center for Transportation Research, The University of Texas, Austin
  132. Mahmassani HS, Hawas Y, Abdelghany K, Abdelfatah A, Chiu Y-C, Kang Y, Chang GL, Peeta S, Taylor R, Ziliaskopoulos A (1998) DYNASMART-X, vol II: Analytical and Algorithmic Aspects. Center for Transportation Research, The University of Texas, Austin
  133. Mahmassani HS, Hawas Y, Hu T-Y, Ziliaskopoulos A, Chang G-L, Peeta S, Taylor R (1998) Development of Dynasmart-X Software for Real-Time Dynamic Traffic Assignment. Technical Report ST067-85-Tast E (revised) submitted to Oak Ridge National Laboratory under subcontract 85X-SU565C,
  134. Mahmassani HS, Peeta S (1993) Network Performance under System Optimal and User Equilibrium Dynamic Assignments: Implications for ATIS. *Transp Res Rec* 1408:83–93
  135. Mahmassani HS, Peeta S (1995) System Optimal Dynamic Assignment for Electronic Route Guidance in a Congested Traffic Network. In: Gartner NH, Improta G (eds) *URBAN TRAFFIC NETWORKS: Dynamic Flow Modeling and Control*. Springer, Berlin, pp 3–37
  136. Mahmassani HS, Peeta S, Hu T, Ziliaskopoulos A (1993) Algorithm for Dynamic Route Guidance in Congested Networks with Multiple User Information Availability Groups. In: 26th International Symposium on Automotive Technology and Automation, Aachen
  137. Matsoukis EC (1986) Road Traffic Assignment - A Review Part I: Non-Equilibrium Methods. *Transp Plan Technol* 11:69–79
  138. Matsoukis EC, Michalopoulos PC (1986) Road Traffic Assignment - A Review Part II: Equilibrium Methods. *Transp Plan Technol* 11:117–135
  139. Mekky A (1995) Toll revenue and traffic study of highway 407 in Toronto. *Transp Res Rec* 1498:5–15
  140. Mekky A (1996) Modeling toll pricing strategies in greater Toronto areas. *Transp Res Rec* 1558:46–54
  141. Mekky A (1998) Evaluation of two tolling strategies for Highway 407 in Toronto. *Transp Res Rec* 1649:17–25
  142. Merchant DK, Nemhauser GL (1978) A Model And An Algorithm For The Dynamic Traffic Assignment Problems. *Transp Sci* 12(3):183–199
  143. Merchant DK, Nemhauser GL (1978) Optimality Conditions For A Dynamic Traffic Assignment Model. *Transp Sci* 12(3):200–207
  144. Minderhoud MM, Elefteriadou L (2003) Freeway Weaving: Comparison of Highway Capacity Manual 2000 and Dutch Guidelines. *Transp Res Rec* 1852:10–18
  145. Moskowitz K (1956) California Method for Assigning Directed Traffic to Proposed Freeways. *Bull Highw Res Board* 130:1–26
  146. Munnich LW Jr, Hubert HH et al (2007) L-394 MnPASS high-occupancy toll lanes planning and operational issues and outcomes (lessons learning in year 1). *Transp Res Rec* 1996:49–57
  147. Murchland JD (1970) Braess's Paradox of Traffic Flow. *Transp Res* 4:391–394
  148. Nagel K (1996) Particle Hopping Model and Traffic Flow Theory. *Phys Rev E* 53 (5):4655–4672
  149. Nagel K, Schreckenberg M (1992) Cellular Automaton Model for Freeway Traffic. *J Phys* 2(20):2212–2229
  150. Nagel K, Schreckenberg M (1995) Traffic Jam Dynamics in Stochastic Cellular Automata. US D Energy, Los Alamos National Laboratory, LA-UR-95-2132, Los Alamos
  151. Nakayama S, Kitamura R (2000) Route choice model with inductive learning. *Transp Res Rec* 1725:63–70
  152. Nakayama S, Kitamura R et al (2001) Drivers' route choice rules and network behavior: Do drivers become rational and homogeneous through learning? *Transp Res Rec* 1752:62–68
  153. Newell GF (1965) Approximation Methods for Queues with Application to the Fixed-Cycle Traffic Light. *SIAM Rev* 7: 223–240
  154. Newell GF (1999) Delays caused by a queue at a freeway exit ramp. *Transp Res Part B Methodol* 33(5):337–350
  155. Nguyen S (1969) An Algorithm for the Assignment Problem. *Transp Sci* 8:203–216
  156. Nie Y, Zhang HM et al (2005) Inferring origin-destination trip matrices with a decoupled GLS path flow estimator. *Transp Res Part B Methodol* 39(6):497–518
  157. Noonan J, Shearer O (1998) Intelligent Transportation Systems Field Operational Test: Cross-Cutting Study Advance Traveler Information Systems. US Department of Transportation, Federal Highways Administration, Intelligent Transportation System, Washington, DC
  158. Okutani I (1987) The Kalman Filtering Approaches in Some Transportation and Traffic Problems. In: *Proceedings of the*

- Tenth International Symposium on Transportation and Traffic Theory. Elsevier, New York
159. Park B (2002) Hybrid neuro-fuzzy application in short-term freeway traffic volume forecasting. *Transp Res Rec* 1802:190–196
  160. Park D, Rilett LR (1998) Forecasting multiple-period freeway link travel times using modular neural networks. *Transp Res Rec* 1617:163–170
  161. Park D, Rilett LR (1999) Forecasting freeway link travel times with a multilayer feedforward neural network. *Comput-Aided Civ & Infrastruct Eng* 14(5):357–367
  162. Park D, Rilett LR et al (1998) Forecasting multiple-period freeway link travel times using neural networks with expanded input nodes. In: *Proceedings of the 1998 5th International Conference on Applications of Advanced Technologies in Transportation*, Newport Beach and *Proceedings of the International Conference on Applications of Advanced Technologies in Transportation Engineering 1998*, ASCE, Reston
  163. Park D, Rilett LR et al (1999) Spectral basis neural networks for real-time travel time forecasting. *J Transp Eng* 125(6): 515–523
  164. Park S, Rakha H (2006) Energy and Environmental Impacts of Roadway Grades. *Transp Res Rec* 1987:148–160
  165. Pavlis Y, Papageorgiou M (1999) Simple decentralized feedback strategies for route guidance in traffic networks. *Transp Sci* 33(3):264–278
  166. Peeta S (1994) System Optimal Dynamic Traffic Assignment in Congested Networks with Advanced Information Systems. University of Texas, Austin
  167. Peeta S, Bulusu S (1999) Generalized singular value decomposition approach for consistent on-line dynamic traffic assignment. *Transp Res Rec* 1667
  168. Peeta S, Mahmassani HS (1995) Multiple user classes real-time traffic assignment for online operations: a rolling horizon solution framework. *Transp Res Part C Emerg Technol* 3C(2):83
  169. Peeta S, Mahmassani HS (1995) System optimal and user equilibrium time-dependent traffic assignment in congested networks. *Ann Oper Res* 60:81–113
  170. Peeta S, Mahmassani HS et al (1991) Effectiveness of real-time information strategies in situations of non-recurrent congestion. In: *Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering*, Minneapolis. ASCE, New York
  171. Peeta S, Paz A (2006) Behavior-consistent within-day traffic routing under information provision. In: *IEEE Intelligent Transportation Systems Conference*, Toronto, pp 212–217
  172. Peeta S, Ramos JL (2006) Driver response to variable message signs-based traffic information. *Intell Transp Syst* 153(1):2–10
  173. Peeta S, Ramos JL, Pasupathy R (2000) Content of Variable Message Signs and On-line Driver Behavior. *Transp Res Rec* 1725:102–108
  174. Peeta S, Yang T-H (2000) Stability of Large-scale Dynamic Traffic Networks under On-line Control Strategies. In: *6th International Conference on Applications of Advanced Technologies in Transportation Engineering*, Singapore, paper no. 11 (eProceedings on CD), p 9
  175. Peeta S, Yang T-H (2003) Stability Issues for Dynamic Traffic Assignment. *Automatica* 39(1):21–34
  176. Peeta S, Yu JW (2004) Adaptability of a Hybrid Route Choice Model to Incorporating Driver Behavior Dynamics Under Information Provision. In: *IEEE Transactions On Systems, Man, And Cybernetics Part A: Systems And Humans* 34(2):243–256
  177. Peeta S, Yu JW (2006) Behavior-based consistency-seeking models as deployment alternatives to dynamic traffic assignment models. *Transp Res Part C Emerg Technol* 14(2): 114–138
  178. Peeta S, Zhou C (1999) On-Line Dynamic Update Heuristics for Robust Guidance. In: *International Conference Modeling and Management in Transportation*, Cracow, October 1999
  179. Peeta S, Zhou C (1999) Robustness of the Off-line A Priori Stochastic Dynamic Traffic Assignment Solution for On-Line Operations. *Transp Res Part C: Emerg Technol* 7C(5):281–303
  180. Peeta S, Ziliaskopoulos AK (2001) Foundations of Dynamic Traffic Assignment: The Past, the Present and the Future. *Netw Spat Econ* 1(3–4):233
  181. Rakha H (1990) An Evaluation of the Benefits of User and System Optimised Route Guidance Strategies. In: *Civil Engineering*. Queen's University, Kingston
  182. Rakha H, Ahn K (2004) Integration modeling framework for estimating mobile source emissions. *J transp eng* 130(2): 183–193
  183. Rakha H, Ahn K et al (2004) Development of VT-Micro model for estimating hot stabilized light duty vehicle and truck emissions. *Transp Res Part D Transport and Environment* 9(1):49–74
  184. Rakha H, Arafeh M (2007) Tool for calibrating steady-state traffic stream and car-following models. In: *Transportation Research Board Annual Meeting*, Washington, 22–25 Jan 2008
  185. Rakha H, Crowther B (2002) Comparison of Greenshields, Pipes, and Van Aerde car-following and traffic stream models. *Transp Res Rec* 1802:248–262
  186. Rakha H, Flintsch AM et al (2005) Evaluating alternative truck management strategies along interstate 81. *Transp Res Rec* 1925:76–86
  187. Rakha H, Kang Y-S et al (2001) Estimating vehicle stops at undersaturated and oversaturated fixed-time signalized intersections. *Transp Res Rec* 1776:128–137
  188. Rakha H, Lucic I (2002) Variable power vehicle dynamics model for estimating maximum truck acceleration levels. *J Transp Eng* 128(5):412–419
  189. Rakha H, Lucic I et al (2001) Vehicle dynamics model for predicting maximum truck acceleration levels. *J Transp Eng* 127(5):418–425
  190. Rakha H, Medina A et al (2000) Traffic signal coordination across jurisdictional boundaries: Field evaluation of efficiency, energy, environmental, and safety impacts. *Transp Res Rec* 1727:42–51
  191. Rakha H, Paramahamsan H et al (2005) Comparison of Static Maximum Likelihood Origin-Destination Formulations. *Transportation and Traffic Theory: Flow, Dynamics and Human Interaction*. In: *Proceedings of the 16th International Symposium on Transportation and Traffic Theory (ISTTT16)*, pp 693–716
  192. Rakha H, Pasumarthy P et al (2004) Modeling longitudinal vehicle motion: issues and proposed solutions. In: *Transport Science and Technology Congress*, Athens, Sep 2004
  193. Rakha H, Pasumarthy P et al (2004) The INTEGRATION framework for modeling longitudinal vehicle motion. *TRANSTEC*, Athens
  194. Rakha H, Snare M et al (2004) Vehicle dynamics model for estimating maximum light-duty vehicle acceleration levels. *Transp Res Rec* 1883:40–49

195. Rakha H, Van Aerde M et al (1989) Evaluating the benefits and interactions of route guidance and traffic control strategies using simulation. In: First Vehicle Navigation and Information Systems Conference - VNIS '89, Toronto. IEEE, Piscataway
196. Rakha H, Van Aerde M et al (1998) Construction and calibration of a large-scale microsimulation model of the Salt Lake area. *Transp Res Rec* 1644:93–102
197. Rakha H, Zhang Y (2004) INTEGRATION 2.30 framework for modeling lane-changing behavior in weaving sections. *Transp Res Rec* 1883:140–149
198. Rakha H, Zhang Y (2004) Sensitivity analysis of transit signal priority impacts on operation of a signalized intersection. *J Transp Eng* 130(6):796–804
199. Rakha HA, Van Aerde MW (1996) Comparison of simulation modules of TRANSYT and integration models. *Transp Res Rec* 1566:1–7
200. Ran B, Boyce DE (1996) A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. *Research Part C Emerg Technol* 4C(1):1–12
201. Ran B, Boyce DE (1996) A link-based variational inequality formulation of ideal dynamic user-optimal route choice problem. *Research Part C (Emerging Technologies)* 4C(1):1–11
202. Ran B, Boyce DE, LeBlanc LJ (1993) A new class of instantaneous dynamic user-optimal traffic assignment models. *Oper Res* 41(1):192–202
203. Ran B, Hall RW, Boyce DE (1996) A link-based variational inequality model for dynamic departure time/route choice. *Transp Res Methodol* 30B(1):31–46
204. Ran B, Shimazaki T (1989) A general model and algorithm for the dynamic traffic assignment problems. In: Fifth World Conference on Transport Research, Transport Policy, Management and Technology Towards, Yokohama, 2001
205. Ran B, Shimazaki T (1989) Dynamic user equilibrium traffic assignment for congested transportation networks. In: Fifth World Conference on Transport Research, Yokohama, 1989
206. Randle J (1979) A Convergence Probabilistic Road Assignment Model. *Traffic Eng Control* 11:519–521
207. Richards PI (1956) Shock waves on the highway. *Oper Res* 4:42–51
208. Rilett L, Aerde V (1993) Modeling Route Guidance Using the Integration Model. In: Proceedings of the Pacific Rim Trans Tech Conference, Seattle, 1993 and Proceedings of the ASCE International Conference on Applications of Advanced Technologies in Transportation Engineering. ASCE, New York
209. Rilett L, Van Aerde M (1991) Routing based on anticipated travel times. In: Proceedings of the 2nd International Conference on Applications of Advanced Technologies in Transportation Engineering, Minneapolis. ASCE, New York
210. Rilett LR, Van Aerde M et al (1991) Simulating the TravTek route guidance logic using the integration traffic model. In: Vehicle Navigation & Information Systems Conference Proceedings Part 2 (of 2). Dearborn, 1991. In: Proceedings - Society of Automotive Engineers n P-253, SAE, Warrendale
211. Rilett LR, van Aerde MW (1991) Modelling distributed real-time route guidance strategies in a traffic network that exhibits the Braess paradox. In: Vehicle Navigation & Information Systems Conference Proceedings Part 2 (of 2). Dearborn, 1991. Proceedings - Society of Automotive Engineers n P-253. SAE, Warrendale
212. Roupail NM (1988) Delay Models for Mixed Platoon and Secondary Flows. *J Transp Eng* 114(2):131–152
213. Roupail NM, Akcelik R (1992) Preliminary model of queue interaction at signalised paired intersections. In: Proceedings of the 16th ARRB Conference, Perth, 9–12 November 1992. Congestion Management Proceedings - Conference of the Australian Road Research Board, Australian Road Research Board, Nunawading
214. Schofer AJKFSKJL (1993) Stated preferences for investigating commuters' diversion propensity. *Transportation* 20(2): 107–127
215. Sheffi Y (1985) *Urban Transportation Networks: Equilibrium Analysis with Mathematical Programming Methods*, Prentice Hall, Englewood Cliffs
216. Sheffi Y, Powell W (1981) A Comparison of Stochastic and Deterministic Traffic Assignment over Congested Networks. *Transp Res* 15B:65–88
217. Shen W, Nie Y et al (2006) Path-based System Optimal Dynamic Traffic Assignment Models: Formulations and Solution Methods. In: IEEE Intelligent Transportation Systems Conference. IEEE, Toronto, pp 1298–1303
218. Sherali HD, Arora N, Hobeika AG (1997) Parameter optimization methods for estimating dynamic origin-destination trip tables. *Transp Res Part B Methodol* 31B(2):141–157
219. Sherali HD, Desai J et al (2006) A discrete optimization approach for locating Automatic Vehicle Identification readers for the provision of roadway travel times. *Transp Res Part B* 40:857–871
220. Simon H (1957) Models of Man, Social and Rational. *Adm Sci Q* 2(2)
221. Simon HA (1947) Administrative Behavior. *Am Political Sci Rev* 41(6)
222. Simon HA (1955) A Behavioral Model of Rational Choice. *Q J Econ* 69(1):99–118
223. Sivanandan R, Dion F et al (2003) Effect of Variable-Message Signs in Reducing Railroad Crossing Impacts. *Transp Res Rec* 1844:85–93
224. Smock R (1962) An Iterative Assignment Approach to Capacity-Restraint on Arterial Networks. *Bulleton Highw Res Board* 347:226–257
225. Srinivasan KK, Mahmassani HS (2000) Modeling inertia and compliance mechanisms in route choice behavior under real-time information. *Transp Res Rec* 1725:45–53
226. Srinivasan KK, Mahmassani HS (2000) Modeling inertia and compliance mechanisms in route choice behavior under real-time information. *Transp Res Rec* 1725:45–53
227. Steinberg R, Zangwill WI (1983) The Prevalence of Braess' Paradox. *Transp Sci* 17:301–318
228. Stewart N (1980) Equilibrium versus System-Optimal Flow: Some Examples. *Transp Res* 14A:81–84
229. Talaat H, Abdulhai B (2006) Modeling Driver Psychological Deliberation During Dynamic Route Selection Processes. In: 2006 IEEE Intelligent Transportation Systems Conference, Toronto, pp 695–700
230. Tarko A, Roupail N et al (1993) Overflow delay at a signalized intersection approach influenced by an upstream signal. An analytical investigation. *Transp Res Rec* 1398:82–89
231. Van Aerde M (1985) Modelling of Traffic Flows, Assignment and Queueing in Integrated Freeway/Traffic Signal Networks. In: Civil Engineering. Ph D, University of Waterloo, Waterloo
232. Van Aerde M, Rakha H (1989) Development and Potential of System Optimized Route Guidance Strategies. In: IEEE Ve-



- hicle Navigation and Information Systems Conference. IEEE, Toronto, pp 304–309
233. Van Aerde M, Rakha H (2007) INTEGRATION © Release 2.30 for Windows: User's Guide - vol I: Fundamental Model Features. M Van Aerde & Assoc, Ltd, Blacksburg
  234. Van Aerde M, Rakha H (2007) INTEGRATION © Release 2.30 for Windows: User's Guide - vol II: Advanced Model Features. M Van Aerde & Assoc, Ltd, Blacksburg
  235. Van Aerde M, Rakha H et al (2003) Estimation of Origin-Destination Matrices: Relationship between Practical and Theoretical Considerations. *Transp Res Rec* 1831:122–130
  236. Van Aerde M, Yagar S (1988) Dynamic Integrated Freeway/Traffic Signal Networks: A Routeing-Based Modelling Approach. *Transp Res* 22A(6):445–453
  237. Van Aerde M, Yagar S (1988) Dynamic Integrated Freeway/Traffic Signal Networks: Problems and Proposed Solutions. *Transp Res* 22A(6):435–443
  238. Van Aerde M, Hellinga BR et al (1993) QUEENSOD: A Method for Estimating Time Varying Origin-Destination Demands For Freeway Corridors/Networks. In: 72nd Annual Meeting of the Transportation Research Board, Washington DC, 1993
  239. Van Der Zijpp NJ, De Romph E (1997) A dynamic traffic forecasting application on the Amsterdam beltway. *Int J Forecast* 13:87–103
  240. Van Vliet D (1976) Road Assignment. *Transp Res* 10:137–157
  241. Van Vliet D (1982) SATURN - A Modern Assignment Model. *Traffic Eng Control* 12:578–581
  242. Van Zuylen JH, Willumsen LG (1980) The most likely trip matrix estimated from traffic counts. *Transp Res* 14B:281–293
  243. Walker N, Fain WB et al (1997) Aging and Decision Making: Driving-Related Problem Solving. *J Hum Factors Ergon Soc* 39(3):438–444(7)
  244. Waller ST (2000) Optimization and Control of Stochastic Dynamic Transportation Systems: Formulations, Solution Methodologies, and Computational Experience. Ph D, Northwestern University, Evanston
  245. Waller ST, Ziliaskopoulos AK (2006) A chance-constrained based stochastic dynamic traffic assignment model: Analysis, formulation and solution algorithms. *Transp Res Part C Emerg Technol* 14(6):418–427
  246. Wardrop J (1952) Some Theoretical Aspects of Road Traffic Research. Institute of Civil Engineers, pp 325–362
  247. Webster F (1958) Traffic Signal Settings. HMsSO Road Research Laboratory, London
  248. Webster FV, Cobbe BM (1966) Traffic Signals. HMsSO Road Research Laboratory, London
  249. Wie BW (1991) Dynamic Analysis Of User-Optimized Network Flows With Elastic Travel Demand. *Transp Res Rec* 1328: 81–87
  250. Willumsen LG (1978) Estimation of an O-D matrix from traffic counts: A review. Institute for Transport Studies, Working paper no 99, Leeds University, Leeds
  251. Wilson AG (1970) Entropy in Urban and Regional Modelling. Pion, London
  252. Wu J, Chang G-L (1996) Estimation of time-varying origin-destination distributions with dynamic screenline flows. *Transp Res Part B Methodol* 30B(4):277–290
  253. Yagar S (1971) Dynamic Traffic Assignment by Individual Path Minimization and Queueing. *Transp Res* 5:179–196
  254. Yagar S (1974) Dynamic Traffic Assignment by Individual Path Minimization and Queueing. *Transp Res* 5:179–196
  255. Yagar S (1975) CORQ - A Model for Predicting Flows and Queues in a Road Corridor. *Transp Res* 553:77–87
  256. Yagar S (1976) Measures of the Sensitivity and Effectiveness of the CORQ Traffic Model. *Transp Res Rec* 562:38–48
  257. Yang Q, Ben-Akiva ME (2000) Simulation laboratory for evaluating dynamic traffic management systems. *Transp Res Rec* 1710:122–130
  258. Yang T-H (2001) Deployable Stable Traffic Assignment Models for Control in Dynamic Traffic Networks: A Dynamical Systems Approach. Ph D, Purdue University, West Lafayette
  259. Zhou X, Mahmassani HS (2006) Dynamic origin-destination demand estimation using automatic vehicle identification data. In: *IEEE Transactions on Intell Transp Syst* 7(1):105–114
  260. Zhou Y, Sachse T (1997) A few practical problems on the application of OD-estimation in motorway networks. *TOP* 5(1):61–80
  261. Ziliaskopoulos A, Wardell W (2000) Intermodal optimum path algorithm for multimodal networks with dynamic arc travel times and switching delays. *Eur J Oper Res* 125(3):486–502
  262. Ziliaskopoulos AK (2000) A linear programming model for the single destination system optimum dynamic traffic assignment problem. *Transp Sci* 34(1):37–49
  263. Ziliaskopoulos AK, Waller ST (2000) An Internet-based geographic information system that integrates data, models and users for transportation applications. *Transp Res Part C Emerg Technol* 8C:1–6

## Traffic Networks, Optimization and Control of Urban

NATHAN H. GARTNER, CHRONIS STAMATIADIS  
University of Massachusetts, Lowell, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Traffic Signal Control](#)

[The Basic Model – Single Intersection](#)

[Arterial Systems – Progression Models](#)

[Delay-Based Models – Cyclic Flow Profiles](#)

[Demand-Responsive Models](#)

[Multi-Level Traffic Control Strategies](#)

[Control System Performance – Analysis of Complexity](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Assignment** (traffic) The allocation of traffic volumes to routes.

**Arterial** A road with two or more intersections.

**Bandwidth** Time interval in an arterial progressive system during which vehicles can travel unimpeded.



- Controller** A device which controls the sequence and duration of indications displayed by traffic signals.
- Coordination** The establishment of a definite timing relationship between adjacent traffic signals.
- Cycle time** The time required for one complete sequence of signal indications.
- Delay** (traffic) The time lost by vehicle(s) due to traffic interference or control devices.
- Detector** (traffic) A device to detect the presence or passage of a vehicle in the roadway.
- Flow** (traffic) The rate at which vehicles cross a given line on the road; the number of vehicles per unit of time.
- Headway** The time interval between successive vehicle crossings of a given line on the road.
- Intelligent transportation system** The addition of information and communications technology to enhance performance of transport infrastructure and of vehicles.
- Offset** The time difference between the start of green at one intersection and the start of green at a subsequent intersection, or with respect to a system time base.
- Phase** The time interval of a cycle time allocated to any combination of movements receiving the right-of-way simultaneously.
- Performance index/disutility index** A numerical quantity used to measure the performance of a traffic in a network in an optimization or a simulation model; also objective function.
- Platoon** A tight group of vehicles traveling along an arterial.
- Progression** A timing plan enabling a platoon to travel unimpeded along an arterial.
- Saturation flow** The rate of flow at which vehicles are discharged from a queue stopped at a signal; the maximum rate of flow on a street.
- Synchronization** A condition under which traffic signals operate with the same cycle time.
- Volume** (traffic) The number of vehicles crossing a line on the road per hour; traffic flow.

## Definition of the Subject

Much of our economic and social life is dependent on communication and transportation systems. Travel in urban networks represents an interaction between the demand for transportation and the supply of transportation means and facilities which includes the vehicles, the road networks, as well as the control systems that govern them. The critical elements within these systems are junctions, or intersections, which are controlled by means of traffic signals. Public authorities, which are responsible for the

operation of those systems, develop control policies that have a dominant impact on the quality of travel and the level of service provided by the networks. An understanding of the basic models involved in urban traffic control is essential for the development of optimal operating policies that would lead to the effective performance of urban traffic networks.

## Introduction

The critical elements within an urban traffic network are the junctions, or the intersections, many of which are controlled by traffic signals. Therefore, traffic control signals are a key determinant for the efficient operation of urban street networks and an essential element of Intelligent Transportation Systems.

In this chapter we focus on the complex array of urban street networks and their control by means of traffic signals. We start by delineating the basic models involved in traffic flow at single intersections. We then present control models on arterial streets and networks, both progression models and delay-based models. Next we introduce demand-responsive systems and we conclude by outlining a strategy for multi-level control and an analysis of the complexity involved in the performance of the different systems.

## Traffic Signal Control

Traffic control signals are defined as “power-operated traffic devices which alternately direct traffic to stop and to proceed.” More specifically, traffic signals are used to control the assignment of right-of-way at locations where conflicts exist or where passive devices, such as signs and markings, do not provide the necessary flexibility of control to properly move traffic in a safe and efficient manner.

Traffic control signals are usually described as either *pre-timed* or *traffic actuated*. Each type may be used in either an independent (isolated) or interconnected (system) application. *Pre-timed control* assigns the right of way at an intersection according to a predetermined schedule. The length of the time interval for each signal indication in the cycle is fixed, based on historic traffic patterns. *Actuated control* differs from pre-timed in that signal phases are not of fixed length. Through the use of vehicle detectors, this type of control assigns the right of way on the basis of actual traffic conditions (demands) within given limitations. The full range of actuated control capabilities depends on the type of equipment employed and the operational requirements.

Traffic signal timings are divided into time phases. A signal phase may be defined as that part of the cycle

length allocated to a traffic movement receiving the right of way or to any combination of traffic movements receiving the right of way simultaneously. A traffic movement is a single vehicular movement, a single pedestrian movement, or a combination of vehicular and pedestrian movements. The sum of all traffic phases at an intersection is equal to the cycle length.

### The Basic Model – Single Intersection

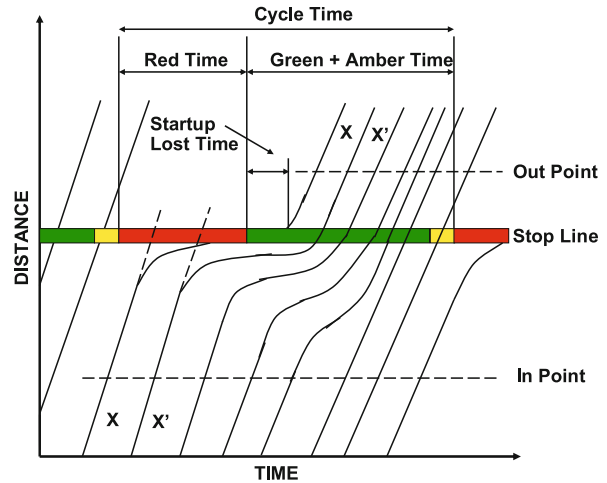
The material in this section is based on models developed by F.V. Webster at the British Road Research Laboratory [1] for estimating delays to vehicles at fixed-time traffic signals and for computing optimum control settings for such signals. The methods developed in this study can be applied both to fixed-time signals and to vehicle-actuation.

Webster's classical results provide a more definite basis for setting fixed-time signals than any which have so far been published. They are also applicable to those vehicle-actuated signals where the green periods owing to heavy traffic demands are frequently running to maximum, giving in effect fixed-time control. In addition, linked systems of traffic signals often work on a fixed cycle which has been set to meet the requirements of the main intersection of the system and the desired speed of traffic along the road.

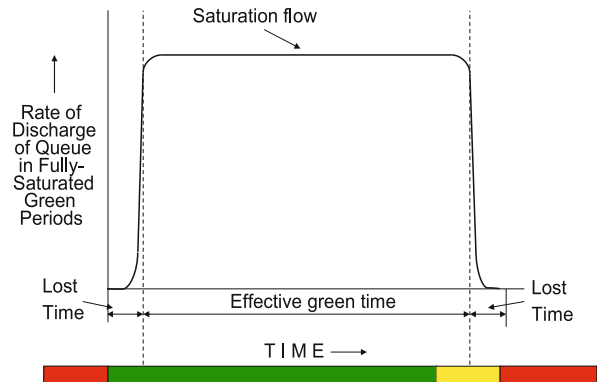
### Flow Model

The first step in the analysis of traffic control signals is to determine performance, most often measured by the delay caused to traffic. Traffic may be considered as arriving at random provided that the point at which it is observed is some distance from a disturbing factor such as a controlled intersection. Total delay to traffic is computed by the area under the queue-length curve. The average delay is obtained by dividing the total delay by the number of arrivals. This is illustrated in Fig. 4 below.

When the signal is red, arriving traffic queues at the signal stop line. After a green period commences, a certain time elapses while vehicles are accelerating to normal running speed, but after a few seconds the queue discharges at a constant rate, called the saturation flow. This is illustrated in the time-space (or distance) diagram of Fig. 1. If there is still a queue at the end of the green period, some vehicles will make use of the yellow interval to cross the intersection. In these circumstances traffic moves on both green and yellow signals but the discharge rate is less than the saturation flow both at the beginning and at the end of the right-of-way period, as shown in Fig. 2. The green and yellow together ( $k + a$ ) may be replaced by an 'effective' green ( $g$ ) and a 'lost' time ( $l$ ), such that the product of the



Traffic Networks, Optimization and Control of Urban, Figure 1  
Distance/time diagram for flow through intersection



Traffic Networks, Optimization and Control of Urban, Figure 2  
Variation of discharge rate of queue with time in a fully-saturated green period

effective green and the saturation flow is equal to the correct number of vehicles (say,  $b$ ) discharged from the queue on the average in a saturated green period (i. e. a green period during which the queue never clears).

Thus,  $k + a = g + 1$  and  $b = g \cdot s$ , where  $s$  is the saturation flow rate.

In the basic model presented in this section the saturation flow is assumed to be constant. The signal sequence on any approach is thus reduced to an 'effective' green period and a 'red' period which comprises all the times when traffic cannot run, i. e. red periods plus 'lost' time. The delay to traffic on a single approach to an intersection controlled by fixed-time traffic signals is computed over a range of values of green times, cycle times, traffic flows and saturation flows covering most practical conditions.

### Signalized Intersections – Fluid Traffic Model

There are several models that may be used in the investigation of queues and delays at signalized intersections. In this section a continuum (i. e., deterministic) or fluid model is first considered for which basic measures of queue length and delay are developed [2]. In the next section the stochastic nature of traffic is taken into consideration. In estimating delay at intersections, traffic flow is considered to consist of identical passenger car units (pcu's). A truck, for example, may be considered as 1.5 or 2 equivalent pcu's and a turning vehicle may also be assigned some value depending on the type of maneuver that is made.

The following notation is used:

- $c$  = the cycle time (sec);
- $g$  = the effective green time (sec);
- $r$  = the effective red time (sec);
- $q$  = the average arrival rate of traffic on the approach (pcu's/sec);
- $s$  = the saturation flow on the approach (pcu's/sec);
- $d$  = the average delay to a pcu's on the approach (sec);
- $Q_0$  = the overflow (pcu's);
- $\lambda = g/c$  (the proportion of the cycle that is effectively green);
- $y = q/s$  (the ratio of average arrival rate to saturation flow); and
- $x = q \cdot c/g \cdot s$  (the ratio of average number of arrivals/cycle to the maximum number of departures/cycle).

Thus,  $r + g = c$  and  $\lambda \cdot x = y$ . The ratio  $x$  is called the degree of saturation of the approach and  $y$  is called the flow ratio of the approach.

The effective green time is the portion of the cycle time during which pcu's are assumed to pass the signal at a constant rate  $s$ , provided vehicles are waiting on the approach. Greenshields [3] observed that the times to cross the signal stop line are 3.8, 3.1, 2.7, 2.4, 2.2, 2.1, 2.1 seconds for the consecutive queued vehicles. The cumulative time for a queue of  $n$  stopped vehicles to pass a signal can then be given by:

$$\text{Cumulative time} = 14.2 + 2.1(n - 5) \text{ sec} \quad \text{for } n \geq 5.$$

Had all the vehicles departed at the saturation rate  $s = 1/2.1 \text{ veh/sec} = 1714 \text{ veh/hr}$ , the first five vehicles would have required 10.5 sec; that is, the effective green is the signal green time less 3.7 sec. In most studies it is also assumed that a waiting queue of vehicles will take advantage of the yellow clearance interval, although the effective green time may be adjusted to reflect particular operating conditions.

Delay can be calculated by illustrating the meaning of arrival time and departure time for a pcu on an approach. They are demonstrated by reference to Fig. 3, in which distance-time curves are plotted for each of four vehicles. AB represents the passage of an un-delayed vehicle, where the line PQ represents the stopline at which the first vehicle waits when there is a queue. CDEF represents the trajectory of the first vehicle that is delayed by a signal. The straight portions of CD and EF are parallel to AB and projected to meet PQ and X and Y so that the length XY is the delay to the first vehicle. Similarly,  $X'Y'$  and  $X''Y''$  represent the delays for the next two vehicles.  $XX'$  and  $X'X''$  represent the arrival headways.

### Continuum Model for a Pre-timed Signal

Representation of a continuum model of traffic flow at a signalized intersection is given in Fig. 4. The vertical axis represents the number of vehicle arrivals of the stop line and the horizontal axis the time  $t$ . The figure illustrates the behavior when the capacity of the green interval exceeds the arrival during the green + red time i. e., an under-saturated period. The top figure illustrates the queue length as a function of time ( $t$  is the queue clearance time, measured from the start of green). The bottom figure illustrates the cumulative arrival and departure functions. In Fig. 4 the vertical distance  $ca$  represents the number of vehicles that have accumulated since the signal entered the red phase. The horizontal distance  $ab$  represents the total time from arrival to departure for any given vehicle. The shaded areas, in each figure, represent the total vehicle delay.

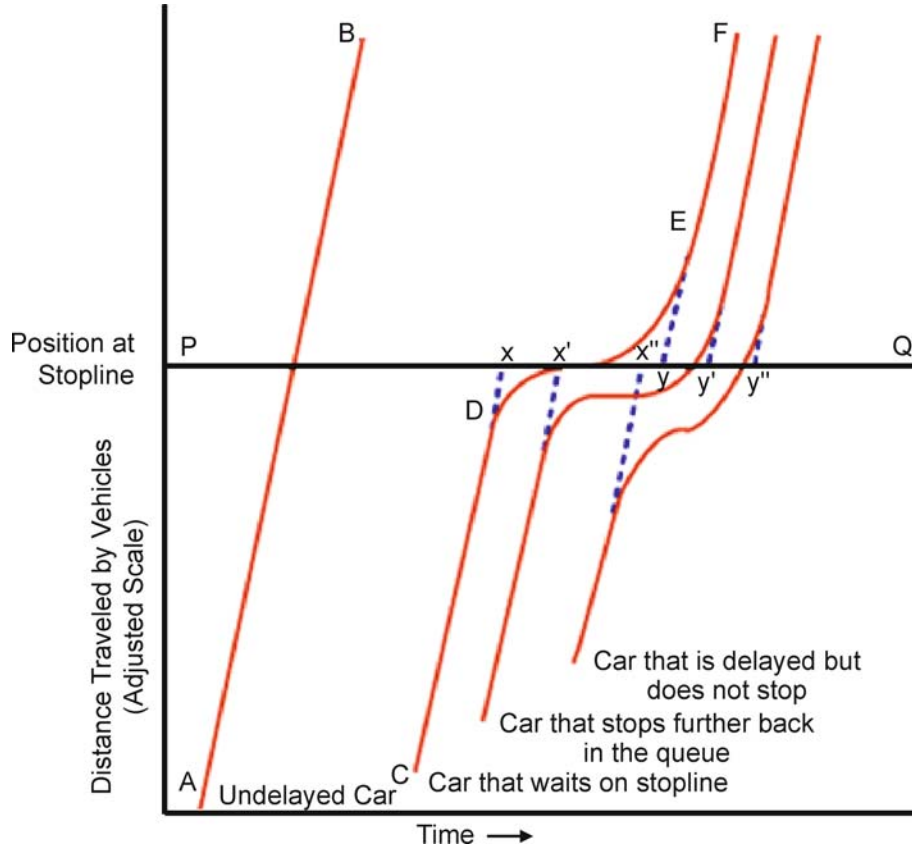
The following measures of queue behavior are developed:

1. Time after start of green that queue is dissipated ( $t_0$ );
2. Proportion of cycle with queue ( $P_q$ );
3. Proportion of vehicles stopped ( $P_s$ );
4. Maximum number of vehicles in queue ( $Q_m$ );
5. Average number of vehicles in queue ( $\bar{Q}$ );
6. Total vehicle-time of delay per cycle ( $D$ );
7. Average individual vehicle delay ( $d$ );
8. Maximum individual vehicular delay ( $d_m$ );

The formulas for these measures are developed from simple geometric relationships:

1. For any given cycle it is evident that at time  $t_0$  after the start of green, the total number of arrivals must equal the number discharged:  $q \cdot (r + t_0) = s \cdot t_0$ . Letting  $y = q/s$  we obtain:

$$t_0 = y \cdot r / (1 - y).$$



**Traffic Networks, Optimization and Control of Urban, Figure 3**  
Vehicle arrival-departure times and delay of the stop-line

2. Proportion of vehicles stopped is equal to queue time/ cycle length

$$P_q = (r + t_0)/c.$$

3. Proportion of vehicles stopped is equal to vehicles stopped/total vehicles per cycle

$$P_s \cdot q \cdot (r + t_0)/q \cdot (r + g) = t_0/(y \cdot c).$$

4. The maximum number of vehicles in queue can be seen by inspection to be the height of the triangle at  $t = r$ , i. e., at the end of the red period.

$$Q_m = q \cdot r.$$

5. The average number of vehicles in the queue, over the total length of cycle  $c$  is

$$\bar{Q} = \frac{(qr/2)r + (qr/2)t_0 + o(g - t_0)}{r + t_0 + g - t_0}.$$

which yields:  $\bar{Q} = (r + t_0) \cdot q \cdot r/(2 \cdot c)$ .

6. The total vehicle-time of delay during the cycle is given by the area of the triangle

$$D = (q \cdot r/2)(r + t_0) = (q \cdot r/2)(r/(1 - y)) \\ = q \cdot r^2/(2(1 - y)).$$

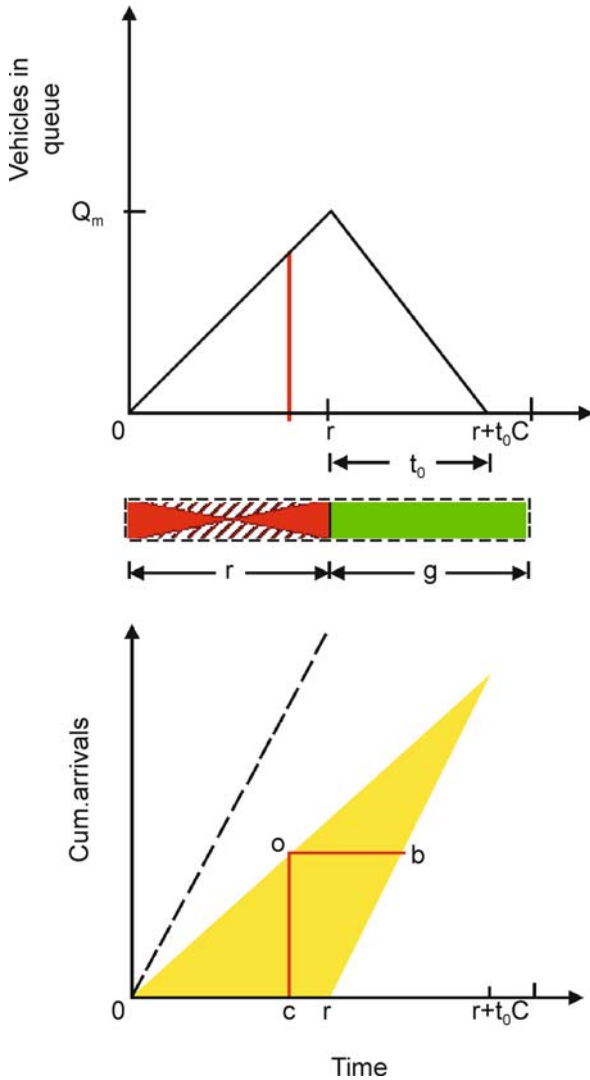
7. The average individual delay is given by dividing the total delay by the number of vehicles, or

$$d = \left[ \frac{qr^2}{2(1 - y)} \right] \frac{1}{qc} = \frac{r^2}{2c(1 - y)}.$$

8. The maximum individual vehicular delay will be seen from Fig. 4 to be

$$d_m = r.$$

If the possible departures during the cycle  $sg$  (i. e., the capacity flow) are less than the arrivals  $qc$ , the queue grows with each successive cycle and the foregoing formulas are no longer applicable. We have, then, over-saturation.



**Traffic Networks, Optimization and Control of Urban, Figure 4**  
Cumulative arrival-departure diagrams and queuing at a signal  
(continuum model)

### Average Delay per Vehicle

Webster found that the results of his computations using a stochastic simulation model could be expressed to a close approximation by the equation

$$d = \frac{c(1-\lambda)^2}{2(1-\lambda x)} + \frac{x^2}{2q(1-x)} - 0.65 \left( \frac{c}{q^2} \right)^{\frac{1}{3}} x^{(2+5\lambda)} \quad (1)$$

where:

$d$  = average delay per vehicle on the particular arm of the intersection

$c$  = cycle time

$\lambda$  = proportion of the cycle which is effectively green for the phase under consideration (i. e.  $g/c$ )

$q$  = flow

$s$  = saturation flow

$x$  = the degree of saturation – this is the ratio of the actual flow to the maximum flow which can be passed through the intersection from this arm, and is given by:  $x = q/\lambda s$ .

An estimate of the average delay per vehicle is required to evaluate the performance of traffic control signals. To enable the expected delay to be estimated more easily, Eq. (1) can be rewritten as

$$d = cA + \frac{B}{q} - C \quad (2)$$

where  $A = (1-\lambda)^2/2(1-\lambda x)$ ,  $B = x^2/2(1-x)$  and  $C$  is a correction term.

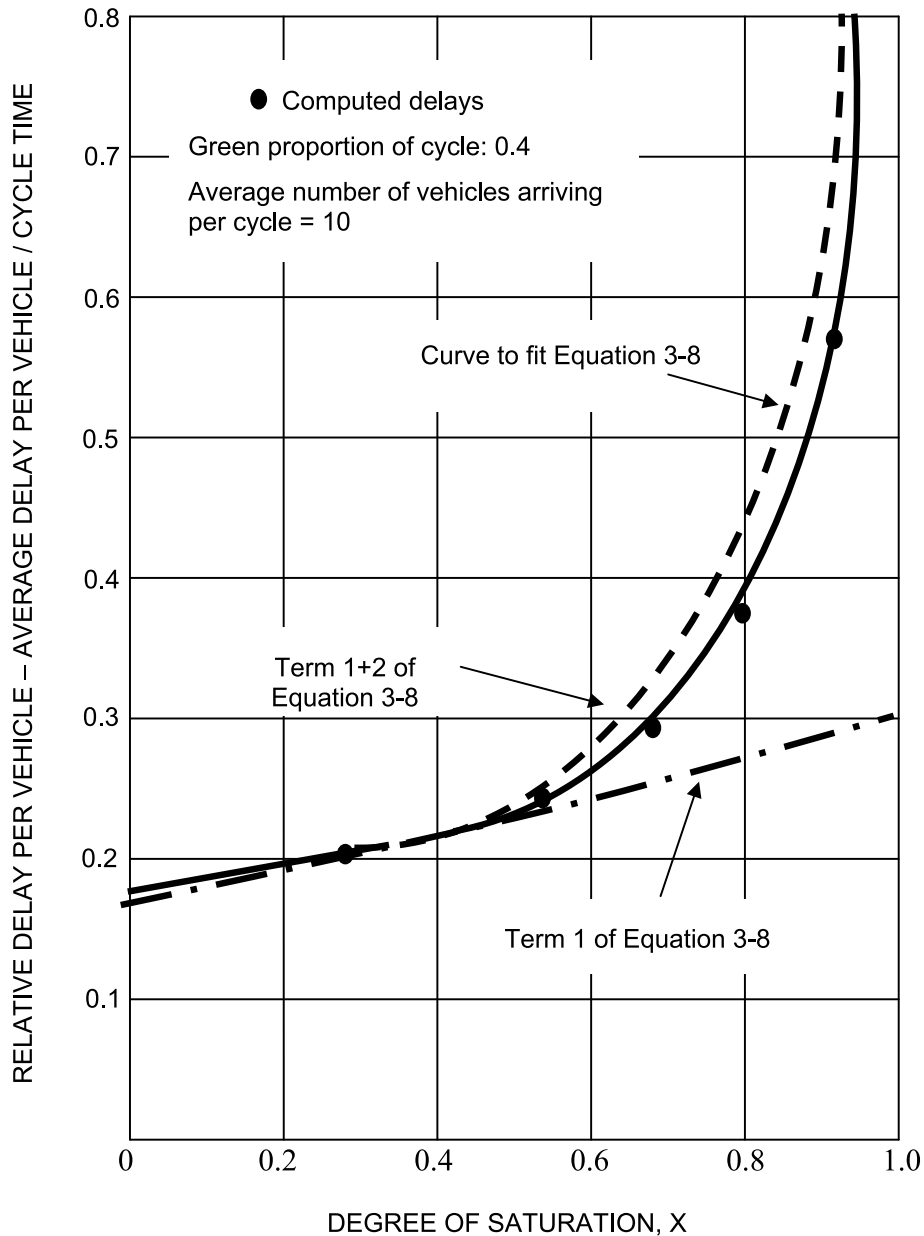
The first two terms in Eq. (1) have a theoretical underpinning but the last term is purely empirical. The first term is the expression for the delay when traffic can be considered to be arriving at a uniform rate. It is identical to the formula in Item No. 7 above. Although the agreement between computed delays and those derived from this term is fairly good at low flows, as shown by Fig. 5, it is not so at higher values where the computed delays, owing to the random nature of the arrivals, are far in excess of values calculated from this term only. The second term makes some allowance for the random nature of the arrivals. It is an expression for the delay experienced by vehicles arriving randomly in time at a 'bottleneck', queuing up, and leaving at constant intervals. In queuing terminology it is the average delay of an M/D/1 queue. The addition of the second term improves the correspondence with Eq. (1), but overestimates it. To avoid the need to calculate the correction factor  $C$ , a suitable approximation can be obtained by one of the following:

$$d = \left[ cA + \frac{B}{q} \right] \frac{100 - P}{100} \quad (3)$$

where (a)  $P = 10\%$ , or (b)  $P = 15x$ , since the value of the correction term  $C$  is generally in the range 5 to 15 percent of  $d$ .

The 'lost time' illustrated in Fig. 2 plays an important role in determining the efficiency of the traffic signal operation. Information available on lost time suggests a value of about 4.5 seconds per phase plus any all-red periods. This is composed of 2.5 sec start-up lost time plus half the typical 4 sec yellow time. Since lost time depends on gradients, type of traffic, etc. its value will vary from site to site and even at the same site will probably vary with time of day.





**Traffic Networks, Optimization and Control of Urban, Figure 5**  
Typical fixed-time delay curve

### Optimum Settings of Fixed-Time Signals

In general, all approaches belonging to the same phase will have the same green period even though the traffic requirements of the approaches may be different. To simplify calculations, and with negligible loss of accuracy, each phase can be represented by one approach only – the one with the highest ratio of flow to saturation flow. Let this ratio be

denoted by the symbol  $y$ . In deriving the optimum green times and cycle time the empirical correction term of the delay equation is neglected since calculations show that its variation with respect to these quantities is slight.

**Green Times** Least delay to traffic is obtained when the green periods of the phases are in proportion to the corresponding ratios of flow to saturation flow, assuming this

ratio to be the same for all arms of the same phase. This division of the cycle time makes the capacity of the phases proportional to the average flows of the phases. That this is approximately the best division of the cycle time is shown by the examples given below.

The total delay experienced by vehicles at an intersection was calculated using the approximate form of the delay Eq. (2). The lost time was assumed to be 10 seconds per cycle and delays were calculated for a variety of cycle times and ratios of the effective green times. A typical result is shown in Fig. 6, where the ratio of the  $y$  values is 2.0. It can be seen that the best ratios of the effective green times is between 1.88 and 2.17 over a range of cycle times of 35 to 80 seconds.

**Cycle Time** In deriving an expression for the optimum cycle time it is assumed that the effective green times of the phases are in the ratio of their respective  $y$  values. Equation (3) representing the delay to traffic on one arm of an intersection is modified to cover the general case of an intersection with  $n$  phases, and the modified equation is differentiated with respect to cycle time to determine the value of cycle time which gives the least delay of all traffic using the intersection.

The derivation of the optimum cycle-time formula is given in [1]. This formula is too complicated for most purposes and a simple approximation is therefore derived as follows:

$$c_0 = \frac{1.5L + 5}{1 - Y} \text{ sec} \quad (4)$$

where  $Y$  is the sum of the  $y$  values and refers to the intersection as a whole and  $L$  is the total lost time per cycle in seconds. The lost time can be expressed by:  $L = nl + R$ .

Where:

- $n$  is the number of phases
- $l$  is the average lost time per phase (excluding any all-red periods)
- $R$  is the time during each cycle when all signals display red (including red-with-amber) simultaneously.

The possible aspect scheduling for a 2-phase signal is shown in Fig. 7.

Several examples of hypothetical intersections (including some fairly extreme cases) have been studied to show the effect of changes in cycle time on the delay. For a symmetrical intersection, values of delay have been computed by Eq. (1) and are shown in Fig. 8 plotted against the cycle length for several values of the total flow entering the intersection. The cycle time giving the least delay in each case is approximately twice the least possible cycle which will just allow the traffic to pass through (although with

long delays). The latter is called the minimum cycle, and is the vertical asymptote to the delay/cycle-time curve. When traffic is of a truly random character the minimum cycle is associated with infinite delay. For uniform flow it is the cycle such that all traffic arriving during the cycle can just be discharged during the green. We can approximate the optimum cycle as follows:

$$c_{\text{opt}} \cong 2c_{\text{min}} = \frac{2L}{1 - Y}$$

From graphs such as Fig. 8, it was found that in most practical cases the delay for cycle times within the range  $\frac{3}{4}$  to  $1\frac{1}{2}$  times the optimum value is never more than 10 to 20 per cent greater than that given by the optimum cycle. This fact can be used in deducing a compromise cycle time when the level of flow varies considerably throughout the day. It would be better either to change the cycle time to take account of this, or, as is more common, to use vehicle-actuated signals. However, for a single setting of fixed-time signals the simple approximate method outlined below may be used.

1. Calculate the optimum cycle for each hour of the day when the traffic flow is medium or heavy, e.g. between the hours of 8 a.m. and 7 p.m. and average over the day.
2. Evaluate three-quarters of the optimum cycle calculated for the heaviest peak hour.
3. Select whichever is greater for the cycle time.

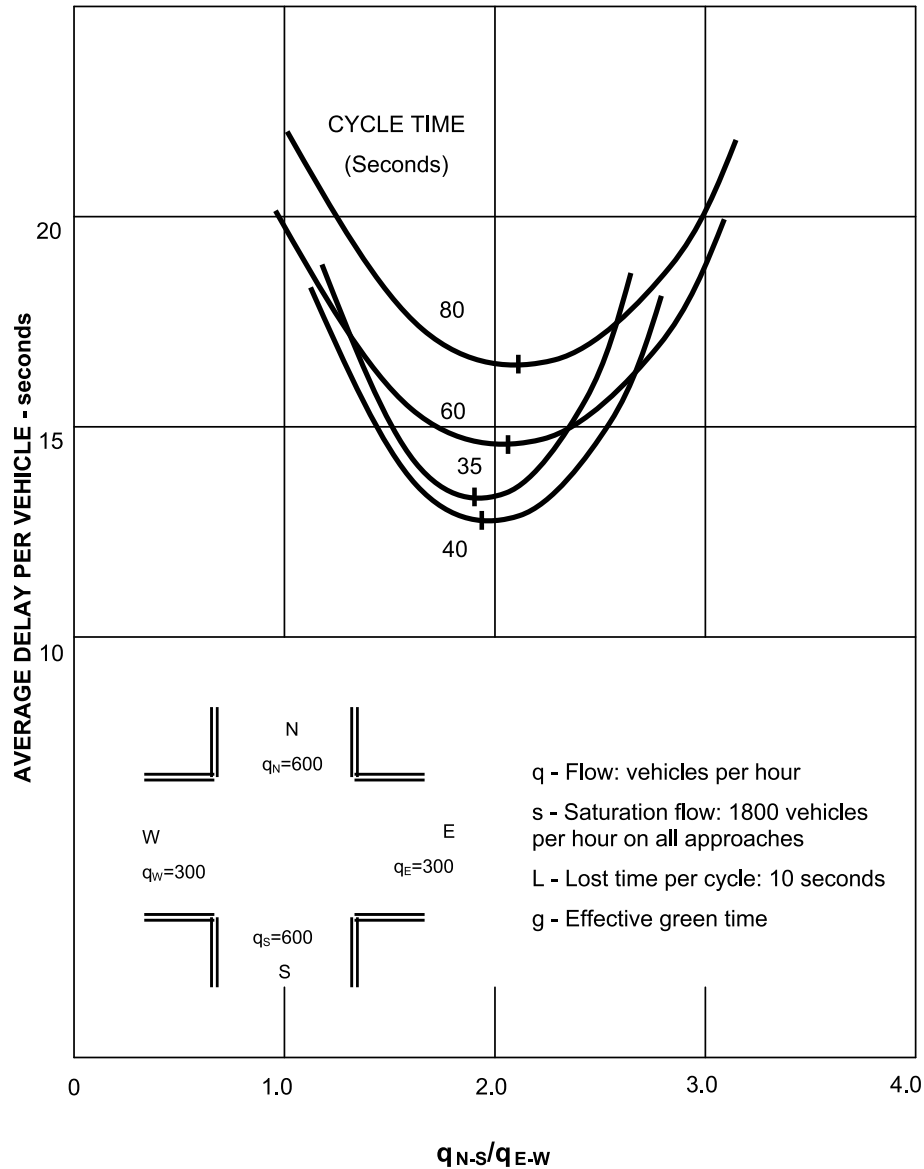
It is suggested as a reasonable procedure that the division of the available green time ( $c_0 - L$ ) should be in proportion to the average  $y$  values for peak periods only, i.e.:

$$\frac{g_1}{g_2} = \frac{(y_1)_{\text{PEAK}}}{(y_2)_{\text{PEAK}}}$$

where  $(y_1)_{\text{PEAK}}$  is the average  $y$  value during the peak periods for phase 1 and  $(y_2)_{\text{PEAK}}$  that for phase 2.

### Arterial Systems – Progression Models

An arterial is a road consisting of two or more intersections. A *signal system* is defined as having two or more individual signal installations which are linked together for coordination purposes. To obtain coordination all signals must operate with the same (common) cycle length, i.e., the signals must be synchronized. In rare instances, some intersections within the system may operate at double or one-half the cycle length of the system in which case there is half- (or double-) synchronization. It is desirable that the arterial for which coordination is being provided have a green plus yellow interval equivalent to at least 50% of



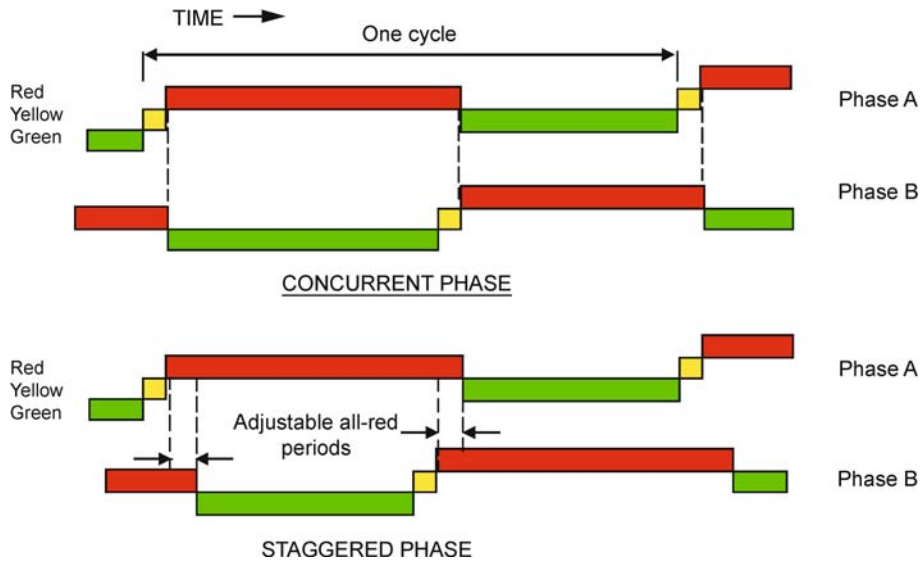
**Traffic Networks, Optimization and Control of Urban, Figure 6**  
Effect on delay of the variation of the ratio of green periods

the cycle length. Two intersecting systems form an *open network* whenever they have only one intersection in common. A *closed network* is formed whenever there are three or more common intersections. A grid network of  $m \times n$  arterials is a closed network with  $m \times n$  common intersections. Closed networks are characterized by having *closed loops*.

The predominant flow of traffic in urban street networks is along arterial streets. Optimal control of the traffic signals along these arteries is essential for the effective op-

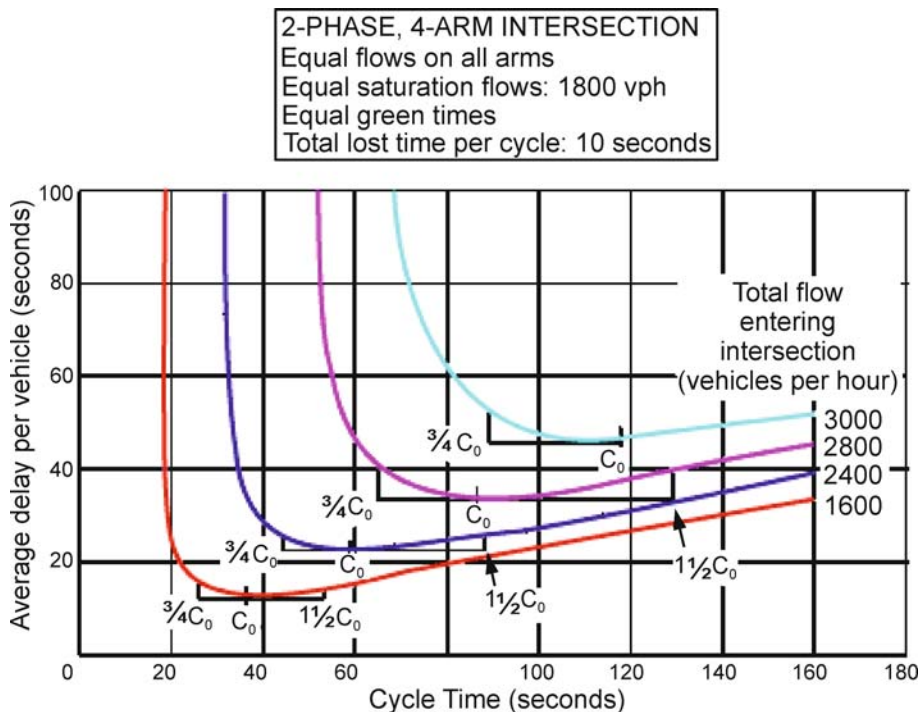
eration of the grid network. Coordination of traffic lights along arterial streets provides numerous advantages:

1. A higher level of traffic service is provided in terms of higher overall speed and reduced number of stops.
2. Traffic flows more smoothly, often with an improvement in capacity due to reduced headways.
3. Vehicle speeds are more uniform because there is no incentive to travel at excessively high speeds to reach a signalized intersection within a green interval that is



Traffic Networks, Optimization and Control of Urban, Figure 7

Possible phase aspect scheduling at a two-phase intersection



Traffic Networks, Optimization and Control of Urban, Figure 8

Effect on delay of variation of the cycle length

- not in step. Also, the slow driver is encouraged to speed up in order to avoid having to stop for a red light.
- There are fewer accidents because the platoons of vehicles arrive at each signal when it is green, thereby reducing the possibility of red signal violations and rear-end collisions.
  - Greater obedience to the signal commands is obtained from both motorists and pedestrians. The motorist tries

to keep within the green interval and the pedestrian stays at the curb because the vehicles are more tightly spaced.

6. Through traffic tends to stay on the arterial streets instead of diverting onto parallel minor streets in search of alternative routes.

There are two basic approaches to computing arterial signal timings: (a) maximize the bandwidth of the progression, or (b) minimize overall delays and stops.

Arterial progression schemes are widely used for signal coordination. The conceptual basis for progression design is that traffic signals tend to group vehicles into a “platoon” with more uniform headways than would otherwise occur. The platooning effect is most prevalent on major streets with signalized intersections at frequent intervals. It is desirable, in these circumstances, to encourage platooning so that continuous movement (or progression) of vehicle platoons through successive traffic lights can be maintained. Signal timings, in this case, are designed to maximize the width of continuous green bands in both directions along the artery at the expected or recommended speed of travel. Although maximizing the bandwidth is only a surrogate performance measure, it provides significant benefits in increased travel speeds and smoothness of traffic flow and is the preferred approach in many countries. Such systems operate best when main-street flow is predominantly through traffic and when the number of vehicles turning onto the main street is small compared with the through volume. The second approach uses direct performance models to minimize delay, stops or other measures of disutility. Due to the complexity of these models and the high degree of non-linearity involved, they are generally not amenable to solution by mathematical programming methods and heuristic approaches have to be used. As a consequence, the resulting solutions are non-optimal. In this section we concentrate on models of the first kind.

The first optimization model used for arterial bandwidth maximization was developed by Morgan and Little using a combinatorial optimization scheme [4]. A more advanced model, using mixed-integer linear programming, was later formulated by Little [5]. Advances in optimization techniques and computational capabilities have steadily increased the sophistication and versatility of the traffic signal optimization models. We introduce, first, single arterial bandwidth optimization models using both uniform and variable-width bandwidth progressions. These models are then extended to the optimization of arterial grid networks. Heuristic decomposition approaches are developed to speed-up the computation and to make it

suitable for on-line implementation as well as for combination with traffic assignment and routing models.

The models presented below use mixed-integer linear programming (MILP) for optimization. They are presented with progressively increasing complexity. MILP-1 is the basic uniform-width bandwidth maximization model for a single arterial. MILP-2 introduces a multi-band, multi-weight approach that generates variable-width two-way progression bands along the arterial. MILP-3 extends the previous models to a grid network of intersecting arterials.

### The Basic Bandwidth Maximization Problem

The geometric relations for the uniform bandwidth model are shown in Fig. 9. Consider an arterial having  $n$  signalized intersections. Let  $S_i$  denote the  $i$ th signal on the arterial and  $L_i$  denote the  $i$ th link (between signals  $i$  and  $i + 1$ ) of the arterial, with  $i = 1, \dots, n_j$ . All time variables are defined in units of the cycle time. The following variables are defined:

$C$  = cycle time (sec);

$b(\bar{b})$  = outbound (inbound) bandwidth;

$r_i(\bar{r}_i)$  = outbound (inbound) red time at  $S_i$ ;

$w_i(\bar{w}_i)$  = interference variables, time from right (left) side of red at  $S_i$  to left (right) side of outbound (inbound) green band;

$t_{hi}(\bar{t}_{hi})$  = travel time from  $S_i$  to  $S_h$  in the outbound (inbound) direction;

$d_{hi}(\bar{d}_{hi})$  = distance from  $S_i$  to  $S_h$  in the outbound (inbound) direction;

$\phi_{hi}(\bar{\phi}_{hi})$  = internode offsets, time from the center of the outbound (inbound) red at  $S_h$  to the center of the outbound (inbound) red at  $S_i$ ;

$\Delta_i$  = directional node phase shift, time from center of  $\bar{r}_i$  to nearest center of  $r_i$ ;

$\tau_i(\bar{\tau}_i)$  = queue clearance time for advancement of outbound (inbound) bandwidth at  $S_i$  to clear turning-in traffic before arrival of main-street platoon;

$V_i(\bar{V}_i)$  = outbound (inbound) progression speed on link  $L_i$  (m/sec).

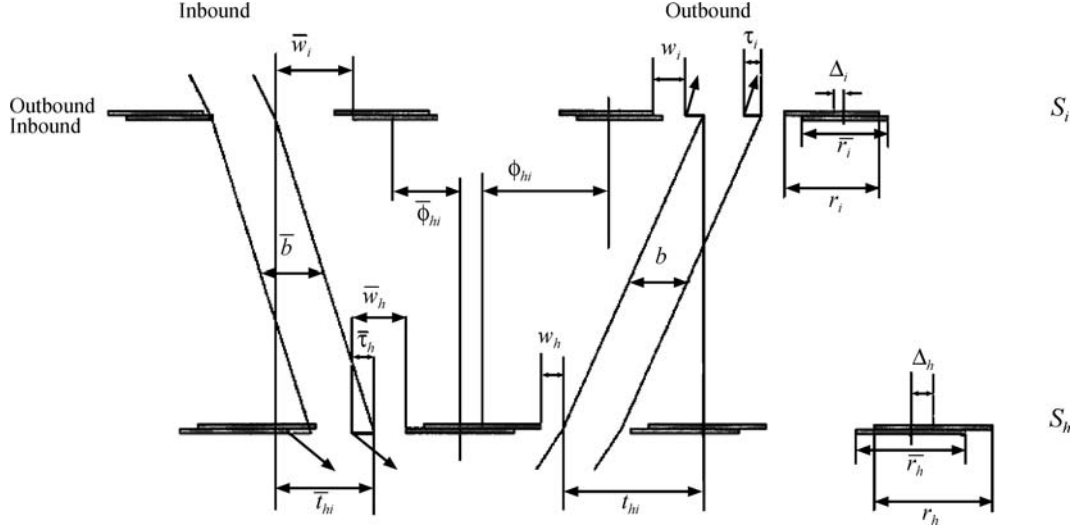
In the case of uniform bandwidths the objective function has the following form:

$$\text{Maximize } b + k \cdot \bar{b} \quad (5)$$

where  $k$  is the target ratio of inbound to outbound bandwidth taken as the ratio of total inbound to total outbound volumes along the arterial.

The first set of constraints that we introduce are the *directional interference constraints*. These constraints make sure that the progression bands use only the available





Traffic Networks, Optimization and Control of Urban, Figure 9

Time space diagram for MILP-1

green time and do not infringe upon any of the red times. From Fig. 9 we see that:

$$w_i + b \leq 1 - r_i \quad (6a)$$

$$\bar{w}_i + \bar{b} \leq 1 - \bar{r}_i. \quad (6b)$$

Next, we calculate the *arterial-loop integer constraint*. This constraint is due to the fact that the signals of the arterial are synchronized, i. e., they operate in a common cycle time. If we start at the center of the outbound red at  $S_j$  and proceed along a loop consisting of the following points:

- Center of outbound red at  $S_i$ —
- Center of inbound red at  $S_i$ —
- Center of inbound red at  $S_h$ —
- Center of outbound red at  $S_h$ ,

we must end up at a point that is removed an integral number of cycle times from the point of departure. Summing algebraically the appropriate internode offsets and the directional node phase shift along the loop, we obtain

$$\phi_{hi} + \bar{\phi}_{hi} + \Delta_h - \Delta_i = v_{hi} \quad (7)$$

where  $v_{ij}$  is the corresponding loop integer variable.

From Fig. 9, we observe the following identities:

$$\phi_{hi} + \frac{r_i}{2} + w_i + \tau_i = \frac{r_h}{2} + w_h + t_{hi} \quad (8a)$$

and:

$$\bar{\phi}_{hi} + \frac{\bar{r}_i}{2} + \bar{w}_i + \bar{\tau}_i = \frac{\bar{r}_h}{2} + \bar{w}_h + \bar{t}_{hi}. \quad (8b)$$

Substituting (8b) into (8a) to eliminate  $\phi$  and  $\bar{\phi}$  gives:

$$t_{hi} + \bar{t}_{hi} + \frac{r_h + \bar{r}_h}{2} + (w_h + \bar{w}_h) - \frac{r_i + \bar{r}_i}{2} - (w_i + \bar{w}_i) - (\tau_i + \bar{\tau}_h) + \Delta_h - \Delta_i = v_{hi}. \quad (9)$$

So far we have assumed that  $S_i$  follows  $S_h$  in the outbound direction, but this restriction is not necessary. Let  $x$  represent any of the variables  $t, \bar{t}, v, \phi, \bar{\phi}$  then the following relationship is satisfied:

$$x_{hj} = x_{hi} + x_{ij}$$

whence, by setting  $h = j$ , we obtain  $x_{hi} = -x_{ih}$ . Thus Eqs. (8a) and (8b) hold for arbitrary  $S_h$  and  $S_i$ .

To simplify notation we number the signals sequentially from 1 to  $n$  in the outbound direction, and we define  $x_i = x_{i,i+1}$ . We can now rewrite Eq. (9) as follows:

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2} - (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i. \quad (10)$$

The common signal cycle time  $C$  (seconds) and the link specific progression speed  $V_i (\bar{V}_i)$  (m/sec) are optimizable variables as well. This introduces considerable flexibility in the calculation of the best arterial progression. Each of these variables is constrained by upper and lower limits. In addition, changes in speed from one link to the next can also be limited. Let the limits be as follows:

$C_1, C_2$  = lower and upper bounds on cycle length;  
 $(e_i, f_i), (\bar{e}_i, \bar{f}_i)$  = lower and upper bounds on outbound  
 (inbound) speed  $V_i(\bar{V}_i)$  (m/sec);  
 $(g_i, h_i), (\bar{g}_i, \bar{h}_i)$  = lower and upper bounds on change in  
 outbound (inbound) speed  $V_i(\bar{V}_i)$  (m/sec).

To obtain constraints that are linear in the decision variables we define the inverse of the cycle time, i.e., the signal frequency  $z = 1/C$  (cycles/second), such that:

$$1/C_2 \leq z \leq 1/C_1$$

The progression speeds and the changes in speeds are also expressed in terms of their reciprocals:

$$\frac{1}{f_i} \leq \frac{1}{V_i} \leq \frac{1}{e_i} \quad \text{and} \quad \frac{1}{\bar{f}_i} \leq \frac{1}{\bar{V}_i} \leq \frac{1}{\bar{e}_i}$$

$$\frac{1}{h_i} \leq \frac{1}{V_{i+1}} - \frac{1}{V_i} \leq \frac{1}{g_i} \quad \text{and} \quad \frac{1}{\bar{h}_i} \leq \frac{1}{\bar{V}_{i+1}} - \frac{1}{\bar{V}_i} \leq \frac{1}{\bar{g}_i}$$

Using the notation  $d_i = d_{i,i+1}$ , we obtain

$$t_i = \frac{d_i}{V_i} z \quad \text{and} \quad \bar{t}_i = \frac{\bar{d}_i}{\bar{V}_i} z$$

Substituting these expressions above and manipulating the variables, we obtain

$$\frac{d_i}{f_i} z \leq t_i \leq \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i} z$$

$$\frac{d_i}{h_i} z \leq \frac{d_i}{d_{i+1}} t_{i+1} - t_i \leq \frac{d_i}{g_i} z \quad \text{and}$$

$$\frac{\bar{d}_i}{\bar{h}_i} z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \dots, n-1$$

Thus,  $t_i, \bar{t}_i$  and  $z$  are decision variables which, once known, determine the progression speeds. The complete MILP-1 is given below:

**MILP-1.** Find  $b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i, v_i$  to

**Maximize**  $b + k \cdot \bar{b}$  **subject to**

$$1/C_2 \leq z \leq 1/C_1$$

$$w_i + b \leq 1 - r_i \quad \text{and} \quad \bar{w}_i + \bar{b} \leq 1 - \bar{r}_i$$

$$i = 1, \dots, n$$

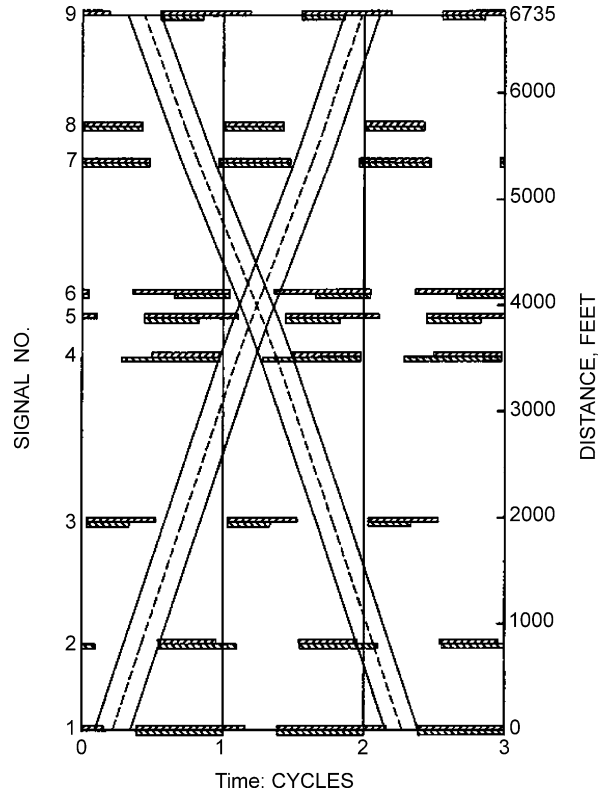
$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2}$$

$$- (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i$$

$$i = 1, \dots, n-1$$

$$\frac{d_i}{f_i} z \leq t_i \leq \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i} z$$

$$i = 1, \dots, n-1$$



**Traffic Networks, Optimization and Control of Urban, Figure 10**  
 Time space diagram for a uniform bandwidth progression scheme generated by MILP-1

$$\frac{d_i}{h_i} z \leq \frac{d_i}{d_{i+1}} t_{i+1} - t_i \leq \frac{d_i}{g_i} z \quad \text{and}$$

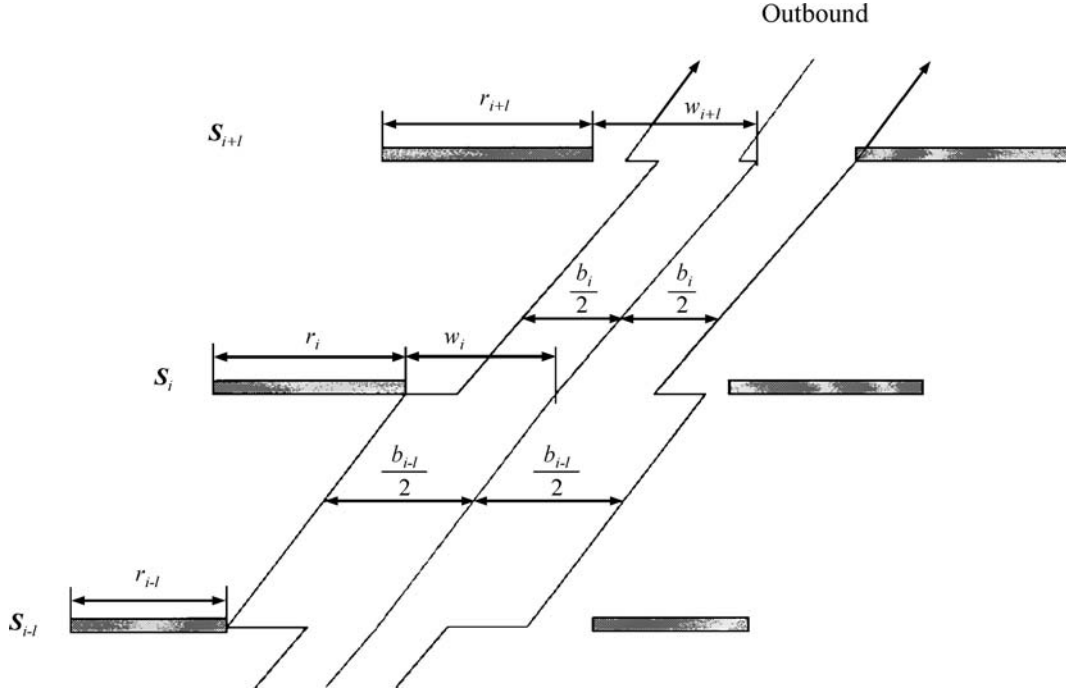
$$\frac{\bar{d}_i}{\bar{h}_i} z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \dots, n-1$$

$$b, \bar{b}, z, w_i, \bar{w}_i, t_i, \bar{t}_i \geq 0 \quad \text{and} \quad v_i \text{ integer}.$$

In addition to bandwidths and interference variables, this program also calculates optimal cycle length ( $1/z$ ) and progression speeds. To calculate the offsets, green splits need to be given or calculated based on flow/capacity considerations. Phase sequencing can also be optimized. MILP-1 is the model used by the MAXBAND program developed by Little et al. [6]. An example of a uniform bandwidth solution obtained by MILP-1 is shown in Fig. 10. The scheme contains signals with multi-phase sequences.

### The Variable Bandwidth Problem

In this section we describe a more versatile multi-band/multi-weight approach for the bandwidth maximization problem. A different bandwidth is defined for each directional road section of the arterial and is individually



**Traffic Networks, Optimization and Control of Urban, Figure 11**  
Geometric relations for MILP-2

weighted with respect to its contribution to the overall objective function. While the band is continuous, its width can vary and adapt to the prevailing traffic flow on each link. In this way the formulation is sensitive to varying traffic conditions and can tailor the progression scheme to the different possible traffic flow patterns. The user can still choose a uniform bandwidth progression if desired, but this is only one of the many possible options. Referring to the geometry in Fig. 11, we define the following variables:

$b_i(\bar{b}_i)$  = outbound (inbound) bandwidth of link  $i$ ; there is now a specific band for each link  $L_i$ ;

$w_i(\bar{w}_i)$  = the time from right (left) side of red at  $S_i$  to the centerline of the outbound (inbound) green band; the reference point at each signal has been moved from the edges to the centerline of the band.

The following constraints apply in the outbound directions at signal  $S_i$ :

$$w_i + \frac{b_i}{2} \leq 1 - r_i \quad \text{and} \quad w_i + \frac{b_i}{2} \geq 0.$$

The pair of constraints can be combined as follows:

$$\frac{b_i}{2} \leq w_i \leq (1 - r_i) - \frac{b_i}{2}.$$

A similar relation must be observed at  $S_{i+1}$ , since band  $b_i$  must be constrained at both ends:

$$\frac{b_i}{2} \leq w_{i+1} \leq (1 - r_{i+1}) - \frac{b_i}{2}.$$

Corresponding relationships exist in the inbound direction (marked by a bar on all variables):

$$\frac{\bar{b}_i}{2} \leq \bar{w}_i \leq (1 - \bar{r}_i) - \frac{\bar{b}_i}{2}$$

$$\frac{\bar{b}_i}{2} \leq \bar{w}_{i+1} \leq (1 - \bar{r}_{i+1}) - \frac{\bar{b}_i}{2}.$$

These constraints are referred to as the *directional interference constraints*. The time reference points are redefined to the *centerline* of the bands (or, the *progression line*), rather than the edges. The *loop integer constraint* given by Eq. (10) remains unchanged and so do the travel time and speed-change constraints.

The most important change occurs in the objective function. Since the bands are link-specific, they can be weighted disaggregately to reflect user-defined traffic performance objectives which can be individually weighted for each link of the arterial. The objective function now

has the following form:

$$\text{Maximize} \quad \frac{1}{n-1} \sum_{i=1}^n a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \quad (11)$$

where  $a_i$  and  $\bar{a}_i$  are the link specific weighting coefficients for the outbound and inbound directions respectively. There is a multitude of possible options for choosing the weighting coefficients in Eq. (11). Some of the more common expressions are as follows:

$$a_i = \left( \frac{q_i}{s_i} \right)^p \quad \text{and} \quad \bar{a}_i = \left( \frac{\bar{q}_i}{\bar{s}_i} \right)^p$$

where

$q_i(\bar{q}_i)$  = outbound (inbound) directional flow rate on link  $L_i$ ; either the total or the through volume can be used;

$s_i(\bar{s}_i)$  = saturation flow rate outbound (inbound) directional volume on link  $V_{ij}$ ; either the total flow rate or the through flow rate can be used;

$p$  = an integer exponent; the values 0, 1, 2 and 4 were used.

To obtain an objective function value that is consistent with those used previously, we normalize the weighting coefficients to obtain

$$\sum_{i=1}^{n-1} a_i = n-1 \quad \text{and} \quad \sum_{i=1}^{n-1} \bar{a}_i = n-1$$

The multi-band/multi-weight optimization program is summarized in MILP-2.

**MILP-2.** Find  $b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i, v_i$  to

$$\text{Maximize} \quad \frac{1}{n-1} \sum_{i=1}^n a_i \cdot b_i + \bar{a}_i \cdot \bar{b}_i \quad \text{subject to}$$

$$1/C_2 \leq z \leq 1/C_1$$

$$\frac{b_i}{2} \leq w_i \leq (1-r_i) - \frac{b_i}{2}, \quad \frac{\bar{b}_i}{2} \leq \bar{w}_{i+1}$$

$$\leq (1-\bar{r}_{i+1}) - \frac{\bar{b}_i}{2},$$

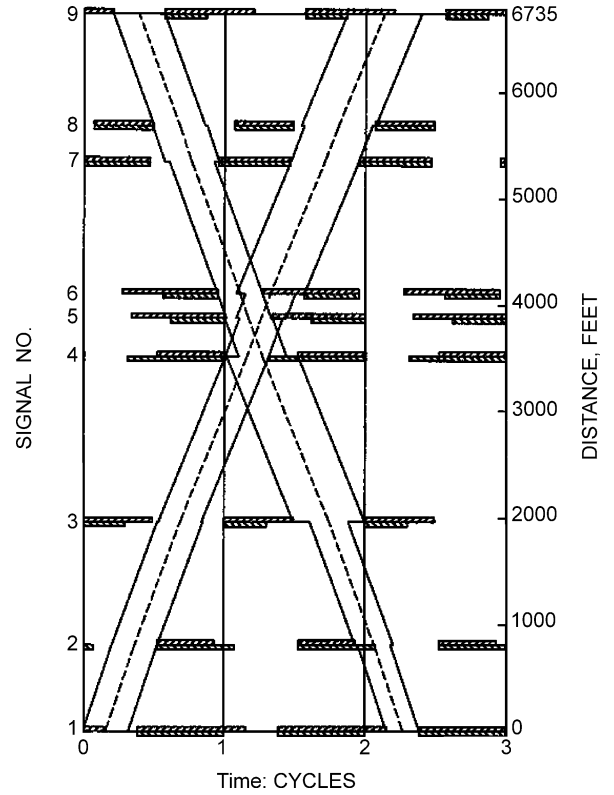
$$\frac{\bar{b}_i}{2} \leq \bar{w}_i \leq (1-\bar{r}_i) - \frac{\bar{b}_i}{2} \quad \text{and}$$

$$\frac{\bar{b}_i}{2} \leq \bar{w}_{i+1} \leq (1-\bar{r}_{i+1}) - \frac{\bar{b}_i}{2} \quad i = 1, \dots, n-1$$

$$t_i + \bar{t}_i + \frac{r_i + \bar{r}_i}{2} + (w_i + \bar{w}_i) - \frac{r_{i+1} + \bar{r}_{i+1}}{2}$$

$$- (w_{i+1} + \bar{w}_{i+1}) - (\tau_{i+1} + \bar{\tau}_i) + \Delta_i - \Delta_{i+1} = v_i$$

$$i = 1, \dots, n-1$$



**Traffic Networks, Optimization and Control of Urban, Figure 12**  
Time space diagram for a variable bandwidth progression scheme calculated by MILP-2

$$\frac{d_i}{f_i} z \leq t_i \leq \frac{d_i}{e_i} z \quad \text{and} \quad \frac{\bar{d}_i}{\bar{f}_i} z \leq \bar{t}_i \leq \frac{\bar{d}_i}{\bar{e}_i} z$$

$$i = 1, \dots, n-1$$

$$\frac{d_i}{h_i} z \leq \frac{d_i}{d_{i+1}} t_{i+1} - t_i \leq \frac{d_i}{g_i} z \quad \text{and}$$

$$\frac{\bar{d}_i}{\bar{h}_i} z \leq \frac{\bar{d}_i}{\bar{d}_{i+1}} \bar{t}_{i+1} - \bar{t}_i \leq \frac{\bar{d}_i}{\bar{g}_i} z \quad i = 1, \dots, n-1$$

$$b_i, \bar{b}_i, z, w_i, \bar{w}_i, t_i, \bar{t}_i \geq 0 \quad \text{and} \quad v_i \text{ integer.}$$

MILP-2 involves  $(18n - 20)$  constraints and  $(6n - 3)$  continuous variables, and  $n - 1$  unrestricted integer variables, not counting slack variables. Green or red splits are determined the same as in MILP-1. The increased decision capabilities of MILP-2 require a corresponding increase in the size of the mathematical program. The formulation given in MILP-2 corresponds to the MULTIBAND optimization program developed by Gartner et al. [7]. The variable bandwidth solution for the same arterial as for MILP-1 (Fig. 9) is shown in Fig. 12.

### The Network Problem

The MILP-2 problem can be extended for controlling traffic flow in a network of intersecting arterials. This represents a significantly more challenging problem: progressions must be provided on all the arterials of the network, simultaneously. Consider a network of  $m$  arterials with each arterial having  $n_j$  signalized intersections. Let  $S_{ij}$  denote the  $i$ th signal on the  $j$ th arterial of the network and  $L_{ij}$  denote the  $i$ th link (between signals  $i$  and  $i + 1$ ) of the  $j$ th arterial, with  $j = 1, \dots, m$  and  $i = 1, \dots, n_j$ . The variables presented in MILP-1 and MILP-2 are redefined as follows:

- $b_{ij}(\bar{b}_{ij})$  = outbound (inbound) bandwidth of link  $i$  on arterial  $j$ ;
  - $w_{ij}(\bar{w}_{ij})$  = interference variables, the time from right (left) side of red at  $S_{ij}$  to the centerline of the outbound (inbound) green band;
  - $r_{ij}(\bar{r}_{ij})$  = outbound (inbound) red time at  $S_{ij}$ ;
  - $t_{ij}(\bar{t}_{ij})$  = travel time on link  $i$  of arterial  $j$  in the outbound (inbound) direction;
  - $d_{ij}(\bar{d}_{ij})$  = length of link  $i$  of arterial  $j$  in the outbound (inbound) direction;
  - $\phi_{(ij),(jl)}(\bar{\phi}_{(ij),(kj)})$  = internode offsets, time from the center of the outbound (inbound) red at  $S_{ij}$  to the center of the outbound (inbound) red at  $S_{kj}$ ;
  - $\Delta_{ij}$  = directional node phase shift, time from center of  $\bar{r}_{ij}$  to nearest center of  $r_{ij}$ ;
  - $\tau_{ij}(\bar{\tau}_{ij})$  = queue clearance time for advancement of outbound (inbound) bandwidth at  $S_{ij}$  to clear turning-in traffic before arrival of main-street platoon;
  - $V_{ij}(\bar{V}_{ij})$  = outbound (inbound) progression speed on link  $L_{ij}$  (m/sec).
  - $e_{ij}, f_{ij}, (\bar{e}_{ij}, \bar{f}_{ij})$  = lower and upper bounds on outbound (inbound) speed  $V_{ij}(\bar{V}_{ij})$  (m/sec);
  - $g_{ij}, h_{ij}, (\bar{g}_{ij}, \bar{h}_{ij})$  = lower and upper bounds on change in outbound (inbound) speed  $V_{ij}(\bar{V}_{ij})$  (m/sec).
- In addition to the above variables, the *intranode offset* is defined as:

$\omega_{(ij),(kl)}(\bar{\omega}_{(ij),(kl)})$  = intranode offsets, the time from the center of the inbound (outbound) red at  $i$ th node of the  $j$ th arterial ( $S_{ij}$ ), to the center of the inbound (outbound) red in the crossing direction at the same node which is also the  $k$ th node of the  $l$ th arterial ( $S_{kl}$ ).

The directional interference constraints, the arterial loop integer constraint and the travel time and speed-change constraints remain unchanged and are applied, in turn, for each arterial. A new set of constraints, though, is required for the synchronization of all signals. Similar to the arterial loop-integer constraints Eq. (7), for any closed loops of the network consisting of more than 2 links, the

summation of *internode* and *intranode offsets* around the loop of intersecting arterials must be an integer multiple of the cycle time. This results in the formulation of the *network loop-integer constraints*. The number of network loop constraints and the choice of a fundamental set of loops is given by Gartner [8].

The network optimization problem is summarized in MILP-3:

**MILP-3.** Find  $b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij}, v_{ij}, \mu_i$  to

$$\text{Maximize } \sum_{j=1}^m \frac{1}{n-1} \sum_{i=1}^n a_{ij} \cdot b_{ij} + \bar{a}_{ij} \cdot \bar{b}_{ij} \text{ subject to}$$

$$1/C_2 \leq z \leq 1/C_1$$

$$\frac{b_{ij}}{2} \leq w_{ij} \leq (1 - r_{ij}) - \frac{b_{ij}}{2},$$

$$\frac{b_{ij}}{2} \leq w_{i+1,j} \leq (1 - r_{i+1,j}) - \frac{b_{ij}}{2},$$

$$\frac{\bar{b}_{ij}}{2} \leq \bar{w}_{ij} \leq (1 - \bar{r}_{ij}) - \frac{\bar{b}_{ij}}{2} \quad \text{and}$$

$$\frac{\bar{b}_{i,j}}{2} \leq \bar{w}_{i+1,j} \leq (1 - \bar{r}_{i+1,j}) - \frac{\bar{b}_{ij}}{2}$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$t_{ij} + \bar{t}_{ij} + \frac{r_{ij} + \bar{r}_{ij}}{2} + (w_{ij} + \bar{w}_{ij}) - \frac{r_{i+1,j} + \bar{r}_{i+1,j}}{2} - (w_{i+1,j} + \bar{w}_{i+1,j})$$

$$- (\tau_{i+1,j} + \bar{\tau}_{i,j}) + \Delta_{i,j} - \Delta_{i+1,j} = v_{i,j}$$

$$i = 1, \dots, n_j - 1; \quad j = 1, \dots, m$$

$$\phi_{(ia),(i+1,a)} + \omega_{(i+1,a),(jb)} + \phi_{(jb),(j+1,b)} + \omega_{(j+1,b),(kc)} + \phi_{(kc),(k+1,c)} + \omega_{(k+1,c),(ld)} + \phi_{(ld),(l+1,d)} + \omega_{(ld),(ia)} = \mu_N$$

$$i = 1, \dots, n_a - 1; \quad j = 1, \dots, n_b - 1;$$

$$k = 1, \dots, n_c - 1; \quad l = 1, \dots, n_d - 1;$$

$$a, b, c, d \text{ arterials forming a fundamental network loop};$$

$$\frac{d_{ij}}{f_{ij}} z \leq t_{ij} \leq \frac{d_{ij}}{e_{ij}} z \quad \text{and} \quad \frac{\bar{d}_{ij}}{\bar{f}_{ij}} z \leq \bar{t}_{ij} \leq \frac{\bar{d}_{ij}}{\bar{e}_{ij}} z$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$\frac{d_{ij}}{h_{ij}} z \leq \frac{d_{ij}}{d_{i+1,j}} t_{i+1,j} - t_{ij} \leq \frac{d_{ij}}{g_{ij}} z \quad \text{and}$$

$$\frac{\bar{d}_{ij}}{h_{ij}} z \leq \frac{\bar{d}_{ij}}{d_{i+1,j}} \bar{t}_{i+1,j} - \bar{t}_{ij} \leq \frac{\bar{d}_{ij}}{g_{ij}} z$$

$$i = 1, \dots, n_j - 1; j = 1, \dots, m$$

$$b_{ij}, \bar{b}_{ij}, z, w_{ij}, \bar{w}_{ij}, t_{ij}, \bar{t}_{ij} \geq 0 \quad \text{and}$$

$$v_{ij}, \mu_i \text{ integer}.$$



For an  $m \times n$  closed grid network ( $m \times n$  intersections and  $m + n$  arterials), MILP-3 may involve up to  $(23mn - 14m - 14n + 1)$  constraints and  $(12mn - 6m - 6n)$  continuous variables, and  $(3mn - 2m - 2n + 1)$  unrestricted integer variables, not counting slack variables.

The principal difficulty in solving mixed-integer problems, such as the ones described above, is the number of integer variables and their range, i. e., the size of the integer feasible set. Virtually all MILP codes use branch-and-bound strategies for calculating the optimal values. Solving the primary multi-band problem by general purpose branch-and-bound is still a rather formidable task. For this reason, heuristic methods which quickly lead to a good solution have been developed [9].

Delay-Based Models – Cyclic Flow Profiles

TRANSYT

TRANSYT, the traffic network study tool, is a computer model to optimize traffic signal timings and perform traffic signal simulation. TRANSYT has two main elements – the traffic model which is used to calculate the performance index of the traffic network for a given set of signal timings and an optimizing process that makes changes to the

settings and determines whether they improve the performance index or not.

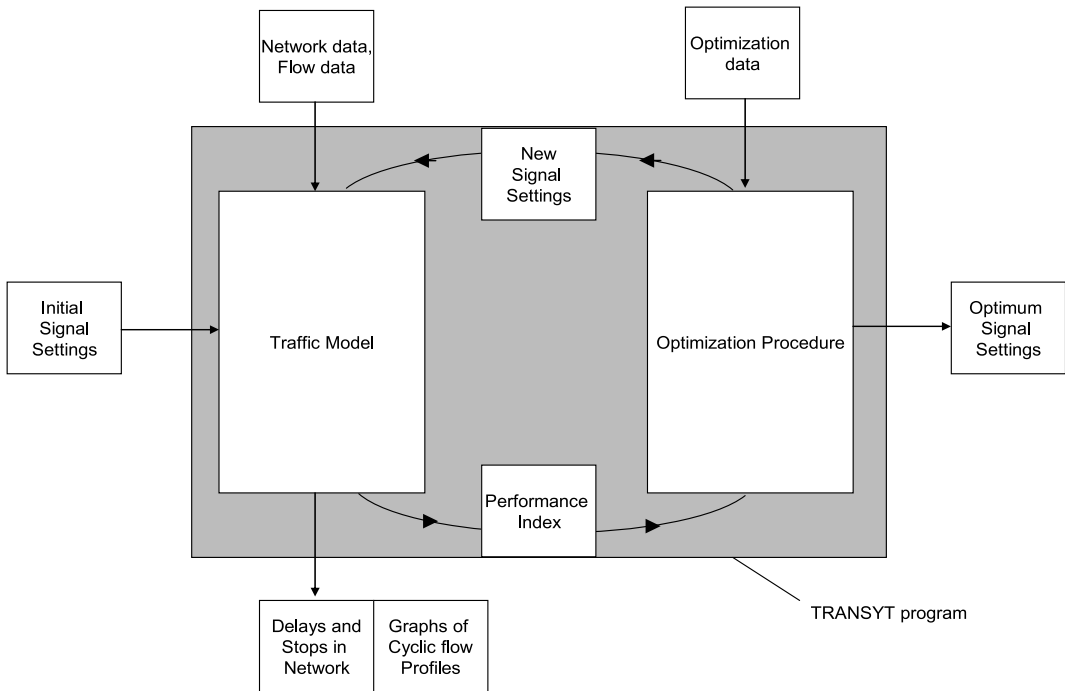
Simulation Model

The traffic simulation model used in TRANSYT is a macroscopic traffic simulation model. In a macroscopic model platoons of vehicles are considered rather than individual vehicles. TRANSYT simulates traffic flow macroscopically, but in a step-wise manner. The cycle length is divided into small, equal time increments, called steps. All TRANSYT calculations are made on the basis of flow rates (such as vehicles per hour). The average flow pattern past a point in the road network is represented by a histogram also known as Cyclic Flow Profile. There are three traffic flow patterns considered while simulating behavior of a signalized intersection: the “IN”, “GO”, and “OUT” patterns (see figure below). [10].

The IN-Pattern or Arrival Flow Pattern

The IN-Pattern is the pattern of traffic arriving at the stop line. The arrival flow pattern is expressed mathematically as follows:

$$IN_{it} = \sum_j^n F_{ij}(P_{ij} \cdot OUT_{jt'})$$



Traffic Networks, Optimization and Control of Urban, Figure 13  
Structure of TRANSYT program

where,

$IN_{it}$  = the IN-pattern on link  $i$  for time step  $t$

$F_{ij}$  = smoothing process related to platoon dispersion for flow to link  $i$  from link  $j$

$P_{ij}$  = the proportion of the feeding link OUT-pattern that feeds the subject link

$OUT_{jv}$  = the OUT-pattern of link  $j$  for step  $t'$

$t'$  = the time step  $t$  minus the travel time for flow to link  $i$  from link  $j$

$n$  = the number of links ( $j$ ) that feed link  $i$

The IN-pattern is computed for each step,  $t$ , in the cycle, thus forming a pattern, or flow profile.

### The GO-Pattern or Saturation Flow Pattern

The GO-pattern is the flow rate at each step that would leave the stop line when the traffic signal turns green if there were enough traffic to saturate the green. There is a certain Startup Lost Time (SLT) involved due to reaction of driver etc.

### The OUT-Pattern or Departure Flow Pattern

The OUT-pattern is the profile of traffic actually leaving the stop line. It is usually equal to the GO-pattern as long as there is a queue. After the queue dissipates, it is equal to the IN-pattern for the remainder of the effective green.

The queue (or the number of vehicles held at the stop line during any time interval,  $t$ ) is first determined to determine the OUT-pattern

$$m_t = \max\{(m_{t-1} + q_t - s_t), 0\}$$

where,

$m_t$  = number of vehicles in the queue in time interval  $t$  on a given link (and similarly for  $m_{t-1}$ )

$q_t$  = number of vehicles arriving in interval  $t$ , given by the IN-pattern

$s_t$  = number of vehicles allowed to leave in interval  $t$ , given by the GO-pattern

The OUT-pattern is given for link  $i$  during time interval  $t$  by the expression:

$$OUT_{it} = m_{i,t-1} + q_{it} - m_{it}$$

The traffic model further utilizes a platoon dispersion algorithm that simulates the normal dispersion (i. e., the "spreading out") of platoons as they travel downstream. The start-and-stop operation of signals tends to create platoons of vehicles that travel along a link. TRANSYT models the dispersion of these platoons as they progress along a link.

On internal links, for each time interval (step),  $t$ , the downstream arrival flow is determined by the following recurrence equation:

$$v'_{(t+\beta T)} = F.v_t + [(1 - F).v'_{(t+\beta T-1)}]$$

where,

$v'_{(t+\beta T)}$  = predicted flow rate in time interval  $t + \beta T$  of the predicted platoon

$v_t$  = flow rate of the initial platoon during step  $t$

$\beta$  = factor to take into account vehicles traveling faster than speed limit

$T$  = the cruise travel time on the link, in steps

$F$  = a smoothing factor, where:

$$F = (1 + \alpha \beta T)^{-1},$$

where  $\alpha$  = an empirically derived constant, called the platoon dispersion factor (PDF) takes into account site-specific factors such as grade, curvature, parking, opposing flow interference, and other sources of impedance.

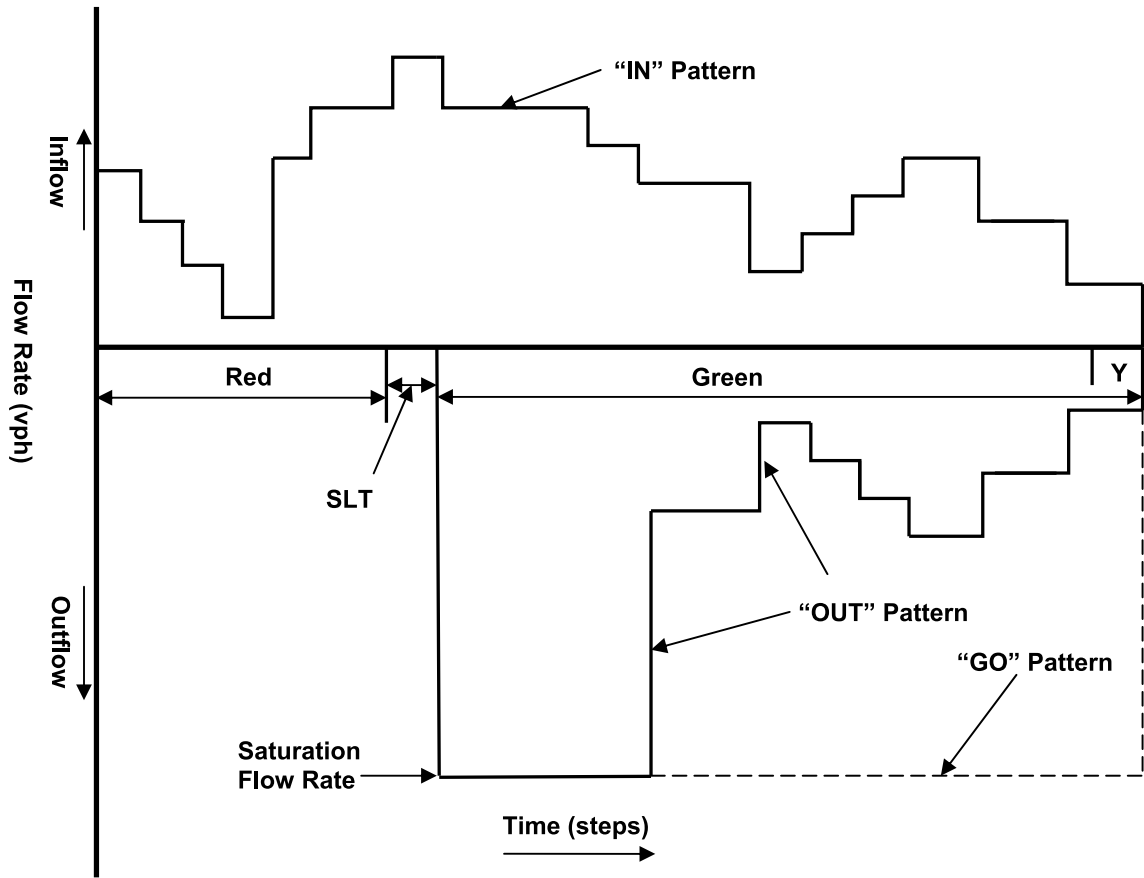
### Optimization Model

TRANSYT uses a Performance Index (PI) (also the objective function) allowing the user to define their preferences regarding performance of the traffic network. TRANSYT develops a signal timing plan that produces an optimal value of the PI. TRANSYT-7F (the US-adopted version) offers numerous PI's to choose from, which can reflect anything that the user desires including delay, progression, stops, fuel consumption, queuing, and throughput. These PI options are listed below:

$$PI = \left[ \begin{array}{l} \text{DI only} \\ \text{PROS only} \\ \text{PROS \& DI} \\ \text{PROS/DI} \\ \text{DI } \sum \frac{\text{Average back of queue on link } i}{\text{Queuing capacity on link } i} \\ \text{Throughput only} \\ \text{Throughput \& DI} \\ \text{Throughput/DI} \end{array} \right]$$

In general, TRANSYT-7F always attempts to maximize the PI, unless the disutility index (DI) has been selected as the PI. Since the DI can be a combination of delay, stops, queuing, and fuel consumption, this value must be minimized. The other objective functions that involve progression or throughput must be maximized.

The DI is a combination of vehicle delay, stops, and fuel consumption. Weighting factors are available in order



Traffic Networks, Optimization and Control of Urban, Figure 14  
TRANSYT Flow Patterns

to place more emphasis on any of these three DI components if desired. Delay-only optimization results in excessive stops and fuel consumption. Excess fuel consumption (minimization) is considered to be a good compromise between bandwidth and delay-based optimization.

The “standard” delay and stops DI optimization objective function is defined as follows:

$$DI = \sum_{i=1}^n \{ (w_{d_i} d_i + K w_{S_i} S_i) + QP \}$$

where,

DI = disutility index, analogous to the original TRANSYT performance index

$d_i$  = delay on link  $i$  (of  $n$  links) and on an optional user-specified upstream input link  $i - 1$

$K$  = user-coded stop penalty factor to express the importance of stops relative to delay

$S_i$  = stops on link  $i$  per second

$w_{xi}$  = link-specific weighting factors for delay ( $d$ ) and stops ( $s$ ) on link  $i$

$U_i$  = binary variable that is ‘1’ if link-to-link weighting has been established, zero otherwise

QP = queuing penalty

A queuing penalty is used to minimize the possibility of spillback.

$$QP = QB_i W_q (q_i - qc_i)^2$$

where,

$Q$  = a binary variable, ‘1’ if the queue penalty is included in the DI ‘0’ otherwise

$B_i$  = a binary variable, ‘1’ if the maximum back of queue ( $q_i$ ) exceeds the queuing capacity ‘0’ otherwise

$W_q$  = a network-wide penalty applied to the excess queue spillback

$q_i$  = computed maximum back of queue on link  $i$

$qc_i$  = queuing capacity for link  $i$

Progression opportunities (PROS) represent the ability of vehicles to progress through multiple intersections without stopping. Points are scored every time a vehicle is able to traverse additional intersections. Therefore it is possible to achieve PROS without having a wide bandwidth stretching from the beginning of an arterial street to the end.

It is also possible to select a combination of PROS and DI in order to achieve a compromise between the two objectives. For example, if the PI is defined as PROS/DI, the program attempts to simultaneously maximize the numerator (PROS) and minimize the denominator (DI). For optimizing bandwidth and progression speeds, PROS/DI is recommended. PROS-only is rarely acceptable, because although it often produces a good bandwidth, the minor movements are allowed to fail.

### Demand-Responsive Models

A number of demand-responsive, or adaptive signal control strategies have been developed. Two such strategies are described herein: SCOOT and OPAC.

#### SCOOT

Adaptive traffic signal systems adjust signal timings based on measured flow. Reacting to these flow variations results in reduced delay, shorter queues and decreased travel time. SCOOT (Split Cycle Offset Optimization Technique) is one of the earliest real time adaptive traffic control systems [11]. It coordinates the operation of all the traffic signals in an area to give good progression to vehicles through the network.

The operation of the SCOOT model is illustrated in Fig. 15. SCOOT obtains information on traffic flows from detectors. As an adaptive system, SCOOT depends on good traffic data so that it can respond to changes in flow. Detectors are normally required on every link. Their location is important and they are usually positioned at the upstream end of the approach link.

When vehicles pass the detector, SCOOT receives the information and converts the data into its internal units and uses them to construct “Cyclic flow profiles” for each link. The sample profile shown in the diagram is color coded green and red according to the state of the traffic signals when the vehicles will arrive at the stop line at normal cruise speed. Vehicles are modeled down the link at cruise speed and join the back of the queue (if present). During the green, vehicles discharge from the stop line at the validated saturation flow rate.

The data from the model is then used by SCOOT in three optimizers which are continuously adapting three

key traffic control parameters – the amount of green for each approach (Split), the time between adjacent signals (Offset) and the time allowed for all approaches to a signaled intersection (Cycle time). These three optimizers are used to continuously adapt these parameters for all intersections in the SCOOT controlled area, minimizing wasted green time at intersections and reducing stops and delays by synchronizing adjacent sets of signals. This means that signal timings evolve as the traffic situation changes without any of the harmful disruption caused by changing fixed time plans on more traditional urban traffic control systems.

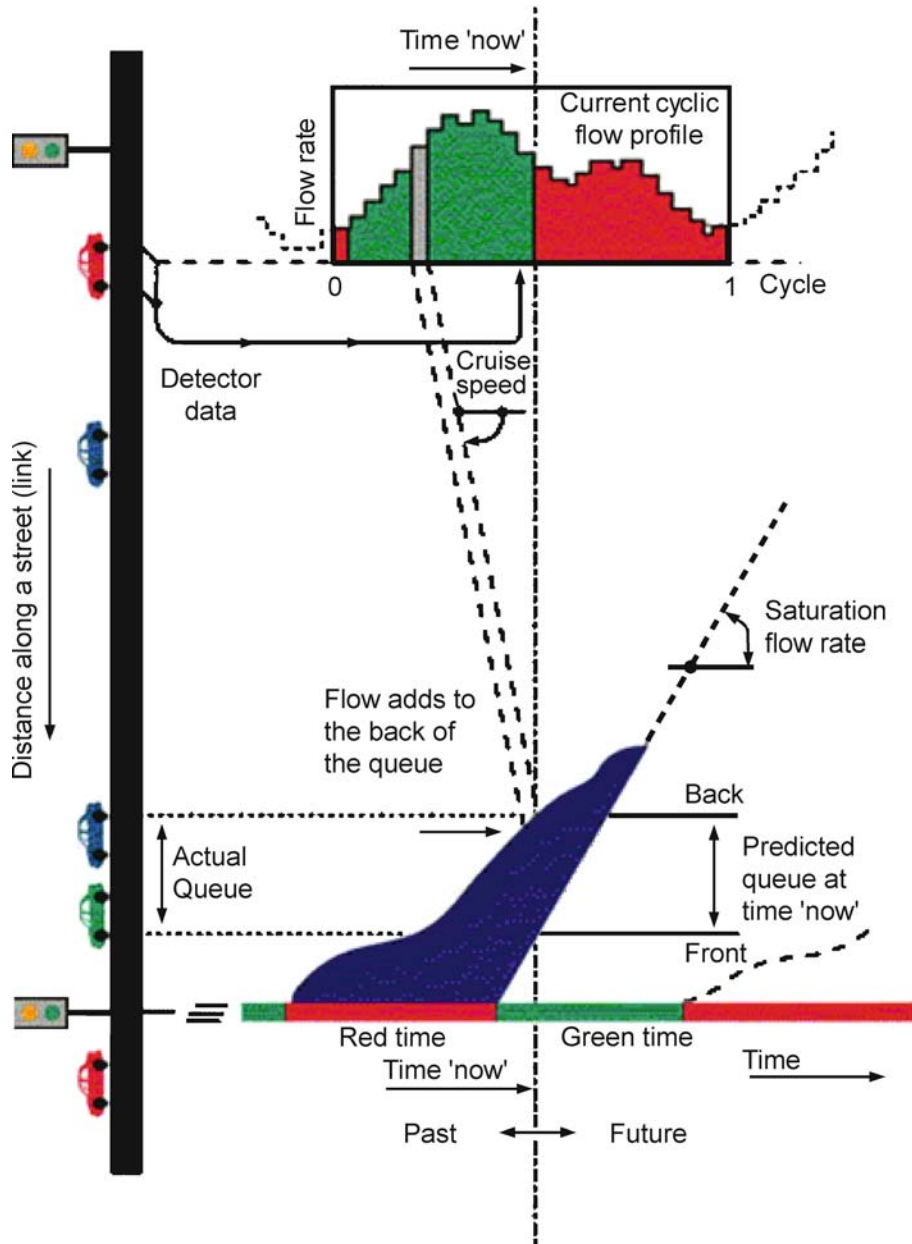
#### OPAC – Optimization Policies for Adaptive Control

This section presents an adaptive control strategy for coordinating and synchronizing signals in a network using the virtual-fixed-cycle concept and is labeled *VFC-OPAC*. The strategy is based on the single intersection dynamic-programming-based *OPAC* (optimization policies for adaptive control) adaptive controller developed by Gartner [12]. *VFC-OPAC* consists of a distributed control strategy featuring a dynamic optimization algorithm that calculates signal timings to minimize a performance function of total intersection delays and stops. The algorithm uses a combination of measured and modeled demand to determine, in a distributed manner, phase durations at each signal that are constrained by minimum and maximum green times and, when running in a coordinated mode, by coordination and synchronization parameters that can be updated based on real-time data.

*OPAC* was developed from the outset as a distributed strategy featuring a dynamic optimization algorithm for traffic signal control without requiring a rigid, fixed cycle time. Signal timings are calculated to directly minimize performance measures, such as vehicle delays and stops, and are only constrained by minimum and maximum phase lengths and, if running in a coordinated mode, by a virtual cycle length and by a virtual offset. Development of the strategy has progressed through several versions, each one serving as a base for a subsequent version. The principal features of each version are outlined below.

#### OPAC I: Dynamic Programming

The first version, designated *OPAC I*, was designed to serve as a basis for subsequent *opac* strategy development. *OPAC I* utilizes a dynamic programming (DP) model for the determination of the traffic control parameters. Since DP is a global optimization strategy for multistage decision processes, it provides a standard against which all other strategies can be compared [13,14].



Traffic Networks, Optimization and Control of Urban, Figure 15  
Schematic diagram showing operation of SCOOT model

The optimization process is decomposed into  $N$  stages, where each stage corresponds to a discrete time interval in which the arrivals are measured, (2 to 5 sec intervals). The total number of stages  $N$  corresponds to the *horizon length* (HL) of the input predictions. For exploratory purposes the horizon length was assumed to be several minutes long, e. g., 5, 10, or 15 mins. A typical stage  $i$  is illustrated in Fig. 16.

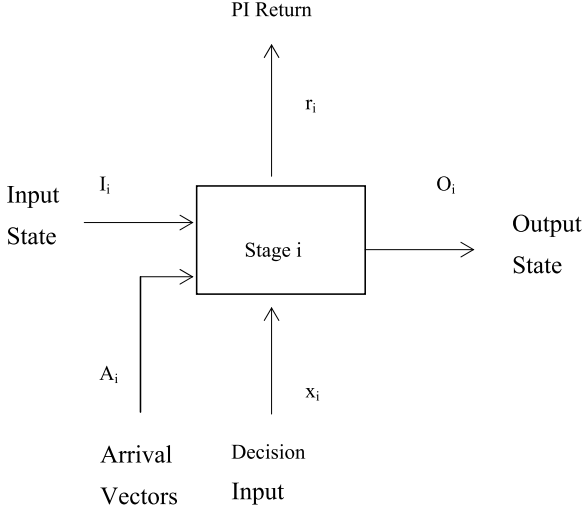
At stage  $i$  we have an input state vector  $I_i$ , an arrival vector  $A_i$ , output state vector  $O_i$ , input variable  $x_i$ , economic return (cost) output  $r_i$ , and a set of transformations:

$$O_i = T_i(I_i, A_i, x_i)$$

$$r_i = R_i(I_i, A_i, x_i).$$

The state of the intersection is characterized by the state of the signal (green or red) and by the queue-length





**Traffic Networks, Optimization and Control of Urban, Figure 16**  
Dynamic Programming stage in OPAC I

on each of the approaches. The input decision variable indicates whether the signal is to be switched at this stage or to remain in its present state. The return cost output is the intersection's index of performance (e. g., total delay time and /or number of stops), which has to be minimized. The functional relationships between the input and output variables are based on the queuing-discharge process occurring at the intersection, i. e., the vehicle inflow and outflow as a function of the signal settings.

We define the following variables (all corresponding to stage  $i$ ):

$a$  = approach designation, by direction,  $a = N, S, E, W$   
 $A^a$  = number of arrivals during the stage (for a four-leg intersection)

$D^a$  = number of departures (discharges) during the stage

$Q^a$  = queue-length on approach at beginning of stage

$S^a$  = status of signal at start of stage (green/red)

$x$  = stage input decision variable (change/no-change)

The input state vector (transposed) is:  $I^t = [I^N, I^S, I^E, I^W]$ , where for each approach the state  $I$

$$I^a = [S^a, Q^a]^t.$$

The output state vector contains the same elements as the input state vector and equals to the input state of the succeeding stage, i. e.,  $O_i = I_{i+1}$ . The signal status for each approach  $a$ , is a binary variable:

$$S^a = \{0 \text{ for Green; } 1 \text{ for Red}\}$$

The input decision variable is also binary

$$x = \begin{cases} 0 & \text{no change in signal status} \\ 1 & \text{change current signal status} \end{cases}$$

The transformation of input to output, at stage  $i$ , is as follows:

$$S_{i+1}^a = (S_i^a + x_i) \bmod 2$$

$$Q_{i+1}^a = Q_i^a + A_i^a - D_i^a$$

The mod 2 operator ensures that  $S_{i+1}^a$  will always be 0 or 1. The arrivals at each stage  $i$ ,  $A_i^a$  are an observed input (e. g., from detectors); the departures are a function of the state and decision variables:

$$D = \begin{cases} 0 & \text{if } S = 1 \\ \min(Q + A, d_{\max}) & \text{if } S = 0 \end{cases}$$

where  $d_{\max}$  is the saturation discharge rate (in veh/int). The DP algorithm goes backward in time, i. e., starting from the last interval and back-tracking to the first, at which time an optimal switching policy for the entire horizon is determined. The switching policy consists of the sequence of phase switch-ons and switch-offs throughout the horizon. The recursive optimization functional is:

$$f_i^*(I_i) = \min_{x_i} \{R_i(I_i, A_i, x_i) + f_{i+1}^*(I_i, A_i, x_i)\}.$$

The Performance Index (return) at stage  $i$  is the queuing delay and number of stops  $N_s$  incurred at this stage:

$$r_i = R_i(I_i, A_i, x_i) = \sum_a (Q_i^a + A_i^a - D_i^a) + \sum_a N_s^a.$$

When the optimization is terminated at stage  $i = 1$  we have,

$$f_1^*(I_1) = \min_{x_1} \left\{ \sum_{i=1}^N R_i(I_i, A_i, x_i) \right\} = \sum_{i=1}^N R_i(I_i, A_i, x_i^*) \quad (12)$$

which is the minimized Performance Index over the horizon period for a given initial input state  $I_1$ . Since the initial conditions at stage 1 are specified (i. e., the queue-lengths on all approaches are given as well as the initial signal status), we can retrace the optimal policy by taking a forward pass through the arrays of  $X_i^*(I_1)$ . The pol-

icy consists of the optimal sequence of switching decisions  $\{x_i^*, i = 1, \dots, N\}$  at all stages of the optimization process.

While this procedure assures globally optimal controls for the given horizon length, it requires complete information of arrivals over the entire control period. It cannot be used for real-time implementation due to both the (excessive) amount of processing involved and due to the lack of a practical method to gather real-time information for such a length of time. Much of the output generated by the procedure is never implemented because optimized policies are calculated for all possible combinations of initial conditions at each stage of the control period. In practice, only one 'optimum policy' is implemented. Nonetheless, *OPAC-I* serves an important function as a standard for the evaluation of the relative effectiveness of other, more practical strategies.

### OPAC II: Sequential Optimization

*OPAC II* breaks up the horizon into sequential optimization stages to speed up the optimization process. The model is a reformulation of the *OPAC I* algorithm. The purpose is to create a building block for a distributed on-line strategy. *OPAC II* has the following features:

- The control period is divided into successive (back-to-back) horizon lengths of  $T$  seconds each ( $T$  may encompass one or more cycle lengths).
- Each horizon is divided into an integral number of intervals ' $t$ ' seconds long; typically,  $t = 2 - 5$  sec.
- During each horizon there must be a sufficient number of phase changes to guarantee that no optimal solution is missed. The phase change (switching) times are measured from the start of the horizon in time units of  $t$ .
- For any given switching sequence in a horizon, the performance function for each approach computes the total delay and/or stops.

The optimization problem in *OPAC II* can be stated as follows: For each horizon length, given the initial queues on each approach and the arrivals for each interval of the horizon, determine the sequence of switching times, in terms of intervals, which yield the least delay and/or stops to vehicles over the entire horizon.

The procedure used for solving the problem consists of an intelligent search over the set of all possible combinations of feasible switching times within the horizon to determine the optimum sequence. Valid switching times are constrained by minimum and maximum phase durations. The problem can be re-formulated as an alternative dynamic programming problem by re-defining the control

variable  $x_j$  to denote the amount of green plus yellow time allocated to stage  $s_j$ . The stages, in this case, correspond to the phase lengths during the horizon. The recursive optimization functional (forward DP), is:

$$f_j(s_j) = \text{Min}_{x_j} \{R_j(s_j, x_j) + f_{j-1}^*(s_{j-1})\} . \quad (13)$$

The return (or Performance Index) is now

$$R = \sum_a \int_{s_{j-1}}^{s_j} A^a(t)dt + \sum_a N_s^a . \quad (14)$$

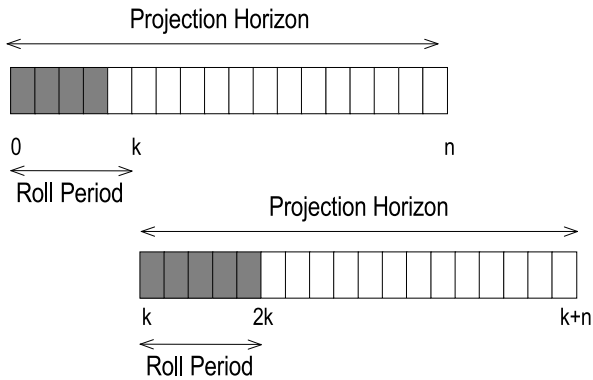
This calculation results in the optimal sequence of signal phase (stage) lengths during the horizon. By reformulating the DP model, computation is considerably more economical than in *OPAC I*. *OPAC II* lends itself more readily to operation in real-time than does *OPAC I*; however, it still requires information on arrivals (flows) over the entire horizon length.

### OPAC III: A Rolling Horizon Approach

A typical horizon length is 1–2 minutes long. Obtaining accurate arrival predictions for this length of time is not feasible with current technology. To use only readily available flow data without degrading the performance of the optimization procedure, a 'rolling horizon' strategy is applied to the *OPAC II* algorithm. In this version, the horizon length, or the *Projection Horizon* is the period for which traffic patterns are projected and optimum phase change information is calculated. The key feature is that real-time data are required for only a small portion of the horizon.

Figure 17 is an illustration of the rolling horizon procedure. From detectors placed upstream of each approach actual arrival data for  $k$  intervals can be obtained for the beginning, or head, portion of the horizon. For the remaining  $n - k$  intervals, the tail of the horizon, flow data may be obtained from a model. A simple model consists of a moving average of all previous arrivals on the approach. An optimal switching policy is calculated for the entire horizon, but only those changes which occur within the head portion are actually being implemented. In this way, *OPAC III* can dynamically revise the switching decisions as more recent (i. e., more accurate) real-time data continuously become available.

By placing the detectors well upstream of the intersection (10 to 15 sec. travel time) one can obtain actual arrival information for the head period. This allows for a more reliable estimation of delay for any given phase change decision. At the conclusion of the current head pe-



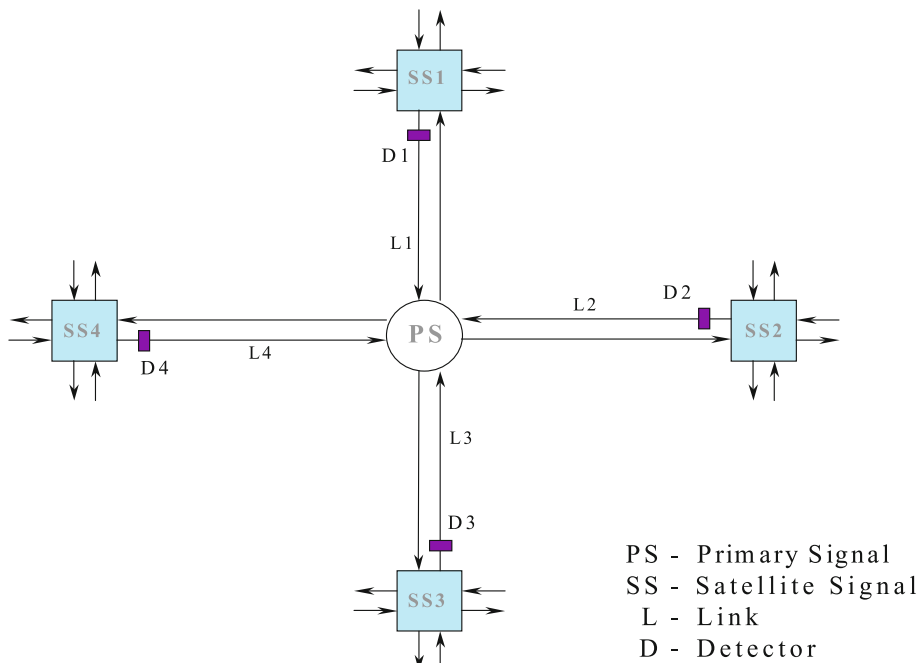
**Traffic Networks, Optimization and Control of Urban, Figure 17**  
Implementation of the rolling horizon approach in OPAC

riod, a new projection horizon containing new head and tail periods is defined with the new horizon beginning at (rolled to) the termination of the old head period. Calculations are then repeated for the new projection horizon. The roll period can be any multiple number of steps, including one. A shorter roll period implies more frequent calculations and, generally, closer to optimum (i.e., ideal) results. Extensive simulation and field tests have demonstrated the effectiveness of this strategy in achieving performance that approaches the theoretical optimum estab-

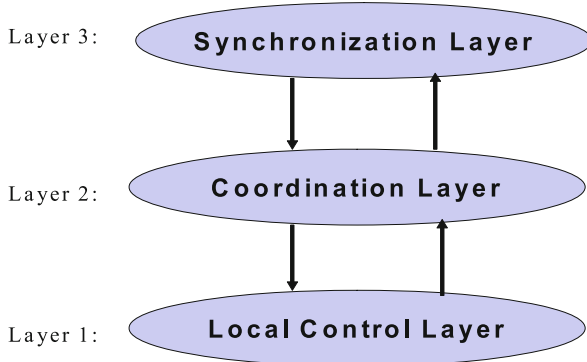
lished by *OPAC I*. The placing of the detectors and the information flow among the intersection controllers is illustrated in Fig. 18.

#### OPAC IV: The Virtual Fixed Cycle Strategy

Rhythmic, or cyclic operation is essential for coordination of signals. Whereas an independent *opac* controller is cycle-free, linking of adaptive signals in a network configuration requires coordinated operation to facilitate opportunities for unimpeded progression. This is achieved in the *vfc-opac* model by re-introducing the concepts of cycle time and offsets in a benign manner that allows for increased flexibility of signal timing selection and control strategy implementation. *Vfc-opac* controlled intersections interact with neighboring intersections (fixed-time, or other *vfc-opac* controlled) in response to projected arrival flows from upstream feeder signals. The model offers, at the option of the user, a coordination-synchronization strategy that is suitable for implementation in arterials and in networks. The strategy is referred to as *virtual-fixed-cycle* because from cycle to cycle the yield point, or local cycle reference point, is allowed to range about the fixed yield points dictated by the virtual cycle length and the virtual offset. This allows the synchronization phases to terminate early or extend later to better manage dy-



**Traffic Networks, Optimization and Control of Urban, Figure 18**  
Information processing at an OPAC controlled intersection



**Traffic Networks, Optimization and Control of Urban, Figure 19**  
Control architecture in VFC-OPAC

namic traffic conditions. *Vfc-opac* consists of a three-layer control architecture as shown in Fig. 19.

**Layer 1:** The *Local Control Layer* implements the *OPAC III* rolling horizon procedure using the dynamic programming model of *OPAC II*. It continuously calculates optimal switching sequences for the Projection Horizon, subject to the VFC constraint communicated from Layer 3.

**Layer 2:** The *Coordination Layer* optimizes the offsets at each intersection (once per cycle). This is done by searching for the best offset of the *PS* (primary signal) within the mini-network shown in Fig. 4. A choice of three offset increments is being considered: 0 (no change), +2 – sec (move right one interval), –2 – sec (move left one interval). The cyclic flow profile associated with each incoming link is being discharged by the model through the intersection and projected to the downstream intersections, *SS* (satellite signals). All other parameters are being kept at their latest values in a relaxation mode. Since the coordination process is carried out in a distributed fashion at each intersection, each *SS*, in its turn, is also considered a *PS* of its own mini-network once during each cycle.

**Layer 3:** The *Synchronization Layer* calculates the network-wide *virtual-fixed-cycle* (once every few minutes, as specified by the user) in order to maintain a rhythmic operation of the signals in the network. The objective is to provide maximum leeway of phase switching timings as dictated by local conditions, yet maintain a capability for coordination with neighboring intersections by maintaining synchronicity of the signals which are linked in the network. The *virtual-fixed-cycle* (VFC) is calculated in a way that provides sufficient capacity at the most heavily loaded intersection while, at the same time, maintaining suitable

progression opportunities among adjacent intersections. The VFC is calculated as follows:

- Check if pre-set time period (3–5 min), or number of cycles counted, has elapsed
- Identify the dominant intersection (based on flow/saturation flow ratios)
- Establish bounds on VFC to satisfy the following requirements:
  - a. Provide sufficient capacity

$$C_a \geq \sum_j (G_j)_{\min} + L$$

- b. Keep the degree of saturation under a preset maximum  $k_m$  (e. g.,  $k_m = 0.90$ )

$$C_b \geq \frac{k_m L}{k_m - \sum_j y_j}$$

- c. Do not exceed upper phase length limits

$$C_c \leq \sum_j (G_j)_{\max} + L$$

where,  $G_j$  = length of (green) phase  $j$

- $y_j = (q/s)_j$  = ratio of flow/sat-flow on phase  $j$
- $L$  = total lost time at the intersection

The virtual-fixed-cycle is chosen as the lowest value satisfying these requirements;

$$\text{VFC} = \min(C_a, C_b, C_c) .$$

In addition, there may be exogenously determined limits on the cycle time

$$C_{\min} \leq \text{VFC} \leq C_{\max} .$$

The three-layer architecture is an effective means for implementing the distributed dynamic programming (DDB) algorithm. The VFC can be calculated separately for groups of intersections, as desired. The position of the marker line also determines the “virtual offset” of the signal. Over time the flexible cycle length and offsets are updated as the system adapts to changing traffic conditions. Due to the embedded flexibility this approach can provide improved local adaptability while maintaining network coordinability. Further information on the virtual-fixed-cycle model is given in [15].

### Multi-Level Traffic Control Strategies

The performance of the transportation system is the result of a complex interaction between the physical supply of transportation facilities, and the individual trip-makers' decisions. The supply includes the management policies and operational controls that are applied in these facilities. The flows in a transportation network are the result of an "equilibration" process between the demand for transport services and the supply of transport capacity. This was the basis for the development of "static" models for the design and planning of TSM actions [16,17].

The extension of the static framework to the quasi-dynamic and dynamic cases, where on-line control and guidance is being provided in response to real-time traffic information. The application of advanced technologies in sensing, communications and computation in intelligent transportation systems, coupled with advanced modeling concepts, provides great opportunities to improve the performance of traffic networks under both recurrent and non-recurrent conditions. Current research and development efforts are directed toward development of a Dynamic Traffic Assignment (DTA) capability to predict future traffic conditions and a Real Time Traffic Adaptive Control System (RT-TRACS) for the generation of signal control strategies. Although these models are intimately connected, their development has proceeded independently of each other so far. In the framework described herein the two models are integrated into a combined system.

### The Traffic Management Problem

The traffic management problem can be viewed as the interaction between the urban traffic manager and the individual trip-maker and their (sometimes differing) perceptions of the performance of, or the supply provided by, the transportation system. A schematic illustration of the interactions in a transportation network is shown in Fig. 20. The physical transportation system and the socioeconomic activity system interact, via the equilibration process, to produce a set of flows on the links of the network. Two major feedback loops affect this process. In Loop I, the traffic manager assesses the system's performance according to his measures-of-effectiveness and intervenes in the physical transportation system to achieve his desired objectives (or, more accurately, the objectives of the community which he represents), such as reducing total travel times, minimizing pollution, increasing safety, etc. The trip-makers, on the other hand, assess the flows according to their own perceptions, which may be different from those of the traffic manager, and propagate an adjustment

in the travel demand pattern via Loop II. Performance of the system is a result of the combined interaction of the demand and the supply feedback loops.

The problem can be set in the following context: given (fixed) demands for travel in an urban area and a (fixed) supply of transportation facilities, the traffic manager must consider a variety of management strategies to induce a traffic flow pattern that will meet, in an optimal way, the overall objectives of the community. The manager's criteria for evaluating his actions may include public interest measures of performance (MOP) such as travel time, energy consumption, noise and pollutant emissions, etc. On the other hand, individual trip-makers are assumed to minimize only their own travel costs (usually the travel time), subject to the constraints imposed by the physical system supply, the manager's actions, and the interactions resulting from the other trip-makers' decisions. The problem can be cast as *system-optimization* (the manager's objective) for flows that result from *user-optimization* (the trip-makers' objective). This leads to a compound mathematical optimization formulation, as follows:

$$\min_M \left\{ Z_1 \left[ M, \arg \min_F Z_2(F) \right] \right\} \quad (15)$$

subject to flow conservation constraints:

$$\sum_{p \in P(i,k)} h_p = r_{(i,k)} \quad \text{with } (i,k) \in I \quad (16a)$$

and non-negativity constraints:

$$h_p, r_{(i,k)} \geq 0 \quad (16b)$$

where:

$M$  = the set of management variables under the control of the traffic manager

$I$  = set of all OD pairs  $(i,k)$

$P(i,k)$  = set of all simple paths between OD pair  $(i,k)$

$f$  = set of all link flows

$f_j$  = flow on link  $j$

$r_{(i,k)}$  = rate of trip interchange (demand in vehicles) between origin node  $i$  and destination node  $k$ , with  $(i,k) \in I$

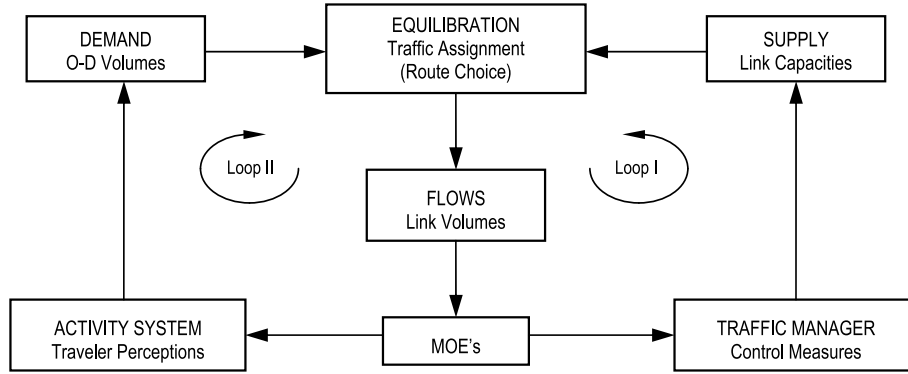
$h_p$  = flow on path  $p$ , with  $p \in P(i,k)$ .

The compound objective function in Eq. (15) consists of the following components:

$$Z_1 = \sum_j f_j \cdot w_j(f_j, M) \quad (17a)$$

$$Z_2 = \sum_j \int_0^{f_j} c_j(x) \cdot dx \quad (17b)$$





**Traffic Networks, Optimization and Control of Urban, Figure 20**  
Static interactions in a transportation system

where  $c_j(f_j)$  is the average user-perceived travel cost function on link  $j$  and  $w_j(f_j, M)$  is a more general performance function reflecting the multiplicity of objectives pursued by the traffic manager in the public's interest. The traffic management problem stated by Eq. (15) consists of a primary optimization program (the *system-optimization*), and a secondary optimization program (the *user-optimization*). The “argmin” of a mathematical program is the optimal solution of the program; therefore,  $Z_1(\cdot)$  has to be evaluated at the optimal solution of the secondary optimization program and represents aggregate traffic performance in the network with user-optimal flows.

An alternative mathematical representation of the traffic management problem is given by Tan et al. [18] and is labeled “hybrid optimization”. In this case the objective function of the traffic management problem coincides with the system optimization objective described in Eq. (17a), while the user optimization objective acts as a constraint. Tan et al. show that Wardrop's First Principle describing the user optimization problem can be fully expressed mathematically as follows:

1. Every path  $p$  must carry non-negative flows:

$$h_p \geq 0$$

2. Path flows of the same O-D pair sum to the required flow  $\sum_{p \in P(i,k)} h_p = r_{(i,k)}$  with  $(i, k) \in I$
3. Utilized paths have the same cost  $c_q$  which is equal to the minimum cost and paths with a greater cost than the minimum carry no flow. This is expressed by the following:

$$c_q(h_q, w_q) \geq \sum_{p \in P(i,k)} h_p \cdot c_p(h_p, w_p) \cdot c_p(h_p, w_p) / r_{(i,k)}$$

$$\text{with } (i, k) \in I \text{ and } p \in P_{(i,k)}.$$

Thus the hybrid optimization formulation of the traffic

management problem is:

$$\text{Minimize } Z_1 = \sum_j f_j \cdot w_j(f_j, M)$$

subject to conditions (1.), (2.) and (3.) above.

Decision variables in the primary program are the set of management variables  $M$  under the control of the traffic manager. Decision variables in the secondary program are the link flows  $F$ , given the OD demands  $r_{(i,k)}$ . Major categories of actions that can be considered are: (a) actions to ensure the efficient use of existing road space; (b) actions to reduce vehicle use in congested areas; and (c) actions to improve transit service. For example, representative variables that are particularly amenable to formal optimization include traffic signalization variables (cycle time, green splits, offsets, signal phasing and phase sequencing) traffic operations variables (e.g., channelization, one-way streets, metering, variable message signs, reversible lanes) and preferential treatment variables (e.g., reserved or preferential lanes, exclusive transit lanes, transit vehicle preemption of traffic signals, special turning lanes, truck routes, parking control). The compound formulation and the hybrid-optimization are similar models; however, the latter has been more thoroughly investigated and, therefore, is more amenable to computation.

The framework given above for the static traffic management problem can be extended to the dynamic case. In principle, the same model applies; however, both the OD demands and the control actions  $M$  are now time-dependent. Further information is given in [19].

### Signal Control and Traffic Assignment

Although signal control and route choice are intimately intertwined on the street, there is little practical experience with the integration of signal control and traffic as-

signment from a modeling standpoint. As indicated previously, most modeling efforts were done independently, i. e., assignment separate from control.

Smith [20] proposed a combined assignment-control model to take into consideration the effect of signal settings on route choice. The model is based on the  $P_0$  control policy which is designed to maximize the “travel capacity” of the road network, under the assumption that all drivers seek their own best route. From computations conducted by Ghali and Smith [21] it was concluded that the  $P_0$  control policy performs overall better than two standard control policies, (a) the local delay minimization for current traffic patterns at each intersection and (b) Webster’s equisaturation method, where green times are selected to equalize the degree of saturation on competing approaches. Gartner and Al-Malik [22] presented a combined network model which simultaneously accounts for both the route choices made by motorists and the desired signal controls to match these choices. An important feature of this approach is that offset optimization can also be incorporated.

### Real Time Traffic-Adaptive Control System (RT-TRACS)

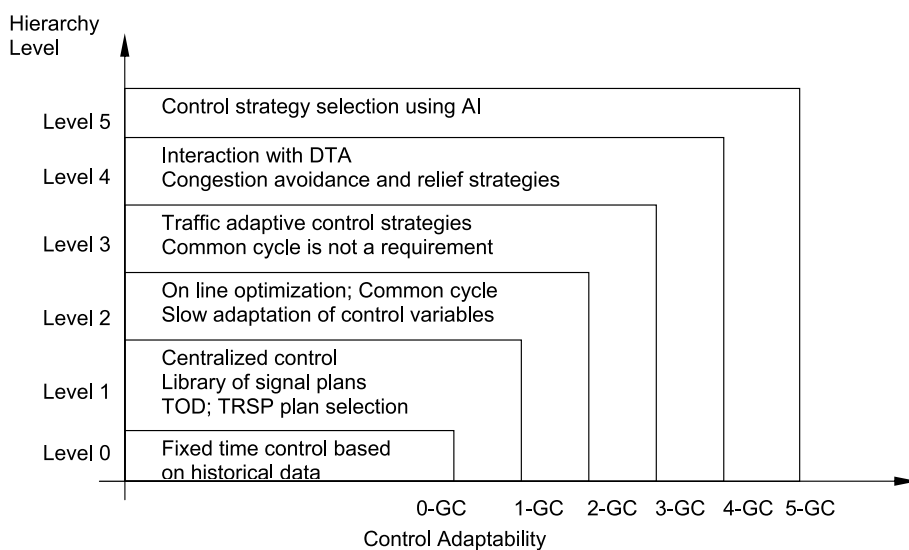
RT-TRACS consists of a multi-level hierarchy of traffic control strategies and is illustrated in Fig. 21. This system provides a modern approach to adaptive traffic signal control and has been specifically designed to facilitate integration with traffic prediction [23]. The section below

describes an RT-TRACS version that can also be incorporated within the framework of a dynamic traffic management system which is a dynamic version of the system illustrated in Fig. 20.

The degree of adaptability corresponds to the generation of the control strategy development, where  $n$ -GC is the  $n$ th generation control. Each level encompasses the capabilities of the lower levels in a nested fashion. The different levels and their interaction with predicted traffic data are described below.

**Level 0** is the most basic type of signal control. The traffic engineer determines timing plans manually, or by some PC-based optimization program, using historical volume data. This level is suitable for recurrent traffic patterns, i. e., when there is little need to predict traffic flows, and there is infrequent updating of plans.

**Level 1** corresponds to the 1-GC system: centralized control with limited availability of traffic surveillance and communications. Timing plans are calculated off-line using a multitude of scenarios to create a library of plans that can be activated in response to changing traffic and network conditions. Time of day (TOD) and traffic responsive (TRSP) options are available. Update interval is 15 minutes or higher. This level is suitable for application in the quasi-dynamic case. Integration with DTA is achieved in the following way: given a predicted O-D matrix, the system selects the most suitable timing plan in the library for the upcoming period. The plan should be created by a combined traffic assignment/signal optimization procedure and

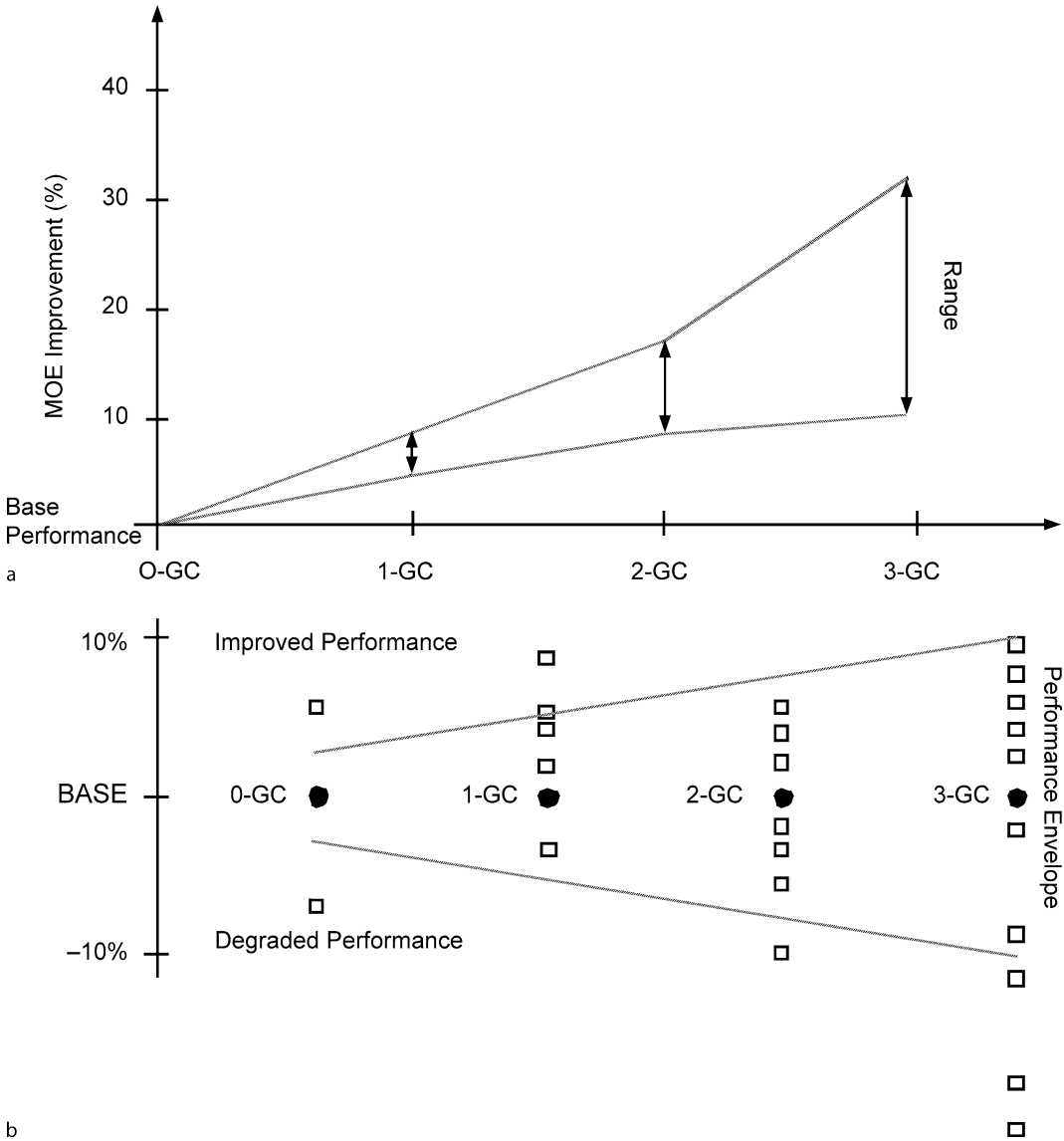


Traffic Networks, Optimization and Control of Urban, Figure 21  
Nested hierarchy of RT-TRACS control levels

should match the current demand and supply conditions.

**Level 2 (2-GC)** is centralized control with on-line optimization using a fixed, common cycle. Similar to Level 1, time frames for optimization are shorter (5 to 10 minutes), plans are being generated on-line. This level of control can respond more rapidly to changes in travel demands and/or capacities. For example, progression schemes for priority routes can be generated on-line in conjunction with the traffic prediction module.

**Level 3 (3-GC)** is a fully-adaptive traffic signal control system that may also incorporate the capabilities of the previous levels if warranted. A common, fixed cycle time is not required for coordination. Capability of phase sequence optimization is available at selected locations. Example of a 3-GC strategy: the OPAC network version. This level of control is capable of reacting to rapidly changing traffic conditions and can also be used in conjunction with assignment capabilities to provide optimal control and rerouting of traffic. The latter capability is incorporated in the next level.



Traffic Networks, Optimization and Control of Urban, Figure 22

a Expected relative performance of control generation. b Reported relative performance of control generation

**Level 4** includes all the capabilities of 3-GC with additional intelligence so that the system can interact with the DTA module. It can provide dynamic priority control on selected routes and implement congestion avoidance and relief strategies. This level is most suitable for handling incident and accident conditions.

**Level 5** is a super level that makes the most efficient use of the available array of control strategies based on accumulated experience under local conditions. Selection of the appropriate control strategy for a particular condition can be invoked by Artificial Intelligence technology. Practical experience has shown that no single strategy is optimal for the entire spectrum of demand and supply variations. Therefore, it is beneficial to have available an arsenal of strategies and to be able to choose the most appropriate one when needed.

RT-TRACS provides a new design that is different from the traditional signal control generation strategies that are still state-of-the-art. In the traditional strategies no consideration is given to re-assignment of traffic, whereas here assignment is an integral part of the signal control generation process. The mutual interdependence between route choice and signal control is embodied at all levels of the hierarchy.

### Control System Performance – Analysis of Complexity

The hierarchical framework described above offers the possibility to tailor the system's capabilities to particular needs and means. Advanced strategies can be deployed gradually and can coexist with lower level strategies. The full spectrum of capabilities can be built-up gradually in terms of deployment of sensors, surveillance equipment, controllers and communications, as well as central control hardware and software. It has been common wisdom that increasing responsiveness will contribute to improved traffic performance. Experience indicates that this has not always been the case. The more complex a system is the wider is the dispersion of the performance results [23]. The argument can be best illustrated with reference to Fig. 22. It shows the spreads in performance that are being perceived to exist among the different control generations (Fig. 22a), as well as the spreads in performance that were actually measured in the field (Fig. 22b). The points shown in the figure were collected from published reports in the literature. These figures illustrate an interesting phenomenon: the more "advanced" (i.e., responsive) strategies do not lead to improved performance all the time, notwithstanding the common perception that they do. More likely, they lead to a wider spread in per-

formance results compared with a common basis. In some cases, due to reasons that are not completely explained, traditional off-line methods perform better than responsive methods. Therefore, one of the principal objectives of any intelligent control system would be to invoke a particular control strategy that will be most suitable for existing conditions so that overall performance of the system is optimized.

An expert system can be used to make sure that we implement a particular generation of control strategies when it is likely to yield the best performance. Development of the system requires careful characterization of signal networks and identification of the particular traffic flow patterns that are most amenable to benefit from a particular control strategy. The underlying proposition is that lower level strategies may often be as good, or better than higher level, more advanced strategies. By recognizing the conditions under which the particular strategies perform best, we can optimize overall system performance. This is the task of the 5th generation control level.

### Future Directions

With the continuing developments in technology, one can expect to have gradual improvements in the technology of detection, communication and optimization which would lead to better real-time information on the status of the transportation system and, thus, to improved online control and performance. Specific areas in which advances can be sought and are likely to be achieved are: traffic-adaptive control of signal systems and the combined optimization of controls and routing. In particular, the latter would benefit from the proliferation of in-vehicular route guidance devices that are tied-in to advanced traveler information systems.

### Bibliography

#### Primary Literature

1. Webster FV (1958) Traffic signal settings. HMSO, London
2. Gerlough DL, Huber MJ (1975) Traffic flow theory. Special report 165. Transportation Research Board, Washington DC
3. Greenshields BD, Schapiro D, Ericksen EL (1947) Traffic performance at urban street intersections. Technical Report, No 1. Yale Bureau of Highway Traffic
4. Morgan JT, Little JDC (1964) Synchronizing traffic signals for maximal bandwidth. Oper Res 12:896–912
5. Little JDC (1966) The synchronizing of traffic signals by mixed-integer linear programming. Oper Res 14:568–594
6. Little JD, Kelson MD, Gartner NH (1982) MAXBAND: A program for setting signals on arteries and triangular networks. Transp Res Res 795:40–46

7. Gartner NH, Assmann SF, Lasaga F, Hou DL (1991) A multi-band approach to arterial traffic signal optimization. *Transp Res* 25B(1):55–74
  8. Gartner NH (1972) Constraining relations among offsets in synchronized signal networks. *Transp Sci* 6:88–93
  9. Gartner NH, Stamatiadis C (2002) Arterial-based control of traffic flow in urban grid networks. *Math Comput Model* 35: 657–671
  10. Robertson DI (1969) TRANSYT: A traffic network study tool. TRRL Report No LR 253. Transportation and Road Research Laboratory, Crowthorne
  11. Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – A traffic responsive method of coordinating signals. TRRL Report No LR 1014. Transportation and Road Research Laboratory, Crowthorne
  12. Gartner NH (1983) OPAC: A demand-responsive strategy for traffic signal control. *Transp Res Rec* 906:75–81
  13. Bellman R, Dreyfus S (1962) Applied dynamic programming. Princeton University Press, Princeton
  14. Grafton RB, Newell GF (1967) Optimal policies for the control of an undersaturated intersection. In: Edie LC, Herman R, Rothery R (eds) *Vehicular traffic science*. Elsevier, New York, pp 239–257
  15. Gartner NH, Pooran FJ, Andrews CM (2002) Optimized policy for adaptive control strategy in real-time adaptive control systems: Implementation and field testing. *Transp Res Rec* 1811. Transportation Research Board, Washington DC, pp 148–156
  16. Gartner NH, Gershwin SB, Little JDC, Ross P (1980) Pilot study of computer-based urban traffic management. *Transp Res* 14(1/2):203–217
  17. Gershwin SB, Ross P, Gartner NH, Little JDC (1978) Optimization of large traffic systems. *Transp Res Rec* 682:8–15
  18. Tan HN, Gershwin SB, Athans M (1979) Hybrid optimization in urban traffic networks. Report No DOT-TSC-RSPA-79-7. Transportation Systems Center, US Dept of Transportation
  19. Gartner NH, Stamatiadis C (1998) Integration of dynamic traffic assignment with real-time traffic adaptive signal control. *Transp Res Rec* 1644:150–156
  20. Smith MJ (1980) A local traffic control policy which automatically maximizes the overall travel capacity of an urban network. *Traffic Eng Control* June 1980:298–302
  21. Ghali MO, Smith MJ (1994) Comparisons of the performances of three responsive traffic control policies, taking drivers' day-to-day route choices into account. *Traffic Eng Control* October 1994:555–560
  22. Gartner NH, Al-Malik M (1996) A combined model for signal control and route choice in urban networks. *Transp Res Rec* 1554:27–35
  23. Gartner NH, Stamatiadis C, Tarnoff PJ (1995) Development of advanced traffic signal control strategies for ITS: A multi-level design. *Transp Res Rec* 1494:98–105
- Organization for economic cooperation and development (1987) *Dynamic traffic management in urban and suburban road systems*. OECD, Paris
- Papageorgiou M (ed) (1991) *Concise encyclopedia of traffic & transportation systems*. Pergamon Press, Oxford
- RiLSA (2002) *Richtlinien für Lichtsignalanlagen für den Strassenverkehr*. Forschungsgesellschaft für Strassen-und Verkehrsweisen, Köln

## Traffic Prediction of Congested Patterns

HUBERT REHBORN<sup>1</sup>, SERGEY L. KLENOV<sup>2</sup>

<sup>1</sup> Group Research GR/ETF, HPC: G021, Daimler AG, Sindelfingen, Germany

<sup>2</sup> Department of Physics, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Congested Traffic Pattern Features on Freeways Relevant for Prediction](#)

[Congested Patterns in Urban Areas and Their Prediction](#)

[Reconstruction of Freeway Congested Traffic Patterns](#)

[Based on Measured Data](#)

[Methods for Traffic Prediction: Spatiotemporal Pattern](#)

[Analysis Based on Kerner's Three-Phase Traffic Theory](#)

[Versus Other Approaches](#)

[Long-Term Traffic Prediction](#)

[Applications of Traffic Prediction](#)

[Future Directions](#)

[Acknowledgment](#)

[Bibliography](#)

### Glossary

**Measurements of traffic flow variables** Traffic variables like flow rate (normally given as number of vehicles per hour), vehicle speed and occupancy (percentage of time in which a road location (i. e., normally a detector) is occupied by a vehicle in a certain time interval) are measured with local detectors (single or double induction loop detectors) in the road network. The locally measured traffic variables are averaged and transferred from the roadside detectors to online traffic management centers in usually fixed time intervals (commonly 30 s in the USA, 60 s in Germany). Individual vehicles can measure travel times on road segments, which can be aggregated to travel times.

### Books and Reviews

- Gartner NH, Improtta G (eds) (1995) *Urban traffic networks: Dynamic flow modeling and control*. Springer, Berlin
- Gordon RL, et al (1996) *Traffic control systems handbook*. Federal highway administration. Report FHWA-SA-95-032. US Dept. of Transportation, Washington DC
- Homburger WS (ed) (1982) *Transportation and traffic engineering handbook*, 2nd edn. Prentice-Hall, Englewood Cliffs
- Kerner BS (2004) *The physics of traffic*. Springer, Berlin



**Spatiotemporal congested traffic pattern** Spatiotemporal congested traffic patterns are traffic patterns on freeways that emerge and develop in space and time as a result of the onset of congestion in an initially free traffic flow. Spatiotemporal congested traffic patterns occur mostly at freeway bottlenecks associated with on- and off-ramps, roadwork, decreasing freeway lane numbers, etc.

**Kerner's three-phase traffic theory** In Kerner's three-phase traffic theory, empirical (measured) features of spatiotemporal congested traffic patterns are explained. Three traffic phases are identified and the features of spatiotemporal dynamics within these traffic phases, as well as the phase transitions between them are examined. These three traffic phases are (i) free traffic, (ii) synchronized flow and (iii) wide moving jam. The synchronized flow and wide moving jam traffic phases are associated with a congested traffic state. These two traffic phases are defined through the following empirical criteria. The definition of the wide moving jam phase [*J*]: a wide moving jam is a moving jam that propagates through any bottleneck while maintaining the mean velocity of the downstream jam front. The definition of the synchronized flow phase [*S*]: the downstream front of a synchronized flow region does not exhibit the above characteristic jam feature; in particular, the downstream front of a synchronized flow is often fixed at the location of the bottleneck.

**Onset of congestion** Onset of congestion (traffic breakdown) is defined as congested traffic pattern emergence in an initial free flow. The onset of congestion is observed mostly at freeway bottlenecks. In Kerner's three-phase traffic theory, the onset of congestion is explained by a phase transition from free flow to synchronized flow at a bottleneck ( $F \rightarrow S$  transition for short).

**Recurring and non-recurring congestion** The congestion emerging by the routine presence of large numbers of vehicles on roads is called recurring congestion. Unexpected disruptions of traffic caused by lane blockings, incidents from work zones, weather events etc. are called non-recurring congestion. Spatiotemporal congested patterns are caused by both recurring and non-recurring congestion. After recognition of a congested pattern the further prediction is mainly independent from its original reason. The methods described in this contribution are valid for both kinds of congestion reasons.

**Wide moving jam emergence** In empirical observations, wide moving jams emerge spontaneously in synchro-

nized flow. As explained by Kerner's three-phase traffic theory, this wide moving jam emergence is associated with a first-order phase transition from synchronized flow to wide moving jam ( $S \rightarrow J$  transition for short). Thus, wide moving jams emerge in free flow as a result of a sequence of two phase transitions: Firstly, a phase transition at a certain freeway location from free flow to synchronized flow occurs. Later and usually at another road location this synchronized flow can transform into a spatiotemporal congested pattern with propagating wide moving jams.

#### **Prediction of spatiotemporal congested traffic pattern**

The function of the prediction approach is to forecast the whole congestion picture in space and time. Parameters of the congested patterns, such as size and form, at specific bottlenecks can provide a "fingerprint" of the road network, because they are typical, recurring and specific to the related bottleneck. These features facilitate the predictability of traffic congestion.

**Time series of traffic variables** Traffic variables can be archived and combined as daily or hourly time series. For example, a flow rate time series could be used as the traffic demand estimation for vehicles driving into congestion. A time series of travel time for a road network could be used for dynamic route guidance applications. Important parameters for the choice of the most probable time series used for prediction are at least the characteristics of the specific day (e. g., Monday morning rush-hour) and the probability of occurrence of each of the time series of traffic variables for days with the same characteristics.

**Short-term prediction with time series** A short-term prediction with time series can be performed via the comparison of a historical database of traffic variables and the current traffic data measurements. The choice of the appropriate traffic data curve that is then used for traffic prediction is called "matching". This choice is performed from the historical database finding the closest match to the current traffic data. The best matched time series is not necessarily the one with the highest probability among the different time series in the database. Therefore, matching can increase the prediction quality considerably.

**Long-term prediction with time series** Long-term prediction is defined to be independent from the current situation, i. e., traffic variables solely from a historic database are used. This is the established method for traffic prediction with larger time horizons, e. g., planning tomorrow's roadwork. In contrast, short term traffic prediction is influenced by the current traffic

situation. Without this influence the most probable time series from the database can be chosen directly to perform a specific traffic variable prediction.

**Data mining or physics of traffic – two different approaches to traffic prediction** In “data mining” approaches reproducible features of measured traffic data are identified and analyzed through the use of various learning systems like neural networks, regression techniques, time series, wavelet analysis, Bayesian networks, etc. The nature of the traffic phenomena does not play a key role for this reproducible traffic feature learning. Instead, reproducible phenomena of the input data are learned by those methods. In contrast, “physics of traffic” approaches understand, explain and model these reproducible features of measured traffic data. Then features of traffic phenomena are used as the basis for traffic prediction methods.

### Definition of the Subject

Empirical congested traffic patterns and their predictions are an interesting field of complex systems because of their complex spatiotemporal behavior. The physics of traffic with its non-linear phenomena, self-organization processes based on individual driver behavior and phase transitions from free to congested traffic states are a great challenge for traffic prediction approaches. The road infrastructure approaches its capacity with severe congestion problems because of the high and increasing mobility demand in many countries of the world: a better understanding of traffic congestion patterns, their emergence and their dissolution as well as their prediction may improve both the complex individual and collective management of traffic networks. Due to advancing traffic data measurement techniques, the variety and complexity of available traffic data is steadily increasing. The amount of available traffic data can be used as archive information for the time series of traffic variables. Traffic predictions in the long term need historical knowledge about the traffic variables. The available historical traffic data varies from single vehicle trajectory data and large amounts of local induction loop detector data up to traffic message databases from service broadcasters. Specifically, the traffic flow rate and/or the travel times are needed for an estimation of future traffic situation. In the long-term, traffic patterns are recurring and regular (like a morning peak traffic hour with commuters). This regularity of traffic makes long-term time series based traffic predictions possible and feasible. On the other hand, a comparison with a current situation is a basis of short-term traffic prediction.

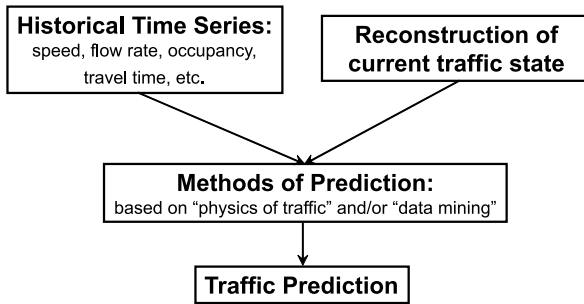
Relevant applications of traffic predictions are, from the collective point of view, the network predictions of traffic states as basic information for traffic management and control as well as the planning of the probable congestion consequences of roadwork. From an individual driver's point of view, traffic prediction is important for efficient route choice and the estimation of arrival times. Additionally, precise congestion information could be used for safety and comfort applications in vehicles like the adaptation of cruise control systems.

### Introduction

Every morning, the commuters in urban areas of the world suffer from time delays caused by traffic congestion. The congestion is regular, recurring and permanent at certain locations of the freeway road network. Therefore, from a better understanding of the underlying physical processes in traffic, the wish for precise and reliable traffic prediction arises for multiple traffic management purposes, both collective (e.g., traffic control) or individual (e.g., dynamic route guidance). To reach these goals, empirical congestion patterns must first be analyzed and understood correctly.

Throughout the modern world, traffic congestion has enormous associated costs in terms of delay, fuel consumption and emissions, etc. One way to alleviate this problem is to build new transportation facilities and expand the transportation networks which is very cost intensive and not always possible. Another cost-effective method would be the implementation of real-time and predicted traffic generated and disseminated to the traveling public for both pre-trip planning and en-route usage. Having accurate and up-to-date traffic information, travelers would be able to have the opportunity to plan their trip and adjust their routes accordingly.

Infrastructure planning needs information on the traffic demand for its long-term construction processes. Additionally, the planning of roadwork must ensure efficiency, minimizing the effect on the current traffic demand while also considering requirements regarding costs and working necessities. The individual driver is interested in long-term traffic information for his pre-trip planning which can range from tomorrow's trip to work to his next holiday journey further in the future. Long-term time-series of traffic based on historical data are an opportunity to fulfil such individual or collective wishes for forecasted traffic information. The goal of the prediction approach is the precise prediction of a traffic variable like flow rate or travel time. Next, the probability of traffic breakdown, i.e., the possible onset of congestion at a certain location in the



**Traffic Prediction of Congested Patterns, Figure 1**  
Possible qualitative scheme of traffic prediction

road network, should be predicted. After this congestion emergence, the spatiotemporal behavior of the congested region must be forecasted.

Figure 1 illustrates a possible qualitative scheme to perform a traffic prediction: historical time series of traffic variables and the reconstruction of the current traffic state are the “input data” for traffic prediction methods including “matching”, i. e., a choice of the most probable traffic pattern, which are used then for the traffic prediction of traffic variables.

There are a huge number of various approaches to reconstruction of current traffic state and short-term prediction (Fig. 1). Related approaches can be subdivided into two main categories, distinguished by their principal concept: driven by “physics of traffic” or “data mining” (Fig. 2). Some approaches, as presented in this article, combine both “physics of traffic” and “data mining”.

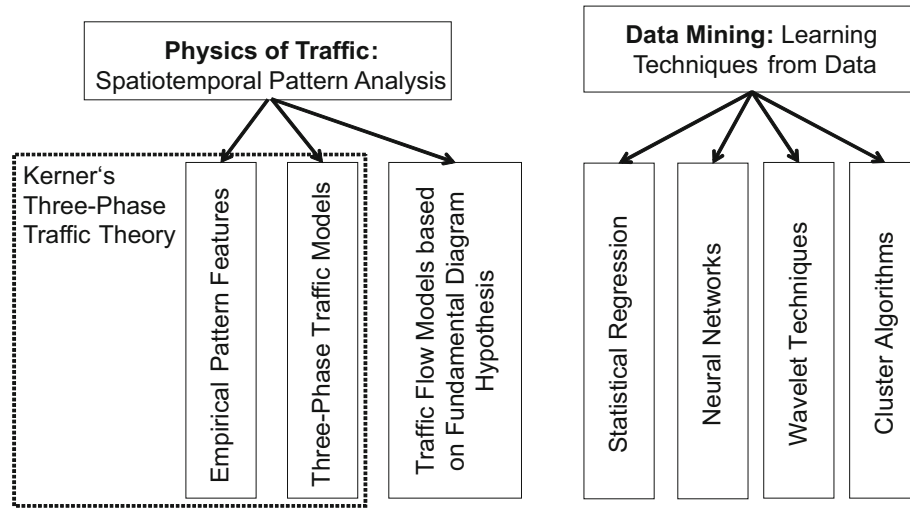
In the article, the data mining approaches are illustrated with regards to their prediction capability; in particular, we discuss the following methods for traffic prediction: “statistical regression”, “neural networks”, “wavelet techniques” and “cluster algorithms”. It should be noted, that data mining technologies may be used in addition for the generation of databases of historical time series.

The approaches associated with “physics of traffic” are devoted to the understanding of empirical (measured) spatiotemporal features of traffic as well as to modeling of these features. Then these features are used for traffic prediction. Although there were a huge number of studies of measured data on freeways (see classic works, e. g., by Edie and Foote [36], Treiterer [143] and Koshi et al. [86] and references in the book [69]), only recently Kerner solved the puzzle of empirical spatiotemporal features of freeway traffic congestion [64,67,69,72,75,76,77], ► [Traffic Congestion, Modeling Approaches to](#). Kerner’s three-phase traffic theory has been developed to explain the puzzle of freeway measured data.

It turns out that earlier known modeling approaches to freeway traffic congestion (e. g., [24,25,26,51,98,102,132,142]), which are based on the fundamental diagram hypothesis (Fig. 2), cannot explain and predict the empirical probabilistic features of the onset of congestion and resulting congested patterns occurring on freeways that are required for traffic prediction (see [70,72], ► [Traffic Congestion, Modeling Approaches to](#) for more detail). These traffic flow models are standard ones for validation of freeway traffic prediction. Thus the related simulations or other applications of these models cannot be used for a reliable freeway traffic prediction. In other words, at time Kerner’s three-phase traffic theory is the only traffic theory, which can explain the observed spatiotemporal phenomena of congested freeway traffic. For this reason, we will use in this article often Kerner’s three-phase traffic theory [69] as the basic approach associated with “physics of traffic” application for traffic prediction on freeways.

Empirical features of freeway traffic that make the prediction difficult and complex are the probabilistic nature of the traffic breakdown and the probabilistic features of the spatiotemporal patterns of freeway traffic [64,67,69,72,75,76,77], ► [Traffic Congestion, Modeling Approaches to](#). As one result to be explained later, the travel times could be predicted based on measured and archived travel times in principle precisely only in free traffic situations which are the less interesting from the driver perspective because his personal trip would not be influenced by congestion. Possible applications of freeway traffic predictions are the use in navigation systems, for safety and comfort functions in vehicles and on the infrastructure side for the traffic assignment in networks and the choice of appropriate traffic control strategies.

The article starts with a description of the variety of congested traffic pattern both on freeways and in urban areas with roads interrupted by intersections (Sects. “[Congested Traffic Pattern Features on Freeways Relevant for Prediction](#)” and “[Congested Patterns in Urban Areas and Their Prediction](#)”, respectively). FOTO and ASDA models for reconstruction, tracking, and prediction of spatiotemporal congested pattern dynamics are the subject of Sect. “[Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data](#)”. These methods are also used in Sect. “[Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner’s Three-Phase Traffic Theory Versus Other Approaches](#)” in which various approaches to reconstruction of current traffic state and short-term prediction (Fig. 2) are discussed. A long-time traffic prediction based on historical time series derived from measured traffic data are considered in Sect. “[Long-Term Traffic Prediction](#)”. In Sect. “[Applica-](#)



**Traffic Prediction of Congested Patterns, Figure 2**

General overview on approaches to reconstruction of current traffic state and short-term prediction

tions of Traffic Prediction”, currently known applications of traffic prediction are explained. Finally, we discuss some future directions and remaining tasks in the field followed by an extended bibliography.

### Congested Traffic Pattern Features on Freeways Relevant for Prediction

After extensive traffic data analyses of available stationary measurements spanning several years Kerner discovered that in congested freeway traffic two different traffic phases must be differentiated: “synchronized flow” and “wide moving jam” [64,67,68,69,75,76,77]. Experimental features of traffic on freeways were investigated and interpreted intensively in the late 1990ies ([67,75,76,77] and references in [69]). Empirical macroscopic spatiotemporal objective criteria for the phases as important elements of Kerner’s three-phase traffic theory (see the book [69] for comprehensive explanations) are as follows:

- (i) A wide moving jam is a moving jam that maintains the mean velocity of the downstream jam front, even when the jam propagates through any other traffic states or freeway bottlenecks.
- (ii) In contrast, the downstream front of the “synchronized flow” phase is often fixed at a freeway bottleneck. Within the front, vehicles accelerate from lower speeds in synchronized flow to higher speeds in free flow.

It must be noted that the observation of speed synchronization in congested traffic is no criterion for the phase

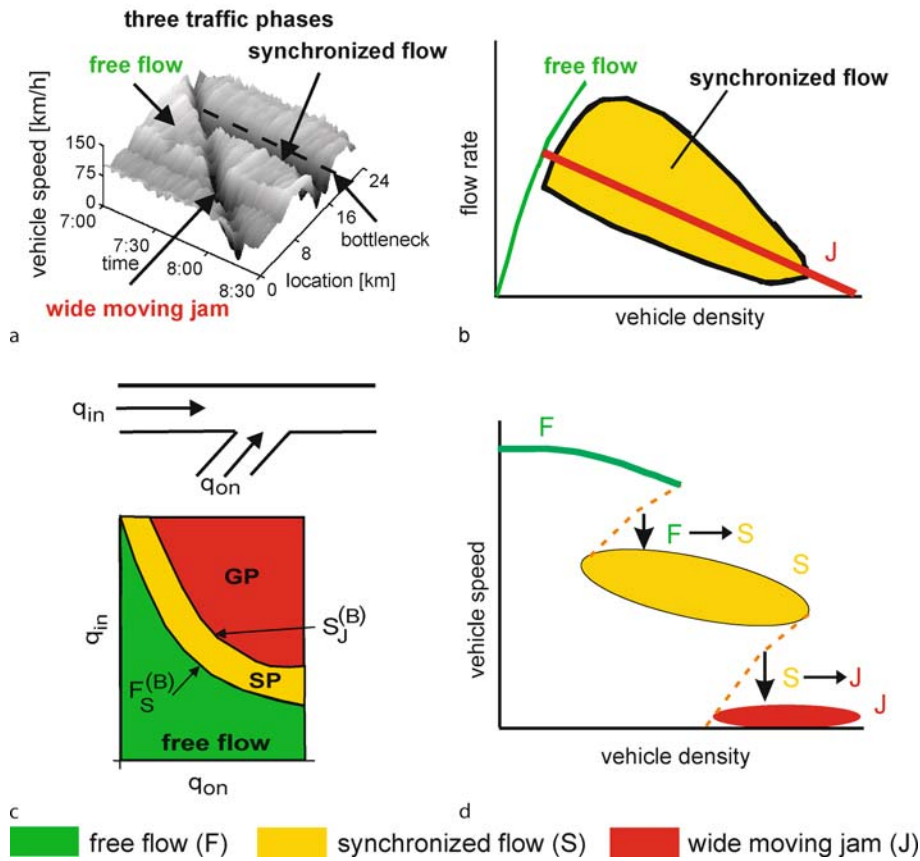
differentiation as well as some other relationships and features of congested traffic measured at a freeway location (e.g., in the flow-density plane). The clear differentiation between the synchronized flow and wide moving jam phases can be made on these above objective criteria (i) and (ii) *only*. These objective criteria define the features of the two congested traffic phases [69]. Traffic phase definition in congested traffic cannot be made definitely and consistently via differences in vehicle speeds, densities or vehicle flow rates or flow-density criteria in measured traffic variables.

As a consequence, the identification of the location and of the parameters of both traffic phases on the freeway is the basis for successful traffic prediction: the objective criteria have to be used to identify the traffic phases. After this identification, both traffic phases have predictable features and interdependences among each other which will be used in the prediction method. Some elements and issues from Kerner’s three-phase traffic theory will introduce those predictable features. The aspects described here focus on the relevant features for traffic prediction.

Figure 3a illustrates a vehicle speed profile over time and location with measured traffic data and shows the three traffic phases. A wide moving jam propagates as a “low speed valley” through the freeway stretch. In contrast, a second speed valley is almost fixed at the bottleneck location in Fig. 3a: this congested traffic phase belongs to the synchronized flow phase.

The fundamental hypothesis of Kerner’s three-phase traffic theory [69] is as follows: steady states of synchronized flow cover a two-dimensional (2D) region in

### Kerner's Three-Phase Traffic Theory



**Traffic Prediction of Congested Patterns, Figure 3**

Explanations to Kerner's three-phase traffic theory: **a** phase definition in measured data, **b** two-dimensional region of hypothetical steady states of synchronized flow, **c** simplified traffic phase diagram at an on-ramp bottleneck and **d** qualitative representation of the double Z-characteristic for phase transitions between free flow, synchronized flow and wide moving jams (with low speed traffic states within a wide moving jam [69])

the flow-density plane (so-called "Kerner's 2D-states") (Fig. 3b). This means that in contrast with the fundamental hypothesis of all earlier traffic flow theories there is no fundamental diagram for steady states of traffic flow in this theory. The red line  $J$  (so-called "Kerner's line  $J$ ", [96]) represents the velocity of the downstream front of a wide moving jam (Fig. 3b).

Some other hypotheses and results of the theory: at any density of free flow at which both an  $F \rightarrow S$  transition and an  $F \rightarrow J$  transition are possible, a critical speed (density) disturbance needed for the  $F \rightarrow S$  transition is considerably smaller than that needed for the  $F \rightarrow J$  transition. This should explain why empirical traffic breakdown is associated with a  $F \rightarrow S$  transition. This explains also why despite of free flow metastability with respect to a  $F \rightarrow J$

transition empirically observed, this  $F \rightarrow J$  transition does not occur spontaneously in empirical observations.

One of the effects of the fundamental hypothesis of the three-phase traffic theory is as follows. When a vehicle approaches the preceding vehicle that moves with a slower time-independent speed and cannot pass it, the vehicle adapts its own speed to the speed of the preceding vehicle at a space gap, which is only one possible space gap from the infinite amount of space gaps within the above 2D-region at this speed (Fig. 3b): there is no desired (or optimal) space gap at a time independent, i. e., constant speed in the car-following behavior in this theory. This feature is called the speed adaptation effect in synchronized flow.

Traffic breakdown ( $F \rightarrow S$  transition) is explained in Kerner's three-phase traffic theory as a competition be-



tween the speed adaptation effect and a well-known over-acceleration effect, i. e., a tendency of a vehicle to accelerate to a free flow speed when it seems to be possible.

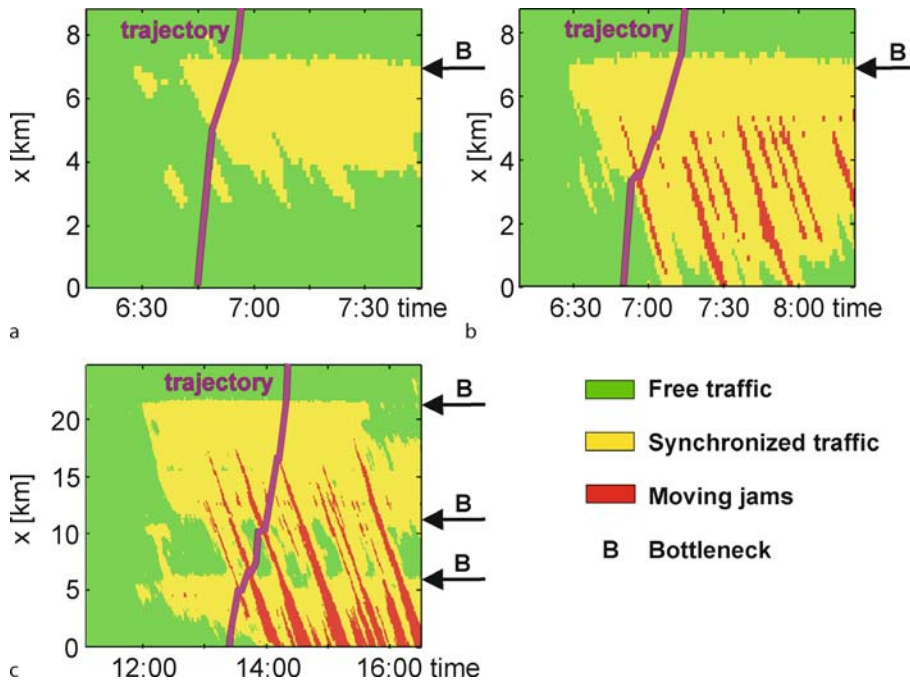
All steady states on the line  $J$  (an infinite number) are threshold states for wide moving jam existence and emergence. The line  $J$  separates all steady states of synchronized flow in the flow-density plane into two different classes: All states below the line  $J$  are stable with respect to wide moving jam existence and emergence. All states on and above the line  $J$  are metastable states with respect to wide moving jam existence and emergence (Fig. 3b).

The simplified phase diagram at an on-ramp bottleneck illustrates the phase transitions. While increasing  $q_{on}$  (the flow rate to the on-ramp) at a same level of  $q_{in}$  (the flow rate on the main road) in free flow (Fig. 3c) the following occurs: the onset of congestion in an initial free flow at a freeway bottleneck is associated with a phase transition from free flow to synchronized flow. Firstly, synchronized flow emerges ( $F \rightarrow S$  transition) at the bottleneck. The boundary  $F_S^{(B)}$  is associated with this  $F \rightarrow S$  transition. Wide moving jams can emerge spontaneously only in synchronized flow ( $F \rightarrow S$  transition). The boundary  $F_J^{(B)}$  is associated with the  $S \rightarrow J$  transition (Fig. 3c). Thus, wide moving jams emerge due to a sequence of two phase transitions called  $F \rightarrow S \rightarrow J$  transitions.

A double Z-characteristic in the speed-density plane illustrates the sequence of the  $F \rightarrow S \rightarrow J$  transitions and explains these phase transitions [69]. The first Z-characteristic related to higher speeds explains the transition (arrow labeled  $F \rightarrow S$  in Fig. 3d) and includes the traffic states  $F$  and  $S$ . The second Z-characteristic, which is related to lower speeds, explains the  $S \rightarrow J$  transition (arrow labeled  $S \rightarrow J$  in Fig. 3d) and includes the traffic states  $S$  and  $J$ . Dashed curves in the double Z-characteristic are associated with the so-called “critical speeds” within local disturbances in traffic flow required for a phase transition (see for more detailed the book [69]). It must be noted that the double Z-characteristic of the three-phase traffic theory (Fig. 3d) can usually be correctly analyzed only after spatiotemporal features of the phase transitions and congested patterns have been studied. This is because  $F \rightarrow S$  and  $S \rightarrow J$  transitions occur at different locations.

Considering the case of an isolated effectual bottleneck (a bottleneck, which is far enough from other bottlenecks), the following two main congested pattern types can emerge [69]:

1. “Synchronized Pattern” (SP): this consists exclusively of an area of synchronized flow upstream of the effectual bottleneck, i. e., no wide moving jams develop



**Traffic Prediction of Congested Patterns, Figure 4**

Spatiotemporal traffic pattern classification: **a** empirical SP (Synchronized Pattern), **b** empirical GP (General Pattern), **c** empirical EP (Expanded Pattern) [69]. Solid lines labeled “trajectory” illustrate some vehicle trajectories through these congested patterns

in this pattern (Fig. 4a). In its variants, SP could be localized at the bottleneck (“LSP”), moving upstream (“MSP”) or increasing over time (“WSP”).

2. “General Pattern” (GP): this consists of an area of synchronized flow upstream from the effectual bottleneck and wide moving jams, which develop in this synchronized flow spontaneously. GP covers thus both phases of congested traffic. It is the most frequently arising pattern at isolated bottlenecks on freeways (Fig. 4b).

If several effectual bottlenecks are closely neighboring, an expanded traffic pattern can develop (EP: Expanded Pattern) with areas of synchronized flow affecting several adjacent effectual bottlenecks (Fig. 4c). For each effectual bottleneck or for each set of several neighboring effectual bottlenecks the spatiotemporal traffic patterns possess predictable, i. e., characteristic, and recurring characteristics, e. g., the most frequently arising traffic patterns or the average expansion of synchronized flow in a GP or EP. These characteristics can be almost alike on different days and years. The existence of traffic patterns also remain preserved within a large range of the flow rate. The prediction approach based on three-phase traffic theory uses this pattern definition and pattern features for the traffic pattern database concept. The pattern classification scheme is the key element of the related prediction approach. One of the main differences to other known approaches is the definition and use of different traffic phases including their specific features and not based on aggregated information from freeways (e. g., “traffic states”) or drivers (e. g., “travel times”).

### Congested Patterns in Urban Areas and Their Prediction

In contrast to freeway sections, in urban areas a large part of the network is built up by short road sections, which are separated by each other through traffic regulations at intersections (signalized or non-signalized). The emergence and propagation of spatial congested traffic patterns is, therefore, hindered and limited by these intersections rather than self-organized spatiotemporal congested traffic patterns associated with driver interaction effects can emerge. In other words, the urban traffic prediction is mostly determined by traffic light signals at road intersections: traffic flow is interrupted during the red phase.

The recognition, tracking and prediction of those queues built at road intersections and queue parameters as number of vehicles in the queue, waiting time of the vehicles, etc. are the key elements of an urban prediction method for short road sections. Based on parameters

like red and green cycle times, vehicle speeds, inflows into the network and the section length the prediction approaches calculate a traffic state on the urban road sections. Because driver interaction effects are not important for urban traffic with *short* road sections, most models and tools of the fundamental diagram approach, which cannot be applied for an adequate prediction of freeway traffic congestion, can nevertheless be used for such an urban traffic (e. g., [41,52,101,105,123,157]); examples for tools useful for *short* road sections: VISSIM (e. g., [39,153]), PARAMICS (e. g., [13]), AIMSUN (e. g., [5,44]), INTEGRATION (e. g., [122]), DynaMIT e. g., [8]), OPAC [42], TRANSYT [130], SCOOT [57]. A review of optimization and control of urban traffic networks, in particular based on OPAC, TRANSYT and SCOOT appears in this Encyclopedia ► [Traffic Networks, Optimization and Control of Urban](#).

It must be noted that for traffic management purposes the results of the prediction should be much faster than real-time, i. e., a 60 min prediction should be calculated several hundred times in less than a system cycle of, e. g., 5 min: then the predictions can be calculated fast enough to support the choice of the optimal traffic management strategy by predicting the effects of different possible control strategies. For that purpose, Kerner introduced a model UTA (Urban Traffic Analysis) [65] (see section 22.4 in [69]) that has been further developed and implemented for large networks by Kerner, Klenov and Aleksic [74]. Goal of the model UTA is a traffic management oriented macroscopic fast network prediction in a time interval which is much smaller than the prediction horizon and calculates thousands times faster than microscopic models like the above mentioned ones.

In the model UTA, for dynamical traffic prediction in urban areas [65,72], some effective macroscopic traffic variables (the flow rates, the waiting queue’s lengths, the number of vehicles in the net’s links, etc.) are considered. For example, the effective number of vehicles in a queue  $N_q(t)$  is connected with the actual number of vehicles in a queue  $N_q^{(actual)}$  corresponding to the formula:

$$N_q(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} N_q^{(actual)}(t') dt' \quad (1)$$

All other traffic variables are averaged over time corresponding to the above formula. Thus, the effective traffic variables are related to those changes in traffic where the time scale is considerably longer than the duration of the period of light signal  $T$ . This approximation allows us fore-

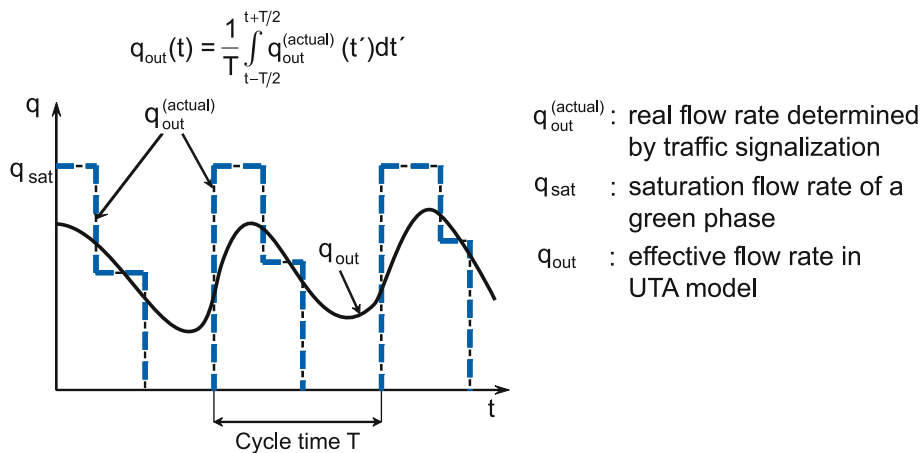
casting of traffic variables in large urban traffic nets much easier and quicker than a calculation of the actual traffic variables, which can be very complex functions of time due to the interruption of traffic during the red phases at intersections of the net. In particular, in contrast to the actual flow rates out from the links of the net  $q_{out}^{(actual)}$ , the effective flow rates  $q_{out}$  are continuous functions of time because they are not interrupted during the red phase (Fig. 5). The durations of the red and green phases  $T_R$ ,  $T_G$  influences the related effective flow rates.

Dashed curves in Fig. 6 illustrate an urban definition of congestion at an intersection with signalization: if during the cycle time  $T$  within the green phase  $T_G$  the waiting queue of the number of vehicles  $N_q^{(actual)}$  dissolves, the state of the related section will be undersaturated. If after the green phase  $T_G$  there are still vehicles in the queue, the traffic state is oversaturated (i.e., it could be called

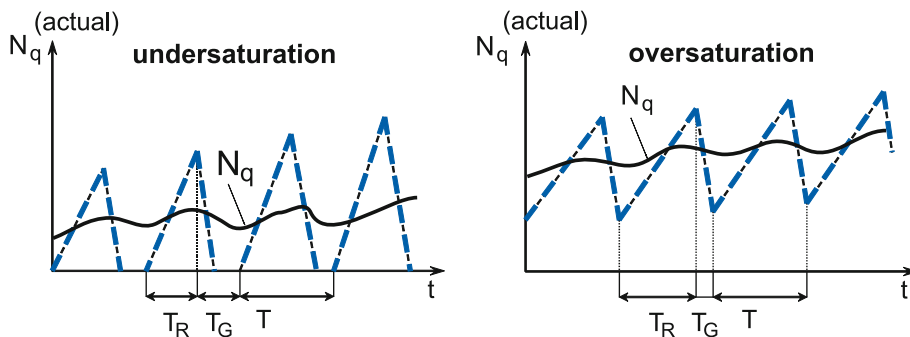
“congested”). Solid curves in Fig. 6 illustrate the effective number of vehicles in the queue  $N_q(t)$  used in the model UTA.

The model UTA [65] calculates the future values for the lengths of a queue  $L_q$  on a link and the related number of vehicles in the queue  $N_q$ , the number of vehicles inside the link  $N$ , the travel time across the link, the delay time in the queue, the flow rate to the queue  $q_{in,q}$  and all other macroscopic traffic variables in the net as functions of time.

As mentioned, the calculation time of the model UTA can be about several thousand times shorter than the one of microscopic models. It is linked to two reasons: (i) instead of a calculation of the behavior of each vehicle only the numbers of vehicles on the links are considered in UTA; (ii) the time scale considered in UTA is considerably higher than the one in microscopic models because



**Traffic Prediction of Congested Patterns, Figure 5**  
Qualitative explanation of effective flow rate  $q_{out}$  in UTA model [65]



**Traffic Prediction of Congested Patterns, Figure 6**  
Definition of under- and over-saturation at a traffic light signal on an urban road section (“urban congestion”) [65]. Dashed curve – real number of vehicles in a queue. Solid curve – effective number of vehicles in the queue used in the model UTA

UTA goes well beyond the duration of the period of light signals.

The required input information of the UTA model contains [65,69]:

- Description of the network (lanes, intersections, location of light signals),
- Traffic light signal programs (cycle time, red and green phases),
- Historical time series of saturation flow rates (per section),
- Percentages of long vehicles,
- Turning ratios at each intersection,
- Incoming flows at the boundary of the network.

Then model UTA produces for future time intervals of the prediction horizon:

- Predicted length of queues (number of vehicles in the queue),
- Travel times associated with the links of the network and with any route through the network,
- Waiting times within queues.

In contrast to other models, UTA require only parameters which can be measured directly (no “origin-destination (OD)-matrices” are necessary) The macroscopic UTA model is computational efficient because of the concept of effective continuous in space and time variables.

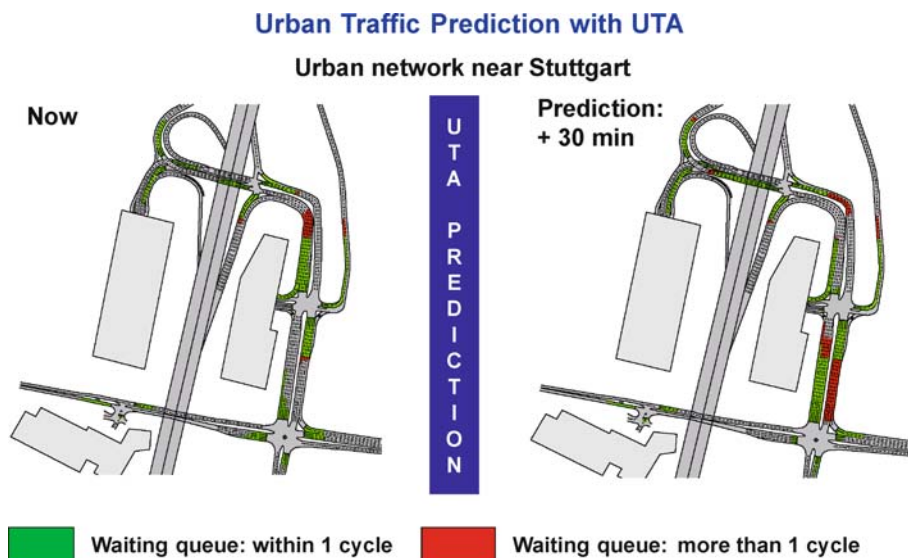
In a laboratory simulation environment developed, large road networks (e.g., Tokyo) have been calculated some thousand times faster than any microscopic simulation tools known for us [74]. This feature allows proving various control strategies for vehicle route guidance, traffic management and control and traffic assignment before their real execution. An empirical example with measured and historical traffic data as well as real light signal control has been developed in the city of Stuttgart area (Fig. 7 shows a small part of this area).

### Reconstruction of Freeway Congested Traffic Patterns Based on Measured Data

#### FOTO and ASDA Models

One of the engineering applications of this three-phase traffic theory are the FOTO and ASDA models proposed by Kerner and further developed by Rehborn, Aleksic and Haug for recognition and tracking of spatiotemporal congested traffic patterns occurring at freeway bottlenecks [66,69,78,79,81], ► [Traffic Congestion, Modeling Approaches to](#). The FOTO and ASDA models recognize and track the synchronized flow and wide moving jam traffic phases in congested traffic based on traffic measurements.

Within the FOTO model, traffic phase identification is performed first. Secondly, FOTO performs the recognition of the locations of the upstream and downstream fronts



**Traffic Prediction of Congested Patterns, Figure 7**

Empirical urban prediction example with UTA [74] for a small network part near Stuttgart: current state (left) and predicted in 30 min (right)

$x_{\text{up}}^{(\text{syn})}$ ,  $x_{\text{down}}^{(\text{syn})}$  of synchronized flow. While the downstream front of synchronized flow is fixed at the bottleneck, the position of the upstream front  $x_{\text{up}}^{(\text{syn})}$  with respect to the bottleneck is calculated by a “cumulative flow” approach with the inflowing vehicles  $q_0(t)$  [vehicles/h] and the vehicles escaping downstream from the synchronized flow  $q_n(t)$  [vehicles/h]

$$x_{\text{up}}^{(\text{syn})}(t) = \mu \frac{1}{n} \int_{t_{\text{syn}}}^t (q_n(t) - q_0(t)) dt, \quad t \geq t_{\text{syn}} \quad (2)$$

with  $20 < \mu < 40$  [m/vehicles] as parameter  $n$  as the number of freeway lanes and  $t_{\text{syn}}$  as time at which the synchronized flow is first registered.

Then, the ASDA model recognizes the upstream and downstream fronts  $x_{\text{up}}^{(\text{jam})}$ ,  $x_{\text{down}}^{(\text{jam})}$  of wide moving jams. The positions of these fronts are calculated by:

$$x_{\text{up}}^{(\text{jam})}(t) = \int_{t_0}^t v_{\text{gl}}(t) dt \approx - \int_{t_0}^t \frac{q_0(t) - q_{\text{min}}}{\rho_{\text{max}} - (q_0(t)/v_0(t))} dt, \quad t \geq t_0 \quad (3)$$

$$x_{\text{down}}^{(\text{jam})}(t) = \int_{t_1}^t v_{\text{gr}}(t) dt \approx - \int_{t_1}^t \frac{q_n(t) - q_{\text{min}}}{\rho_{\text{max}} - (q_n(t)/v_n(t))} dt, \quad t \geq t_1 \quad (4)$$

with  $q_0(t)$ ,  $q_n(t)$ , as the upstream (downstream) flow rates,  $v_0(t)$ ,  $v_n(t)$  as the upstream (downstream) vehicle speeds,  $q_{\text{min}}$  as the flow rate in the wide moving jam (normally set to zero) and the parameter  $\rho_{\text{max}}$  as the density inside the wide moving jam,  $t_0$  and  $t_1$  are time moments at which upstream and downstream fronts are first registered at the downstream position, respectively. The densities  $\rho_{\text{max}}$  and  $\rho_{\text{min}} = q_n(t)/v_n(t)$  in case when free flow is formed downstream of the jam are the two distinct characteristic densities that determine the velocity of the downstream front of a wide moving jam ► [Traffic Congestion, Modeling Approaches to](#).

The formulas (3), (4) are associated with the classic Stokes formula for the shockwave velocity  $v_s$  as follows:

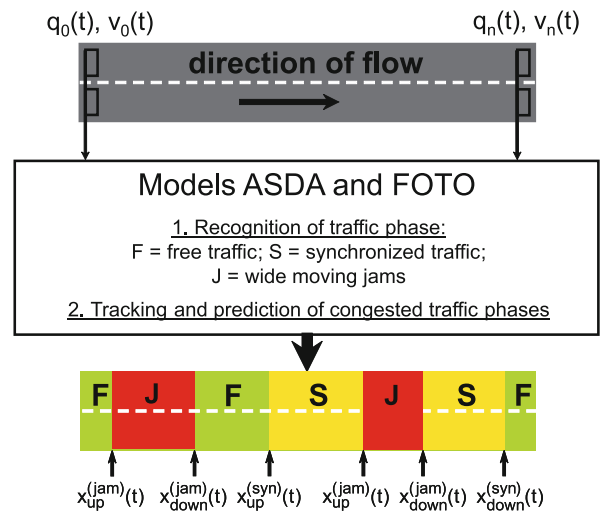
$$v_s = \frac{q_2 - q_1}{\rho_2 - \rho_1} \quad (5)$$

with  $q_1, \rho_1$  as the related flow rate and density downstream of the shockwave and  $q_2, \rho_2$  as the related flow rate and density upstream of the shockwave. It must be stressed that this Stokes formula Eq. (5) and the related

ASDA formulas Eq. (3) and (4) are fundamentally different from the well-known formula for shockwaves in the classic Lighthill–Whitham–Richards (LWR)-theory [152]:

$$v_s^{(\text{LWR})} = \frac{Q(\rho_2) - Q(\rho_1)}{\rho_2 - \rho_1} \quad (6)$$

where  $Q(\rho_2)$  and  $Q(\rho_1)$  are flow rates on the fundamental diagram that gives a single correspondence between the density and the flow rate. In contrast, in Eq. (5) and ASDA formulas Eq. (3) and (4) there is no given correspondence between the density  $\rho_2$  and the flow rate  $q_2$ , as well as between the density  $\rho_1$  and the flow rate  $q_1$ . ASDA solves the problem to find  $q_2 = q_0$  in Eq. (3) through the use of measured data from an upstream detector. In turn,  $q_1 = q_n$  in Eq. (4) is found through the use of measured data at the downstream detector. For each time interval (e.g., 1 min) based on measured data, ASDA formulas Eqs. (3) and (4) find the density as a ratio of the flow rate and the speed. As shown in empirical observations in synchronized flow there is no single relationship between flow rate and density. Instead there are an infinite number of such relations covering a two-dimensional region in the flow-density plane (“2D states” are associated with Kerner’s hypothesis). In other words, in contrast to the LWR-theory, there is no fundamental diagram of congested traffic that gives a single flow rate for a given density or vice versa a single density for a given flow rate. For this reason, as we mentioned in the fundamental hypotheses of Kerner’s three-phase traffic theory, there is a 2D-region of steady states in synchronized flow. Both the “cumu-



**Traffic Prediction of Congested Patterns, Figure 8**

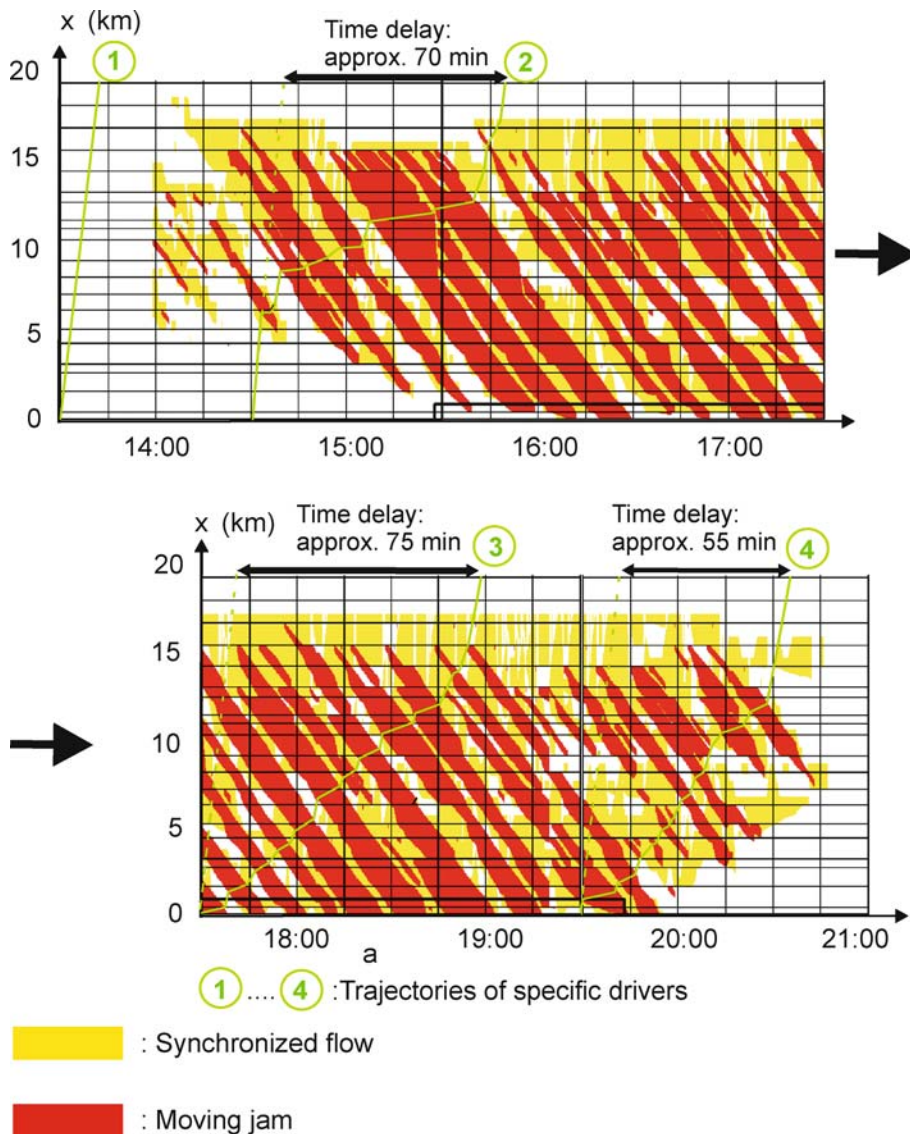
Illustration of the FOTO and ASDA model approach [66,69,78, 79,81], ► [Traffic Congestion, Modeling Approaches to](#)



lative flow” approach and Stokes shockwave formula are well-known in traffic science. The main feature of FOTO and ASDA models is that the models identify firstly two different traffic phases in congested traffic (synchronized flow and wide moving jam), and then synchronized flow is tracked with the “cumulative flow” approach, while in contrast wide moving jams are tracked with the Stokes shockwave formula.

The front locations  $x_{\text{up}}^{(\text{syn})}$ ,  $x_{\text{down}}^{(\text{syn})}$  and  $x_{\text{up}}^{(\text{jam})}$ ,  $x_{\text{down}}^{(\text{jam})}$  define the spatial size and location of the related “synchro-

nized flow” and “wide moving jam” objects, respectively. Finally, the FOTO and ASDA models track these object fronts  $x_{\text{up}}^{(\text{jam})}(t)$ ,  $x_{\text{down}}^{(\text{jam})}(t)$ ,  $x_{\text{up}}^{(\text{syn})}(t)$ ,  $x_{\text{down}}^{(\text{syn})}(t)$  in time and space (see Fig. 8). Note, that through the use of the FOTO and ASDA models the tracking of congested traffic objects is also carried out between detectors, i.e., when the object fronts cannot be measured at all. Additionally, the FOTO and ASDA models perform without any validation of model parameters in different environmental and traffic conditions. Model applications are not limited to station-



**Traffic Prediction of Congested Patterns, Figure 9**

Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram with vehicle trajectories 1–4 and related delay times from freeway A5-North in Hessen, 14th June 2006

ary detector measurements which could measure the necessary flow rates and vehicle speed directly; the use of more advanced measurement technologies like floating car data (vehicles acting as moving traffic sensors) or phone probes (phones acting as moving traffic sensors) is also possible.

Tracking of spatiotemporal congested patterns at freeway bottlenecks is done for 1200 km of freeway network in the state Hessen (Germany) through the use of the FOTO and ASDA models in an online application in the traffic control center and evaluated comprehensively in a three-month investigation of all measured congestion on the freeway A5 [85]. Since April 2004 measured data of nearly 2500 detectors are automatically analyzed by FOTO and ASDA models [69,79]. The resulting spatiotemporal traffic patterns are illustrated in a space-time diagram showing congested pattern features.

Figure 9 shows the space-time diagram for very large traffic congestion at the A5-North in Hessen reconstructed from detector data by FOTO and ASDA models. On the 14th June, 2006, for about 18 km and nearly seven hours an “Expanded Pattern” exists on the freeway. The numbers 1 to 4 illustrated in Fig. 9 represent the trajectory for individual drivers at different times of the measurement period with different travel times for each driver on the 20 km section: the first one starts at 13:30 h in free traffic and passes in 10 min, the second starts one hour later and needs approx. 80 min (in heavy congested traffic), the third starts at 17:30 h and it takes approx. 85 min to pass this section and the fourth driver starts at 19:30 h and needs approx. 65 min to drive through the congested pattern in the decreasing congestion. The delay times of more than one hour are caused by the reduced vehicle speeds in both the “synchronized flow” and “wide moving jam” traffic phases, respectively.

The traffic control center in Hessen with its model application of FOTO and ASDA models is the laboratory environment for first realization of freeway traffic prediction based on three-phase traffic theory, because of online and archived detector measurements, historical time series and the recognition of the current situation by FOTO and ASDA models.

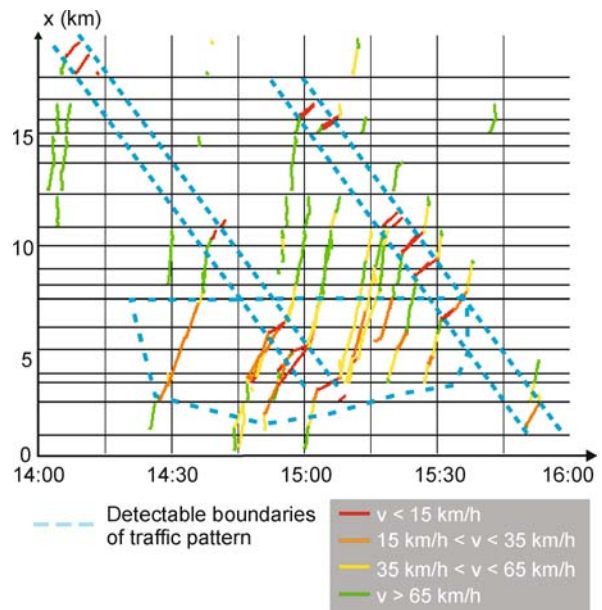
### Detection of Current Traffic States with Floating Cars

Floating cars can send information on their individual drive based on pre-defined protocols. A specific onboard-algorithm interprets the vehicle data and decides if an event criterion like a slow speed for a certain time interval of a freeway is recognized [6,38]. Then, a part of the vehicle trajectory is communicated to the traffic control center. The vehicle can act as a moving sensor in traffic in addition

or even as a replacement to stationary roadside sensors of the infrastructure (e. g., [9]).

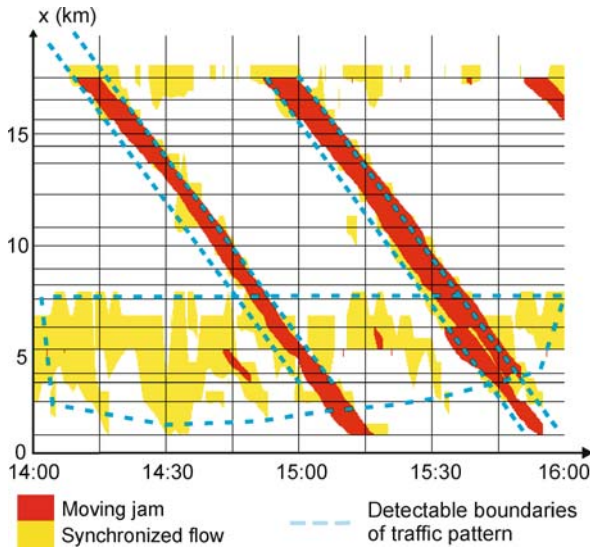
The empirical Figs. 10 to 12 show the vehicle data solely and in combination with FOTO and ASDA results from the traffic control center based on stationary detectors (horizontal lines in Figs. 10 to 12) (see [125] for further examples). Several vehicles have sent their messages on this 20 km freeway stretch. The colors of the vehicle trajectories illustrate the different speeds. Solely based on this vehicle information, the blue dotted locations of congested patterns, i. e., a region of synchronized flow and two propagating wide moving jams, would have been assumed and registered in the traffic control center. In Fig. 10 the existence of spatiotemporal traffic patterns in congested traffic has been detected solely based on moving detectors, i. e., moving vehicles. The vehicles detect in this congested situation one traffic phase with a fixed downstream location (blue region between 2–7 km in Fig. 10) and two examples of a propagating traffic phase in congested traffic (blue bars diagonal on the whole freeway stretch in Fig. 10). As a consequence, the vehicles have measured empirically two different traffic phases in congested traffic as it is stated in Kerner’s three-phase traffic theory [69].

The number of vehicles, which have sent messages in this example, has been about 20 per hour which correlates to a penetration rate of about 0,5% (estimated 4000 vehicles/h on the two lane freeway). These quantitative results

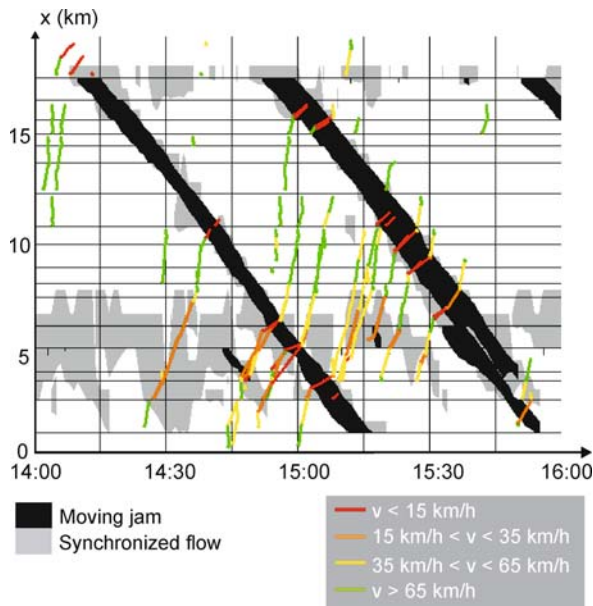


**Traffic Prediction of Congested Patterns, Figure 10**

Vehicle trajectories (approx. 20 per hour) from equipped vehicle (event messages per vehicle have been sent to the center) (see [125])



**Traffic Prediction of Congested Patterns, Figure 11**  
Congested pattern reconstruction with FOTO and ASDA models (see [125])



**Traffic Prediction of Congested Patterns, Figure 12**  
Overview of FOTO and ASDA traffic reconstruction and vehicle trajectories (see [125])

give the impression that such a small penetration rate of equipped vehicles could be sufficient for congested traffic state recognition on freeways without any stationary sensors.

The same traffic situation has been reconstructed by FOTO and ASDA models solely based on local detectors (Fig. 11). Qualitatively the same regions of the two propa-

gating wide moving jams and the synchronized flow have been recognized (i. e., the blue dotted regions).

The overview of all available data sources, i. e., local detectors with FOTO and ASDA reconstruction and vehicle trajectories prove the technical feasibility of detecting congested traffic patterns with different measurement techniques, both local and moving like a vehicle. It has to be taken into account that based on only one vehicle it is impossible to reconstruct the whole picture of the congested traffic. To perform such a congested traffic pattern reconstruction inside a vehicle, the vehicle has to have additional information from other drivers, beacons or traffic centers.

### Methods for Traffic Prediction: Spatiotemporal Pattern Analysis Based on Kerner's Three-Phase Traffic Theory Versus Other Approaches

As explained in Introduction, from a methodological point of view two main traffic prediction concepts can be distinguished (see Fig. 2):

- (i) “data mining”: statistical methods based on regression (e. g., [29,139,155,156,160]), wavelet (e. g., [61,140]), or filtering models as well as neural networks (or in hybrid combinations) (e. g., [28,116])
- (ii) “physics of traffic”: approaches with the use of empirical (measured) spatiotemporal features of traffic congestion as well as traffic flow modeling (e. g., [2,7,22,24,25,26,102,142,150,151]).

We have also stated in Introduction that the first kind of approaches more or less take techniques from other disciplines like mathematics, information technology or artificial intelligence and try to transfer successful concepts into traffic science. The nature of empirical traffic process does not play a key role in the data mining approaches. The “data mining” techniques have the principal problem of determining which data features they should focus on and which are less or not relevant for the accurate prediction.

The “physics of traffic” approaches are based on the nature of traffic phenomena. They begin with the understanding of measurements of traffic variables made in space and time as well as with simulations of driver's behavior, roads with possible bottlenecks, etc. The key element for a successful prediction concept here lies in the correct and reliable understanding of the empirical features of the traffic process and their reproducible characteristics. Furthermore, this understanding of real traffic should be incorporated in a traffic flow model, which should explain and predict the empirical traffic features.

It is well-known that traffic flow is subject to occasional abrupt disturbances (e. g., incidents, congestion) that can

potentially change the underlying dynamics and the stability of the data [34]. For example, when a traffic accident occurs, the average speed may suddenly drop to a new level. Not only discrete events like accidents can alter the underlying dynamics but also the level of congestion. Nevertheless, reliable predictions are especially valuable and needed on occasions when unexpected events occur [92]. In fact, researchers also acknowledge the occurrence of larger prediction errors in their models whenever traffic congestion sets in and dissolves [145,146] and when accidents occur [165].

Many researchers have investigated the field of traffic science (see [23,27,51,52,69,91,98,115,132,142]). However, due to a lack of understanding of the underlying physical processes and the availability of the relevant high-quality traffic data as well as its interpretation, applications of traffic prediction are today still in their infancy. In Kerner's three-phase traffic theory [69] empirical phenomena on freeways have been explained well enough to develop reliable methods for traffic prediction. In this paper, traffic prediction methods are based on recurring and predictable features of traffic congestion as they are empirically proven in the three-phase traffic theory. The combination of empirical features of freeway traffic and its correct interpretation with a pattern database is the key for successful and precise traffic prediction.

Successful traffic prediction should be focused more or less on a congestion prediction. In the industrialized world with heavy vehicle traffic on roads, the congestion is still of much less duration than the free traffic: a very simple traffic prediction citing "free traffic" at all times on all roads would be correct in large parts of the networks. The time basis for traffic prediction evaluation must not cover the whole day, but the duration of the relatively seldom congestion events. The traffic prediction quality should be evaluated and focused on for the congested time periods of the day. Prediction errors of about 30% are reported for those heavily congested situations [59]. The deviations of the "real" congested situation on the road and the predicted situation have to be compared later on in the laboratory to evaluate the prediction approach. It has to be taken into account, that in today's measurement systems the detection of the reality in freeway networks is still imperfect, but those results of a detection system have to be used as a "gold standard" for traffic prediction evaluation.

### Definition of "Short-Term" and "Long-Term" Traffic Predictions

Freeway traffic prediction could additionally be differentiated based on the time horizon: the short-term predic-

tion in a horizon of hours can be influenced by the current situation on the road. If larger time horizons have to be forecasted in traffic, the traffic prediction should be based more on historical assumptions and less on the current situation.

In literature, there is no single definition which allows distinguishing between short-term and long-term predictions. Sometimes, traffic prediction on a current day is considered as a short-term and traffic prediction on the next days as long-term. In other publications, short term is a prediction during a chosen time interval, e. g., that is equal or shorter than 60 min [137]. Overviews on short-term prediction approaches can be found in, e. g., [4,90,93,99,135,136,148,163,165]. Short and long term prediction can also be distinguished through the use of current measured data: if in the traffic prediction the current measured data are taken into account, then the prediction is considered as a short-term, because this prediction has a sense for the current day only. Traffic prediction in which only historical time series are used is then long-term prediction. The prediction for the next days is automatically a long-term prediction, because current measured data have usually no influence on prediction for the next days. In general distinguishing between long and short term predictions made in the article is used for clarity of consideration rather than for analysis of the related prediction methods. Indeed, many of the methods discussed below can be used both for short time and long time predictions.

A possible definition of "short-term" could be based on the physical features of the congested patterns found in Kerner's three phase traffic theory and the methodology of the prediction approach. If the prediction of the congested pattern is influenced by the current situation on the freeway, i. e., there is already a congested traffic pattern emerging, this situation is called "short-term prediction". If there is no use of the current traffic situation in the traffic pattern prediction method, the traffic prediction will be called "long-term prediction", because it is performed on any kind of historical assumption of breakdown probabilities or traffic pattern correlations and not the current situation. With this definition, a short-term prediction in contradiction to a solely time-based definition as in [137] could last several hours (e. g., for more than five hours in Fig. 9 after the pattern emergence).

### Traffic Flow and Travel Time Prediction

Many of the proposed forecasting models are focused on the traffic flow (e. g., [1,17,18,33,46,48,61,89,95,106,113,120,134,135,136,137,138,154,155,156]). Furthermore, most of these approaches predict the traffic flow condi-



tions at a location given previous observations at that point only. This neglects the spatiotemporal behavior of traffic with upstream moving structures. As proposed by [155], an efficient forecasting approach should involve the use of traffic flow data from upstream detectors to improve the forecasting at downstream locations. The traffic is flowing downstream, but the congestion is propagating upstream, as will be explained later in more detail; therefore the flow rate upstream is influenced by the congestion emergence downstream which is very relevant for successful predictions. Although travel times have been identified as an important parameter for the real-time performance of transportation management systems, most elder studies have been focused on predicting traffic flows. This is mainly because most traffic surveillance systems, especially in the USA, use single inductive loop detectors. These have point-measured information regarding traffic flow and occupancy. For many applications, the travel time is the most useful prediction variable and therefore much research effort has been dedicated to predicting travel times [18,21,28,30,31,37,56,63,83,87,88,90,93,97,103,104,106,108,110,112,118,119,127,129,133,145,146,166]. Those travel times are very relevant as static input data for trip planning, i. e., the reliable dynamic adaptation of navigation systems [19].

### Statistical Approaches and Time Series

Other known methods [108] use approaches from mathematical statistics to “learn” coefficients of traffic prediction in the form of correlations in time and space. These ideas try to learn implicitly spatiotemporal features of the traffic process and try to find correlations between heterogeneous traffic data for time series prediction. For example, the velocity of the propagating wide moving jam of 15 km/h against the direction of the flow as characteristic parameter of this traffic phase [69] can be “learned” statistically: such a wide moving jam registered downstream could be predicted with high probability to be 3 km upstream in about 12 min.

Another widely used approach is the use of time series of traffic variables (e. g., [33,120,144,149]). Average flow rates and speeds vary over time, but are regular and recurring for the same network location. Therefore, clustering methods based on time series similarities are used for traffic prediction. Such average flow rates can be useful for an estimation of the expected congestion due to tomorrow’s roadwork on a freeway section. Here the average flow rate is taken as traffic demand estimation. One of the principal problems of these clustering methods is the fact that the congested situations are part of the measurements which

are used for the traffic time series. Therefore, if congestion is part of a flow rate time series at a specific location or section, the demand is underestimated because no vehicles have passed the detector but have had to wait in the congested area further upstream: for prediction the expected flow rate is systematically underestimated.

Several authors have used statistical approaches (e. g., [20,53,100,106,107,108,116,120,133,156]) and time series for traffic prediction. The spikes of travel time curves caused by (sudden) traffic breakdowns are a problem for these kinds of linear approaches, because they are under-represented in the data set and the regression filters and smoothes the traffic breakdowns. Finding a proper prediction model defining the prediction variable and the number of necessary data cycles is a question of finding adequate parameters. Therefore, the success of regression techniques for short-term congestion predictions seems questionable.

### Neural Networks

One widely used “data driven” approach was the use of Artificial Neural Networks (ANN) for traffic prediction. As a relatively new mathematical model, ANN were first introduced in the 1940s (e. g., [47,49,131]). The basic idea of the neural network is to emulate the human brain, which consists of a very large number of inter-connected neurons. Each of the neurons (“cells”) have similar behavior and there is the possibility of having several layers of neurons in an ANN. ANNs try to adopt features of a given traffic data set in a learning phase and then perform a traffic prediction based on a given data set (i. e., the current traffic situation). Many researchers have used ANN for data completion or short-term forecasting of traffic variables which very often focus on travel times [1,16,31,32,35,37,55,58,60,82,83,94,97,117,118,129,134,144,145,158,159,161,162,163,164]. ANNs try to learn and predict spatiotemporal relationships in traffic data. They are capable of dealing with the spatiotemporal relationships implicitly without understanding the underlying complex processes. The ANN predictions are a “black-box”-process and, however, depend strongly on the completeness of the input data set. Therefore, the success of the ANN approach rises and falls with the choice of the training data set, which makes the approach difficult in prediction realization.

### Pattern Matching: Fitting Traffic Pattern from Database

As it is said before, the congested traffic patterns SP, GP, and EP with its variants on freeways are composed of two basic elements (“traffic phases”), i. e., the regions of



wide moving jams and of synchronized flow. The prediction will be done by forecasting the current spatiotemporal traffic objects, i. e., the regions of the wide moving jams and synchronized flow [69]. Position and width have to be predicted: this is performed based on the most similar traffic pattern for the related effectual bottleneck in the pattern database. The traffic prediction is calculated as follows: firstly, the current situation is reconstructed by FOTO and ASDA models. Next, a filtered current (“generalized”) traffic pattern is used as a search pattern for the traffic pattern database. The clustering is carried out to generalize the traffic pattern into some basic “puzzle” pieces of wide moving jam and synchronized flow objects. The reason for this is the dependence of the traffic pattern reconstruction quality on the detector system, which often has limitations in precision and reliability. Therefore, a congested traffic pattern measured by inductive loop detectors and interpreted by FOTO and ASDA models like in Fig. 9 is clustered and combined into larger regions of synchronized flow and wide moving jams, respectively. The similarity between the current clustered pattern and the pattern database is determined based on a distance measure which allows deviations in space and time (e. g., some hundreds of meters and some minutes as parameters). Then, the traffic pattern from the database is linked to the current pattern. The aggregated predicted delay time of the congested pattern consists of the delay time contributions of (i) the propagation of the current wide moving jams already detected, (ii) the predicted regions of the synchronized flow from the traffic pattern database, and (iii) the predicted new wide moving jams emerging in the future as information from the traffic pattern database.

If traffic congestion has been detected, how can the prediction pattern be identified from the database? Criteria for this “matching” process, i. e., finding the most probable traffic pattern for the future, are freeway, location of the effectual bottleneck, time, size of the current synchronized flow region, location of the first wide moving jam (if GP or EP) and the number and frequency of the wide moving jams (if GP or EP).

For applications it should be noted, that the matching process could use the current traffic situation from a traffic control center (e. g., based on online traffic reconstruction with FOTO and ASDA models) or from autonomous vehicles based on the vehicle internal measurements (Fig. 13).

### Prediction of Propagation of the Fronts of Wide Moving Jam

An approach for wide moving jam short-term prediction is as follows: from three-phase traffic theory [69] the

upstream stable propagation of the downstream front of a wide moving jam is a characteristic parameter of traffic. The upstream front is influenced by the incoming traffic flow into the wide moving jam. A short-term prediction of a wide moving jam can be based on this stable propagation feature: if a wide moving jam is detected and registered by ASDA model, the mean velocities of the fronts can be used for short-term predictions of the wide moving jam. After calculation of the mean velocities of the fronts, those values are used to predict the wide moving jam’s position in the future. Therefore, for the wide moving jams stable movement in space and time can be calculated based on the registration times at the local detectors.

A short-term prediction of wide moving jams is suggested in [81]. If both fronts of a wide moving jam have been registered, the clearing up of the jam can be obtained. This could be useful to predict the time, when traffic control methods regarding the wide moving jam could be eliminated. The time  $t_{\text{clear}}$  is obtained as [81]:

$$t_{\text{clear}} = t_{\text{reg}} - \frac{\int_{t_0}^{t_{\text{reg}}} v_{\text{gl}}(t) dt}{v_{\text{gl}}(t_{\text{reg}}) - v_{\text{gr}}(t_{\text{reg}})} \quad \text{at } |v_{\text{gl}}| < |v_{\text{gr}}| \quad (7)$$

with  $v_{\text{gl}}$  and  $v_{\text{gr}}$  as the related velocities of the up- and downstream front of the wide moving jam, respectively.  $t_0$ ,  $t_{\text{reg}}$  are the registration times of the upstream and downstream front of the wide moving jam at the downstream position. Figure 14 illustrates scheme and empirical example for a wide moving jam clearing up.

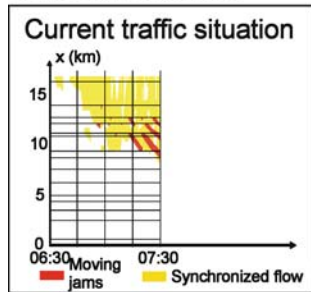
A short-time travel time prediction can be established in the case of a propagating wide moving jam, if the downstream front of the wide moving jam has been registered at a downstream location. As illustrated in Fig. 15, during the first drive section until the upstream front reached at  $t_{\text{up}}$ , the drive line is based on an average free speed of  $v_0$  the vehicle. For the subsequent drive section within the wide moving jam, the average vehicle speed of  $v_{\text{jam}}$  is used to generate the drive line. The last drive section after the downstream front of the wide moving jam is then calculated with the speed  $v_n$ . Therefore, the individual vehicle travel time for this section can be obtained as [81]:

$$t_{\text{R}} = \frac{L + v_n t_{\text{down}} - v_0 t_{\text{up}} - v_{\text{jam}}(t_{\text{down}} - t_{\text{up}})}{v_n} \quad \text{at } t_0 = 0 \quad (8)$$

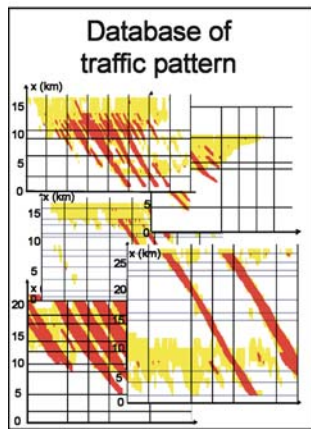
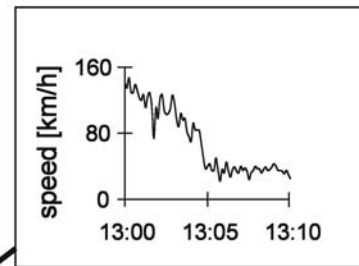
In many cases, the remaining speed within the wide moving jam  $v_{\text{jam}}$  can be neglected, so that for the travel time the following approximation formula is obtained [81]:

$$t_{\text{R}} = \frac{L + v_n t_{\text{down}} - v_0 t_{\text{up}}}{v_n} \quad (9)$$

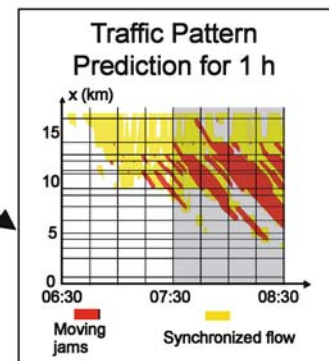
Traffic state reconstruction  
e.g. by ASDA/FOTO  
based on detectors



Traffic state reconstruction  
by vehicle measurements

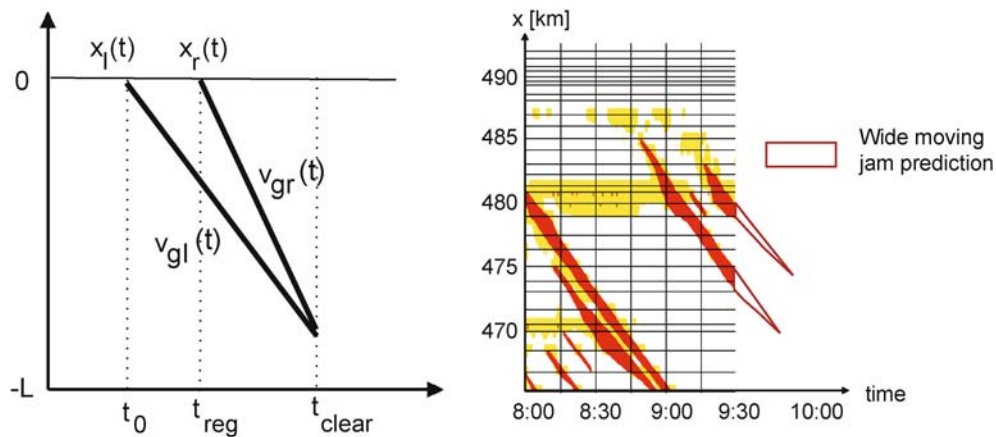


Pattern -  
Choice :  
„Matching“



Traffic Prediction of Congested Patterns, Figure 13

Use of spatiotemporal traffic pattern database for traffic prediction with "matching" [69]

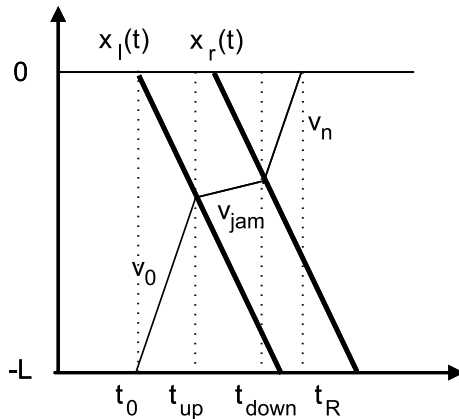


Traffic Prediction of Congested Patterns, Figure 14

Scheme for a wide moving jam clearing-up (left: scheme, right: empirical example) [81]

This method can also be used when several wide moving jams occur between two measuring points. In this case, assuming that no entry or exit roads exist in the monitoring area, the plausible assumption used is that the flow and

average speed of vehicles corresponds both upstream and downstream of the wide moving jam at the time after the downstream front has passed the downstream measuring point [81].



**Traffic Prediction of Congested Patterns, Figure 15**  
**Scheme for a travel time prediction for a vehicle [81]**

### Traffic Pattern Prediction with Traffic Flow Models

Traffic flow models offer a possibility to perform traffic predictions in different scenarios (e.g., [22,24,25,26,96,102,151]): through the use of a model, different (possibly future) influences on traffic like e.g., incidents, roadwork, accidents, etc. could be simulated. Therefore, the consequences of such future events could be estimated and used for current decisions, e.g., the choice of the best traffic management strategy. In this context, a traffic prediction based on traffic modeling could be a valuable instrument.

Many authors have used traffic flow models for traffic prediction, e.g., [22,24,25,26,102,151]. One of the key requirements for successful modeling is the correspondence with the empirical features of freeway congestion. In [73] is shown that all earlier traffic flow model approaches like the above mentioned ones based on the so-called “fundamental-diagram” approach are not able to explain those empirical features. As a consequence, the related traffic prediction results are doubtful.

On the other hand, traffic prediction and scenarios modeled based on KKW-model [69] could be a reliable instrument for decision support with traffic simulation: the consequences of different influences on traffic flow could be simulated in advance. More detail information on traffic simulation can be found in ► [Traffic Congestion, Modeling Approaches to](#).

### Long-Term Traffic Prediction

Many empirical studies of traffic have shown that the time-dependence of traffic flow possesses some predictable features (e.g., [50,52,69,80,98,109,120]) because they are recurring and regular. These features depend on character-

istics of the day (for example working day or weekend) or whether there is an event like a soccer match. The regularity and recurrence is even stronger if a certain time shift of the congestion emergence is allowed, the morning traffic peak with the same traffic congestion 10 min later is not an “anomaly”, but still a regular and similar traffic pattern.

If the characteristics of a day are known, there is a certain set of dependences of the flow rates over time for a specific freeway section. Each of these dependences is one of the historical time series for flow rates. A historical archive of traffic flow rates is therefore the basis for a time series database structured by the characteristics of the different days. At least two parameters of the databases are the day characteristics and an event which influences the traffic. The historical database contains a related probability of the occurrence for each of the time series for each of the different parameters. A cluster analysis of the historical data is a common approach to produce the time series database.

For urban areas time series based predictions have been proposed (e.g., [54,69,162]). Those methods should predict the traffic state for different time horizons. The database should use data from the infrastructure (road net, lanes, traffic lights), basic traffic engineering data (e.g., capacities, speed limits), historical traffic data as expectation values of, for example, the flow rates and current measurements.

Two types of prediction based on long-term time series are possible. Firstly, the choice of the most probable time series for a similar event or day in the future (e.g., soccer event, weekday or weekend) could be applied. Based on archived traffic data the database contains time series with additional parameters like day, time of day, event, weather, etc. The prediction for the next day is then performed based on a fulfilment of such criteria. Secondly, the prediction can include “matching of time series”, i.e., the prediction is not necessarily the most probable curve for the future from the historical database, but based on time series with minimal deviation from the current traffic variables.

Long term time series of traffic variables have two aspects: (i) processing of the raw data to produce a time series database and (ii) their later usage in a predictive application. For the processing of raw traffic data, approaches from signal processing have been proposed [45,121,140,141,147,158] to extract and detect the relevant events in the data. Others take different and multiple methodologies from statistics (e.g., [10,11,15,33,40,43]) like cluster analysis [50,120], ARIMA (Auto-Regressive Integrated Moving Average) approaches for stationary linear processes ([14,144,156]), stochastic methods [3,107,108] or Bayes’

Theorem (e. g., [20,53]). Additionally, neural network approaches from artificial intelligence have been used for unsupervised learning transforming raw traffic data into time series [144,161].

In general, the methods for long-term time series prediction try to reproduce the traffic phenomena by their different aggregation methods. The resulting time series from the raw data should model a realistic curve of the traffic variable to be useful for a predictive application. The mentioned approaches have a commonality in their usage of a large variety of data analysis techniques. Some of them do not take into account the known physical and predictable features of the traffic process [69] while others even neglect the principle feature of non-linearity and non-stationarity of traffic (e. g., ARIMA [144,156]). Without proper understanding of the relevant features of traffic congestion it is doubtful that those methods could be optimal and successful.

Also microscopic traffic simulation has been used to estimate long-term effects in traffic in combinations of time series and modeling of traffic process [22,62,84,102,142,151]. The goal is a network wide traffic prediction in a combination of measurements and modeling techniques.

### Predictable Features of Traffic Time Series

Long-term time series of flow rates are used for infrastructure planning and modeling of the traffic demand (e.g. [50,52,120]). For the long-term the typical minute values of traffic data measurements are aggregated in different time intervals. For usage as traffic demand time series the influence of traffic congestion has to be minimized or filtered. Data aggregation is one key technique to extract predictable features like the flow rate curve from raw measurement data [111].

An example of typical detector measurements for double induction loop detectors is shown in Fig. 16. Three loops count the number of vehicles ("flow rate") and their average speed in one minute intervals on each lane. Here, the aggregated values for the freeway direction A5-North are illustrated for Friday, 23rd March, 2001. The minute values show several local minima of flow rates in the afternoon, both clearly visible in the flow and speed time series (Fig. 16, top). The emergence of the congestion has occurred at a downstream location, and the local detector on the freeway has registered the local minima. These local minima maybe associated with a part of the whole congested traffic pattern, e. g., some wide moving jams propagating upstream as part of a General Pattern (see [69] for explanation): in the dynamic spatiotemporal traffic on freeways, the local detector only registers a part

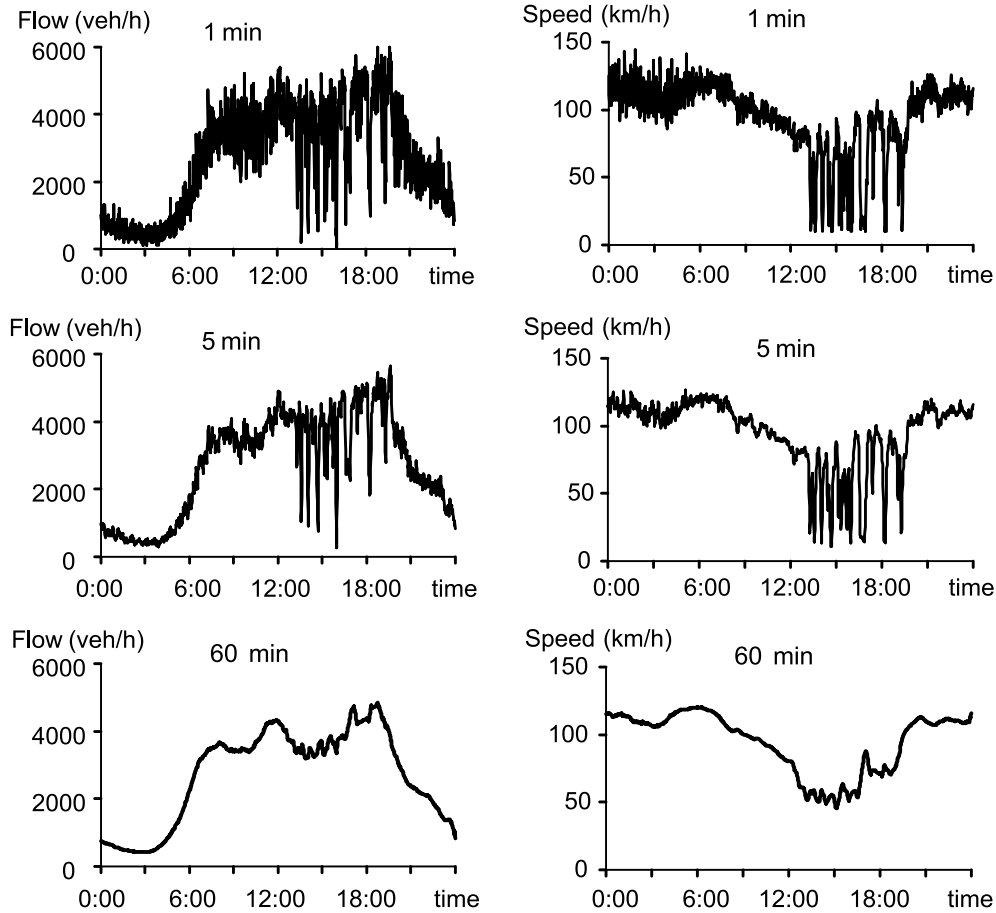
of the whole congested situation). A maximum flow rate of approximately 6000 veh/h on this day is a good "rule of thumb" for a three lane freeway, i. e., 2000 vehicles/h per lane is a good estimation for the maximum possible throughput (including long vehicles).

The moving average of 5 min (Fig. 16, middle) smoothes the fluctuations of the minute interval and shows clearly similar local minima, because they usually last longer than 5 min. If the moving averaging interval is set to one hour (Fig. 16, bottom), the resulting time series is smoothed considerably. For long-term applications, e. g., a traffic prediction of the next day, those aggregated hourly traffic data is sufficient. The form of the 60 min-time series is typical for a freeway: an increasing flow rate in the morning, higher flow rates during the day, an afternoon peak and decreasing flows during the evening hours.

### Predictable Features of Effectual Bottlenecks

In addition to the temporal aggregation of the previous chapter, a spatial aggregation would enhance the long-term time series prediction to a long-term congested traffic pattern prediction with use of time and location as key predictive parameters. According to three-phase traffic theory with the defined two phases in congested traffic, i. e., synchronized flow and wide moving jam, spatiotemporal congested traffic patterns emerge after a first phase transition from free flow to synchronized flow that explains traffic breakdown [68,69]. A bottleneck at which traffic breakdown is observed is called "effectual bottleneck". It has two empirical features: (i) the phase transition from free to synchronized flow, i. e., traffic breakdown, occurs more frequently at this location in comparison to other freeway locations and (ii) the downstream front of the synchronized flow region where the vehicles accelerate from the congested region to free flow is almost fixed at this effectual bottleneck (see  $B_1$ - $B_3$  in Fig. 17). The reason for the existence of an effectual bottleneck is a non-homogeneity of the freeway infrastructure caused by, e. g., on- and off-ramps, a decrease in the number of lanes, a higher slope and so on which influences the undisturbed homogeneous traffic flow. From a study of congested pattern formation from 1995–2003, it has been shown that the phase transitions from free to synchronized flow at the freeway A5-South have occurred mainly in the vicinity of the effectual bottlenecks  $B_1$  (intersection "Nordwestkreuz"),  $B_2$  (intersection "Bad Homburg") and  $B_3$  (intersection "Friedberg") (see Fig. 17).

The resulting congested patterns emerging after the phase transition at the effectual bottlenecks have pre-



**Traffic Prediction of Congested Patterns, Figure 16**

Minute values (*top*), moving averages of this data for 5 min (*middle*) and 60 min (*bottom*) of flow rates (*left*) and vehicle speed (*right*) from a typical detector from three-lane freeway A5-North between intersections “Nordwestkreuz” and “Bad Homburger Kreuz” on Friday, 23rd March 2001

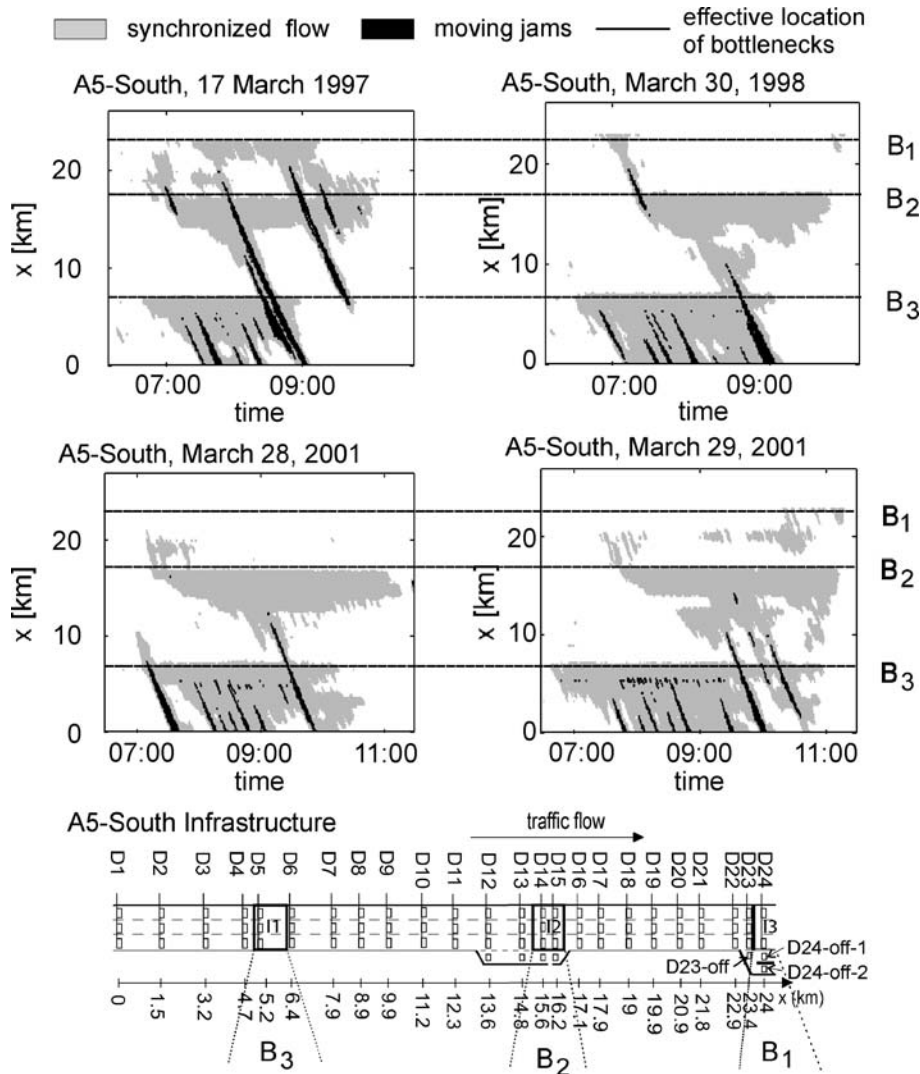
dictable and characteristic features: the pattern features are the same for different days and years (see Fig. 17 with space over time diagrams). From a global analysis many similarities among the patterns can be identified. The size, form and duration of the synchronized flow regions, wide moving jam emergence and their propagation through different bottlenecks, as well as the time dependences of phase transitions at different bottlenecks (i. e., phase transition from free to synchronized flow always earlier at  $B_3$  than at  $B_2$ ) are very similar on different days in different years. This supports predictability in a long-term congested traffic pattern database: prediction at a certain time on a certain day could be performed based on the current traffic conditions (current congestion, flow rate, weather, time of day, etc.) and a correlation to the pattern parameters of the database. Furthermore, the data analysis shows that the congested patterns are like “fingerprints” of the road network: each effectual bottleneck may have its spe-

cific characteristic parameters of the congested patterns, i. e., type, form and duration of the congested pattern are linked to the specific bottleneck at which the congestion emerges [69].

For long-term time series of congested traffic patterns, the possible degree of predictability has to be taken into account. According to probabilistic features of the congested patterns, even under the same traffic parameters (e. g., flow rates, weather, road conditions, etc.) the pattern features could be different for different situations (days) with different degrees of predictability. Some deterministic features with a high degree of predictability are as follows [69]:

- (i) Location of the effectual bottleneck,
- (ii) Emergence of wide moving jams in synchronized flow regions upstream of bottlenecks in a distance of about 3 km or more,





**Traffic Prediction of Congested Patterns, Figure 17**

Empirical examples of predictable congested patterns in different years at the effectual bottlenecks  $B_1$ – $B_3$  on the freeway A5-South (bottom: infrastructure). Overview on regions of free flow (white), synchronized flow (light grey) and wide moving jams (black) in space over time diagrams [69]

- (iii) Stable propagation of wide moving jams through any other traffic states,
- (iv) Considerably lower flow rate in the outflow of the wide moving jam if free flow emerges after the jam (about two thirds of the maximum flow rate in free traffic).
- (ii) The type of congested pattern occurring at an effectual bottleneck. Some of the pattern types [69] may not occur at a specific set of adjacent effectual bottlenecks.

Thirdly, pattern features with a low degree of predictability are:

Congested pattern features with a middle degree of predictability are:

- (i) The instant of a phase transition from free to synchronized flow at the effectual bottleneck at the same traffic demand (“probabilistic nature of speed breakdown”),
- (ii) The instant of wide moving jam emergence in synchronized flow.

A long-term pattern database therefore should contain the empirical congested pattern features and store the predictable information with related probabilities. An approach for development of such a congested traffic pattern database is described later on.

### Approaches for Producing Long-Term Traffic Time Series

**Cluster Analysis for Flow Rates** One kind of detector on freeways and major roads in Germany allows hourly measurements of the flow rates at certain locations in the freeway network (“Dauerzählstellen”). Performing a cluster analysis of this hourly traffic data, typical recurring and regular time series of the flow rates are extracted from the traffic data (Fig. 18). The typical bimodal structure for weekdays (e. g., Monday in Fig. 18) and the unimodal structure for weekend days has been identified clearly [50,52,120]. In [120] seven types of time series for weekdays based on a statistical cluster analysis have been identified which can be characterized by their features:

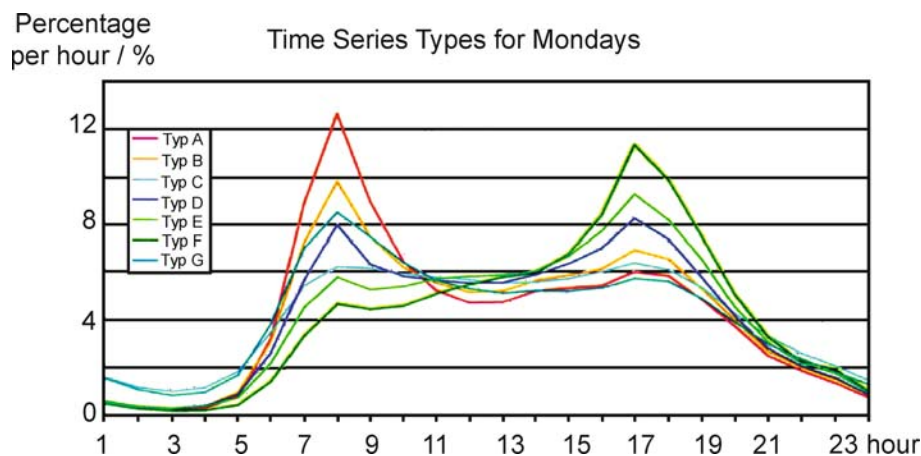
- Type A: strong morning peak
- Type B: morning peak and lower afternoon peak
- Type C: relatively balanced distribution during the daytime
- Type D: double peak
- Type E: afternoon peak, low morning peak
- Type F: strong afternoon peak
- Type G: more than average morning, then decreasing steadily.

According to [120] Type A and B are dominant on freeways with commuting traffic with the opposite direction of Type E and F.

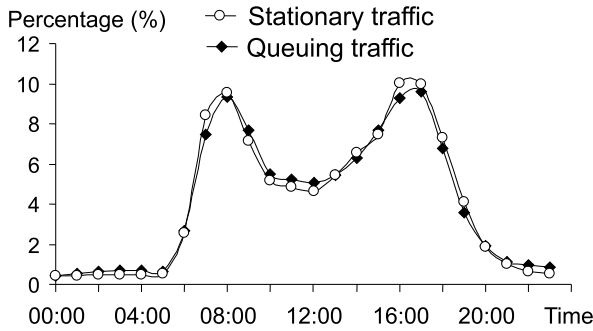
These time series represent an estimation of the traffic demand for the freeway and major road network, they do not take traffic congestion into account. In a comparison of the averaged detector measurements of Fig. 16 (bottom, left) and the hour values and the clustered results of Fig. 18, the freeway stretch of the A5-North between intersections “Nordwestkreuz” and “Bad Homburger Kreuz” on Friday, 23rd March, 2001 shows a “Type C” form with a relatively balanced distribution during the daytime. In a database, only this type classification and some parameters like the maximum flow rate must be stored for a long-term flow rate time series.

As an additional source of information for long-term time series the radio traffic messages via RDS/TMC (Radio Data System/Traffic Message Channel) protocol can be used as archived traffic data [124]. The main events of the broadcasted information are the “stationary” and “queuing” traffic which are coded in the broadcast data stream. In the year 2000, approximately one million messages were broadcasted in Germany. A daily distribution of those messages with relative percentages per hour shows in Fig. 19 a double peak in the morning and afternoon hours similar to Type D in Fig. 18. These time series illustrate the time distribution of the traffic congestion; for the whole network, the evening peak is wider than the morning one and approximately 10–12% of all congestion occur in the freeway network during the two peak hours of the day (from 7:30–8:30 h and from 15:30–16:30 h, respectively).

The comparison of Figs. 18 and 19 illustrate the correlation between the flow rate time series and the congestion message distribution: from a global perspective high traffic flows correlate with highly congested traffic states.



Traffic Prediction of Congested Patterns, Figure 18  
Daily time series (Weekday Mondays, Type A – Type G) [120]



**Traffic Prediction of Congested Patterns, Figure 19**

Relative percentage of RDS/TMC radio messages ("stationary traffic" and "queuing traffic") per hour for all days of year 2000 in Germany [124]

The interpretation of the RDS/TMC messages of "stationary" and "queuing" traffic is related to speed and flow parameter values as definitions for the TMC protocol, e. g., all measured speeds below 60 km/h are defined as being "queuing" traffic and speeds below 30 km/h are defined as "stationary" traffic. This is in contradiction with empirical features of congested patterns like in Fig. 17 and therefore those time series cannot be used for successful traffic prediction. The TMC protocol definitions of today make it impossible to have information on the real congested traffic situation, but this information is broadcasted by radio stations and all radio listeners in most European (and recently USA) regions are experienced with this common form of traffic information. For dynamic route guidance applications the "stationary" and "queuing" traffic could be transformed into travel delay times and then be used in navigation systems. A specific high degree of predictability based on the real empirical traffic phases has been lost on this aggregated level of information.

**Wavelet Analyses** The extracting and producing of traffic data time series is cumbersome and time-consuming, as well as computer storage intensive. The frequency domain analyses, using Fourier series [11], offer a possible alternative, especially regarding storage capacity. The Fourier analysis is appropriate for stationary time series models without spikes. The local minima in traffic variables are on the contrary very important for traffic data analyses because of traffic congestion: they can occur suddenly and are relatively seldom events in comparison to all traffic data measurements. The wavelet analyses try to solve these problems [12,45]. Wavelets have been applied to traffic pattern analysis [61], traffic prediction [140,158], incident detection [141], and determining data aggregation intervals [121] in transportation technology. Wavelet analysis

consists of breaking a given signal into other waves of different frequencies, thus forming both a high pass and low pass filter at the same time. The ability to analyze the coefficients of the wavelets significantly reduces the dimension of the data vector without losing too much information. This is an efficient approach considering the computational storage costs of the database as well as the performance of the pattern database.

Traffic patterns have some regularity or repetition of some basic underlying structure [45]. An example of this regularity is the typical structure of the bimodal daily pattern of weekday traffic with peaks in the morning and in the afternoon [147] (see Figs. 16 and 18).

For long-term traffic prediction the traffic data time series are averaged over longer time intervals (see Fig. 16). Therefore, it seems questionable that the mentioned dominant advantages of wavelet transformation are relevant for long-term time series.

**Cluster Analysis for a Database of Congested Traffic Patterns** In an approach supporting Kerner's three-phase traffic theory concepts [69], a database of congested traffic patterns can be used for long-term predictions. In a first step the empirical congested pattern has to be clustered. The clustering builds the congested regions into larger spatiotemporal regions and suppresses smaller congested regions. Therefore, a minimum duration of 30 min of the congestion and a width of at least 1 km distance between upstream and downstream front of a congested region are chosen for the clustering approach. Additionally, smaller spatiotemporal gaps have to be filled. Those occur typically in an online detection system: if at a specific location the detector malfunctions or in one cycle the average speed increases a little, there will be one cycle of free traffic within a congested pattern. As a chosen parameter, time gaps up to a 20 min interval will be filled in during the clustering process. The method for both traffic phases is single linkage clustering based on the minimum distance from two clusters A and B

$$\min_{a \in A, b \in B} \{d(a, b)\} < X, \quad (10)$$

where  $X$  is a parameter for cluster merging. The distance between  $a$  and  $b$  for two traffic flow cluster is calculated by Euclidian distance:

$$d(a, b) = \sqrt{\sum_{k=1}^p (a_k - b_k)^2}, \quad (11)$$

where  $k$  and  $p$  are the index and the length of arrays  $a$  and  $b$ , respectively. The given data points  $a$  and  $b$  have

to be compared for the calculation of the spatiotemporal distance. Figure 20 illustrates the result: from the empirical pattern based on minute values from a detection system and the reconstruction with FOTO and ASDA models [69] on the left side the number of wide moving jams is reduced in the clustered traffic pattern on the right. Similarly the regions of the synchronized traffic flow are clustered. The resulting clustered congested pattern could be stored in the database with parameters for both synchronized flow and wide moving jam. In comparison, the clustered pattern can reduce the amount of storage capacity for a congested traffic pattern by about 90% which is very computationally efficient for larger road networks.

In the database of congested traffic patterns the patterns are stored with some basic parameters which are relevant for each of the traffic phases (“puzzle pieces”). From three-phase traffic theory it is known that the downstream boundary of the synchronized flow is almost fixed at the bottleneck location. For a synchronized flow region, the starting location and starting time of the synchronized flow region is stored. Secondly, the duration and the expanse are relevant for each of those regions. Two speed parameters describe the form of emergence and dissolution of the synchronized flow regions. The puzzle pieces for synchronized flow have, therefore, four edges in the database.

For the wide moving jams in the traffic pattern database relevant parameters are the length of the wide moving jams and the maximum time of its propagation life cycle. Additionally, for the frequency of wide moving jams following one after the other, the distance between two following upstream jam fronts is given. To define the angle in the space-time diagram, the speed of the front is stored. The latest wide moving jam of a pattern has to take into account the dissolution of the synchronized flow re-

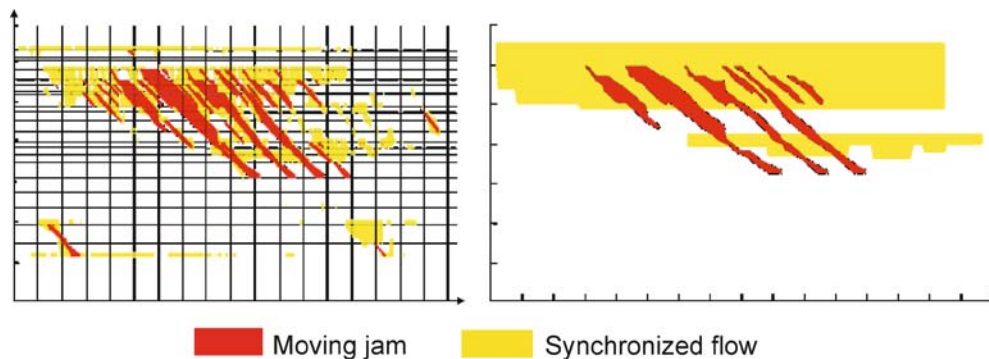
gion, i. e., there is no wide moving jam emerging without a synchronized flow region further downstream. This simplified database approach is the first step into online traffic prediction application. Furthermore, detailed information of the congested traffic patterns could be stored in prediction applications [69]:

- the average speed in synchronized flow region,
- the mean velocity and time dependence of the upstream front of the synchronized flow,
- transformations between different types of pattern (SP, GP, EP) after pattern emergence,
- additional information about EP characteristics depending on bottleneck locations,
- probabilities of pattern occurrence dependent on bottleneck location and traffic demand.

The probability of the first pattern occurrence (e. g., from Fig. 21 at a typical bottleneck) is in itself distributed over time. After the traffic breakdown the further development of the congested region is more deterministic than the first phase transition (occurring from free to synchronized flow in Fig. 21 at approx. 15 km and 14:20 h).

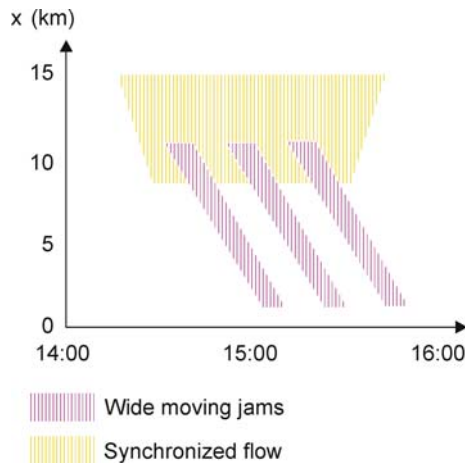
### Matching with Current Traffic Situation for Prediction

If a structured database with sets of time series of traffic variables exists, the current traffic situation can be used as criterion for choice of the “best fit” prediction (Fig. 22). Measurements in the vehicle or at the traffic center support the search in the archive: a spatiotemporal criterion determines the prediction output. The choice of the “best fit” from the historical archive should classify a “reference” pattern from an anomaly. A time-shift component has to be established to ensure that for example, a morning peak 10 min later is not considered as an anomaly, but

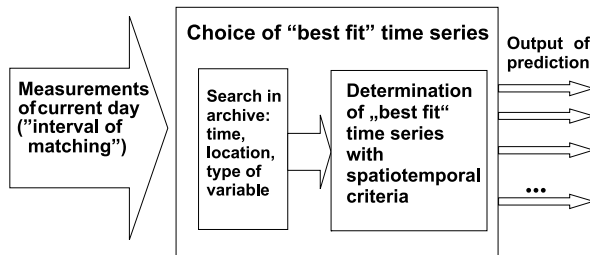


**Traffic Prediction of Congested Patterns, Figure 20**

Empirical congested traffic pattern at a bottleneck reconstructed by FOTO and ASDA models (*left*) and clustered traffic pattern for the database (*right*)



**Traffic Prediction of Congested Patterns, Figure 21**  
Time series of congested traffic pattern database (General Pattern)



**Traffic Prediction of Congested Patterns, Figure 22**  
Matching of time series with current measurements

still a regular pattern [69,147]. Because of the features of the traffic process, the acceptable tolerance defining similarity among patterns has to integrate spatiotemporal distances. “Best fit” time series must therefore be found in the environment of the specific location, i. e., neighboring detectors or road sections, as well as with time shifts of the current prediction time.

### Applications of Traffic Prediction

Currently known applications of traffic prediction are split into several areas. They could be implemented on the infrastructure or the vehicle side. The Internet offers a broadcast medium for distributing traffic prediction to the user without individualizing the prediction information which is dependent on current position and time. One other use of prediction on the infrastructure side is the planning of roadwork: the prediction system estimates the optimal time period of a planned roadwork or estimates the consequences of such a roadwork. The online applica-

tion of FOTO and ASDA models offers traffic prediction information which can be used in the traffic control center for traffic management purposes. An application concept for prediction usage in vehicles closes this overview.

### Online Internet Applications of Traffic Prediction

From the currently known online prediction applications with use of long-term time series a model-driven approach (in Germany: North-Rhine Westphalia) and a data-driven approach (in the USA) are illustrated. They are available via the Internet as collective network short-term predictions and provide dynamic section oriented prediction information. The applications give a global overview similar to a weather prediction approach. Predictions are not individualized for a driver at a certain location at a certain time, for that, the individual route travel time has to be calculated which is dependent on position and time of the traffic prediction.

In addition, both are principally lacking in interpretation of the underlying process in traffic: the traffic state as well as the travel time aggregates the traffic phases on the road, i. e., the same traffic state or travel time value could be calculated possibly from different regions of different traffic phases and its different features. Therefore the precise predictability has been lost in principle on this aggregated information level. In contrast, traffic prediction must be done based on the basic elements of traffic, which are the regions of the traffic phases, because the traffic phases have different deterministic and predictable behavior. Again, traffic prediction methods based on aggregated information levels will not reach a higher accuracy in principle, because deterministic predictable features of traffic phases have been neglected. One example of this problem is the fact, that one traffic phase propagates through the freeway network independent of intersections, on- and off-ramps, etc., while another traffic phase remains almost fixed at one location. However, both traffic phases might be interpreted as the same traffic state or the same travel time value for a road section could be given.

Based on a cellular automata model [102] and the detector measurements, a network traffic state prediction for 30 min and 60 min is performed in a federal state of Germany (<http://www.autobahn.nrw.de/nrw.html>). It has been developed for fast computation performance in larger road networks. The basic element of this approach is long-term time series of the detector network of North-Rhine Westphalia. The prediction for the next 30 or 60 min is a combination of the model calculations and the estimated speed and flow values from the time series database. It is stated [22], that the real-world mea-



surements adapt in each online cycle of simulation results: therefore, the measurements and the time-series “over-rule” the cellular automaton simulation. Strictly speaking, the system shows interpolated detector measurements and long-term assumptions of traffic states generated from archived speed and flow data. The traffic state prediction of this approach makes it impossible to evaluate the prediction quality of the traffic model itself because of the strong combination with the real measurements. The overall system prediction quality could be evaluated with comparisons of predicted and real travel times.

In the USA, nationwide traffic prediction applications have only recently been offered (e.g., [53], <http://www.inrix.com>). They collect all available detector measurements (flow and occupancy sensor from public detector networks) and dynamic traffic data (incident data, GPS based (Global Positioning System) vehicle data, etc.). Their methodological approach uses Bayesian networks [53]: all factors that combine to result in a congestion event occurring are attempted to be determined. Obvious factors for the approach are the day of the week, the prevailing weather, the occurrence of an accident or some other incident, sporting or other types of event and road construction. The approach views predictive and causal relationships in the context of the effect that one variable has on the probability distribution of outcomes of the other: those probabilistic relationships are in the framework of Bayesian statistics. Historical time series are integrated in these Bayesian networks as traffic congestion are recurring and regular. This approach neglects physical features of the traffic process and Bayes’ theorem in itself is not a prediction model, but a probabilistic formula for dependences among variables. The prediction quality in comparison to the reality is currently unknown. The representation of the prediction results is similar to the German approach at <http://www.autobahn.nrw.de/nrw.html>: the current as well as estimated predicted traffic states are colored for each road segment in up to four different states.

### Roadwork Planning

Long-term time series of flow rates are used in the planning of road construction work and estimate how much higher than the remaining reduced capacity of the network the traffic demand will be [109,128]. Free flow capacities of different kinds of roadwork (depending of the number of (reduced) lanes, the reduced lane width, the form of way in the roadwork) have been measured [52,126] and could be used as parameters for the planning process. If the estimated traffic demand from the long-term flow rate is higher than the expected remaining free flow capacity



**Traffic Prediction of Congested Patterns, Figure 23**

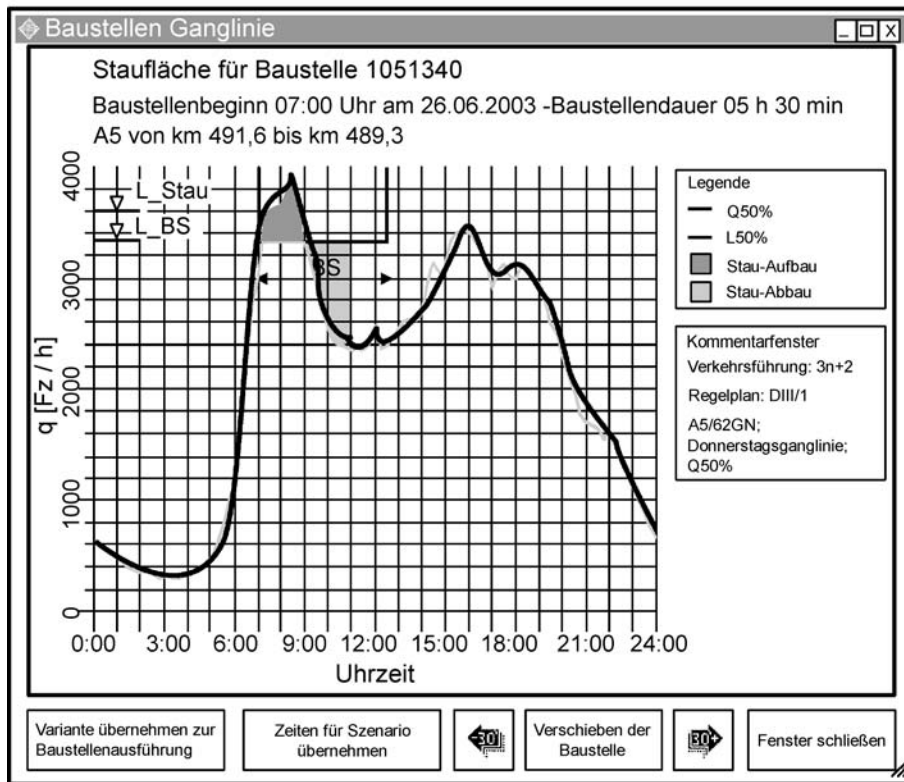
Seattle traffic map illustrating a sample user interface, highlighting currently congested road segments in red and smooth flowing traffic in green and predictive information (clocks for duration of the events). Taken from: <http://www.inrix.com/predictive.asp>

of the planned roadwork, the number of vehicles in the “queue” for the predicted congestion is counted (dark grey region in Fig. 24). If the estimated flow decreases below the estimated capacity later, a time delay for the dissolution of the predicted “queue” is modeled. This simple model neglects dynamic processes of traffic congestion, probabilistic features and uses a simple queuing of vehicles upstream of the roadwork. Nevertheless, the traffic control center can move the time interval of the planned roadwork to a period where no congestion is predicted, because the traffic demand is assumed to be below the reduced capacity due to the roadwork. For the management of roadwork in a freeway network, this approach is used in the German federal state of Hessen [128]. Because the dynamic processes in real traffic could not be modeled by this approach, i.e., the flow rate is not the only sufficient parameter for congestion emergence [69], the results should be taken only as a guide for public roadwork planning and not as a precise traffic congestion prediction with correct time intervals.

### Traffic Control Centers

For the control of freeway traffic, traffic control centers have some instrumentation like speed advisory, lane blockings, collective route choice, etc. To decide the best possible control strategy, a traffic prediction is relevant in-

- selection of different risk scenarios for capacity
- variation of the beginning time of the road work
- variation of the road work duration
- manual simulation of the road work impact by activating discrete time shifts (interactive shift of the beginning time of the road work by +/-30 minutes).



**Traffic Prediction of Congested Patterns, Figure 24**

Congestion prediction for planned roadwork with long-term flow time series from the German online system in Hessen [128]

formation. In the control center of the German federal state of Hessen, online traffic prediction based on FOTO and ASDA models has been installed and used.

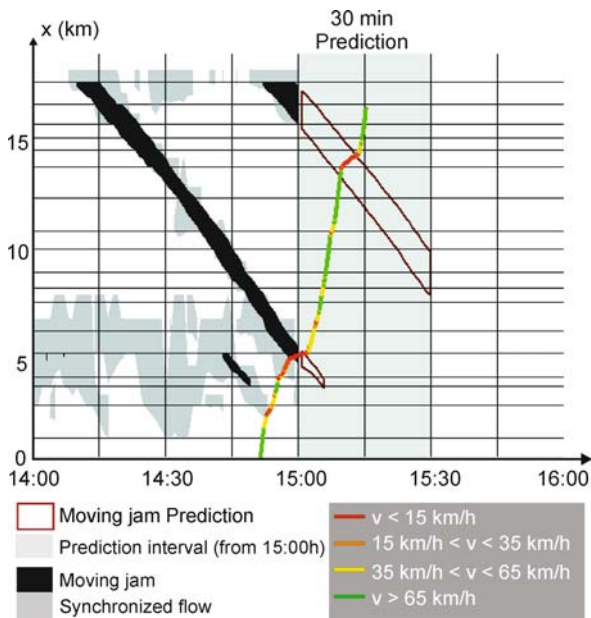
The online application – called FOTOWin integrating both FOTO and ASDA models – offers the possibility to predict the movement of a wide moving jam (see above Subsect. “[Pattern Matching: Fitting Traffic Pattern from Database](#)”). In case of wide moving jam recognition, the online system predicts the subsequent propagation of the related wide moving jam. This information could be very useful for vehicle applications of local danger warnings: the sudden traffic breakdown of vehicle speed could be predicted, if such a wide moving jam approaches the individual vehicle.

An empirical example is illustrated in Fig. 25. At 15:00 h the future positions of a wide moving jam have been predicted. The test vehicle has met the upstream front at 14 km at about 15:08 h. A local danger warning of

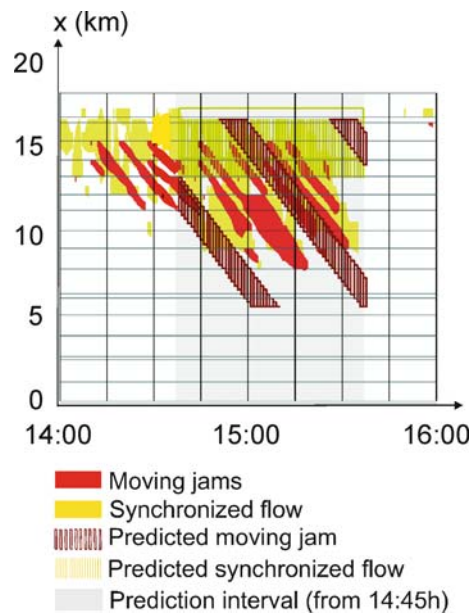
a speed breakdown could have been given about 8 min and 9 km in advance. Inside the vehicle this could be used for driver information, safety and comfort functions. If this information could be given with enough quality in larger road networks, the use for vehicle assistance systems like adaptive cruise control becomes technically feasible.

The online FOTOWin system in Hessen is able to “match” the current traffic pattern with a pattern database. Therefore, the infrastructure gets information on the subsequent development of the congested traffic pattern. This is relevant for decision support for the current traffic management strategy.

One empirical result of a short-term pattern prediction with matching is represented in the following space-time diagram (Fig. 26). The horizontal lines represent the positions of the detectors on the freeway. The light grey area denotes the prediction period with red and yellow vertical lines signifying wide moving jams and synchronized flow,



**Traffic Prediction of Congested Patterns, Figure 25**  
Empirical wide moving jam short term prediction with FOTO and ASDA models for a vehicle (colored trajectory)



**Traffic Prediction of Congested Patterns, Figure 26**  
Traffic pattern prediction, freeway A3-North, 06th June 2006, prediction at 14:38h

respectively. The related matched traffic database elements from the pattern database are also shown. Together with the prediction from 14:38 h the reality is shown, which has been reconstructed later.

The online FOTowin system has been installed recently for North-Rhine Westphalia freeways [114]. In principle, the traffic control center located there has similar detector loop data available as in Hessen. These raw traffic data are transferred to WDR, the major public radio broadcasting station from North-Rhine Westphalia in Cologne, who offers traffic messages to the end customer (e. g., radio listener or driver) via broadcast channel RDS (Radio Data System). The application FOTowin covers a part of the whole freeway network with 1900 km of freeway and more than 1000 double loop detectors.

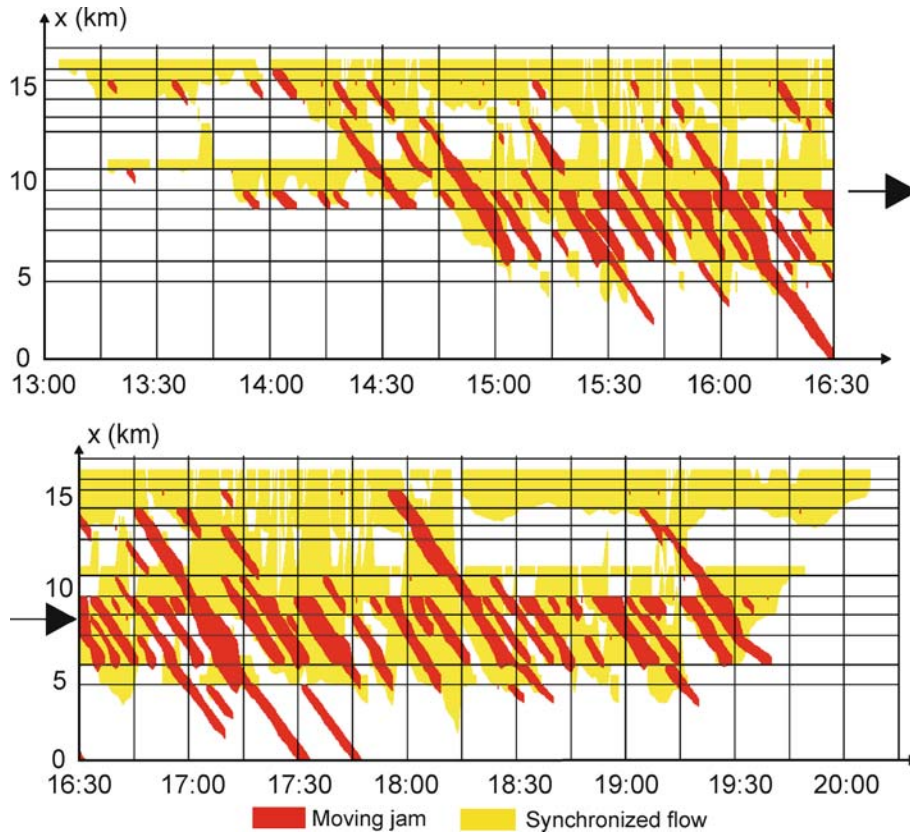
As an empirical example, a large congested pattern on freeway A45-North is illustrated in Fig. 28. It is clearly visible, that such a pattern is very similar to congested traffic patterns in Hessen (see Fig. 9): emergence of wide moving jams and their stable propagation through many other bottlenecks does eventually occur on every freeway with bottlenecks.

In this example, for almost 7 h on the 17 km stretch a large number of propagating wide moving jams have emerged. At the locations approximately at 10 and 16 km phase transitions from free to synchronized flow occur at



**Traffic Prediction of Congested Patterns, Figure 27**  
Freeway road network observed by FOTowin in North-Rhine Westphalia [114]

different time moments. Later on, wide moving jams propagate from the downstream bottleneck at approximately 16 km to the upstream bottleneck in the vicinity of 10 km:



**Traffic Prediction of Congested Patterns, Figure 28**

Congested traffic pattern reconstructed by FOTO and ASDA models: space-time diagram from freeway A45-North in North-Rhine Westphalia, 31st August 2007 [114]

the occurring empirical Expanded Pattern illustrates the complex interactions of congested traffic patterns which can be explained based on Kerner three-phase traffic theory. Because of the dense freeway network in North-Rhine Westphalia, there are many possible bottlenecks on the freeways due to a large variety of on- and off-ramps with the occurrence of EPs.

### In-Vehicle Traffic Prediction

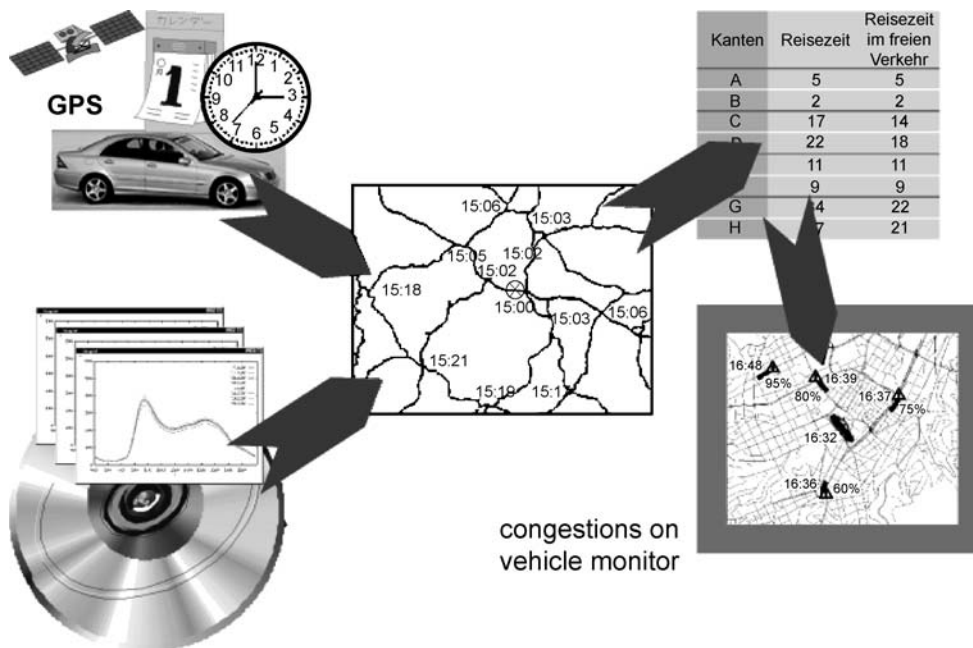
The concept of autonomous traffic prediction in a vehicle is presented in [80]. Figure 29 illustrates the approach: a vehicle has a position unit, a digital map, and the current time. Additionally, the vehicle has an on board memory unit for historical travel time series (e. g., stored on CD/DVD). The linkage of the map information with stored traffic congestion information leads to an adequate choice of travel time prediction associated with the current time and the specific vehicle location. Subsequently, the

driver can see this traffic prediction indicated in a vehicle monitor. Furthermore, the traffic prediction is useful for dynamic route guidance in navigation systems (e. g., [19]). Results of traffic prediction can also be presented in different visualization forms [80].

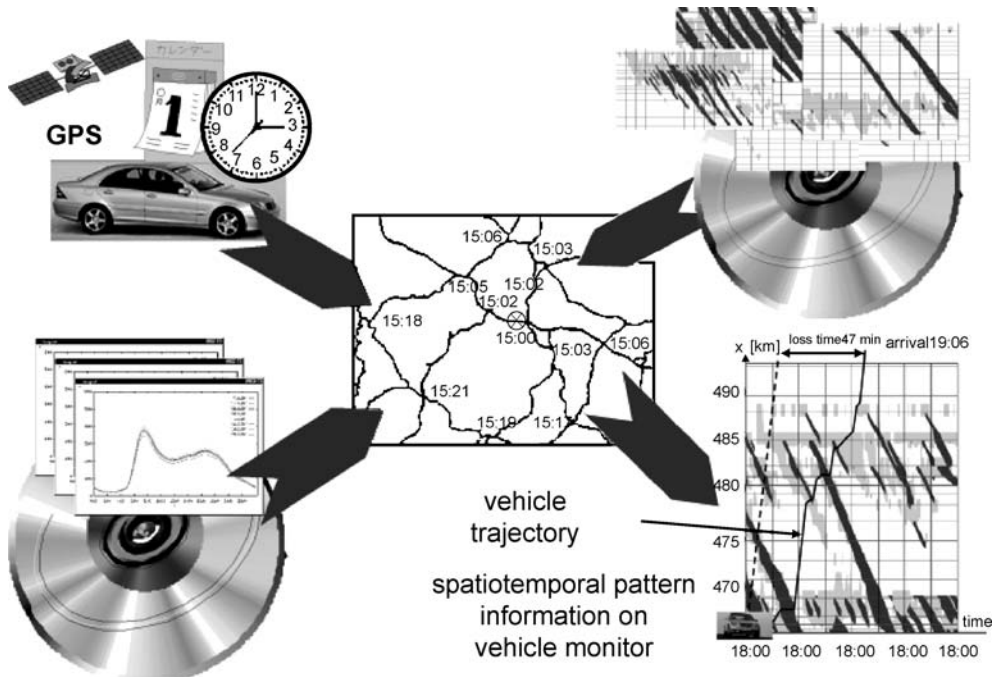
A possible extension of the above concept is the integration of historical spatiotemporal patterns [69] to also make short range traffic prediction available (Fig. 30). To reach this goal, a historical spatiotemporal pattern database should be stored in the vehicle, in addition to historical travel times (or flow rates). In this case, based on the current vehicle location and speed measured by the vehicle, an appropriate choice of a spatiotemporal pattern can be made in the vehicle, which the vehicle should meet in the near future (see vehicle trajectory in Fig. 30, bottom, right).

As it is shown (e. g., [33]), traffic is highly regular with recurring and typical characteristics. Therefore the historical patterns are relatively stable within proposed update





**Traffic Prediction of Congested Patterns, Figure 29**  
Concept of autonomous traffic prediction [80]



**Traffic Prediction of Congested Patterns, Figure 30**  
Concept of autonomous traffic prediction with use of spatiotemporal pattern database in vehicle [80]



cycles of 1–2 years because of long-term changes in traffic demand and infrastructure: the pattern database in the vehicles could be relatively steady. Communication channels into the vehicle could update the spatiotemporal pattern database.

### Future Directions

It must be stated that in the field of traffic prediction a large number of different concepts are coexisting. The proof of reliability especially in times of congested traffic is a still open question to the traffic prediction approaches. Traffic predictions in online information systems and used for collective traffic management or individual navigation system in established applications are in its infancies because of the heterogeneous and often unreliable prediction methods. Therefore, quality definitions and measurement techniques are a key issue for the future evaluation of traffic prediction approaches.

One of the missing elements is a quality definition for precise traffic predictions, because the step beforehand – an area covering precise measuring of the current traffic situation – is still not realized. The measurements of the current situation could be enhanced by using the vehicle as a moving sensor (FCD: Floating Car Data) or using any mobile device in the vehicle (FPD: Flow Phone data). Both technological ideas have been understood conceptually, but are still not introduced in the market because of the lack of successful business models. A definition of prediction quality tested in a laboratory environment should be based on deviation of the predicted values from real measured and archived travel times of the congested time periods.

Microscopic simulation is a very helpful instrument to predict consequences of both traffic control as well as individual traffic assistance systems (e.g., automatic cruise control) before introducing them into traffic control center or into the vehicles. Especially microscopic modeling of vehicle behavior in the case of automatic vehicle assistance systems and their influence of the global traffic system is a complex field of vehicle dynamics. A quality assured network wide traffic prediction with microscopic simulation analogous to weather prediction is still a task of the future.

The traffic pattern database could be enhanced for planned roadwork and accidents which both show characteristic predictable features. The traffic pattern database then has the characteristic content of long-term time series information.

A short-term precise prediction could be used as input for driver assistance systems in vehicles, e.g., for an

adapted automatic speed control of the vehicle in the case of approaching congestion.

Traffic simulation in good accordance with the reality on the roads will be a valuable instrument for network traffic predictions. Benchmarking criteria for evaluation of model results should be defined to have quality measures for their application.

One relevant further issue is the enhancement and combination of long-term time series with time series of breakdown probabilities. Anticipation effects will arise with further distribution of dynamic traffic information: people will react on the information in different ways adapting their routing decisions. As consequence, the problem of dynamic traffic assignment under more and more dynamic current traffic information increases: how should the long-term traffic prediction information be given to have an efficient use of the available network resources and not the problem of emerging new congestion in regions where people use roads trying to avoid the already congested regions.

Looking on the network perspective, the effects of prediction on the driver behavior must be taken into account if larger information penetration rates are realized. In case of more available information, the dynamic traffic assignment problem arises, i.e., which part of the driver collective should be advised on which route, finding a system optimum for the road network.

### Acknowledgment

We would like to thank Paul Mathias for his critical remarks and fruitful discussions as well as Andreas Haug, Mario Aleksic, Boris Kerner, Jörg Weber and Andrea Köhler for their valuable support. In addition, we thank Gerd Riegelhuth, Ulrich Steiger, Katrin Beckroth and Hendrik Zurlinden from the Hessen Landesamt für Straßen- und Verkehrswesen in Frankfurt for the possibility of research projects and the help with the preparation of traffic data and Susanne Breitenberger from BMW in Munich for the support with vehicle data.

### Bibliography

#### Primary Literature<sup>1</sup>

1. Abdulhai B, Porwal H, Recker W (1999) Short term freeway traffic flow prediction using genetically-optimized time-delay-based neural networks. In: Proceedings 78th Annual Meeting Transportation Research Board, National Academies Press, Washington DC

<sup>1</sup>Patents information at [www.depatistnet.de](http://www.depatistnet.de)

2. Acha-Daza JA, Hall FL (1993) A graphical comparison of the predictions for speed given by catastrophe theory and some classic models. *Transp Res Rec* 1398:119–124
3. Ahmed MS, Cook AR (1979) Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques. *Transp Res Rec* 722:1–9
4. Arem BV, Kirby HR, Van Der Vlist MJM, Whittaker JC (1997) Recent Advances and Applications in the Field of Short-Term Traffic Forecasting. *Int J Forecast* 13:1–12
5. Barceló J, Ferrer JL, García D, Florian M, Le Saux E (1998) Parallelization of Microscopic Traffic simulation for ATT Systems Analysis. In: Marcotte P, Nguyen S (eds) *Equilibrium and Advanced Transportation Modeling*. Kluwer, Dordrecht
6. Becker M, Fastenrath U (1998) Method for transmitting local data and measurement data from a terminal, including a telematic terminal, to a central traffic control unit. German Patent Publication DE19755875A1, USA: US6426709B1
7. Ben-Akiva M, Cuneo D, Hasan M, Jha M, Yang Q (2003) Evaluation of freeway control using a microscopic simulation laboratory. *Transportation Research, Part C. Emerg Technol* 11(1):29–50
8. Ben-Akiva M, Bierlaire M, Koutsopoulos H, Mishalani R (1998) DynaMIT: a simulation-based system for traffic prediction. In: *Proceedings of the DACCORD Short-Term forecasting workshop*. Delft University, Delft
9. Boker G, Lunze J (2001) State estimation in freeway traffic with floating car data. *Automatisierungstechnik* 49(11): 497–504
10. Box GEP, Jenkins GM (1976) *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco
11. Brockwell PJ, Davis RA (2002) *Introduction to Time Series and Forecasting*, 2nd edn. Springer, New York
12. Burrus CS, Gopinath RA, Guo HT (1998) *Introduction to Wavelets and Wavelet Transforms: A Primer*. Prentice Hall, Upper Saddle River
13. Cameron GDB, Duncan GID (1996) PARAMICS: Parallel Microscopic Simulation of Road Traffic. *J Supercomput* 10(1):25–53
14. Cetin M, Comert G (2006) Short-Term Traffic Flow Prediction with Regime Switching models. *Transp Res Rec* 1965:23–31
15. Chatfield C (2001) *Time-Series Forecasting*. Chapman and Hall/CRC, London
16. Chen H, Grant-Muller S, Mussone L, Montgomery F (2001) A study of hybrid neural network approaches and the effects of missing data on traffic forecasting. *Neural Comput Appl* 10:277–286
17. Chen H, Grant-Muller S (2001) Use of sequential learning for short-term traffic flow forecasting. *Transp Res C* 9:319–336
18. Chen M, Chien SIJ (2001) Dynamic freeway travel-time prediction with probe vehicle data – link based versus path based. *Transp Res Rec* 1768:157–161
19. Chen Y, Bell MGH, Bogenberger K (2007) Reliable Multipath Planning and Dynamic Adaptation for a Centralized Road Navigation System. *IEEE Trans ITS* 8(1):14–20
20. Chickering DM, Heckerman D, Meek C (1997) A Bayesian approach to learning Bayesian networks with local structure. In: *Proceedings 13th Conference on Uncertainty in Artificial Intelligence*. Rhode Island, USA, pp 80–89
21. Chien SIJ, Kuchipudi CM (2003) Dynamic travel time prediction with real-time and historic data. *ASCEJ, Transp Eng* 129(6):608–616
22. Chrobok R, Wahle J, Schreckenberg M (2001) Traffic Forecast Using Simulations of Large Scale Networks, In: Stone B, Conroy P (eds) *Broggi A4th International IEEE Conference on Intelligent Transportation Systems*. IEEE, Oakland, pp 434–439
23. Cremer M (1979) *Traffic Flow on Freeways* (in German). Springer, Berlin
24. Daganzo CF (1994) The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transp Res B* 28(4):269–287
25. Daganzo CF (1995) The cell transmission model, Part II: Network Traffic. *Transp Res B* 29(2):79–93
26. Daganzo CF (1999) The Lagged Cell-Transmission Model. In: Ceder A (ed) *Proceedings of the 14th International Symposium on Transportation and Traffic Theory*. Jerusalem, Israel, pp 81–104
27. Daganzo CF (1997) *Fundamentals of Transportation and Traffic Operations*. Elsevier Science, Oxford
28. D'Angelo MP, Al-Deek HM, Wang MC (1999) Travel-time prediction for freeway corridors. *Transp Res Rec* 1676:184–191
29. Davis GA, Nihan NL (1991) Nonparametric regression and short-term freeway traffic forecasting. *J Transp Eng* 117(2):178–188
30. de Rham C, Lange R (2000) Short Term Forecast and Evaluation for Intelligent VMS Settings. In: *Proceedings of the 7th World Congress on ITS*. ERTICO ITS Congress Association, Torino, Italy
31. Dharia A, Adeli H (2003) Neural network model for rapid forecasting of freeway link travel time. *Eng Appl Artif Intell* 16(7–8):617–613
32. Dia H (2001) An object oriented neural network approach to short term traffic forecasting. *European J Oper Res* 131: 253–261
33. Ding A, Zhao X, Jiao L (2002) Traffic flow time series prediction based on statistics learning theory. In: *IEEE 5th International Conference on Intelligent Transportation Systems*. Singapore, pp 727–730
34. Disbro JE, Frame M (1989) Traffic Flow Theory and Chaotic Behaviour. *Transp Res Rec* 1225:109–125
35. Dougherty M (1995) A review of neural networks applied to transport. *Transp Res C* 3(4):247–260
36. Edie LC, Foote RS (1960) Effect of Shock Waves on Tunnel Traffic Flow. In: *Highway Research Board Proceedings 39 National Research Council*. Washington DC, pp 492–505
37. Fallah-Tafti M (2001) The application of artificial neural networks to anticipate the average journey time of traffic in the vicinity of merges. *Knowledge-Based Syst* 14:203–211
38. Fastenrath U (1998) Method for determining traffic data and traffic information exchange. German Patent Publication DE19737440A1, USA: US6329932B1
39. Fellendorf M, Vortisch P (2001) Validation of the microscopic traffic model VISSIM in different real-world situations. In: *80th Annual Meeting Transportation Research Board*. National Academies Press, Washington DC
40. Fuller WA (1996) *Introduction to Statistical Time Series*, 2nd edn. Wiley, New York
41. Gartner NH (1973) *Highway Research Record* 445:12–23
42. Gartner NH (1983) OPAC – A Demand-Responsive Strategy for Traffic Signal Control. *TRB, Transp Res Rec* 906:75–81
43. Gazis D, Knapp C (1971) Online Estimation of Traffic Densities From Time Series of Traffic and Speed Data. *Transp Sci* 5: 283–301

44. Gipps PGA (1981) A Behavioural Car-Following Model for Computer Simulation. *Transp Res B* 15:105–111
45. Grenander U (1996) Elements of Pattern Theory. Johns Hopkins University Press, Baltimore
46. Hamed MM, Al-Masaeid HR, Bani Said ZM (1995) Short-Term Prediction of Traffic Volume in Urban Arterials. *J Transp Eng* 121(3):249–254
47. Haykin S (1999) Neural Networks: A Comprehensive Foundation. Prentice Hall, Upper Saddle River
48. Head LK (1995) Event-based Short-term Traffic Flow Prediction Model. *Transp Res Rec* 1510:45–52
49. Hecht-Nielsen R (1990) Neurocomputing. Addison-Wesley, Reading
50. Heidemann D, Wimber P (1982) Types of traffic flow rate time series based on clustering methods, vol 26. BAST, Straßenverkehrszählungen (in German)
51. Helbing D (1997) Traffic Dynamics: New Modeling Concepts in Physics (in German). Springer, Berlin
52. Highway Capacity Manual 2000 (2000) Transportation Research Board. National Research Council, Washington DC
53. Horvitz E, Apacible J, Sarin R, Liao L (2005) Prediction, Expectation, and Surprise: Methods, Designs, and Study of a Deployed Traffic Forecasting Service. In: Proceedings of the Conference on Uncertainty and Artificial Intelligence. AUAI Press, Edinburgh, pp 275–283
54. Hoyer R, Chrobok R, Feldges M, Folkerts G, Friedrich B, Huber W, Kates R, Kemper C, Kirschfink H, Lange R, Listl G, Mathias P, Offermann F, Pinkofsky L, Rehborn H, Schlichting B, Stieler P, Thiemann O, Vortisch P (2003) Advice for data completion and data aggregation in traffic management applications (in German). Hinweispapier der Forschungsgesellschaft für Straßen- und Verkehrswesen, FGSV-Papier, vol 382
55. Huang SH, Ran B (2003) An application of neural network on traffic speed prediction under adverse weather condition. In: 82nd TRB Annual Meeting. National Academies Press, Washington DC
56. Huisken G, Van Berkum EC (2003) A comparative analysis of short-range travel time prediction methods. In: 82nd TRB Annual Meeting Transportation Research Board. National Academies Press, Washington DC
57. Hunt PB, Robertson DI, Bretherton RD, Winton RI (1981) SCOOT – A traffic responsive method of coordinating signals. TRRL report No. LR1014. Transport and Road Research Laboratory, Crowthorne
58. Innama S (2001) Short term prediction of highway travel time using MLP neural networks. In: 8th World Congress on Intelligent Transportation Systems. Sydney, Australia, pp 1–12
59. Ishak S, Al-Deek H (2002) Performance evaluation of short term time series traffic prediction model. *ASCEJ, Transp Eng* 128(6):490–498
60. Ishak S, Alecsandru C (2004) Optimizing traffic prediction performance of neural networks under various topological, input, and traffic condition settings. *J Transp Eng* 130:452–465
61. Jiang X, Adeli H (2004) Wavelet Packet-Autocorrelation Function Method for Traffic Flow Pattern Analysis. *Comput Aided Civ Infrastruct Eng* 19:324–337
62. Kaumann O, Froese K, Chrobok R, Wahle J, Neubert L, Schreckenberg M (2000) Online Simulation of the Freeway Network of NRW. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) Traffic and Granular Flow '99. Springer, Berlin, pp 351–356
63. Kaysi I, Ben-Akiva M, Koutsopoulos H (1993) An integrated approach to vehicle routing and congestion prediction for real-time driver guidance. *Transp Research Rec* 1408:66–74
64. Kerner BS (1998) Experimental features of self-organization in traffic flow. *Phys Rev Lett* 81:3797
65. Kerner BS (1999) Traffic prediction method for road network with traffic controlled network nodes (in German). German Patent DE19940957C2
66. Kerner BS (1999) Method for monitoring the condition of traffic for a traffic network comprising effective narrow points. German Patent DE19944075C2, USA Patent: US6813555B1, Japan Patent: JP2002117481
67. Kerner BS (1999) Congested traffic flow: Observations and theory. *Transp Res Rec* 1678:160–167
68. Kerner BS (2002) Empirical macroscopic features of spatial-temporal traffic patterns at highway bottlenecks. *Phys Rev E* 65:046138
69. Kerner BS (2004) The Physics of Traffic. Springer, Berlin, New York
70. Kerner BS (2007) On-ramp metering based on three-phase traffic theory. *Traffic Eng Control* 48(1):28–35
71. Kerner BS, Aleksic M, Deneller U (1999) Traffic condition supervision in traffic network, undertaking inquiry of current position and/or prognosis of future position of flank between area of free traffic and area of synchronized traffic continuously. German Patent DE19944077C1
72. Kerner BS, Herrtwich RGH (2001) Traffic Forecasting. *Automatisierungstechnik* 49:505–511
73. Kerner BS, Klenov SL (2006) Deterministic microscopic three-phase traffic flow models. *J Phys A Math Gen* 39:1775–1809
74. Kerner BS, Klenov SL, Aleksic M, Rehborn H: Development and Implementation of UTA model for urban traffic prediction. (Unpublished)
75. Kerner BS, Rehborn H (1996) Experimental properties of complexity in traffic flow. *Phys Rev E* 53:R4257
76. Kerner BS, Rehborn H (1996) Experimental features and characteristics of traffic jams. *Phys Rev E* 53:1297
77. Kerner BS, Rehborn H (1997) Experimental properties of phase transitions in traffic flow. *Phys Rev Lett* 79:4030
78. Kerner BS, Rehborn H (1998) Traffic surveillance method and vehicle flow control in a road network. German Patent Publication DE19835979A1, USA Patent: US6587779B1
79. Kerner BS, Rehborn H, Aleksic M, Haug A (2004) Recognition and Tracing of Spatial-Temporal Congested Traffic Patterns on Freeways. *Transp Res C* 12:369–400
80. Kerner BS, Rehborn H, Haug A, Aleksic M (2005) Traffic Prediction in Vehicles. In: Proceedings 8th IEEE Conference on Intelligent Transportation Systems. Vienna, pp 251–256
81. Kerner BS, Rehborn H, Kirschfink H (1998) Method for the automatic monitoring of traffic including the analysis of back-up dynamics. German Patent DE19647127C2, Dutch Patent: NL1007521C, USA Patent US5861820
82. Kirby HR, Watson SM, Dougherty MS (1997) Should we use neural networks or statistical models for short-term motorway traffic forecasting? *I J Forecast* 13:43–50
83. Kisgyorgy L, Rilett LR (2002) Travel time prediction by advanced neural network. *Period Polytech Ser Civ Engin* 46(1):15–32
84. Kitamura K, Kuwahara M (eds) (2005) Simulation Approaches in Transportation Analysis: Recent Advances and Challenges.

- Operations Research/Computer Science Interfaces Series, vol 31. Springer, New York
85. Kniss HC (2000) Evaluation of ASDA/FOTO in traffic control centre Hessen (internal report, in German)
  86. Koshi M, Iwasaki M, Ohkura I (1983) Some Findings and an Overview on Vehicular Flow Characteristics. In: Proceedings 8th International Symposium on Transportation and Traffic Theory, p 403
  87. Kuchipudi CM, Chien SIJ (2003) Development of a hybrid model for dynamic travel time prediction. In: 82nd Annual Meeting Transportation Res Board. Transportation Res. Board, Washington DC
  88. Kwon J, Coifman B, Bickel P (2000) Day-to-day travel-time trends and travel-time prediction from loop-detector data. *Transp Res Rec* 1717:120–129
  89. Lan CJ, Miaou SP (1999) Real-time prediction of traffic flows using dynamic generalized linear models. *Transp Res Rec* 1678:168–178
  90. Lee S, Kim D, Kim J, Cho B (1998) Comparison of models for predicting short-term travel speeds. In: 5th World Congress on Intelligent Transportation Systems. ERTICO ITS Congress Association, Seoul
  91. Leutzbach W (1988) Introduction to the theory of traffic flow. Springer, Berlin
  92. Lieu HC (2000) Traffic estimation and prediction system. *Transp Res News* 208:3–6
  93. Lindveld CDR, Thijs R, Bovy PHL, Van der Zijpp NJ (2000) Evaluation of online travel time estimators and predictors. *Transp Res Rec* 1719:45–53
  94. Lingras P, Sharma S, Zhong M (2002) Prediction of recreational travel using genetically designed regression and time-delay neural network models. *Transp Res Rec* 1805:16–24
  95. Lu J (1990) Prediction of Traffic Flow by an Adaptive Prediction System. *Transp Res Rec* 1287:13–20
  96. Maerivoet S, De Moor B (2005) Cellular Automata models of road traffic. *Phys Rep* 419:1–64
  97. Matsui H, Fujita M (1998) Travel time prediction for freeway traffic information by neural network driven fuzzy reasoning. In: Himanen V, Nijkamp P, Reggiani A, Raito J (eds) *Neural networks in transportation applications*. Ashgate Publishers, Burlington, Vermont, pp 355–364
  98. May AD (1990) *Traffic Flow Fundamentals*. Prentice Hall, Upper Saddle River
  99. Middelham F (2001) Predictability: Some thoughts on modeling. *Future Gener Comput Syst* 17(5):627–636
  100. Moorthy CK, Ratcliffe BG (1998) Short Term Traffic Forecasting Using Time Series Methods. *Transp Plan Technol* 12(1):45–56
  101. Miyata S, Noda M, Usami T (1995) STREA. Proceedings of the 2nd World Congress on Intelligent Transport Systems. Yokohama 1:289–297
  102. Nagel K, Schreckenberg M (1992) A cellular automaton model for freeway traffic. *J Phys I France* 2:2221–2229
  103. Nam DH, Drew DR (1996) Traffic dynamics: Method for estimating freeway travel times in real time from flow measurements. *J Transp Eng* 122(3):186–191
  104. Nanthawichit C, Nakatsuji T, Suzuki H (2003) Application of probe vehicle data for real-time traffic state estimation and short term travel time prediction on a freeway. In: Proceedings 82nd Annual Meeting Transportation Research Board. Transportation Res. Board, Washington DC
  105. Newell GF (1982) *Applications of Queuing Theory*. Chapman Hall, London
  106. Nicholson H, Swann CD (1974) The Prediction of Traffic Flow Volumes Based on Spectral Analysis. *Transp Res* 8:533–538
  107. Nihan NL, Holmesland KO (1980) Use of the Box and Jenkins Time Series. Technique in Traffic Forecasting. *Transportation* 9:125–14372
  108. Nikovski D, Nishiuma N, Goto Y, Kumazawa H (2005) Univariate Short-Term Prediction of Road Travel Times. In: International IEEE Conference on Intelligent Transportation Systems. IEEE 2005, Vienna, pp 1074–1079
  109. Ober-Sundermeier A, Zackor H (2001) Prediction of Congestion due to Road Works on Freeways. In: Proceedings IEEE Intelligent Transportation Systems. Oakland, pp 240–244
  110. Oda T (1990) An algorithm for prediction of travel time using vehicle sensor data. In: IEEE 3rd International Conference on Road Traffic Control, pp 40–44
  111. Oh C, Ritchie SG, Oh JS (2005) Exploring the relationship between data aggregation and predictability toward providing better predictive traffic information. *Transp Res Rec* 1935: 28–36
  112. Ohba Y, Koyama T, Shimada S (1997) Online learning type of traveling time prediction model in expressway. In: IEEE Conference on Intelligent Transportation Systems. Boston, Massachusetts, pp 350–355
  113. Okutani I, Stephanedes YI (1984) Dynamic prediction of traffic volume through Kalman Filtering theory. *Transp Res B* 18B(1):1–11
  114. Palmer J, Rehborn H (2008) ASDA/FOTO based on Kerner's Three-Phase Traffic Theory in North-Rhine Westfalia (in German). *Straßenverkehrstechnik* 8:463–470
  115. Papageorgiou M (1983) *Application of Automatic Control Concepts in Traffic Flow Modeling and Control*. Springer, Berlin, New York
  116. Pancratz A (1991) *Forecasting with dynamic regression models*. Wiley, New York
  117. Park B, Messer CJ, Urbanik TII (1998) Short term traffic volume forecasting using radial basis function neural network. *Transp Res Rec* 1651:39–47
  118. Park DJ, Rilett LR, Han G (1999) Spectral basis neural networks for real-time travel time forecasting. *J Transp Eng* 125(6): 515–523
  119. Petty KF, Bickel P, Ostland M, Rice J, Schoenberg F, Jiang J, Ritov Y (1998) Accurate estimation of travel times from single loop detectors. *Transp Res A* 32(1):1–17
  120. Pinkofsky L (2002) Types of Time Series (in German). In: *Verkehrsentwicklung auf Bundesfernstraßen, Bericht der Bundesanstalt für Straßenwesen, Reihe Verkehrstechnik, vol V99*. Bergisch Gladbach
  121. Qiao F, Wang X, Yu L (2003) Optimizing Aggregation Level for ITS data based on Wavelet Decomposition. In: Proceedings 82nd Annual Meeting Transportation Research Board, National. Academies Press/Transportation Res. Board, Washington DC
  122. Rakha H, Crowther B (2003) Comparison and Calibration of FRESIM and INTEGRATION Steady-State Car-Following Behaviour. *Transp Res A* 37:1–27
  123. Ran R, Boyce D (1996) *Modeling Dynamic Transportation Networks*. Springer, Berlin
  124. Rehborn H, Haug A, Aleksic M, Kerner BS, Fastenrath U (2002) Statistical analysis of traffic message archives as decision sup-



- port for road construction up to traffic management (in German). *Straßenverkehrstechnik* 9:478–485
125. Rehborn H, Haug A, Kerner BS, Aleksic M, Fastenrath U (2003) Floating Car Data and methods for recognition and tracking of spatiotemporal traffic patterns (in German). *Straßenverkehrstechnik* 9:461–468
  126. Ressel W (1994) Investigation of traffic at roadworks in the region of maximum capacity (in German). *Informationen Verkehrsplanung und Straßenwesen*, Universität der Bundeswehr München, vol 7. Munich
  127. Rice J, Van Zwet E (2001) A simple and effective method for predicting travel times on freeways. In: *Proceedings of the IEEE Conference on Intelligent Transportation Systems*. Oakland, pp 227–232
  128. Riegelhuth G, Kirschfink H (2003) Management with decision support of road works for traffic flow optimization on freeways. In: *Proceedings of ITS World Congress*, paper No. 2255T. ERTICO Congress Association, Madrid
  129. Rilett LR, Park D (2001) Direct forecasting of freeway corridor travel times using spectral basis neural networks. *Transp Res Rec* 1752:140–147
  130. Robertson DI (1969) TRANSYT: A traffic network study tool. TRRL Report No LR 253, Transportation and Road Research Laboratory. Crowthorne
  131. Rumelhart DE, McClelland JL (1986) *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*. MIT press, Cambridge
  132. Schönhof M, Helbing D (2007) Empirical features of Congested Traffic States and Their Implications for Traffic Modeling. *Transp Sci* 41(2):135–166
  133. Schrader CC, Kornhauser AL, Friesen LM (2004) Using historical information in forecasting travel times. In: *82nd Annual Meeting Transportation Research Board*, National Academies Press, Washington DC
  134. Smith BL, Demetsky MJ (1994) Short term traffic flow prediction: Neural network approach. *Transp Res Rec* 1453:98–104
  135. Smith BL, Demetsky MJ (1997) Traffic flow forecasting: Comparison of modeling approaches. *J Transp Eng* 123(4):261–266
  136. Smith BL, Williams BM, Oswald KR (2002) Comparison of parametric and non-parametric models for traffic flow forecasting. *Transp Res C* 10(4):303–321
  137. Smith BL, Oswald KR (2003) Meeting real-time traffic flow forecasting requirements with imprecise computations. *Comput-Aided Civ Infrastruct Eng* 18:201–213
  138. Stathopoulos A, Karlaftis MG (2003) A multivariate state-space approach for urban traffic flow modeling and prediction. *Transp Res C* 11:121–135
  139. Sun H, Liu HX, Xiao H, He RR, Ran B (2003) Short-term traffic forecasting using the local linear regression model. *J Transp Res Board* 1836:143–150
  140. Sun H, Xiao HX, Yang F, Ran B, Tao Y, Oh Y (2004) Wavelet Pre-processing For Local Linear Traffic Prediction. In: *83rd Transportation Research Board Annual Meeting*. Transportation Res. Board, Washington DC
  141. Teng H, Qi Y (2003) Application of wavelet technique to freeway incident detection. *Transp Res Part C* 11(3–4):289–308
  142. *Traffic Flow Theory 2006* (2006) Monograph with 22 papers on the subject of traffic flow theory. *Transp Res Rec* 1965
  143. Treiterer J (1975) *Investigations of Traffic Dynamics by Aerial Photogrammetry*. Ohio State University Technical, Report PB 246 094. Columbus, Ohio
  144. Van der Voort M, Dougherty M, Watson S (1996) Combining KOHONEN maps with ARIMA time series models to forecast traffic flow. *Transp Res Part C* 4:307–318
  145. Van Lint JWC, Hoogendoorn P, Van Zuylen HJ (2002) Freeway travel time prediction with state-space neural networks-modeling state-space dynamics with recurrent neural networks. *Transp Res Rec* 1811:30–39
  146. Van Lint JWC, Van der Zijpp NJ (2003) Improving a travel time estimation algorithm by using dual loop detectors. *Transp Res Rec* 1855:41–48
  147. Venkatanarayana R, Smith BL, Demetsky MJ (2005) Traffic Pattern Identification using Wavelets Transforms. In: *84th Transportation Research Board Annual Meeting*. Washington DC
  148. Vlahogianni EI, Golias JC, Karlaftis MG (2004) Short-term Traffic Forecasting: Overview of Objectives and Methods. *Transp Res* 24(5):533–557
  149. Vlahogianni EI, Karlaftis MG, Golias JC (2006) Statistical Methods for Detecting Non-linearity and Non-stationarity in Univariate Short-term Time-series of Traffic Volume. *Transp Res C* 14(5):351–367
  150. Wang Y, Papageorgiou M (2005) Real-Time Freeway Traffic State Estimation based on Extended Kalman Filter: A General Approach. *Transp Res B* 39:141–167
  151. Wahle J, Bazzan A, Klügl F, Schreckenberg M (2000) Anticipatory Traffic Forecast Using Multi-Agent Techniques. In: Helbing D, Hermann HJ, Schreckenberg M, Wolf DE (eds) *Traffic and Granular Flow '99*. Springer, Berlin, pp 87–92
  152. Whitham G (1974) *Linear and Nonlinear Waves*. Wiley, New York
  153. Wiedemann R (1974) *Simulation of Traffic Flow* (in German). Schriftenreihe des Instituts für Verkehrswesen der Universität Karlsruhe, No 8
  154. Wild D (1997) Short-Term Forecasting Based on A Transformation and Classification of Traffic Volume Time Series. *Int J Forecast* 13:63–72
  155. Williams BM (2001) Multivariate Vehicular Traffic Flow Prediction: An Evaluation of ARIMAX Modeling. *Transp Res Rec* 1776:194–200
  156. Williams BM, Hoel LA (2003) Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical results. *J Transp Eng* 129(6):664–672
  157. Williams JC, Mahmassani HS, Herman R (1987) Urban Network Flow Models. *Transp Res Rec* 1112:78–88
  158. Xiao H, Sun H, Ran B, Oh Y (2003) Fuzzy-Neural Network Traffic Prediction with Wavelet Decomposition. *Transp Res Rec* 1836:16–20
  159. Yang F, Sun H, Tao Y, Ran B (2004) Temporal difference learning with recurrent neural network in multi-step ahead freeway speed prediction. In: *83rd Transportation Research Board Annual Meeting*. Washington DC
  160. Yang F, Lin Z, Liu H X, Ran B (2004) Online Recursive Algorithm for Short-Term Traffic Prediction. *Transp Res Rec* 1879:1–9
  161. Yasdi R (1999) *Prediction of Road Traffic using a Neural Network*. Neural Computing and Applications, vol 8. Springer, Berlin, pp 135–142
  162. Yin H, Wong SC, Xu J (2002) Urban Traffic Prediction Using a Fuzzy-Neural Approach. *Transp Res C* 10:85–98
  163. Zhang G, Patuwo E, Hu MY (1998) Forecasting with artificial neural networks: The state of the art. *Int J Forecast* 14:35–62



164. Zhang HM (2000) Recursive Prediction of Traffic Conditions with Neural Networks. *J Transp Eng* 126(6):472–481
165. Zhang X, Rice J (2003) Short term travel time prediction. *Transp Res C* 11:187–210
166. Zwahlen HT, Russ A (2002) Evaluation of the accuracy of a real-time travel time prediction system in a freeway construction work zone. *Transp Res Rec* 1803:87–93

### Books and Reviews

- Kalman R (1960) A new approach to linear filtering and prediction problems. *ASME Basic Eng J*
- Kants H, Schreiber T (2004) *Nonlinear Time Series Analysis*. Cambridge University Press, Cambridge

## Travel Behavior and Demand Analysis and Prediction

KONSTADINOS G. GOULIAS  
University of California Santa Barbara,  
Santa Barbara, USA

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Dynamic Planning Practice  
Sustainable and Green Visions  
New Research and Technology  
The Evolving Modeling Paradigm  
Examples of Mathematical Models  
Summary  
Future Directions  
Bibliography

### Glossary

**Activity-based approach** A modeling method that accounts for the interdependent relationships among activities and persons to derive travel demand equations.

**Dynamic planning** The incorporation of trends, cycles, and feedback mechanisms into a process of actively shaping our future. Desired futures are first defined in terms of performance measures and a combination of forecasting and backcasting methods are used to identify the right paths to follow in achieving these futures.

**Microsimulation** A method to represent the movement in space and time of the most elementary units of a phenomenon. When applied in traffic engineering the units are vehicles. When applied in travel behavior the units are persons and households. Multi-agent

microsimulation allows to also represent human interaction with each person modeled as an agent.

**Travel demand** The amount of travel within a time interval such as number of trips in a day, total amount of distance and total amount of travel time, the locations (destinations) visited, the means used to reach these locations, departure time and arrival time of trips, routes followed in reaching these locations, the sequencing and assembly of trips in groups, and the purpose or activity engaged in at the end of each trip.

### Definition of the Subject

Transportation modeling and simulation aims at the design of an efficient infrastructure and service to meet our needs for accessibility and mobility. At its heart is good understanding of human behavior that includes the identification of the determinants of behavior and the change in human behavior when circumstances change either due to control (e. g., policy actions), trends (e. g., demographic change), or unexpectedly (e. g., disasters). This is the key ingredient that drives most decisions in transportation planning and traffic operations. Since transportation systems are the backbone connecting the vital parts of a city (a region, a state or an entire country), in-depth understanding of transportation-related human behavior is essential to the planning, design, and operational analysis of all the systems that make a city function.

Understanding human nature requires us to analyze and develop synthetic models of human agency in its most important dimensions and the most elemental constituent parts. This includes, and it is not limited to, understanding of individual evolution along a life cycle path (from birth to entry in the labor force to retirement to death) and the complex interaction between an individual and the anthropogenic environment, natural environment, and the social environment. Travel behavior research is one aspect of analyzing human nature and aims at understanding how traveler values, norms, attitudes, perception and constraints lead to observed behavior. Traveler values and attitudes refer to motivational, cognitive, situational, and disposition factors determining human behavior. Travel behavior refers primarily to the modeling and analysis of travel demand, based on theories and analytical methods from a variety of scientific fields. These include, but are not limited to, the use of time and its allocation to travel and activities, methods to study this in a variety of time contexts and stages in the life of people, and the arrangement or artifacts and use of space at any level of social organization such as the individual, the household, the community, and other formal or informal

groups This includes the movement of goods and the provision of services having strong interfaces and relationships with the engagement in activities and the movement of persons.

Travel behavior analysis and synthesis can be examined from both objective (observed by an analyst) and subjective (perceived by the human) perspectives in an integrated manner among four dimensions of time, geographic space, social space, and institutional context. In a few occasions the models reviewed here include and integrate time and space as conceived in science with perceptions of time and space by humans in their everyday life. For this reason research includes theory formation, data collection, modeling, inference, and simulation methods to produce decision support systems for policy assessment and evaluation that combine different views of time and space. Another objective of understanding human behavior is conceptual integration. Explanation of facts from different perspectives can be considered jointly to form a comprehensive understanding of people and their groups and their interactions with the natural and built environment. In this way, we may see explanations of human behavior fusing into the same universal principles. These principles eventually will lead to testable hypotheses from different perspectives offering Wilson's, 1998, famous concisence among, for example, psychology, anthropology, economics, the natural sciences, geography, and engineering. Unavoidably this is a daunting task with many model propositions in the research domain and very few ideas finding fertile ground in applications. The analysis-synthesis path in travel behavior gave us methods that help us understand and predict human (travel) behavior only partially leaving many gaps [163]. However, policy questions are becoming increasingly impossible to address with old tools, a large pool of researchers is actively working on new methods, and many public agencies commenced a variety of tool development projects to fill the travel behavior analysis gaps. To capture these trends, we see modeling examples with ideas from a transdisciplinary viewpoint and contributors to modeling and simulation from a variety of merged backgrounds (e.g., see the evolution of ideas in a sequence of the International Association for Travel Behavior Research conferences – [www.public.asu.edu/~rpendyal/iatbr/iatbr\\_index.htm](http://www.public.asu.edu/~rpendyal/iatbr/iatbr_index.htm)).

In the next sections the evolving paradigm of modeling and simulation is reviewed in detail and three of its fundamental sources are presented. Through the lens of contemporary planning practice the analytical requirements for modeling and simulation are discussed. Then, these same requirements are refined by examining contemporary visions about the world surrounding us and

the theories and technologies we can use to build policy analysis models. This article ends with a section describing the emerging modeling and simulation paradigm, a brief section of mathematical models and closes with a summary.

## Introduction

The impressive movement forward of transportation modeling and simulation emerges from three related but distinct sources. The first source is a fundamental change in planning practice that one could name *dynamic planning practice* to indicate the existence of bi-directional time (from the past to the future and from the future to today), as well as, assessment cycles and adjustments taking place within the short term, medium term, and long term horizons. These cycles are also bidirectional in time. This source contains three fundamental directions of practice that are *inventory creation and maintenance, strategy measurement and evaluation, forecasting and backcasting*. The second source is a vision that generates the substantive problems that we need to solve and the specific policies we need to examine. It is named *sustainable and green visions*. Problems and solutions in this general area motivate and inspire contemporary substance and content of policies throughout the world. One can identify three complementary and mutually strengthening directions in the *economy, environment, and society* that are the three fundamental pillars of sustainability. The third source is the never ending research for improved understanding of the world surrounding us. This source is named *new research and technology* to capture the most important elements of new discovery and new techniques enabling new discovery but also modeling and simulation. Key directions of inquiry within research and technology are *theory building, modeling and simulation, and enabling technologies*.

## Dynamic Planning Practice

Dynamic thinking means that time and change are intrinsic in the thought processes underlying planning activities. In the past, assumptions about the existence of a tenable and general equilibrium and our ability to build the infrastructure needed to meet demand did not require careful orchestration of actions. This was radically changed in the industrialized world to meet specific goals using available finite resources to maximize benefits. Together with our inability to build at will and a tendency to the preservation of non-renewable resources (e.g., land and open space, fossil fuels, time) we are much more motivated to think strategically and to consider in a more careful way

the performance of the overall anthropogenic system as we plan, design, operate, and manage transportation systems. Any action of this type, however, requires that we have a detailed and accurate picture of our facilities, their interconnectedness, their status within the hierarchy of movements, their conditions, and their evolving role. An accurate and more complete picture like this is called an *inventory* herein.

Many planning activities at all geographical levels are preceded by data gathering steps of identifying all the sources of data and information about the specific study area's transportation system and its relationship with the rest of the world. These inventories include the typical information about the resident population – demographics and employment, land available and land uses, economic development and growth, and so forth. It is worth pointing out the inventory contains data and relationships within the geographic area of interest (region) but also the region's relationship with other areas with which substantial flow of people, goods, and communication takes place. Inventories may also include data and information about cultural and historical factors. For example, statewide plans identify a variety of corridors as buffers of land and communities around major routes of the movement of people and goods. Some of these routes were created centuries ago when pioneers were still exploring uncharted lands. These routes experienced a major change when waterways were the main links among economic and military centers, and they are still evolving. Today these same routes contain as backbones railways, freeways, rivers, and often they surround major distribution locations such as ports and airports. Their nature is heavily influenced by their historical and cultural context.

Travel behavior analysts are familiar with inventories created for the regional long range plans, which subdivide the study area in traffic analysis zones with data from the Decennial Census suitably reformatted and packaged for use in a specific application (i. e., the long range regional plan). Then, additional data are assigned to these same subdivisions to build a richer context for modeling and simulation. Thus, the inventory for a typical long range plan is an electronic map of where people live and work, the network(s) that connect different locations, availability of different modes on each segment of the network, as well as information about travel network performance (e. g., link capacities, speeds on links, congestion, and connectivity). Today the tool of choice for data storage and visualization is a Geographic Information System (GIS).

One of the thorniest problems within this context is maintaining an up to date inventory (e. g., characteristics

of the population in each zone, presence of certain types of businesses, location and characteristics of intermodal facilities). This is a particularly important issue for periods in between decennial censuses. Year to year updates are very often required to provide “fresh” data. Many of these updates are becoming widely available and much less expensive than in the past. For example, the inventory of the highway network, with suitable additions and improvements, is available from the same private providers of in-vehicle navigation systems. In a similar way, inventories of businesses and residences can also be purchased from vendors. Census data, however, are required even when one uses data from private providers because they contain complementary data (e. g., the age distribution of the resident population) and they tend to provide wider coverage of a country. Although the need for inventories is undoubtedly extremely important many important issues are yet to be resolved. This is the core issue of two Transportation Research Board (TRB) conference proceedings on the National Household Travel Survey <http://www.trb.org/Conferences/NHTS/Program.pdf> and the US Census and the Census American Community Survey <http://www.trb.org/conferences/censusdata/>). Examples of unresolved issues include levels of detail we should use in updating the data we have, treatment of errors in the data and model sensitivity to these errors, frequency of data updates and treatment of missing data, and questions about merging different databases. Obviously, the answers to these questions are in the form of “it depends”. It depends on the budget (time and money) available, consequences of errors in the data, and the use of models in decision making. In fact, one particular type of data collection is strategy measurement where some of these questions become even more important. We turn now to the second dimension in the dynamic planning practice which is about strategy and performance.

Strategic planning and performance-based planning changed the way we plan for the future. This has been a 20 year long process in the United States as its transportation policy at the Federal, State, and Metropolitan levels is shaped by three consecutive legislative initiatives (ISTEA, TEA-21, and SAFETEA-LU). Under all three legislative frameworks and independently of role, location and perceived need for investment, the overall goal of funding allocation has been to maximize the performance of the transportation system in its entirety and avoid major new infrastructure building initiatives. As a result, planning practice at the Federal, State, and local levels is becoming heavily performance based and designed in a way that motivates the measurement of policy and program outcomes and judging these outcomes for funding allo-

cation. Two examples of performance-based planning are the Program Assessment Rating Tool (PART) at the federal level and performance-based transportation planning at the state level. PART is used to assess the management and performance of individual programs from homeland security to education, employment, and training. This is a tool that offers assessments about programs based on 25 questions divided into sections. For each program a tailored analysis yields summaries that receive a rating from 0 to 100 ranging from ineffective to effective [172]. In a different way but in the same spirit many states have created long range plans that are strategic and they measure transportation performance. Yearly evaluative updates are also used for a state's strategic transportation plan. After a comprehensive public involvement campaign a few themes capturing the desires of the resident population are first identified. To these themes technical requirements based on planners and agency inputs are added, a large number of objectives are created and then a variety of measures of performance are developed. These measures are given target levels that evolve over time to a desired future performance for the entire state and for a finite number of corridors of statewide significance. Yearly evaluations contain measures of target achievement and they should be used to guide an agency in its investments. The interface with regions is also included in this performance-based framework. Many infrastructure improvement projects in the US are selected from lists of projects that regions (called Metropolitan Planning Organizations) submit to their state to be included in a list of projects in the Transportation Improvement Program (TIP) and become candidates for funding. Under statewide performance-based planning, these projects are evaluated with respect to their contribution in meeting the statewide performance measures and in some states the performance measures of the relevant corridor [122]. Although these examples are far ranging in time and space, they contain operations components and yearly evaluations that: a) require data collection, modeling, and simulation at finer spatial and temporal scales than their counterpart planning feedbacks used in the long range transportation planning practice, and b) need a method that is able to coordinate the short, medium, and long term impacts. Emerging from these considerations are questions about the types of consistency we need among geographic scales for planning and operations actions to perform evaluations, policy requirements for coordination among planning activities to ensure consistency, need for suitable methods to coordinate smaller projects in broader contexts (either of policy assessment or geographical area), development of tools required to perform measurement of impacts and pro-

gram evaluation at the newly defined assessment cycles, and optimal planning activity with evaluation methods. Only a few solutions to the issues above are offered by contemporary projects such as the TRANSLAND project [70]. Within the context of integration between land use and transportation planning and the context of the European Union some of the conclusions include a call to strengthen regional plans, a stronger emphasis on public transport, strategic planning involving all actors, and the packaging of policies aiming at the same objectives. These themes are very similar to statewide and US Federal and European Union levels of planning. Very little, however, is said about the assessment methods and the choices we make in impact estimation. Performance assessment and evaluation of program effectiveness require the use of the inventory discussed before and a battery of models to forecast future expectations as well as to identify the actions required today to achieve desired futures.

As illustrated later in this article a new approach emerges in which models of discrete choice are applied to individual decision makers that are then used to (micro)simulate most of the possible combinations of choices in a day. The result is in essence a synthetic generation of traveling for the entire population. When the microsimulation also includes activities and duration at activity locations it becomes a synthetic schedule. In parallel, for forecasting purposes a synthetic population is first created for each land subdivision with all the relevant characteristics and then models are applied to the residents of each subdivision to represent areawide behavior. Changes are then imposed on each individual as a response to policies and predictive scenarios of policy impacts are thus developed. The evolution of individuals, their groups, and the entire study area can be used for trend analysis that includes details at the level of decision makers (either for passenger travel and/or for freight). In addition, progression in time happens from the present to the future and one could identify paths of change by individuals and groups if the application has been designed in the proper way (e.g., keeping detailed accounting of individuals as they move in time, using models that are designed for transitions over time and so forth). In a forecasting setting progression in time follows calendar time, temporal resolution is most often a year, and the treatment of dynamics is an one-way causal stream to the future.

Within the broader study of futures, forecasting is the method we use to develop *projective scenarios*. Performance-based planning, however, requires tools that can extrapolate from future performance targets the actions required today to reach them. In essence we also need *prospective studies* that start from a desirable future and

**Travel Behavior and Demand Analysis and Prediction, Table 1**  
**Backcasting schema**

Content	Method	
Determine objectives, purpose of the analysis, temporal, spatial and substantive scope of the analysis, decide the number and type of scenarios. Identify endogenous and exogenous variables	Problem orientation with technical representatives and stakeholders	
Specify goals, constraints and targets for each scenario analysis and exogenous variables	Stakeholder creativity workshop and brainstorming sessions	
Describe present system (building and updating of inventories), patterns and trends. Define processes, their actors, and determinants of outcomes. Identify exogenous variables and inputs to scenario analysis.	Scenario development by technical experts	
Scenario analysis. Select suitable approach, analyze system evolution at end time points and intermediate time points, develop scenarios, iterate to make sure all components are consistent/coherent	Scenario assessment by technical experts and stakeholders	
Undertake impact analysis. Consolidate scenario results. Analyze social, economic and environmental impacts. Compare results of the last with targets, iterate analysis with any other step as required to ensure consistency between goals and results	Backcasting workshops and stakeholder consultation (repeat to follow the iterations)	
Implement Policy Actions		

move backwards to identify specific actions that will lead us to that prospect. *Backcasting* was invented in a study of future energy options by [141], to do exactly this through a participatory process. Scenarios in backcasting are the “images” of the future and the possible paths that will take us to that future. A typical application includes the stages shown in Table 1. An open question, however, remains with respect to scenario construction and assessment. This is particularly important when one considers the serious issues we face with inadequate design of experiments/trials in the forecasting setting. Forecasting and backcasting have some important differences in their objectives. On one hand forecasting is employed to identify likely futures and to develop methods to help us identify small changes in our policies. It is also a method to extrapolate past trends into the future and possibly identify paths of changes that are heavily influenced by habit and inertia. Backcasting, on the other hand is designed to discover new ways to build desirable futures. It is perfectly aligned with strategic planning and it is a better suited method for developing a program of conditions to meet targets. Many of the models developed to date are designed for forecasting applications (either to inform the design of forecasting model systems or to create necessary components in the model systems). Yet, planning practice is moving towards strategy development and therefore needs model components that fit within a backcasting scenario building (see the reversed four-step model in Miller and Demetsky [120], and its neural network implementation in Sadek et al. [143] and the participatory tools in California (<http://www.sacregionblueprint.org/> – accessed May 2007).

### Sustainable and Green Visions

Policy actions also view the world surrounding us as an integral ecosystem placing more emphasis on its overall survival by examining direct and indirect effects of individual policy actions and entire policy packages or programs (see the examples in [116]). This trend is not limited to transportation. Lomborg [104], shows that a sustainable and green vision encompasses the entire range of human activity and the entirety of the ecosystem we live in. Although these are good news, because the approach enables analyzes and policies that are consistent in their vision about futures, comprehensive views also reveal that the pace of economic growth and development is in clear conflict with the biological pace of evolution with unknown consequences [162] strengthening the view that more comprehensive analytical frameworks are required.

In fact, one of the most recent studies on research needs, which addresses the transportation and environment relationship by the Transportation Research Board of the National Academies [167,168], expands the envelope to incorporate ecology and natural systems and addresses human health in a more comprehensive way than in the past reiterating the urgency to address unresolved issues about environmental damage. As a result, we also experience a clear shift to policy analysis approaches that have an expanded scope and domain and they are characterized by explicit recognition of transportation system complexity and uncertainty.

Reflecting all this, *sustainable transportation* is now often used to indicate a shift in the mentality of the community of transportation analysts to represent a vision



of a transportation system that attempts to provide services that minimize harm to the environment. In fact, in one of the most comprehensive reviews of policies in North America, Meyer and Miller [116], contrast the non-sustainable to the sustainable approaches. They provide a compelling argument about the change in these policies and pathways toward a more sustainable path. In the US during the past twenty years, the need, to examine these new and more complex policy initiatives, has also become increasingly pressing due to the passage of a series of legislative initiatives (Acts) and associated Federal and State regulations on transportation policy, planning, and programming. The multi-modal character of the new legislation, its congestion management systems and the taxing air quality requirements for selected US regions have motivated many new forecasting applications that in the early years were predominantly based on the Urban Transportation Planning System and related processes but during the last five years motivated a shift to richer conceptual frameworks. In point of fact, air quality mandates motivated impact assessments of the so called transportation control measures and the creation of statewide mobile source air pollution inventories [65,107,154] that require different analytical forecasting tools than in any pre-1990 legislative initiatives [124]. An added motivation is also lack of substantial funding for transportation improvement projects and a shift to charge the firms that benefit the most from transportation system improvements creating a need for impact fee-assessment for individual private developers. These assessments create the need for higher resolution in the three dimensions of geography (space), time (time of day), and social space (groups of people with common interests and missions, households, individuals) used in typical regional forecasting models but also the domain of jurisdictions where major decisions are made. They also create a pressing need for interfaces with traffic engineering simulation tools that are approved and/or endorsed in legislation (for examples see Paaswell et al. [126]). Another push for new tools is the assessment of technologies under the general name of Intelligent Transportation Systems (i. e., bundles of technological solutions in the form of user services attempting to solve chronic problems such as congestion, safety, and air pollution). Natural and anthropogenic tragic recent events are adding requirements for modeling and simulation and urgency in their development and implementation as well as more detail in time and space [75].

As Garrett and Wachs [46], discuss in the context of a lawsuit against a regional planning agency in the Bay Area, traditional four-step regional simulation models [30, 80,125] are outpaced by the same legislative stream of

the past 20 years that defined many of the policies described above. Unlike the “energy crisis” of the 1970s, the urgency and timeliness of modeling and simulation is becoming more urgent, more complex, and requires an “integrated” approach. Under these initiatives, forecasting models, in addition to long-term land use trends and air quality impacts, need to also address issues related to technology use and information provision to travelers in the short and medium terms. Similarly, the European Union focuses on issues such as: increasing citizen participation, intra-European integration, decentralization, deregulation, privatization, environmental concerns, mobility costs, congestion management by population segments, and private infrastructure finance (see van der Hoorn [174]). Table 2 provide an overview of policy tools that are loosely ordered from the longer term of land use and governance to medium and shorter term operational improvements depending on the lag time required for their impacts to be realized.

These policy initiatives place more complex issues in the domain of regional policy analysis and forecasting and amplify the need for methods that produce forecasts at the individual traveler and her/his household levels instead of the traffic analysis zone level. In addition to the long range planning activities and the typical traffic operations management activities, analysts and researchers in planning need to also evaluate the following: a) traveler and transportation system manager information provision and use (e. g., location based services, smart environments providing real time information to travelers, vehicles, and operators); b) combinations of transportation management actions and their impacts (e. g., parking fee structures and city center restrictions, congestion pricing), and c) assessment of combinations of environmental policy actions (e. g., carbon taxes and information campaigns about health effects of ozone).

To perform all this we need tool that also have forecasting and backcasting capabilities that are more accurate and detailed in space and time. In fact, planning initiatives are moving toward parcel by parcel analysis and yearly assessments. It is also conceivable that we need separate analyzes for different seasons of a year and days of the week to capture seasonal and within a week variations of travel. Echoing all this and in the context of the Dutch reality Borgers, Hofman, and Timmermans [21] have identified five information need domains that the new envisioned policy analysis models will need to address and they are (in a modified format from the original list):

1. social and demographic trends that may produce a structural shift in the relationship between places

**Travel Behavior and Demand Analysis and Prediction, Table 2**  
**Examples of Policy Tools**

Type of policy tool	Brief description	Source of information*
Land use growth and management programs	Legislation that controls for the growth of cities in sustainable paths	<a href="http://www.smartgrowth.org">www.smartgrowth.org</a> , <a href="http://www.awcnet.org">www.awcnet.org</a> <a href="http://www.fhwa.dot.gov/planning/ppasg.htm">www.fhwa.dot.gov/planning/ppasg.htm</a> <a href="http://www.compassblueprint.org">www.compassblueprint.org</a>
Land use design and attention to neighborhood design for non-motorized travel	Similar to the previous but with attention paid to individual neighborhoods	<a href="http://www.sustainable.doe.gov/landuse/luothtoc.shtml">www.sustainable.doe.gov/landuse/luothtoc.shtml</a> <a href="http://www.planning.dot.gov/Documents/DomesticScan/domscan2.htm">www.planning.dot.gov/Documents/DomesticScan/domscan2.htm</a>
City annexations and spheres of influence	City boundaries are divided into incorporated, within the sphere of influence, and external to manage growth	<a href="http://countypolicy.co.la.ca.us/BOSPolicyFrame.htm">http://countypolicy.co.la.ca.us/BOSPolicyFrame.htm</a> <a href="http://www.ite.org/activeliving/files/Jeff_Summary.pdf">www.ite.org/activeliving/files/Jeff_Summary.pdf</a>
Accelerated retirement of vehicles programs	Programs to eliminate high emitting and older technology vehicles	<a href="http://ntl.bts.gov/DOCS/SCRAP.html">ntl.bts.gov/DOCS/SCRAP.html</a>
Public involvement and education programs	Programs aiming at defining goals based on the public's desires	<a href="http://www.fhwa.dot.gov/reports/pittd/contents.htm">www.fhwa.dot.gov/reports/pittd/contents.htm</a>
Health promoting programs	Programs that promote physical activity in travel to benefit health	<a href="http://www.activelivingbydesign.org">www.activelivingbydesign.org</a>
Safety measures	A process to incorporate safety considerations in transportation planning	<a href="http://tmip.fhwa.dot.gov/clearinghouse/docs/safety/">tmip.fhwa.dot.gov/clearinghouse/docs/safety/</a> <a href="http://www.fhwa.dot.gov/planning/scp/">www.fhwa.dot.gov/planning/scp/</a> <a href="http://www.safetyanalyst.org/">www.safetyanalyst.org/</a>
Emission control, vehicle miles traveled, and other fee programs (including carbon taxes and trading)	Programs that shift taxation from traditional sources towards pollutant emissions and natural resource depletion agents	<a href="http://www.fresh-energy.org/">www.fresh-energy.org/</a> <a href="http://www.fhwa.dot.gov/environment/">www.fhwa.dot.gov/environment/</a> <a href="http://www.fightglobalwarming.com/">www.fightglobalwarming.com/</a>
Congestion pricing and toll collection programs	A premium is charged to travelers that wish to travel during the most congested periods	<a href="http://www.vtpi.org/london.pdf">www.vtpi.org/london.pdf</a>
Parking fee management	Parking pricing used as a tool to restrict access by space and time	<a href="http://www.gmu.edu/depts/spp/programs/parkingTaxes.pdf">www.gmu.edu/depts/spp/programs/parkingTaxes.pdf</a>
Non-motorized systems	Programs to support walking and biking	<a href="http://www.vtpi.org/tdm/tdm25.htm">www.vtpi.org/tdm/tdm25.htm</a> <a href="http://www.psrc.org/projects/nonmotorized">www.psrc.org/projects/nonmotorized</a>
Telecommuting and Teleshopping	The employment of telecommunications to substitute-complement-enhance travel	<a href="http://www.telework-mirti.org">www.telework-mirti.org</a> <a href="http://www.vtpi.org/tdm/tdm43.htm">www.vtpi.org/tdm/tdm43.htm</a>
Flexible and staggered work programs	Programs that change the workweek of individuals and firms	<a href="http://www.its.dot.gov/JPODOCS/REPTS_PR/13669/section05.htm">www.its.dot.gov/JPODOCS/REPTS_PR/13669/section05.htm</a>
Goods movements (freight) programs to improve operations	A variety of programs to facilitate and minimize the damage for freight movement	<a href="http://ntl.bts.gov/DOCS/harvey.html">ntl.bts.gov/DOCS/harvey.html</a>
Highway system improvements in traffic operations and flow	Improved data collection, monitoring, and traffic management	<a href="http://www.transportation.org">www.transportation.org</a> <a href="http://ite.org/mega/default.asp">ite.org/mega/default.asp</a>
Intelligent Transportation Systems (ITS)	Use of telecommunications and information technology to manage and control travel	<a href="http://www.itsa.org/">www.itsa.org/</a> <a href="http://www.ertico.com/">www.ertico.com/</a> <a href="http://www.its.dot.gov/index.htm/">www.its.dot.gov/index.htm/</a>
Special event planning and associated traffic management	Enhanced procedures to handle the demands of a special event	<a href="http://tmcps.ops.fhwa.dot.gov/cfprojects/new_detail.cfm?id=32xxxnew=0">tmcps.ops.fhwa.dot.gov/cfprojects/new_detail.cfm?id=32xxxnew=0</a>
Security preparedness through metropolitan planning processes	A process to incorporate safety considerations in transportation planning	<a href="http://www.planning.dot.gov/Documents/Securitypaper.htm">www.planning.dot.gov/Documents/Securitypaper.htm</a>
Individualized marketing techniques with improved information and communication with the "customer"	Public programs to provide personal help in changing travel behavior in favor of environmentally friendly modes	<a href="http://www.local-transport.dft.gov.uk/travelplans/index.htm">www.local-transport.dft.gov.uk/travelplans/index.htm</a> <a href="http://www.travelsmart.gov.au/">http://www.travelsmart.gov.au/</a>

\*accessed May 2007

and time allocation by individuals invalidating existing travel behavior model systems;

2. increasing scheduling and location flexibility and degrees of freedom for individuals in conducting their ev-

ery day business leading to the need to consider additional choices (e.g., departure time from home, work at home, shopping by the internet, shifting activities to the weekend) in modeling travel behavior;

3. changing quality and price of transport modes based on market dynamics and not on external to the travel behavior policies (e. g., the effect of deregulation in public transport);
4. shifting of attitudes and potential cycles in the population outlook about travel options; and
5. changing scales/jurisdictions (scale is the original term used to signify the different jurisdictions) – different policy actions in different sectors have direct and indirect effects on transportation and different policy actions in transportation have direct and indirect effects in the other sectors (typical example in the US is the welfare to work program).

The first substantive implication of all these considerations is an expanded envelope of modeling and simulation. Many processes that were left outside the realm of transportation modeling and simulation need to be included as stages of the travel model system. One notable example is the inclusion of *residential location choice*, *work location choice*, and *school location choice* to capture the spatial distribution and relative location of important anchor points on travel behavior and to also capture the impact of transportation system availability and level of service on these choices. In this way when implemented policies lead to improved level of service and the relative attractiveness of locations change, shifts in residential location, work location, and possibly school location can be incorporated as impacts of transportation. A similar treatment is needed for *car ownership and car type choices* of households or *fleet sizes and composition* for firms. These car-related choices are expressed as functions of parking availability, energy and other costs and level of service offered by the transportation system (highway and transit). To account for other resources and facilities available for household travel we also need to consider processes for *driver's licensing*, acquiring of *public transportation subscription (passes)*, and participation in *car sharing programs*. In this way, variables of car availability and public transportation availability in households can be used as determinants of travel behavior. Similar treatment is required for policies that change attitudes, perceptions and knowledge about travel options.

To address some of the policies of Table 2, we need to transition to a domain that contains a variety of outputs that include shares of program participation, sensitivity to accessibility and prices, and the usual indicators of travel on networks using input variables from the processes and behaviors discussed up to this point. Although the number of vehicles per hour per lane is the typical input of traffic operations software, a variety of other variables such

as speeds on network links and types of vehicles are also needed for other models such as emissions estimation.

Ideally longer term social, economic, demographic, resource/facilities, and circumstances of people should be converted into yearly schedules identifying periods of vacation, workdays, special occasions, and so forth. These in turn should lead to weekly schedules separating days during which people stay at home from days during which people go to work and days during which they run errands and/or engage in other non-work and non-school related activities. In this way patterns of working days versus not working days can be derived in a natural (con)sequence. As we will see in a later section, a fundamental leap of faith intervenes in practice and converts all this background information into a representative day that is used to create a more or less complete sequence of activities and trips with their destinations and modes used.

In this way decisions and choices people make are organized along the time scale in terms of the time it takes for these events to occur and their implications. For example, decisions about education, careers and occupation, and residential and job location are considered first and they condition everything that happens next. These should be formulated in terms of one or more life course long projects and not represented by a cross-sectional choice model. Similarly, decisions about yearly school and work schedules that determine work days and vacation days in a year should also be modeled as a stream of interrelated choices. Conditional on all this are the daily schedules of individuals and the myriad of decisions determining a daily schedule, which are modeled in much more detail and paying closer attention to the mutual dependency among the different facets of a within a day schedule. The next section explores this further in the context of research and enabling technology. A section on mathematical models later in this article shows the beginning of a new way in modeling a simulation that emphasizes human interaction.

## New Research and Technology

The planning and policy analysis discussion identified many requirements for modeling and simulation. Planning and policy expanded the context of travel behavior models to entire life paths of individuals and for this reason a more general modeling framework is emerging. In fact, modeling made tremendous progress toward a comprehensive approach to, in essence, build simulated worlds on computer enabling the study of complex policy scenarios. Although, passenger travel received the bulk of the attention, similar contributions to new research

and technology are found in modeling the movement of goods [151,153]. The emerging framework, although incomplete, is rich in the directions taken and potential for scientific discovery, policy analysis, and more comprehensive approaches in dealing with sustainability issues.

There are four dimensions that one can identify in building taxonomies of simulation models. The first is the *geographic space* and its conditional continuity, the second is the *temporal scale* and calendar continuity, the third is interconnectedness of *jurisdictions*, and the fourth and most important is the set of relationships in *social space* for individuals and their communities. The first dimension, *geographic space* here is intended as the physical space in which human action occurs. This dimension has played important roles in transportation planning and modeling because the first preoccupation of the transportation system designers has been to move persons from one location to another (i.e., overcoming spatial separation). Initial applications considered the territory divided into large areas (traffic analysis zones), represented by a virtual center (centroid), and connected by facilities (higher level highways). The centroids were connected to the higher level facilities using a virtual connector summarizing the characteristics of all the local roads within the zone. As computational power increased and the types of policies/strategies required increased resolution, the zone became smaller and smaller. Today, it is not unreasonable to expect software to handle zones that are as small as a parcel of land and transportation facilities that are as low in the hierarchy as a local road (the centroid becomes the building on a parcel and the centroid connector is the driveway of the unit and they are no longer virtual).

In modeling and simulation we are interested in understanding human action. For this reason in some applications geographic space needs to consider more than just physical features (p. 387 in [49]) moving us into the notion of place and social space (see also below). The second dimension is *time* that is intended here as continuity of time, irreversibility of the temporal path, and the associated artificiality of the time period considered in many models. For example, models used in long range planning applications use typical days (e.g., a summer day for air pollution). In many regional long-range models the unspoken assumption is that we target a typical work weekday in developing models to assess policies. Households and their members, however, may not always (if at all) obey this strict definition of a typical weekday to schedule their activities and they may follow very different decision making horizons in allocating time to activities within a day, spreading activities among many days including weekends, substituting out of home with in home activities in some days but doing

exactly the opposite on others, and using telecommunications only selectively (e.g., on Fridays and Mondays more often than on other days). Obviously, taking into account these scheduling activities is by far more complex than what is allowed in existing transportation planning models. The third dimension is *jurisdictions* and their interconnectedness. The actions of each person are “regulated” by jurisdictions with different and overlapping domains such as federal agencies, state agencies, regional authorities, municipal governments, neighborhood associations, trade associations and societies, religious groups, and formal and informal networks of families and friends. In fact, the federal government defines many rules and regulations on environmental protection. These may end up being enforced by a local jurisdiction (e.g., a regional office of an agency within a city). On the one hand, we have an organized way of governance that clearly defines jurisdictions and policy domains (e.g., tax collection in the US). On the other hand, however, the relationships among jurisdictions and decision making about allocation of resources does not follow always this orderly governance principle of hierarchy. A somewhat different and more “bottom up” relationship is found in the social network and for this reason requires a different dimension that is the fourth and final dimension named *social space* and the relationships among persons within this space. For example, individuals from the same household living in a neighborhood may change their daily time allocation patterns and location visits to accommodate and/or take advantage of changes in the neighborhood such as elimination of traffic and the creation of pedestrian zones. Depending on the effects of these changes on the pedestrian network we may also see a shift in the within the neighborhood social behavior. In contrast, increase in traffic to surrounding places may create an outcry by other surrounding neighborhoods, thus, complicating the relationships among the residents.

One important domain and entity within this social space is the household. This has been a very popular unit of analysis in transportation planning recognizing that strong relationships within a household can be used to capture behavioral variation (e.g., the simplest method is to use a household’s characteristics as explanatory variables in a regression model of travel behavior). In this way any changes in the household’s characteristics (e.g., change in the composition due to birth, death, or children leaving the nest or adults moving into the household) can be used to predict changes in travel behavior. New model systems are created to study this interaction within a household looking at the patterns of using time in a day and the changes across days and years. It is therefore very important in modeling and simulation to incorporate

in the models used for policy analysis interactions among these four fundamental dimensions, which bring us to the next major issue that of scale.

The typical long range planning analysis is usually defined for larger geographical areas (region, states, and countries) and addresses issues with horizons from 10 to 50 years. In many instances we may find that large geographic scale means also longer time frames applied to wider mosaics of social entities and including more diverse jurisdictions. On the other side of the spectrum issues that are relevant to smaller geographic scales are most likely to be accompanied by shorter term time frames applied to a few social entities that are relatively homogeneous and subject to the rule of very few jurisdictions. This is one important organizing principle but also an indicator of the complex relationships we attempt to recreate in our computerized models for decision support. In developing the blueprints of these models one can choose from a variety of theories (e.g., neoclassical microeconomics) and conceptual representations of the real world that help us develop these models. At the heart of our understanding of how the world (as an organization, a household, a formal or informal group, or an individual human being) works are models of decision making and conceptual representations of relationships among entities making up this world.

Transportation planning applications are about judgment and decision making of individuals and their organizations. There are different settings of decision making that we want to understand. Three of these settings are the travelers and their social units from which motivations for and constraints to their behavior emerge; the transportation managers and their organizations that serve the travelers and their social units, and the decision makers surrounding goods movement and service provision that contain a few additional actors, Southworth [151]. These include land use markets (see [www.urbansim.org](http://www.urbansim.org)). Travelers received considerable attention in transportation planning and the majority of the models in practice aim at capturing their decision making process. The remaining settings received much less attention and they are poorly understood and modeled.

Conceptual models of this process are transformed into computerized models of a city, a region, or even a state in which we utilize components that are in turn models of human judgment and decision making (e.g., travelers moving around the transportation network and visiting locations where they can participate in activities). Models of this behavior are simplified versions of strategies used by travelers when they select among options that are directly related to their desired activities. In some of these

models we also make assumptions about hierarchies of motivations, actions, and consequences. Some of these assumptions are explicit (e.g., when deriving the functional forms of models as in the typical disaggregate choice models or the rules in a production system) and in other models these assumptions are implicit.

When designing transportation planning model interfaces for transportation planners and managers we also implicitly make assumptions about the managers' ability to understand the input, agent representation, internal functioning, and output of these computerized models. Our objective is therefore not only to understand travel behavior and build models that describe and predict human behavior but also to devise tools that allow transportation managers to understand the assumed behavior in the models, study scenarios of policy actions, and define and explain policy implications to others. This, in essence, implies that we, the model system designers, create a platform for a relationship between planners and travelers. A similar but more direct relationship also exists between travelers and transportation managers when we design the observation methods that provide the data for modeling but also the data used to measure attitudes and opinions such as travel surveys. In fact, this relationship is studied in much more detail in the survey design context and linked directly to the image of the agency conducting the survey and the positive or negative impression of the travelers about the sponsoring agency [33]. Most transportation research for modeling and simulation, however, has emphasized traveler behavior when building surveys and their models neglecting the interface with the planners. The summary of theories below, however, applies to individuals traveling in a network but also to organizations and planners in the sense used by H.A. Simon in his *Administrative Behavior* [150].

Rational decision making is a label associated with human behavior that follows a strategy in identifying the best course of action. In summary, a decision maker solves an optimization problem and identifies the best existing solution to this problem. Within this more general strategy when an operational model is needed and this operational model provides quantitative predictions about human behavior some kind of mathematical apparatus is needed to produce the predictions. One such machinery is the subjective expected utility [146] formulation of human behavior. In developing alternative models to SEU Simon [149] defines four theoretical components:

- A person's decision is based on a utility function assigning a numerical value to each option – *existence and consideration of a cardinal utility function*;



- The person defines an exhaustive set of alternative strategies among which just one will be selected – *ability to enumerate all strategies and their consequences*;
- The person can build a probability distribution of all possible events and outcome for each alternate option – *infinite computational ability*; and
- The person selects the alternative that has the maximum utility – *maximizing utility behavior*.

This behavioral paradigm served as the basis for a rich production of models in transportation that include the mode of travel, destinations to visit as well as the household residence (see the examples in the seminal textbook by Ben-Akiva and Lerman [9]). It served also as the theoretical framework for consumer choice models and for attempts to develop models for hypothetical situations (see the comprehensive book by Louviere, Hensher, and Swait [108]). It has also replaced the aggregate modeling approaches to travel demand analysis as the orthodoxy against which many old and new theories and applications are compared and compete with. SEU can be considered to be a model from within a somewhat larger family of models under the label of weighted additive rule (WADD) models [127]. Real humans, however, may never behave according to SEU or related maximizing and infinitely computational capability models (Simon labels this the Olympian model, [149]). Based on exactly this argument different researchers in psychology have proposed a variety of decision making strategies (or heuristics). For example, Simon created alternate model paradigms under the label of *bounded rationality – the limited extent to which rational calculation can direct human behavior* [149,150] to depict a sequence of a person's actions when searching for a suitable alternative. The modeled human is allowed to make mistakes in this search giving a more realistic description of observed behavior (see also Rubinstein [142]). Tversky is credited with another stream of decision making models starting with the *lexicographic approach* [86, in which a person first identifies the most important attribute, compares all alternatives on the value of this attribute, and chooses the alternative with the best value on this most important attribute. Ties are resolved in a hierarchical system of attributes. Another Tversky model [169] assumes a person selects an attribute in a probabilistic way and influenced by the importance of the attribute, all alternatives that do not meet a minimum criterion value (cutoff point) are eliminated. The process proceeds with all other attributes until just one alternative is left and that one is the chosen. This has been named the *elimination by aspects strategies* (EBA) model. Later, Kahneman and Tversky [170] developed *prospect theory*

and its subsequent version of *cumulative prospect theory* in Tversky and Kahneman [171] in which a simplification step is first undertaken by the decision maker editing the alternatives. Then, a value is assigned to each outcome and *a decision is made based on the sum of values multiplying each by a decision weight*. Losses and gains are treated differently. All these alternatives to SEU paradigms did not go unnoticed in transportation research with early significant applications appearing in the late 1980s. In fact, a conference was organized attracting a few of the most notable research contributors to summarize the state of the art in behavior paradigms and documented in Garling, Laitila, and Westin [45]. One of the earlier examples using another of Simon's inventions, the *satisficing behavior – acceptance of viable choices that may not be optimal* – is a series of transportation-specific applications described in Mahmassani and Herman [110]. Subsequent contributions continue along the path of more realistic models and the most recent example, discussing a few models, by Avineri and Prashker [7], uses cumulative prospect theory giving a preview of a movement toward more realistic travel behavior models. As Garling et al. [45] and Avineri and Prashker [7] point out, these paradigms are not ready for practical applications, contrary to the Mahmassani and colleagues efforts that have been applied, and additional work is required to use them in a simulation framework for applications. Another aspect is contextual *adaptation*. Payne, Bettman, and Johnson [127] provide an excellent review of decision making models and their differentiating aspects. They also provide evidence that decision makers *adapt* by switching between decision making paradigms *to the task and the context of their choices*. They also make mistakes and they may also fail to switch strategies. As Vause [175] discusses to some length transportation applications are possible using multiple decision making heuristics within the same general framework and employing a production system approach [123]. A key consideration, however, that has received little attention in transportation is the definition of context within which decision making takes place. Recent production systems [5] are significant improvements over past simulation techniques. However, travelers are still assumed to be passive in shaping the environment within which they decide to act (action space). This action space is viewed as largely made by constraints and not by their active shaping of their context. Goulias [58,60] reviews another framework from human development that is designed to treat decision makers in their active and passive roles and explicitly accounts for mutual influence between an agent (active autonomous decision maker) and her environment.

Transportation modeling and simulation experienced a few tremendously innovative and progressive steps forward. Interestingly these key innovations are from non-engineering fields but very often transferred and applied to transportation systems analysis and simulation by engineers. These are listed here in a somewhat sequential chronological order merging technological innovations and theoretical innovations. At exactly the time that the Bay Area Rapid Transit system was studied and evaluated in the 1960s, Dan McFadden (the Year 2000 Nobel Laureate in Economics) and a team of researchers produced practical mode choice regression models at the level of an individual decision maker (see <http://emlab.berkeley.edu/users/mcfadden/> – accessed June 2007). The models are based on random utility maximization (of the SEU family) and their work opened up the possibility to predict mode choice rates more accurately than ever before. These models were initially named *behavioral travel-demand models* [155] and later the more appropriate term of *discrete choice models* [9] prevailed. Although restrictive in their assumptions, these models are still under continuous improvement and they have become the standard tool in evaluating discrete choices. Some of the most notable and recent developments advancing the state of the art and practice are:

- Better understanding of the theoretical and particularly behavioral limitations of these models [45,50,115];
- more flexible functional forms that resolve some of the problems raised in Williams and Ortuzar [184] allowing for different choices to be correlated when using the most popular discrete choice regression models [14,16,95];
- combination of revealed preference, stated choices by travelers, with stated preferences and intentions, answers to hypothetical questions by travelers, availability of data in the same choice framework to extract in a more informative way travelers willingness to use a mode and willingness to pay for a mode option [10, 108]. This latter “improvement” enables us to assess situations that are impossible to build in the real world;
- computer-based interviewing and laboratory experimentation to study more complex choice situations and the transfer of the findings to the real world [111]. This direction, however, is also accompanied by a wide variety of research studies aiming at more realistic behavioral models that go beyond mode choice and travel behavior [50]; and
- expansion of the discrete choice framework using ideas from *latent class models* with covariates that were first developed by Lazarsfeld in the 1950s and their estima-

tion finalized by Goodman in the 1970s (see the review in [56], and discrete choice applications in [20]). This family of models was used in Goulias [57] to study the dynamics of activity and travel behavior and in the study of choice in travel behavior [12].

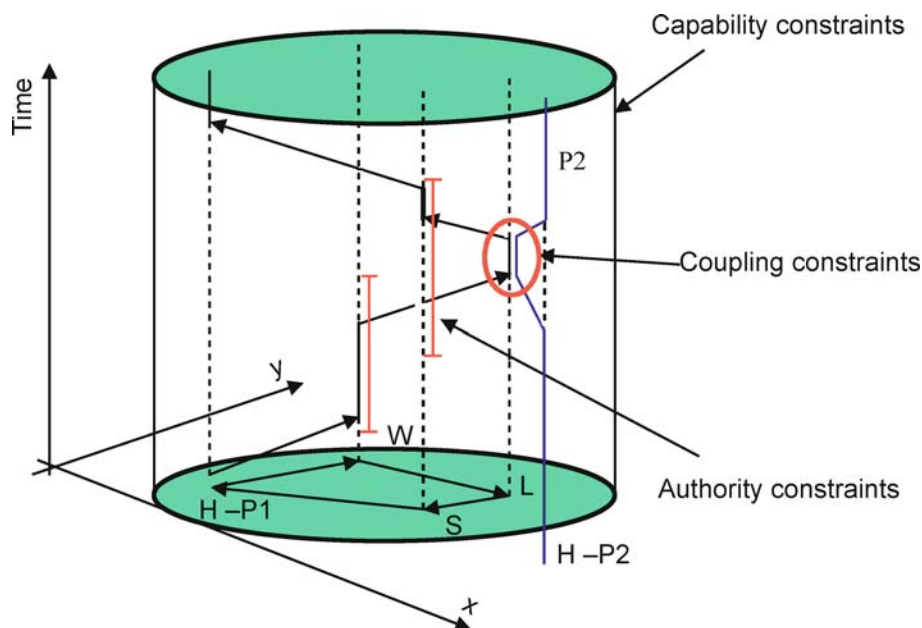
As mentioned earlier the rational economic assumption of the maximum utility model framework (that underlies many but not all of the disaggregate models) is very restrictive and does not appear to be a descriptive behavioral model except for a few special circumstances when the framing of decisions is carefully designed (something we cannot expect to happen every time a person travels on the network). Its replacement, however, requires conceptual models that can provide the types of outputs needed in regional planning applications. A few additional research paths, labeled as *studies of constraints*, are also functioning as gateways into alternate approaches to replace or complement the more restrictive utility-based models. A few of these models also consider knowledge and information provision to travelers. The first aspect we consider is about the choice set in discrete choice models. Choice set is the set of alternatives from which the decision maker selects one. These alternatives need to be mutually exclusive, exhaustive, and finite in number [166]. Identification, counting, and issues related to the alternatives considered have motivated considerable research in choice set formation [77,78,140,158,159]. Key threat to misspecification of the choice set is the potential for incorrect predictions [161]. When this is an issue of considerable threat as in destination choice models where the alternatives are numerous, a model of choice set formation appears to be the additional burden [71]. Other methods, however, also exist and they may provide additional information about the decision making processes. Models of the processes can be designed to match the study of specific policies in specific contexts. One such example and a more comprehensive approach defining the choice sets is the situational approach [25]. The method uses in depth information from survey respondents to derive sets of reasons for which alternatives are not considered for specific choice settings (individual trips). This allows separation of analyst observed system availability from user perceived system availability (e.g., due to misinformation and willingness to consider information). This brings us to the duality between “objective choice attributes” and “subjective choice attributes”. Most transportation applications, independently of the decision making paradigm adopted, assume the analysts (modelers) and the travelers (modeled) measured attributes to be the same. Modeling the process of perceived constraints may be far more complex when one

considers the influence of the context within which decisions are made. Golledge and Stimpson (pp. 33–34 in [49]) describe this within a conceptual model of decision making that has a cognitive feel to it. They also link the situational approach to the activity-based framework of travel extending the framework further (pp. 315–328 in [49]).

Chapin's research [28], providing one of the first comprehensive studies about time allocated to activity in space and time, is also credited for motivating the foundations of activity-based approaches to travel demand analysis. His focus has been on the propensity of individuals to participate in activities and travel linking their patterns to urban planning. In about the same period Becker also developed his theory of time allocation from a household production viewpoint [8] applying economic theory in a non-market sector and demonstrating the possibility of formulating time allocation models using economics reasoning (i. e., activity choice). In parallel another approach was developing in geography and Hagerstrand's seminal publication on time geography [72] presents the foundations of the approach. The idea of constraints in the movement of persons was taken a step further by this time-geography school in Lund. In that framework, the movement of persons among locations can be viewed as their movement in space and time under external constraints. Movement in time is viewed as the one way (irreversible) movement in the path while space is viewed as a three dimen-

sional domain. It provides the third base about *constraints* in human paths in time and space for a variety of planning horizons. These are *capability constraints* (e. g., physical limitations such as speed); *coupling constraints* (e. g., requirements to be with other persons at the same time and place); and *authority constraints* (e. g., restrictions due to institutional and regulatory contexts such as the opening and closing hours of stores). Figure 1 provides a pictorial representation in space and time of a typical activity-travel pattern of two persons (P1 and P2) and the three types of constraints. H indicates home, W indicates work, L indicates leisure, and S indicates shopping.

Cullen and Godson [31] also reviewed by Arentze and Timmermans [5] and Golledge and Stimpson [49] appear to be the first researchers attempting to bridge the gap between the motivational (Chapin) approach to activity participation and the constraints (Hagerstrand) approach by creating a model that depicts a routine and deliberated approach to activity analysis. The Cullen and Dobson study also defined many terms often used today in activity-based approaches. For example, each activity (stay-home, work, leisure, and shopping) is an episode characterized by start time, duration, and end time. Activities are also classified into fixed and flexible and they can be engaged alone or with others. Moreover, they also analyzed sequencing of activities as well as pre-planned, routine, and on the spur of the moment activities. Within this overall theoret-



Travel Behavior and Demand Analysis and Prediction, Figure 1

A two-person (P1 and P2) activity-travel pattern and the time and space limits imposed by constraints [132]

ical framework is the idea of a project which according to Golledge and Stimpson [49], *is a set of linked tasks that are undertaken somewhere at some time within a constraining environment* (pp. 268–269). This idea of the project underlies one of the most exciting developments in activity-based approaches to travel demand analysis and forecasting because seemingly unrelated activity and trip episodes can be viewed as parts of a “big-picture” and given meaning and purpose completing in this way models of human agency and explaining resistance to change behavior.

Most subsequent contributions to the activity-based approach emerge in one way or another from these initial frameworks with important operational improvements (for reviews see [5,17,89,114]). The basic ingredients of an activity based approach for travel demand analysis [5,84] are:

- a) explicit treatment of travel as derived demand [112], i. e., participation in activities such as work, shop, and leisure motivate travel but travel could also be an activity as well (e. g., taking a drive). These activities are viewed as episodes (i. e., they are characterized by starting time, duration, and ending time) and they are arranged in a sequence forming a pattern of behavior that can be distinguished from other patterns (i. e., a sequence of activities in a chain of episodes). In addition, these events are not independent and their interdependency is accounted for in the theoretical framework;
- b) the household is considered to be the fundamental social unit (i. e., decision making unit) and the interactions among household members are explicitly modeled to capture task allocation and roles within the household, relationships at one time point and change in these relationships as households move along their life cycle stages and the individual’s commitments and constraints change and these are depicted in the activity-based model; and
- c) explicit consideration of constraints by the spatial, temporal, and social dimensions of the environment is given. These constraints can be explicit models of time-space prisms [130] or reflections of these constraints in the form of model parameters and/or rules in a production system format [5].

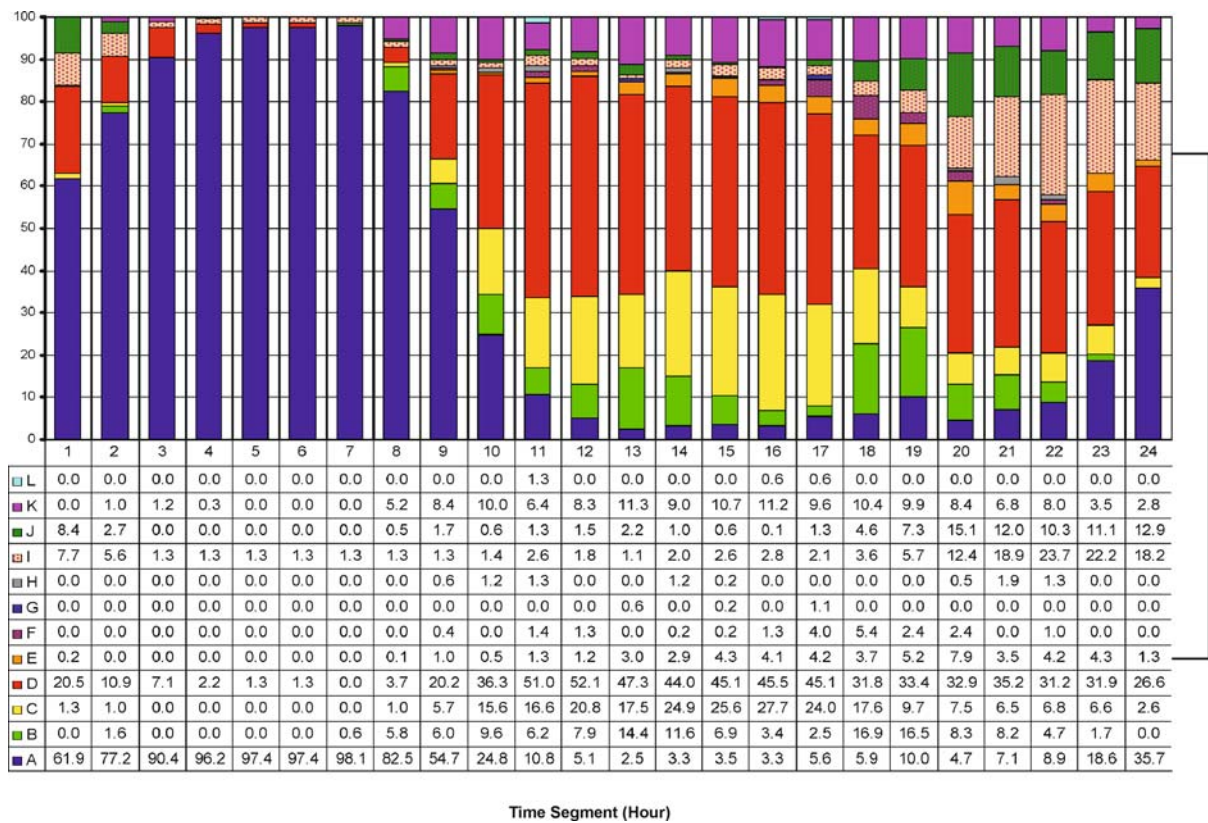
Input to these models are the typical regional model data of social, economic, and demographic information of potential travelers and land use information to create schedules followed by people in their everyday life. The output are detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity (including travel time, mode used, and so forth). This output is very much like a “day-timer” for each per-

son in a given region. Figure 2 provides an example of time allocation to different activities from an application that collected activity participation data [2,3]. It displays time allocation by one segment of the population showing the proportion of persons engaging in each activity by each hour of a day. Figure 3 shows the output from a model that predicts the presence of persons in each building during each hour of a day engaging in each activity type. Combining an activity model with a typical travel demand model produces “volumes” of individuals at specific locations and on the network of a city as shown in Figure 4 (a more detailed description of this study can be found in [67,97,99]).

Many planning and modeling applications, however, aim at forecasting. Inherent in forecasting are the time changes in the behavior of individuals and their households and their response to policy actions. At the heart of behavioral change are questions about the process followed in shifting from a given pattern of behavior to another. In addition to measuring change and the relationships among behavioral indicators that change in their values over time, we are also interested in the timing, sequencing, and staging of these changes. Moreover, we are interested in the triggers that may accelerate desirable or delay undesirable changes and the identification of social and demographic segments that may follow one time path versus another in systematic patterns. Knowledge about all this is required to design policies but it is also required to design better forecasting tools. Developments in exploring behavioral dynamics and advancing models for them have progressed in a few arenas. First, in the data collection arena with panel surveys, repeated observation of the same persons over time that are now giving us a considerable history in developing new ideas about data collection but also about data analysis [55,61] and interactive and laboratory data collection techniques [34] that allow a more in-depth examination of behavioral processes. The second arena is in the development of microeconomic dynamic formulations for travel behavior that challenge conventional assumptions and offer alternative formulations [91]. The third arena, is in the behavior from a developmental viewpoint as a single stochastic process, a staged development process [57], or as the outcome from multiple processes operating at different levels [59]. Experimentation with new theories from psychology emphasizing development dynamics is a potential fourth area that is just beginning to emerge [60]. Behavioral dynamics are also examined using more comprehensive analyzes [68] and models [136].

These models focus more on the paths of persons in space and time within a somewhat short time horizon such as a day, a week, or maybe a month. The consideration of



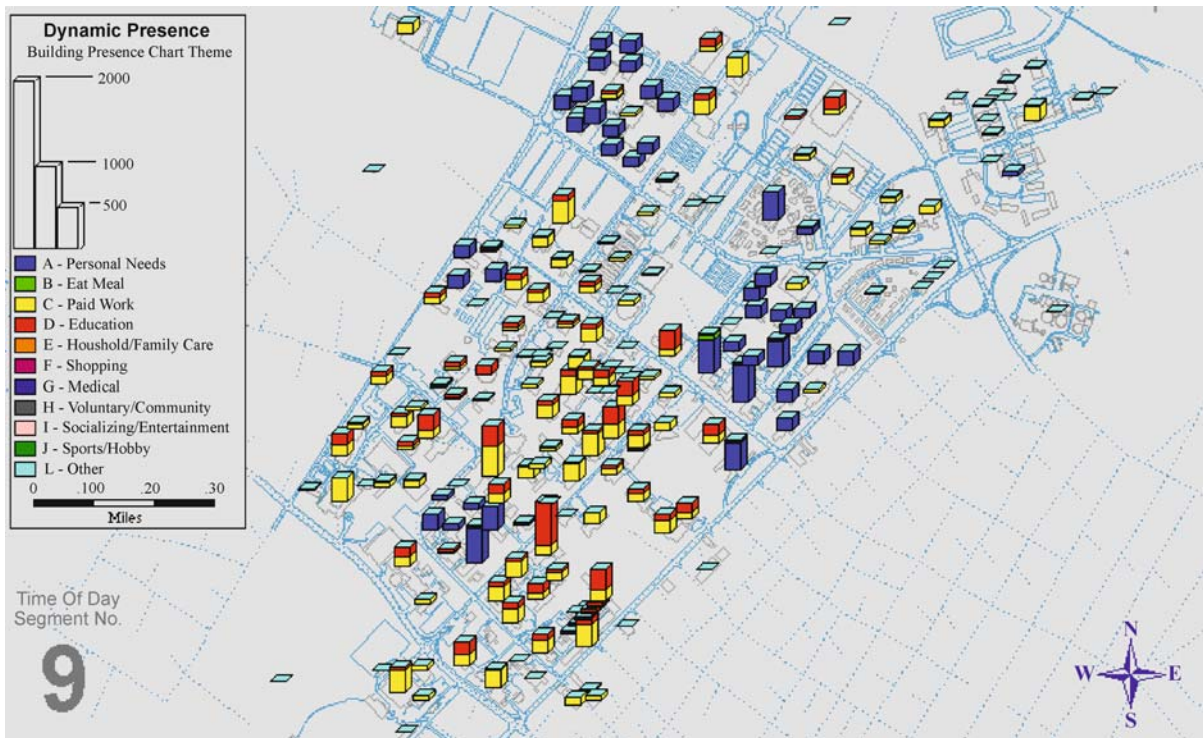


Travel Behavior and Demand Analysis and Prediction, Figure 2  
Time allocation to different activities in a day [2]. A: Personal Needs (includes sleep), B: Eat meal, C: Paid work, D: Education, E: Household and family care, F: shopping, G: medical, H: Volunteering/Community, I: Socializing, J: Sports and Hobbies, K: Travel, L: All other

behavioral dynamics has expanded the temporal horizons to a few years. However, regional simulation models are very often designed for long range plans spanning 25 years or even longer time horizons. Within these longer horizons, changes in the spatial distribution of activity locations and residences (land use) are substantial, changes in the demographic composition and spatial distribution of demographic segments are also substantial, and changes in travel patterns, transport facilities, and quality of service offered can be extreme. Past approaches in modeling and simulating the relationship among land use, demographics, and travel in a region attempted to disengage travel from the other two treating them as mutually exogenous. As interactions among them became more interesting and pressing, due to urban sprawl and suburban congestion, increasing attention was paid to their complex interdependencies. This led to a variety of attempts to develop “integrated model systems” that enable the study of scenarios of change and mutual influence between land use and travel. An earlier review of these models with

heavy emphasis on discrete choice models can be found in Anas [4]. Miller [117] and Waddell and Ulfarsson [180] twenty years later provide two comprehensive reviews of models that have integrated many aspects in the interdependent triad of demographics-travel-land use models. Both reviews trace the history of some of the most notable developments and both link these models to the activity-based approach above. Both reviews also agree that a microeconomic and/or macroeconomic approach to modeling land and transportation interactions are not sufficient and more detailed simulation of the individuals and their organizations “acting” in an time-space domain need to be simulated in order to obtain the required output for informed decision making. They also introduce the idea of simulating interactive agents in a dynamic environment of other agents (multi-agent simulation). The vast literature is reviewed by Timmermans [163] and Miller [118], from different viewpoints about progress made until now. However, they both agree that progress is rapidly made and that integration of land use and transportation models needs





Travel Behavior and Demand Analysis and Prediction, Figure 3  
Persons and activities assigned to buildings [2]

to move forward. Creation of integrated systems is further complicated by the emergence of an entire infrastructural system as another layer of human activity – telecommunication. Today telecommunication and transportation relationships are mostly absent from regional simulation planning and modeling as well from the most advanced land use and transportation integrated models. Considerable research findings, however, have been accumulating since the 1970s [53,66,81,96,111,113,121,128,129,144,182]. Another type of technologies (named enabling herein) helped us move modeling and simulation further.

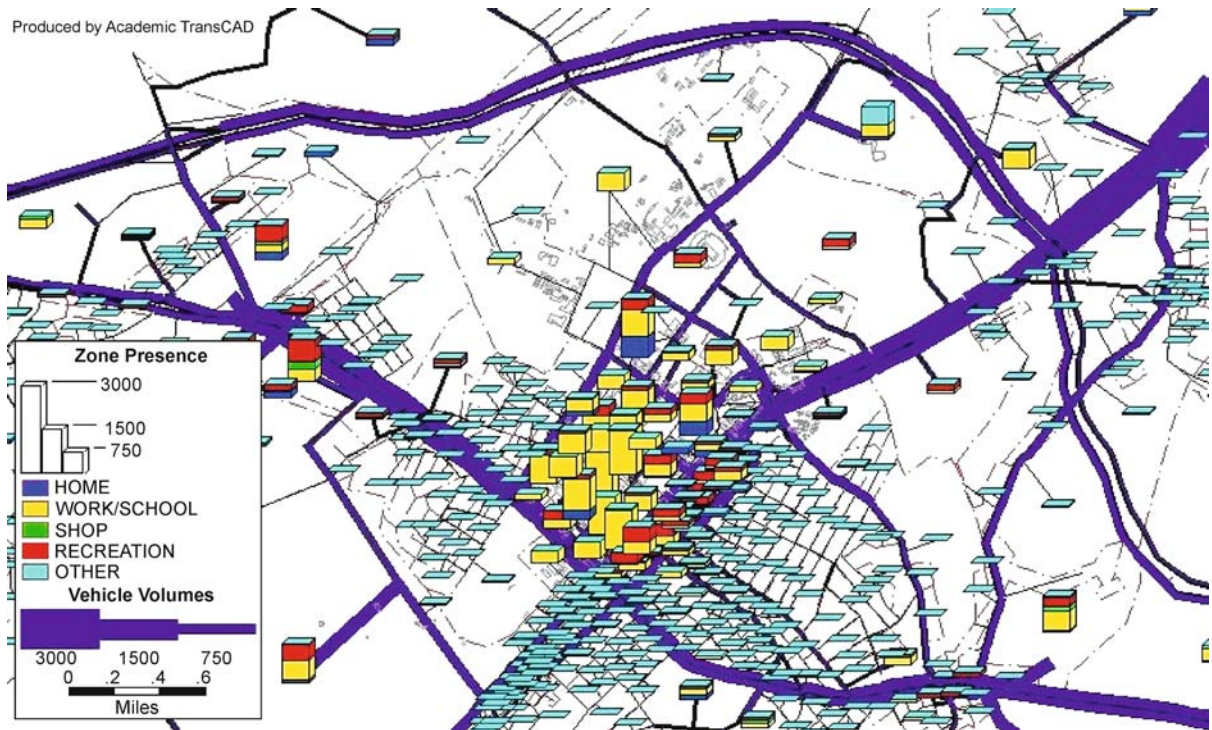
A few of the most important technologies are *stochastic simulation*, *production systems*, *geographic information systems*, *interactive and technology-aided data collection approaches*, and more *flexible data analysis techniques*.

*Stochastic microsimulation*, as intended here, is an evolutionary engine software that is used to replicate the relationships among social, economic, and demographic factors with land use, time use, and travel by people. As discussed above the causal links among these groups of entities are extremely complex, non-linear, and in many instances unknown or incompletely specified. This is the reason that no closed form solution can be created for such

a forecasting model system. An evolutionary engine, then, provides a realistic representation of person and household life histories (e.g., birth, death, marriages, divorces, birth of children, etc.), spatio-temporal activity opportunity evolution, and a variety of models that account for uncertainties in data, models, and behavioral variation (see [59,117], for overviews and [157] for an application).

*Production systems* were first developed by Newell and Simon [123] to explicitly depict the way humans go about solving problems. These are a series of condition-action (note the parallel with stimulus-response) statements in a sequence. From this viewpoint they are search processes that may never reach an absolute optimum and they replicate (or at least attempt to) human thought and action. Models of this kind are called *computational process models* (CPM) and through the use of IF ... THEN... rules have made possible the creation of a variety of new models.

*Geographic information systems* are software systems that can be used to collect, store, analyze, modify, and display large amounts of geographic data. They include layers of data that are able to incorporate relations among the variables in each layer and allow to build relationships in data across layers. One can visualize a GIS as a live map



Travel Behavior and Demand Analysis and Prediction, Figure 4  
Persons and activities assigned to buildings and travel to the network [67]

that can display almost any kind of spatio-temporal information. Maps have been used by transportation planners and engineers for long time and they are a natural interface to use in modeling and simulation. GIS today is moving beyond this relational database definition and is transforming the entire field into GI Science, which is beyond the scope of this article.

*Advanced data collection methods and devices* that are technologies that merit a note, although, not strictly developed for modeling. The first is about data collection and particularly data collection using internet technologies to build complex interviews that are interactive and dynamic [34]. In the same line of development we also see the use of geographic positioning systems (GPS) that allow one to develop a trace of individual paths in time and space [35,186]. Very important development is also the emergence of devices that can record the bulk of environmental data surrounding a person's movement, classify the environment in which the individual moves, and then ask simplified questions [74].

*Soft computing and non-parametric data analysis* is the last innovation mentioned here. In the data analysis we see greater strides in using data mining and artificial intelligence-born techniques to extract travel behavior pat-

terns [134,160] and advanced and less restrictive statistical methods to discover relationships in travel behavior data (e.g., [88]). Soft computing is increasingly finding many applications in activity-based models (see [www.imob.uhasselt.be](http://www.imob.uhasselt.be)). For a more recent and accessible review see Pribyl [133].

### The Evolving Modeling Paradigm

Policies are dictating to create and test increasingly more sophisticated policy assessment instruments that account for direct and indirect effects of behavior, procedures for behavioral change, and to provide finer resolution in the four dimensions of geographic space, time, social space, and jurisdictions. Dynamic planning is also stressing the need to examine trends, cycles, and the inversion of time progression to develop paths from the future visions to today's actions. New model developments are also becoming increasingly urgent. Although, tremendous progress has been observed in the past 20 years, development requires a faster pace to create new policy tools. These policy tools need to disentangle the actions of persons under different policy actions and the impact of policy actions on aggregates to identify conflicts and resolutions. Supporting all

this is a rich collection of decision paradigms that are already used and a few new ideas are starting to migrate to practice as illustrated below.

Early models incorporating activity-based behavioral processes into applications were published in the late 1970s and early 1980s as proof-of-concept and experimental applications. Following Hagerstrand's time-geography approach, PESASP [103] is one of the first models to operationally show the use of a time-space prism and to account for the relationship among activities. The Cullen and Godson [31] study was also the first comprehensive treatment of activities that brought different research findings together. In parallel, models were developed that were utility-maximizing models such as Adler and Ben-Akiva model [1] and much later the Kawakami and Isobe model [87]. Following these studies, BSP [79] and Computational Algorithms for Rescheduling Lists of Activities – CARLA [83] also use the activities within a time-space prism paradigm and define the foundations of data collection for activity-based approaches.

After this period of experimentation three streams in model development emerged. The first is in deriving representative activity patterns (RAPs) and then using regression techniques to correlate RAPs to person and household social and demographic data and then forecasting. The second development refines the methods used to simulate persons and adds to the forecasting repertoire other forecasting tasks via *microsimulation*. The third is a movement that expands the envelope to include cognition and explicit representation of mental processes through CPMs.

The Simulation of Travel/Activity Responses to Complex Household Interactive Logistic Decisions (STAR-CHILD – Recker and McNally [138,139] derived RAPs, employed a utility-based model and incorporated constraints. It is considered a fundamental transition development from research to practical application of an activity-based approach and it is still the foundation of models that first derive representative patterns and then forecast travel behavior. The more recent SIMAP [100] is a direct derivation of STARCHILD. In this line of development, Ma [109] created a model system that combined long term activity patterns (Long-term activity and travel planning – LATP) with a within-a-day activity scheduling and simulation (Daily Activity and Travel Scheduling – DATS) incorporating day-to-day variation and history dependence. Her model system produced very accurate forecasts. However it required panel survey data (the repeated observation of the same persons and households over time) that are rarely collected. In the LATP/DATS system longitudinal statistical models are extracted from longitudinal

records and they capture important aspects of behavioral dynamics such as habit persistence, day-to-day switching behaviors, and account for observed and unobserved heterogeneity contributed by the person, the household, the area of residence, and the area of workplace.

One of the first models to include a microsimulation in its paradigm is ORIENT [152]. This methodology suitably refined was demonstrated in a countrywide model for the Netherlands developed between 1989 and 1991 and named the Microanalytic Integrated Demographic Accounting System (MIDAS – Goulias and Kitamura [63,64]). MIDAS integrates demographic microsimulation, with dynamic car ownership models and a comprehensive suite of travel behavior equations. A cross-sectional version of MIDAS using data from the United States was also developed by Chung and Goulias [29]. MIDAS-USA simulates the evolution of households along with car ownership and travel behavior for Centre County, PA, and it is linked to a model to assign fees for development using GIS. A more ambitious development is the Activity Mobility Simulator – AMOS – by Kitamura et al. [93], which defines a few RAPs as templates. Then, uses a neural network to identify choices and a satisficing rule to simulate schedule changes due to policies. While MIDAS is a strictly longitudinal process econometric model progressing one year at a time, AMOS is constraint-based model designed for much finer temporal resolution. DEMOS, developed by Sundararajan and Goulias [157], is another MIDAS derivative. DEMOS is an object-oriented environment designed to simulate the evolution of people and their households using a variety of external data with the core models based on the Puget Sound Transportation Panel. It also simulates activity participation, travel, and telecommunication market penetration using a few representative patterns that were derived in Ma's LATP/DATS supplemented by telecommunications and travel behavior models.

SCHEDULER (Gärling et al. [43] is the first CPM that adds a psychometric cognitive implementation based on the Hayes-Roth and Hayes-Roth [73] planning model. In SCHEDULER, activities, selected from the long term calendar that represents a person's long term memory, comprise a schedule that is "mentally executed". Models start to combine CPM, microsimulation, and data derived behavioral patterns with random utility models to fill different modeling needs. The Simulation Model of Activity Scheduling Heuristics (SMASH – Ettema et al. [38]) is a CPM and econometric utility-based hybrid model that focuses on the pre-trip planning process predicting sequences of activities. In parallel, COMRADE [37], uses competing risk hazard models for activity scheduling and incorporates duration models in the system. The Model of



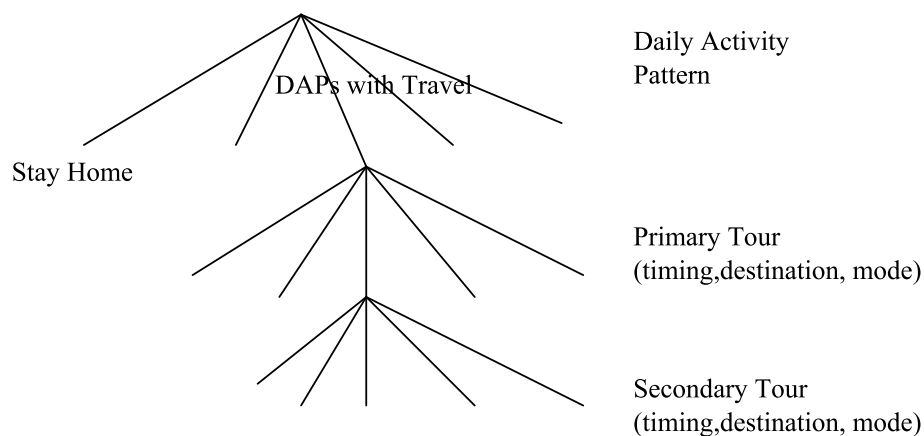
Action Space in Time Intervals and Clusters (MASTIC – Dijkstra and Vidakovic [32]), identifies clusters in the action space to perform and schedule activities. Time-space prisms are also the foundation of the Prism-Constrained Activity-Travel Simulator (PCATS – Kitamura [90], Kitamura and Fujii [92]), which is also a utility-based model. A direct operational derivative of SCHEDULER [44] was developed by Kwan, in her 1994 dissertation [101,102], and named GIS-Interfaced Computational-process modeling for Activity Scheduling (GISICAS). It is a simplified CPM, that uses time-space constraints and GIS to incorporate spatial information into a behavioral model to create individual schedules, starting with activities at higher levels of priority. Other models also attempt to recreate personal schedules such as Vaise's model [175], a CPM that creates a restricted choice set for creating activity patterns, a model by Ettema [39], and VISEM [41], a data-driven model that is a part of PTV Vision, an urban and regional transportation planning system, that creates daily activity patterns for behaviorally homogeneous groups within the population. Stopher et al. [156] also proposed the Simulation Model for Activity Resources and Travel (SMART) using a time geography framework and a taxonomy of activities in a GIS environment. All these use observed patterns to derive behavioral models. In contrast, Recker [137], developed the Household Activity Pattern Problem (HAPP) as a normative model based on the pick up and delivery time window problem to be used as a yardstick model testing optimal behavioral hypotheses.

The model framework that impacted practice the most in the United States is the Daily Activity Schedule model by Ben-Akiva et al. in [11]. This model, was used to create the Portland Daily Activity Schedule Model [23], advocat-

ing modeling lifestyle and mobility decisions on a scale of years. These influence daily activity schedules, which are comprised of primary and secondary tours constrained in time and space. It contains two key elements that simplify activity-based model development and takes advantage of the research surge in developing more general discrete choice models. A similar simplification using conditional probabilities was also developed for Los Angeles by Kitamura et al. [94].

Figure 5 shows this hierarchy of decisions and the scheme used to convert the daily pattern into a system of discrete choices. This framework was used to design new models for the regions around San Francisco, New York, Columbus, Denver, Atlanta, and Sacramento [24].

Arentze and Timmermans [5] designed the most complete CPM named ALBATROSS, which is a multi-agent simulation and predicts the time, location, duration, activity companionship, and travel modes subjecting everything to spatio-temporal, institutional, and household constraints. The theoretical underpinnings of this model are by far wider and all encompassing than any other activity-based model. However, it does not simulate route choice and does not produce data suitable for traffic assignment algorithms. Development of the third version of ALBATROSS is currently underway [76]. This model is also representative of raising the ambitions of travel modelers. The Alam Penn State Emergency Management model (Alam-PSEM, Alam and Goulias [3]) is a building-by-building simulation of activity participation and presence at specific locations of a university campus for each hour of a typical day. In parallel Bhat and his co-workers [15,18] developed the Comprehensive Activity-Travel Generation System for Workers (CATGW), which is a se-



Travel Behavior and Demand Analysis and Prediction, Figure 5  
The Bowman and Ben-Akiva daily activity model formulation [23]

ries of econometric models that replicate a commuter's evening mode choices, number of evening commute stops, and the number of stops after arriving home. The models developed by Bhat and colleagues are characterized by the use of hazard/duration regression models that were specifically developed for activity-based approaches and are by far more flexible than other regression methods. Another econometric model, the Conjoint-Based Model to Predict Regional Activity Patterns (COBRA), developed by Wang and Timmermans in [181], generates general patterns of stops for specific activities using a conjoint-based model with stated preference data instead of travel or activity diary data. The Wen and Koppelman model [183] utilizes three layers of decisions that are influenced by exogenous variables to generate activity patterns.

All these models point to new directions such as spatial choice needs to be dealt in more detail [3], activity choice and duration need to be dealt in a way that recognizes satiation in activity participation (e.g., in the duration models of Bhat [15]), sooner or later we will need to account for unobserved patterns and lack of experimental data (e.g., using conjoint experiments Wang and Timmermans [181]), and relations within the household need to also receive attention and be inserted in the model hierarchy [183].

Spatial aspects of model development were considered in the CentreSIM regional model [67,98,99] that uses time-of-day activity and travel data for different market segments to predict hour-by-hour presence at locations and travel among zones. In 2004, as a part of the Longitudinal Integrated Forecasting Environment (LIFE) framework [58], Pribyl and Goulias [135] developed CentreSIM (medoid simulation) to derive a few representative patterns and simulate daily schedules accounting explicitly for within-household interactions for entire daily patterns. In the Netherlands, PATRICIA (Predicting Activity-Travel Interdependencies with a Suite of Choice-Based, Interlinked Analyses), was developed by Borgers et al. [22] to help assess the performance of ALBATROSS. PATRICIA is a suite of linked models that incorporates an expanded set of activity choices, based on 63 distinct patterns, and activity destinations and describes activity transport modes and sequences. AURORA [82,165], which is a complementary model to ALBATROSS, is a utility-based system that models the dynamics of activity scheduling and rescheduling decisions as a function of many choice facets. AURORA is for short-term adaptation and rescheduling using just a few critical parameters. The model has since been expanded to include many new facets [76]. A much simpler model is PETRA [42] that allows the model to work with a small number of daily travel

patterns with some statistical advantages (see also Henson et al. [76]). Microsimulation software experienced another push forward by the development of a multi-million investment in TRANSPORTATION ANALYSIS SIMULATION SYSTEM. This model system was developed in the decade 1995–2005 and one of its versions is now available via a NASA open source license from TMIP at <http://tmip.fhwa.dot.gov/transims/>. TRANSIMS is a survey data-driven cellular automata microsimulation and was developed by a team at Los Alamos National Laboratory [106]. It was one of the first simulation packages to contain models that create a synthetic population, generate activity plans for individuals using directly observed data in travel surveys, formulate routes on a network based on these, and execute activity plans. Microsimulation models also evolved in the interface between land use and travel behavior. The Integrated Land Use, Transportation and Environment (ILUTE) model [145] model is designed to simulate the evolution of people and their activity patterns, transportation networks, houses, commercial buildings, the economy, and the job market over time. Within this vision, Miller and Roorda [119], developed the Toronto Area Scheduling model for Household Agents (TASHA) that uses *projects* to organize activity episodes into schedules of persons. Schedules for members in a household are simultaneously generated to allow for joint activities. Both ILUTE and TASHA utilize CPMs and econometric utility-based paradigms.

Another microsimulation that uses econometric models to simulate daily activity travel patterns for an individual, is the Comprehensive Econometric Microsimulator for Daily Activity-travel Patterns (CEMDAP) model [19] that is based on land use, socio-demographic, activity system, and level-of-service (LOS) attributes. Key distinctive element of CEMDAP is its reliance on hazard-based regression models to account for the continuous nature of activity duration. It includes population synthesis as well as the activity-pattern generation and scheduling of children, which is missing from many other simulators. Another model that utilizes constraints is the Florida Activity Mobility Simulator (FAMOS) [131]. FAMOS encompasses two modules, the Household Attributes Generation System (HAGS) and PCATS. Together, they comprise a system for modeling the activity patterns of individuals in Florida. The output is a series of activity-travel records. FAMOS is currently being further enhanced to include intra-household interactions and capture task allocation behavior among household members. Most recently, Ettema et al. [40] developed PUMA (Predicting Urbanization with Multi-Agents), a full-fledged multi-agent system of urban processes that represents land use changes



in a behaviorally realistic way. These processes include the evolution of population, businesses, and land use as well as daily activity and travel patterns of people. To simulate activity-travel patterns, an updated version of AURORA by Arentze et al. [6] will be created and also in the model FEATHERS (Forecasting Evolutionary Activity-Travel of Household and their Environmental Repercussions) to simulate activity-level scheduling decisions, within-a-day rescheduling, and learning processes in high resolutions of time and space. Developed as a complement to ALBATROSS, FEATHERS is econometric utility-based microsimulation that utilizes constraints that focuses on the short-term dynamics of activity-travel patterns. Members from this same Dutch team also developed MERLIN [173] and RAMBLAS [176].

Microsimulations have continued to gain in popularity in the activity-based modeling universe as they move from research applications to practice. Besides the Portland Daily Activity Schedule Model mentioned previously, New York's "Best Practice" Model (2002) and the Mid-Ohio Regional Planning Commission (MORPC) Model [179], both developed by Vovsha et al., and the San Francisco model [85] are currently being utilized by their respective MPO. The San Francisco model is currently being updated to implement enhanced destination choice models and being recalibrated using more recent household and census data. Four other models for Atlanta, Sacramento, the San Francisco Bay Area, and Denver are currently in various stages of implementation [24].

Although many past activity-based models have undefined or large time resolutions, STARCHILD already in mid-1980s used 15-min temporal resolution. The most recent models, however, go even further to simulate activities at small time intervals such as 5 min (TASHA) and 10 min intervals (SIMAP), minute by minute (MASTIC, CentreSIM, MASTIC, GISICAS, and RAMBLAS), and second-by-second (TRANSIMS-LANL, ALBATROSS, AURORA, CATGW, CEMDAP, FAMOS, and FEATHERS). Many applications, however, operate with large resolutions of one hour and they are implemented with a target of 30 min to one hour [24]. Spatial resolution of the models is still dominated by the zonal level. ALBATROSS and MORPC both can operate at the subzone level. Alam-PSEM, AURORA, CEMDAP, FEATHERS, GISICAS, ILUTE, PUMA, SIMAP, SMASH, and TRANSIMS-LANL utilize data at essentially the building or point level. Only two applications have spatial resolutions below the zonal level (Denver model that contains a two-stage destination locator to predict the address within a zone and the Sacramento model that operates at the parcel level). Cognitive theories (models of

knowledge and memory as well as behavioral process for planning activities) were used only in SCHEDULER and based on that in ALBATROSS and FEATHERS. Behavior is most often incorporated as intra-household interaction in ALBATROSS, CEMDAP, FAMOS, FEATHERS, ILUTE/TASHA, and CentreSIM as well as some of the applications in regions such as MORPC.

## Examples of Mathematical Models

In this section additional details of two examples of mathematical models for activity and travel behavior analysis are offered. Both examples aim at incorporating human interaction in time allocation models and they are multilevel regression models (based on Goulias [59]) and group decision making utility maximization models (based on Zhang et al. [187]).

### Multilevel Regression Models

These regression models are known by different names in different fields of research such as random coefficient models ([69] and p. 669 in [105]), multilevel models [48], mixed models [147], and hierarchical linear models [26]. They describe the contextual nature of the data and/or the way of accounting for dependent variable variation from multiple sources. Key advantages of these models are: explicit recognition in model formulation of the hierarchical, multiple level and nested structure of the data we analyze, and model specification using three groups of regression components in the same regression model. The first group assumes constant sensitivity to explanatory variables among the units of analysis representing the mean effect of an explanatory variable on the dependent variable. The second group assumes a random deviation around this mean and the third group is the usual random error term(s) of the regression equation. When compared to traditional regression models, which contain only one level, multilevel models do not underestimate the standard errors of coefficient estimates avoiding overstatements about the statistical significance of policy variables (e.g., we do not exaggerate the effect of taxation on car ownership or the effect of time and cost on route choice). A system of multilevel regression models can be written as follows.

$$Y_{tij}^q = \alpha_{tij}^q + \beta_k^q X_{tij} + \gamma_m^q Z_{tij} \quad (1)$$

$$a_{tij}^q = \gamma_0^q + v_j^q + u_{ij}^q + \varepsilon_{tij}^q, \quad \text{where } q = 1, \dots, Q, \quad (2)$$

$$\beta_{k1}^q = \gamma_{k1}^q + u_{k1ij}^q, \quad \beta_{k2}^q = \gamma_{k2}^q + v_{k2j}^q. \quad (3)$$

Equation (1), represents  $Q$  equations that are one for each  $Y_{tij}^q$  variable that we want to explain and use in travel

demand forecasting. They can be the amount of time dedicated to activities and travel or distances to specific destinations or even attributes of routes considered by trip makers. The index  $t$  represents the time at which an observation was made for a person  $i$  from within a household  $j$  (with  $t = 1, 2, 3, \dots, T$ ,  $i = 1, 2, \dots$ , number of people in household  $j$ ,  $j = 1, 2, \dots$ , number of households in sample). In this way we can identify change from one time point to another by an individual and study the relationships among individuals within social units (e.g., households, associations, neighborhoods and so forth).

The time points can be the same for all individuals or they may vary depending on the data collection procedures and willingness of respondents to provide information. Equation 1 is called the level 1 model because it is written at the level of the time point (observation occasion). The first term in the right hand side of Eq. (1) is a random intercept,  $\alpha$ , given by Eq. (2). This component has specific meaning. For example,  $\alpha_{tij}^q$  is the mean value of person  $i$  in household  $j$  at time  $t$  for variable  $q$ . The term  $\varepsilon_{tij}^q$  is a random temporal variation (also called within person variation) and it is the deviation of time expenditure around  $\gamma_0^q$ . The term  $u_{ij}^q$  is a random person to person variation (also called within household variation) and it is also a deviation of around  $\gamma_0^q$ . The term  $v_j^q$  is a random household to household variation and it is also a deviation around  $\gamma_0^q$ . These are also called random error components and they are usually assumed normally distributed with  $E(\varepsilon) = E(u) = E(v) = 0$ , with  $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$ ,  $\text{Var}(u) = \sigma_u^2$ , and  $\text{Var}(v) = \sigma_v^2$  to be estimated. It is worth noting that the system of equations represented by Eq. (1) contain a set of gamma coefficients (associated with a matrix  $Z$  representing explanatory variables) that are defined in a similar way as in typical regression models. The  $\beta$ s, however, that multiply the matrix  $X$  are defined as random with a mean and a variation around the means  $\gamma$ s. This variation can be due to the temporal, personal, and/or household levels. In this way, we can define a variety of equations at each of these levels to represent heterogeneous behavior that is either due to temporal fluctuations, personal variation, or household variation. Equation (3) differentiates between  $\beta$ s that vary within individuals and those that vary within households. In this way, at each level we have a level-specific variance-covariance matrix of all the random terms ( $\varepsilon$ s,  $u$ s,  $v$ s). The significance of the elements in each of these three matrices can be tested using goodness-of-fit measures based on the deviance, which is the difference in the  $-2\text{Log}(\text{likelihood})$  at convergence between two nested (in terms of specification) models. In addition, the  $\gamma$ s can also be tested if they are significantly different than zero using a  $t$ -test. The  $\gamma$ s in Eq. (1) are

called the *fixed effects* and the remaining terms are called the *random effects* at each of the three levels in the hierarchy. Estimation of all the fixed and random parameters can be accomplished either by Full Information Maximum Likelihood, FIML, applied to  $Y$  directly or applied to the least-squares residuals, called Restricted Maximum Likelihood-REML that can be used in tandem with a generalized least squares approach. Longford [105], Bryk and Raudensbush [26] and [48] provide a comprehensive review of estimation techniques, their performance assessment, and detailed algorithms.

### Household Utility Models

The second example is also representative of a movement toward more detailed consideration of within household decision making dynamics. Although the model was specified by Zhang et al. [187] for time allocation to shared ( $j$ ) and non-shared activities ( $s$ ), it is a potentially useful model for other trip making decisions. Each person in a household is assumed to form two utility functions. One utility is for the shared activities (i.e., engagement in activities with other household members) and non-shared activities. These utility functions are given by Eq. (4) (shared activity) and 5 (non-shared activity).

$$u_{is} = \exp \left( \left( \alpha_s + \sum_k \beta_{sk} x_{isk} \right) \ln \left( \sum_m \kappa_{sm} \tau_{ism} \right) + \varepsilon_{is} \right) \ln(t_{is}) \quad (4)$$

$$u_{ij} = \exp \left( \left( \alpha_j + \sum_k \beta_{jk} x_{ijk} \right) \ln \left( \sum_m \kappa_{jm} \tau_{ijm} \right) + \varepsilon_{ij} \right) \ln(t_{ij}), \quad (5)$$

where:

- $\alpha_s$  is the constant term for each shared activity  $s$ .
- $x_{isk}$  is the  $k$ th explanatory variable (and/or attribute) of household member  $i$  for shared activity  $s$ .
- $\beta_{sk}$  is the parameter associated with the  $k$ th attribute of the shared activity.
- $\tau_{ism}$  is the travel time by mode  $m$  for each activity  $s$  by person  $i$ .
- $\kappa_{sm}$  is the parameter associated with travel time by mode  $m$ .
- $\varepsilon_{is}$  is a random error term of the shared activity  $s$  by person  $i$ .
- $t_{is}$  is the amount of time dedicated to activity  $s$  by person  $i$ .

- $\alpha_j$  is the constant term for each non-shared activity.  
 $x_{ijk}$  is the  $k$ th explanatory variable (attribute) of household member  $i$  for non-shared activity  $j$ .  
 $\beta_{jk}$  is the parameter associated with the  $k$ th attribute of non-shared activity.  
 $\tau_{ijm}$  is the travel time by mode  $m$  for each activity  $s$  by person  $i$ .  
 $\kappa_{jm}$  is the parameter associated with travel time by mode  $m$ .  
 $\varepsilon_{ij}$  is a random error term of the non-shared activity  $j$  by person  $i$ .  
 $t_{ij}$  is the amount of time dedicated to activity  $j$  by person  $i$ .

The overall utility of activity participation and travel for each person  $i$  under the assumption of a multi-linear utility is given by Eq. (6).

$$u_i = \sum_{j=1}^{J+S} r_{ij} u_{ij} + \sum_{j=1}^{J+S} \sum_{j' > j} \delta_i r_{ij} r_{ij'} u_{ij} u_{ij'} \quad (6)$$

where,

- $u_{ij}$  is the utility of activity  $j$  for person  $i$ .  
 $r_{ij}$  is the relative interest of person  $i$  for activity  $j$ .  
 $\delta_i$  is parameter of activity dependency for member  $i$ .  
 $J + S$  is the number of non-shared and shared activities for a person within the unit of time under consideration.

In a similar way the household utility function is a multi-linear combination of the individual utilities in Eq. (7).

$$\text{HUF} = \sum_{i=1}^n w_i u_i + \lambda \sum_{i=1}^n \sum_{i' > i} (w_i w_{i'} u_i u_{i'}) \quad (7)$$

where,

- HUF is the household utility combining the utilities of all household members  $n$ .  
 $u_i$  is the utility of household member  $i$ .  
 $w_i$  is the relative influence of each household member  $i$ .  
 $\lambda$  is a parameter of within household interaction.

Under the assumption of maximizing HUF it is possible to create a Lagrangian function that accounts for constraints (i.e., total amount of time available, signs of parameters and so forth) and through a maximization solution derive equations that can be used to estimate the unknown parameters in Eqs. (4–7) (details are provided in Zhang et al. [187] for time allocation). It is worth noting that Zhang et al. [187], derived two alternate model systems by changing the utility functions to represent different intra-household bargaining models (for a detailed review see [13]). Then through a linearization process

they developed a system of linear equations and estimated the parameters using a multiple equations econometric approach (the Seemingly Unrelated Regression Estimation [69]) that is a simplifying alternative to the multilevel models described earlier in this section. A more general review of this type of model formulation is also provided by Timmermans [164].

## Summary

Similarities and differences among the implemented modeling ideas are:

- A hierarchy of decisions by households is assumed that identifies longer term choices determining the shorter term choices. In this way different blocks of variables can be identified and their mutual correlation used to derive equations that are used in forecasting.
- Anchor points (Home location – work location – school location) are inserted in the first choice level and they define the overall spatial structure of activity scheduling.
- Out-of-home activity purposes include work, school, shopping, meals, personal business, recreation, and escort. These expand the original home-based and non-home based purposes in travel behavior and the three activity types in home economics (labor for pay, labor at home, and leisure).
- In-home activities are explicitly modeled or allowed to enter the model structure as a “stay-at-home” choice with some models allowing for activity choice at home (work, maintenance and discretionary). In this way limited substitution between at home and outside home can be reflected in the models.
- Stop frequencies and activities at stops are modeled at the day pattern and tour levels to distinguish between activities and trips that can be rescheduled with little additional efforts versus the activities and trips that cannot be rescheduled (e.g., school trips).
- Modes and destinations are modeled together. In this way the mutual influence – sequential and/or simultaneous relationships can be reflected in the model structure.
- Time is included in a few instances in activity-based models. For example departure time for trips and tour time of day choice are modeled explicitly. Model time periods are anywhere between 30 min and second-by-second and time windows are used to account for scheduling. This modeling component allows to incorporate time-of-day in the modeling suites. It also allows to identify windows of activity and travel opportunities. The presence of departure time also enables models to

trip matrices for any desired periods in a day. In fact, output of time periods depends on route choice and traffic assignment needs and can be adjusted almost at will.

- Human interaction, although limited for now to the within-household interaction, is incorporated by relating the day pattern of one person to the day patterns of other persons within a household, their joint activities and trip making are explicitly modeled (joint recreation, escort trips), and allocation of activity-roles are also modeled.
- Spatial aspects of a region are accounted for using methods that produce spatially distributed synthetic populations using as external control totals averages and relative frequencies of population characteristics.
- Accessibility measures are used to capture spatial interaction among activity locations and the level of service offered by the transportation systems. These are also the indicators used to account for feedback among the lower level in the hierarchy decisions (e. g., activity location choices, routes followed, congestion) and the higher level such as residence location choice.
- Spatial resolution is heavily dependent on data availability and it reached already the level of a parcel and/or building at its most disaggregate level. Outputs of models are then aggregated to whatever level is required by traffic assignment, mode specific studies (nonmotorized and/or transit) and reporting needs and requirements.

Overall, the plethora of advances includes: a) models and experiments to create computerized virtual worlds and synthetic schedules at the most elementary level of decision making using microsimulation and computational process models; b) data collection methods and new methods to collect extreme details about behavior and to estimate, validate, and verify models using advanced hardware, software, and data analysis techniques; and c) integration of models from different domains to reflect additional interdependencies such as land use and telecommunications.

### Future Directions

Much more work remains to be done in order to develop models that can answer more complex questions in policy analysis and for this reason a few steps are outlined here. In policy and program evaluation, transportation analysis appears to be narrowly applied to only one method of assessment that does not follow the ideal of a randomized controlled trial and does not explicitly define what experimental setting we are using for our assessments. Unfor-

tunately this weakens our findings about policy analysis and planning activities. Although we have many laboratory experiments that were done for intelligent transportation systems we lack studies and guidelines to develop experimental and quasi-experimental procedures to guide us in policy development and large scale data collection.

In addition, many issues remain unresolved in the areas of coordination among scale in time and space and related issues. In addition very little is known about model sensitivity and data error tolerance and their mapping to strategy evaluations. This is partially due to the lack of tools that are able to make these assessments but also due to lack of scrutiny of these issues and their implications on impact assessment.

Regarding strategic planning and evaluation, we also lack models designed to be used in scenario building exercises such as backcasting and related assessments. The models about change are usually defined for forecasting and simple time inversion may not work to make them usable in backcasting. This area does not have the long tradition of modeling and simulation to help us develop suitable models. Should more attention be paid to this aspect? Is there room for a combination of techniques including qualitative research methods? What is the interface between this aspect and the experimental methods questions in program evaluation?

In the new research and technology area, since we are dealing with the behavior of persons, it is unavoidable to consider perceptions of time and space. What role should perceptions of time and space [51] play in behavioral models and what is the most appropriate use of these perceptions? The multiple dimensions of time such as tempo, duration, and clock time are neglected in behavioral models – is there a role for them in behavioral models?

Human interaction is considered important and is receiving attention in more recent research Golob and McNally [54], Chandrasekharan and Goulias [27], Simma and Axhausen [148], Gliebe and Koppelman [47], Goulias and Kim [62], Zhang et al. [187], but only partially accounted for in applications as illustrated by Vovsha and Petersen [177]. Future applications will increasingly pay attention to motivations for human interactions and the nature of these interactions within households and in a wider social network context.

### Bibliography

1. Adler T, Ben-Akiva M (1979) A theoretical and empirical model of trip chaining behavior. *Transp Res B* 13:243–257
2. Alam BS (1998) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. MS Thesis. Department of Civil and Environmental

- Engineering, The Pennsylvania State University, University Park
3. Alam BS, KG Goulias (1999) Dynamic emergency evacuation management system using GIS and spatio-temporal models of behavior. *Transp Res Record* 1660:92–99
  4. Anas A (1982) Residential location markets and urban transportation: Economic theory, econometrics and policy analysis with discrete choice models. Academic Press, New York
  5. Arentze T, Timmermans H (2000) ALBATROSS – A learning based transportation oriented simulation system. European Institute of Retailing and Services Studies (EIRASS), Technical University of Eindhoven, Eindhoven
  6. Arentze T, Timmermans H, Janssens D, Wets G (2006) Modeling short-term dynamics in activity-travel patterns: From aurora to feathers. Presented at the Innovations in Travel Modeling Conference, Austin 21–23 May 2006
  7. Avineri E, Prashker Y (2003) Sensitivity to uncertainty: The need for a paradigm shift. CD-TRB ROM Proceedings, Paper presented at the 82nd Annual Transportation Research Board Meeting, 12–16 January 2003, Washington DC
  8. Becker GS (1976) The economic approach to human behavior. The University of Chicago Press, Chicago
  9. Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: Theory and application to travel demand. MIT Press, Cambridge
  10. Ben-Akiva ME, Morikawa T (1989) Estimation of mode switching models from revealed preferences and stated intentions. Paper presented at the International Conference on Dynamic Travel Behavior at Kyoto University Hall, Kyoto
  11. Ben-Akiva M, Bowman JL, Gopinath D (1996) Travel demand model system for the information era. *Transportation* 23: 241–266
  12. Ben-Akiva ME, Walker J, Bernardino AT, Gopinath DA, Morikawa T, Polydoropoulou A (2002) Integration of choice and latent variable models. In: Mahmassani HS (ed) *In perceptual motion: Travel behavior research opportunities and application challenges*. Pergamon, Amsterdam
  13. Bergstrom TC (1995) A survey of theories of the family. Department of Economics, University of California Santa Barbara, Paper 1995D. <http://repositories.cdlib.org/ucsbecon/bergstrom/1995D/>
  14. Bhat CR (2000) Flexible model structures for discrete choice analysis. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 71–89
  15. Bhat C (2001) A comprehensive and operational analysis framework for generating the daily activity-travel pattern of workers. Paper presented at the 78th Annual Meeting of the Transportation Research Board, Washington DC, 10–14 January 2001
  16. Bhat CR (2003) Random utility-based discrete choice models for travel demand analysis. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 10-1–10-30
  17. Bhat CR, Koppelman F (1999) A retrospective and prospective survey of time-use research. *Transportation* 26(2):119–139
  18. Bhat CR, Singh SK (2000) A comprehensive daily activity-travel generation model system for workers. *Transp Res A* 34(1):1–22
  19. Bhat CR, Guo J, Srinivasan S, Sivakumar A (2003) Activity-based travel demand modeling for metropolitan areas in Texas: Software-related processes and mechanisms for the activity-travel pattern generation microsimulator. Research Report 4080-5, Center for Transportation Research, Austin
  20. Bockenholt U (2002) Comparison and choice: Analyzing discrete preference data by latent class scaling models. In: Hagenaars JA, McCutcheon AL (eds) *Applied latent class analysis*. Cambridge University Press, Cambridge, pp 163–182
  21. Borgers AWJ, Hofman F, Timmermans HJP (1997) Activity-based modelling: Prospects. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Pergamon, Oxford, pp 339–351
  22. Borgers AWJ, Timmermans AH, van der Waerden P (2002) Patricia: Predicting activity-travel interdependencies with a suite of choice-based, interlinked analysis. *Transp Res Rec* 1807:145–153
  23. Bowman JL, Bradley M, Shifan Y, Lawton TK, Ben-Akiva M (1998) Demonstration of an activity-based model system for Portland. Paper presented at the 8th World Conference on Transport Research, Antwerp, June 1998
  24. Bradley M, Bowman J (2006) A summary of design features of activity-based microsimulation models for US MPOs. Conference on Innovations in Travel Demand Modeling, Austin 21–23 May 2006
  25. Brög W, Erl E (1980) Interactive measurement methods – Theoretical bases and practical applications. *Transp Res Rec* 765:1–6
  26. Bryk AS, Raudenbush SW (1992) *Hierarchical linear models*. Sage, Newberry Park
  27. Chandrasekharan B, Goulias KG (1999) Exploratory longitudinal analysis of solo and joint trip making in the Puget Sound transportation panel. *Transp Res Rec* 1676:77–85
  28. Chapin Jr FS (1974) Human activity patterns in the city: Things people do in time and space. Wiley, New York
  29. Chung J, Goulias KG (1997) Travel demand forecasting using microsimulation: Initial results from a case study in Pennsylvania. *Transp Res Rec* 1607:24–30
  30. Creighton RL (1970) *Urban transportation planning*. University of Illinois Press, Urbana
  31. Cullen I, Godson V (1975) Urban networks: The structure of activity patterns. *Progr Plan* 4(1):1–96
  32. Dijst M, Vidakovic V (1997) Individual action space in the city. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science Inc, New York, pp 117–134
  33. Dillman DA (2000) *Mail and internet surveys: The tailored design method*, 2nd edn. Wiley, New York
  34. Doherty S (2003) Interactive methods for activity scheduling processes. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 7-1 to 7-25
  35. Doherty ST, Noel N, Lee M-G, Sirois C, Ueno M (2001) Moving beyond observed outcomes: Global positioning systems and interactive computer-based travel behavior surveys. *Transportation Research Circular, E-C026*, March 2001, Transportation Research Board, Washington DC
  36. Ettema DF, Timmermans HJP (1997) *Activity-based approaches to travel analysis*. Elsevier Science Inc, New York, p xiii
  37. Ettema DF, Borgers AWJ, Timmermans HJP (1995) Competing risk hazard model of activity choice, timing, sequencing and duration. *Transp Res Rec* 1439:101–109



38. Ettema DF, Borgers AWJ, Timmermans H (1996) SMASH (Simulation Model of Activity Scheduling Heuristics): Some simulations. *Transp Res Rec* 1551:88–94
39. Ettema DF, Daly A, de Jong G, Kroes E (1997) Towards an applied activity-based travel demand model. Paper presented at the IATBR Conference, Austin 21–25 September 1997
40. Ettema DF, de Jong K, Timmermans H, Bakema A (2006) PUMA: Multi-agent modeling of urban systems. 2006 Transportation Research Board CD-ROM
41. Fellendorf M, Haupt T, Heidl U, Scherr W (1997) PTV vision: Activity based demand forecasting in daily practice. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science Inc, New York, pp 55–72
42. Fosgerau M (2002) PETRA – an activity-based approach to travel demand analysis. In: Lundquist L, Mattsson L-G (eds) *National transport models: Recent developments and prospects*. Royal Institute of Technology, Stockholm, Sweden. Springer, Berlin
43. Gärling T, Brannas K, Garvill J, Golledge RG, Gopal S, Holm E, Lindberg E (1989) Household activity scheduling. In: *Transport policy, management and technology towards 2001. Selected Proceedings of the Fifth World Conference on Transport Research*, vol 4. Western Periodicals, Ventura, pp 235–248
44. Gärling T, Kwan M, Golledge R (1994) Computational-process modeling of household travel activity scheduling. *Transp Res Part B* 25:355–364
45. Gärling T, Laitila T, Westin K (1998) Theoretical foundations of travel choice modeling: An introduction. In: Gärling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 1–30
46. Garrett M, Wachs M (1996) *Transportation planning on Trial. The clean air act and travel forecasting*. Sage Publications, Thousand Oaks
47. Gliebe JP, Koppelman FS (2002) A model of joint activity participation. *Transportation* 29:49–72
48. Goldstein H (1995) *Multilevel statistical models*. Edward Arnold, London, New York
49. Golledge RG, Stimson RJ (1997) *Spatial behavior: A geographic perspective*. The Guilford Press, New York
50. Golledge RG, Gärling T (2003) Spatial behavior in transportation modeling and planning. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 1–27
51. Golledge RG, Gärling T (2004) Cognitive maps and urban travel. In: Hensher D, Button K, Haynes K, Stopher P (eds) *Handbook of transport geography and spatial systems*, vol 5. Elsevier, Amsterdam, pp 501–512
52. Golledge RG, Smith TR, Pellegrino JW, Doherty S, Marshall SP (1985) A conceptual model and empirical analysis of children's acquisition of spatial knowledge. *J Environ Psychol* 5(2):125–152
53. Golob TF (2001) Travelbehaviour.com: Activity approaches to modeling the effects of information technology on personal travel behaviour, in travel behavior research. In: Hensher D (ed) *The leading edge*. Elsevier Science/Pergamon, Kidlington, Oxford, pp 145–184
54. Golob TF, McNally M (1997) A model of household interactions in activity participation and the derived demand for travel. *Transp Res B* 31:177–194
55. Golob TF, Kitamura R, Long L (eds) (1997) *Panels for transportation planning: Methods and applications*. Kluwer, Academic Publishers, Massachusetts
56. Goodman LA (2002) Latent class analysis: The empirical study of latent types, latent variables, and latent structures. In: Hagenaars JA, McCutcheon AL (eds) *Applied latent class analysis*. Cambridge University Press, Cambridge, pp 3–55
57. Goulias KG (1999) Longitudinal analysis of activity and travel pattern dynamics using generalized mixed Markov latent class models. *Transp Res B* 33:535–557
58. Goulias KG (2001) A Longitudinal integrated forecasting environment (LIFE) for activity and travel forecasting. In: Villacampa Y, Brebbia CA, Uso J-L (eds) *Ecosystems and sustainable development III*. WIT Press, Southampton, pp 811–820
59. Goulias KG (2002) Multilevel analysis of daily time use and time allocation to activity types accounting for complex covariance structures using correlated random effects. *Transportation* 29(1):31–48
60. Goulias KG (2003) Transportation systems planning. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 1–1 to 1–45
61. Goulias KG, Kim T (2003) A longitudinal analysis of the relationship between environmentally friendly modes, weather conditions, and information-telecommunications technology market penetration. In: Tiezzi E, Brebbia CA, Uso JL (eds) *Ecosystems and sustainable development*, vol 2. WIT Press, pp 949–958
62. Goulias KG, Kim T (2005) An analysis of activity type classification and issues related to the with whom and for whom questions of an activity diary. In: Timmermans H (ed) *Progress in activity-based analysis*. Elsevier, pp 309–334
63. Goulias KG, Kitamura R (1992) Travel demand analysis with dynamic microsimulation. *Transp Res Rec* 1607:8–18
64. Goulias KG, Kitamura R (1997) Regional travel demand forecasting with dynamic microsimulation models. In: Golob T, Kitamura R, Long L (eds) *Panels for transportation planning: Methods and applications*. Kluwer, Academic Publishers, Massachusetts, pp 321–348
65. Goulias KG, Litzinger T, Nelson J, Chalamgari V (1993) A study of emission control strategies for Pennsylvania: Emission reductions from mobile Sources, cost effectiveness, and economic impacts. Final report to the Low Emissions Vehicle Commission. PTI 9403. The Pennsylvania Transportation Institute, University Park
66. Goulias KG, Kim T, Pribyl O (2003) A longitudinal analysis of awareness and use for advanced traveler information systems. Paper to be presented at the European Commission Workshop on Behavioural Responses to ITS – 1–3 April 2003, Eindhoven
67. Goulias KG, Zekkos M, Eom J (2004) CentreSIM3 Scenarios for the South Central Centre County Transportation Study. CentreSIM3 Report submitted to McCormick Taylor Associates and the Mid-Atlantic Universities Transportation Center, April 2004, University Park
68. Goulias KG, Blain L, Kilgren N, Michalowski T, Murakami E (2007) Catching the next big wave: Are the observed behavioral dynamics of the baby boomers forcing us to rethink regional travel demand models? Paper presented at the 86th Transportation Research Board Annual Meeting, 21–25 January 2007, Washington DC and included in the CD ROM proceedings

69. Greene WH (1997) *Econometric analysis*, 3rd edn. Prentice Hall, New Jersey
70. Grieving S, Kemper R (1999) Integration of transport and land use policies: State of the art. Deliverable 2b of the Project TRANSLAND, 4th RTD Framework Programme of the European Commission
71. Haab TC, Hicks RL (1997) Accounting for choice set endogeneity in random utility models of recreation demand. *J Environ Econ Manag* 34:127–147
72. Hagerstrand T (1970) What about people in regional science? *Pap Reg Sci Assoc* 10:7–21
73. Hayes-Roth B, Hayes-Roth F (1979) A cognitive model of planning. *Cogn Sci* 3:275–310
74. Hato E (2006) Development of behavioral context addressable loggers in the shell for travel activity analysis. Paper presented at the IATBR conference, Kyoto
75. Henson K, Goulias KG (2006) Preliminary assessment of activity and modeling for homeland security applications. *Transportation Research Record: J Transportation Research Board*, No. 1942, Transportation Research Board of the national Academies, Washington DC, pp 23–30
76. Henson K, Goulias KG, Golledge R (2006) An assessment of activity-based modeling and simulation for applications in operational studies, disaster preparedness, and homeland security. Paper presented at the IATBR conference, Kyoto
77. Horowitz JL (1991) Modeling the choice of choice set in discrete-choice random-utility models. *Environ Plan A* 23: 1237–1246
78. Horowitz JL, Louviere JJ (1995) What is the role of consideration sets in choice modeling? *Int J Res Marketing* 12:39–54
79. Huigen PPP (1986) Binnen of buiten bereik?: Een sociaal-geografisch onderzoek in Zuidwest-Friesland, *Nederlandse Geografische Studies* 7, University of Utrecht, Utrecht
80. Hutchinson BG (1974) *Principles of urban transport systems planning*. Scripta, Washington DC
81. JHK & Associates, Clough, Harbour & Associates, Pennsylvania Transportation Institute, Bogart Engineering (1996) *Scranton/Wilkes-barre area strategic deployment plan*. Final Report. Prepared for Pennsylvania Department of Transportation District 4-0. August 1996, Berlin
82. Joh C-H, Arentze T, Timmermans H (2004) Activity-travel scheduling and rescheduling decision processes: Empirical estimation of aurora model. *Transp Res Rec* 1898:10–18
83. Jones PM, Dix MC, Clarke MI, Heggie IG (1983) Understanding travel behaviour. Gower, Aldershot
84. Jones P, Koppelman F, Orfeuil J (1990) Activity analysis: State-of-the-art and future directions. In: Jones P (ed) *Developments in dynamic and activity-based approaches to travel analysis*. A compendium of papers from the 1989 Oxford Conference. Avebury, Gower-Aldershot, pp 34–55
85. Jonnalagadda N, Freedman J, Davidson WA, Hunt JD (2001) Development of microsimulation activity-based model for San Francisco. *Transp Res Rec* 1777:25–35
86. Kahneman D, Tversky A (1979) Prospect theory: An analysis of decisions under risk. *Econometrica* 47(2):263–291
87. Kawakami S, Isobe T (1989) Development of a travel-activity scheduling model considering time constraint and temporal transferability test of the model. In: *Transport policy, management and technology towards 2001: Selected Proceedings of the Fifth World Conference on Transport Research*, vol 4. Western Periodicals, Ventura, pp 221–233
88. Kharoufeh JP, Goulias KG (2002) Nonparametric identification of daily activity durations using Kernel density estimators. *Transp Res B Methodological* 36:59–82
89. Kitamura R (1988) An evaluation of activity-based travel analysis. *Transportation* 15:9–34
90. Kitamura R (1997) Applications of models of activity behavior for activity based demand forecasting. In: Engelke LJ (ed) *Activity-based travel forecasting conference: Summary, recommendations and compendium of papers*. Report of the Travel Model Improvement Program. Texas Transportation Institute, Arlington, pp 119–150
91. Kitamura R (2000) Longitudinal methods. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 113–128
92. Kitamura R, Fujii S (1998) Two computational process models of activity-travel choice. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 251–279
93. Kitamura R, Pas EI, Lula CV, Lawton TK, Benson PE (1996) The sequenced activity simulator (SAMS): an integrated approach to modeling transportation, land use and air quality. *Transportation* 23:267–291
94. Kitamura R, Chen C, Pendyala RM (1997) Generation of synthetic daily activity-travel patterns. *Transp Res Rec* 1607: 154–162
95. Koppelman FS, Sethi V (2000) Closed-form discrete-choice models. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 211–225
96. Krizek KJ, Johnson A (2003) Mapping of the terrain of information and communications technology (ICT) and household travel. *Transportation Research Board annual meeting CD-ROM*, Washington DC, January 2003
97. Kuhnau JL (2001) Activity-based travel demand modeling using spatial and temporal models in the urban transportation planning system. MS Thesis. Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
98. Kuhnau JL, Goulias KG (2002) Centre SIM: Hour-by-hour travel demand forecasting for mobile source emission estimation. In: Brebbia CA, Zannetti P (eds) *Development and application of computer techniques to environmental studies IX*. WIT Press, Southampton, pp 257–266
99. Kuhnau JL, Goulias KG (2003) Centre SIM: First-generation model design, pragmatic implementation, and scenarios. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 16-1–16-14
100. Kulkarni A, McNally MG (2001) A microsimulation of daily activity patterns. Paper presented at the 80th Annual Meeting of the Transportation Research Board, Washington, 7–11 January 2001
101. Kwan M-P (1994) A GIS-based model for activity scheduling in intelligent vehicle highway systems (IVHS). Unpublished Ph D, Department of Geography, University of California Santa Barbara, Santa Barbara
102. Kwan M-P (1997) GISICAS: An activity-based travel decision support system using a GIS-interfaced computational-process model. In: Ettema DF, Timmermans HJP (eds) *Activity-based approaches to travel analysis*. Elsevier Science Inc, New York, pp 263–282
103. Lenntorp B (1976) Paths in space-time environment: A time geographic study of possibilities of individuals. *The Royal*

- University of Lund, Department of Geography. Lund Studies in Geography, Series B Human Geography 44
104. Lomborg B (2001) *The skeptical environmentalist: Measuring the real state of the world*. Cambridge University Press, Cambridge
  105. Longford NT (1993) *Random coefficient models*. Clarendon Press, Oxford
  106. Los Alamos National Laboratory (2003) TRANSIMS: Transportation analysis system (Version 3.1). LA-UR-00-1725
  107. Loudon WR, Dagang DA (1994) Evaluating the effects of transportation control measures. In: Wholley TF (ed) *Transportation planning and air quality II*. American Society of Civil Engineers, New York
  108. Louviere JJ, Hensher DA, Swait JD (2000) *Stated choice methods: Analysis and application*. Cambridge University Press, Cambridge
  109. Ma J (1997) *An activity-based and micro-simulated travel forecasting system: A pragmatic synthetic scheduling approach*. Unpublished Ph D Dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
  110. Mahmassani HS, Herman R (1990) Interactive experiments for the study of tripmaker behaviour dynamics in congested commuting systems. In: *Developments in dynamic and activity-based approaches to travel analysis. A compendium of papers from the 1989 Oxford Conference*. Avebury
  111. Mahmassani HS, Jou R-C (1998) Bounded rationality in commuter decision dynamics: Incorporating trip chaining in departure time and route switching decisions. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam
  112. Manheim ML (1979) *Fundamentals of transportation systems analysis, vol 1: Basic Concepts*. MIT Press, Cambridge
  113. Marker JT, Goulias KG (2000) Framework for the analysis of grocery teleshopping. *Transp Res Rec* 1725:1–8
  114. McNally MG (2000) The activity-based approach. In: Hensher DA, Button KJ (eds) *Handbook of transport modelling*. Pergamon, Amsterdam, pp 113–128
  115. McFadden D (1998) Measuring willingness-to-pay for transportation improvements. In: Garling T, Laitila T, Westin K (eds) *Theoretical foundations of travel choice modeling*. Pergamon, Elsevier, Amsterdam, pp 339–364
  116. Meyer MD, Miller EJ (2001) *Urban transportation planning*, 2nd edn. McGrawHill, Boston
  117. Miller EJ (2003) Land use: Transportation modeling. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 5–1 to 5–24
  118. Miller EJ (2006) Resource paper on integrated land use-transportation models. IATBR, Kyoto, 2006
  119. Miller EJ, Roorda MJ (2003) A prototype model of household activity/travel scheduling. *Transp Res Rec* 1831:114–121
  120. Miller JS, Demetsky MJ (1999) Reversing the direction of transportation planning process. *ASCE J Transp Eng* 125(3):231–237
  121. Mokhtarian PL (1990) A typology of relationships between telecommunications and transportation. *Transp Res A* 24(3):231–242
  122. National Cooperative Highway Research Program (2000) Report 446. Transp Res Board, Washington DC
  123. Newell A, Simon HA (1972) *Human problem solving*. Prentice Hall, Englewood Cliffs
  124. Niemeier DA (2003) Mobile source emissions: An overview of the regulatory and modeling framework. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 13–1 to 13–28
  125. Ortuzar JD, Willumsen LG (2001) *Modelling transport*, 3rd edn. Wiley, Chichester
  126. Paaswell RE, Roupail N, Sutaria TC (eds) (1992) *Site impact traffic assessment. Problems and solutions*. ASCE, New York
  127. Payne JW, Bettman JR, Johnson EJ (1993) *The adaptive decision maker*. Cambridge University Press, Cambridge
  128. Patten ML, Goulias KG (2001) Test plan: motorist survey – Evaluation of the Pennsylvania turnpike advanced travelers information system (ATIS) project, Phase III PTI-2001-23-I. April 2001. University Park
  129. Patten ML, Hallinan MP, Pribyl O, Goulias KG (2003) Evaluation of the Smarttraveler advanced traveler information system in the Philadelphia metropolitan area. Technical memorandum. PTI 2003–33. March 2003. University Park
  130. Pendyala R (2003) Time use and travel behavior in space and time. In: Goulias KG (ed) *Transportation systems planning: Methods and applications*. CRC Press, Boca Raton, pp 2–1 to 2–37
  131. Pendyala RM, Kitamura R, Kikuchi A, Yamamoto T, Fujii S (2005) The florida activity mobility simulator (FAMOS): An overview and preliminary validation results. Presented at the 84th Annual Transportation Research Board Conference and CD-ROM
  132. Pribyl O (2004) *A microsimulation model of activity patterns and within household interactions*. Ph D Dissertation, Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park
  133. Pribyl O (2007) Computational intelligence in transportation: Short user-oriented guide. In: Goulias KG (ed) *Transport science and technology*. Elsevier, Amsterdam, pp 37–54
  134. Pribyl O, Goulias KG (2003) On the application of adaptive neuro-fuzzy inference system (ANFIS) to analyze travel behavior. Paper presented at the 82nd Transportation Research Board Meeting and included in the CDROM proceedings and accepted for publication in the *Transportation Research Record*, Washington DC, January 2003
  135. Pribyl O, Goulias KG (2005) Simulation of daily activity patterns. In: Timmermans H (ed) *Progress in activity-based analysis*. Elsevier Science, Amsterdam, pp 43–65
  136. Ramadurai G, Srinivasan KK (2006) Dynamics and variability in within-day mode choice decisions. Role of state dependence, habit persistence, and unobserved heterogeneity. *Transportation Research Record*, J Transportation Research Board, No. 1977, Transportation Research Board of the National Academies, Washington DC, pp 43–52
  137. Recker WW (1995) The household activity pattern problem: General formulation and solution. *Transp Res B* 29:61–77
  138. Recker WW, McNally MG, Root GS (1986) A model of complex travel behavior: Part I – Theoretical development. *Transp Res A* 20(4):307–318
  139. Recker WW, McNally MG, Root GS (1986) A model of complex travel behavior: Part II – An operational model. *Transp Res A* 20(4):319–330
  140. Richardson A (1982) Search models and choice set generation. *Transp Res Part A* 16(5–6):403–416
  141. Robinson J (1982) Energy backcasting: a proposed method of policy analysis. *Energ Policy* 10(4):337–344

142. Rubinstein A (1998) Modeling bounded rationality. The MIT Press, Cambridge
143. Sadek AW, El Dessouki WM, Ivan JI (2002) Deriving land use limits as a function of infrastructure capacity. Final Report, Project UVMR13-7, New England Region One University Transportation Center. MIT, Cambridge
144. Salomon I (1986) Telecommunications and travel relationships: A review. *Transp Res A* 20(3):223–238
145. Salvini P, Miller EJ (2003) ILUTE: An operational prototype of a comprehensive microsimulation model of urban systems. Paper presented at the 10th International Conference on Travel Behaviour Research, Lucerne, August 2003
146. Savage LJ (1954) The foundations of statistics. Reprinted version in 1972 by Dover Publications, New York
147. Searle SR, Casella G, McCulloch CE (1992) Variance components. Wiley, New York
148. Simma A, Axhausen KW (2001) Within-household allocation of travel-The case of Upper Austria. *Transportation Research Record: J Transportation Research Board*, No. 1752, TRB, National Research Council, Washington DC, pp 69–75
149. Simon HA (1983) Alternate visions of rationality. In: Simon HA (ed) Reason in human affairs. Stanford University Press, Stanford, pp 3–35
150. Simon HA (1997) Administrative behavior, 4th edn. The Free Press, New York
151. Southworth F (2003) Freight transportation planning: Models and methods. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 4.1–4.29
152. Sparmann U (1980) Ein verhaltensorientiertes Simulationsmodell zur Verkehrsprognose. Schriftenreihe des Instituts für Verkehrswesen 20. Universität (TH) Karlsruhe, Karlsruhe
153. Stefan KJ, McMillan JDP, Hunt JD (2005) An urban commercial vehicle movement model for calgary. Paper presented at the 84th Transportation Research Board Meeting, Washington DC
154. Stopher PR (1994) Predicting TCM responses with urban travel demand models. In: Wholley TF (ed) Transportation planning and air quality II. American Society of Civil Engineers, New York
155. Stopher PR, Meyburg AH (eds) (1976) Behavioral travel-demand models. Lexington Books, Lexington
156. Stopher PR, Hartgen DT, Li Y (1996) SMART: simulation model for activities, resources and travel. *Transportation* 23:293–312
157. Sundararajan A, Goulias KG (2002) Demographic microsimulation with DEMOS 2000: Design, validation, and forecasting. In: Goulias KG (ed) Transportation systems planning: Methods and applications. CRC Press, Boca Raton, pp 14-1–14-23
158. Swait J, Ben-Akiva M (1987) Incorporating random constraints in discrete models of choice set generation. *Transp Res Part B* 21(2):91–102
159. Swait J, Ben-Akiva M (1987) Empirical test of a constrained choice discrete model: Mode choice in Sao Paolo, Brazil. *Transp Res Part B* 21(2):103–115
160. Teodorovic D, Vukadinovic K (1998) Traffic control and transport planning: A fuzzy sets and neural networks approach. Kluwer, Boston
161. Thill J (1992) Choice set formation for destination choice modeling. *Progr Human Geogr* 16(3):361–382
162. Tiezzi E (2003) The end of time. WIT Press, Southampton
163. Timmermans H (2003) The saga of integrated land use-transport modeling: How many more dreams before we wake up? Conference keynote paper at the Moving through net: The physical and social dimensions of travel. 10th International Conference on Travel Behaviour Research, Lucerne, 10–15, August 2003. In: Proceedings of the meeting of the International Association for Travel Behavior Research (IATBR). Lucerne, Switzerland
164. Timmermans H (2006) Analyses and models of household decision making processes. Resource paper in the CDROM proceedings of the 11th IATBR International Conference on Travel Behaviour Research, Kyoto, Japan
165. Timmermans H, Arentze T, Joh C-H (2001) Modeling effects of anticipated time pressure on execution of activity programs. *Transp Res Rec* 1752:8–15
166. Train KE (2003) Discrete choice methods with simulation. Cambridge University Press, Cambridge
167. Transportation Research Board (1999) Transportation, energy, and environment. Policies to promote sustainability. Transportation Research Circular 492. TRB Washington DC
168. Transportation Research Board (2002) Surface transportation environmental research: A long-term strategy. Transportation Research Board, Washington DC
169. Tversky (1969) Intransitivity of preferences. *Psychol Rev* 76:31–48
170. Tversky (1972) Elimination by aspects: A theory of choice. *Psychol Rev* 79:281–299
171. Tversky A, Kahneman D (1992) Advances in prospect theory: Cumulative representation of uncertainty. *J Risk Uncertain* 9:195–230
172. US Government (2006) Analytical perspectives. Budget of the United States Government, Fiscal year 2007. US Government printing Office, Washington DC
173. Van Middelkoop M, Borgers AWJ, Timmermans H (2004) Merlin. *Transp Res Rec* 1894:20–27
174. Van der Hoorn T (1997) Practitioner's future needs. Paper presented at the Conference on Transport Surveys, Raising the Standard. Grainau, Germany, May 24–30
175. Vause M (1997) A rule-based model of activity scheduling behavior. In: Ettema DF, Timmermans HJP (eds) Activity-based approaches to travel analysis. Elsevier Science Inc, New York, pp 73–88
176. Veldhuisen J, Timmermans H, Kapoen L (2000) RAMBLAS: a regional planning model based on the microsimulation of daily activity travel patterns. *Transp Res A* 32:427–443
177. Vovsha P, Petersen E (2005) Escorting children to school: Statistical analysis and applied modeling approach. *Transp Res Rec: J Transp Res Board* 1921, Transportation Research Board of the National Academies, Washington DC, pp 131–140
178. Vovsha P, Peterson EJ, Donnelly R (2002) Microsimulation in travel demand modeling: Lessons learned from the New York best practice mode. *Transp Res Rec* 1805:68–77
179. Vovsha P, Peterson EJ, Donnelly R (2003) Explicit modeling of joint travel by household members: Statistical evidence and applied approach. *Transp Res Rec: J Transp Res Board* 1831: 1–10
180. Waddell P, Ulfarsson GF (2003) [Dynamic simulation of real estate development and land prices within an integrated land use and transportation model system](http://www.urbansim.org/papers/). Presented at the 82nd Annual Meeting of the Transportation Research Board, 12–16 January 2003, Washington DC. Also available in <http://www.urbansim.org/papers/> – accessed April 2003



181. Wang D, Timmermans H (2000) Conjoint-based model of activity engagement, timing, scheduling, and stop pattern formation. *Transp Res Rec* 1718:10–17
182. Weiland RJ, Purser LB (2000) Intelligent transportation systems. In: *Transportation in the New Millennium. State of the art and future directions. Perspectives from transportation research board standing committees*. Transportation Research Board. National Research Council. The National Academies, Washington DC, p 6. Also in <http://nationalacademies.org/trb/>
183. Wen C-H, Koppelman FS (2000) A conceptual and methodological framework for the generation of activity-travel patterns. *Transportation* 27:5–23
184. Williams HCWL, Ortuzar JD (1982) Behavioral theories of dispersion and the mis-specification of travel demand models. *Transp Res B* 16(3):167–219
185. Wilson EO (1998) *Consilience, the unity of knowledge*. Vintage Books, New York
186. Wolf J, Guensler R, Washington S, Frank L (2001) Use of electronic travel diaries and vehicle instrumentation packages in the year 2000. Atlanta Regional Household Travel Survey. Transportation Research Circular, E-C026, March 2001, Transportation Research Board, Washington DC
187. Zhang J, Timmermans HJP, Borgers AWJ (2005) A model of household task allocation and time use. *Transp Res B* 39:81–95

## Treasury Market, Microstructure of the U.S.

BRUCE MIZRACH<sup>1</sup>, CHRISTOPHER J. NEELY<sup>2</sup>

<sup>1</sup> Department of Economics, Rutgers University,  
New Brunswick, USA

<sup>2</sup> Research Department, Federal Reserve Bank  
of St. Louis, St. Louis, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Types of Treasury Issues

Treasury Market Participants

Stages of the Treasury Bond Market

The Treasury Futures Market

Seasonality and Announcement Effects

Discontinuities in the US Treasury Market

Order Flow in the US Treasury Market

Modeling the Limit Order Book

Price Discovery

Future Directions

Bibliography

### Glossary

**Algorithmic trading** Algorithmic trading is the practice of automatically transacting based on a quantitative model.

**Broker** A broker is a firm that matches buyers and sellers in financial transactions. An *interdealer broker (IDB)* is an intermediary providing trading services to hedge funds, institutions, and other dealers. IDB's handle the majority of Treasury securities transactions in the secondary market.

**Coupons** Owners of Treasury notes and bonds receive periodic payments called coupons. They are fixed by the Treasury at auction and are typically paid semi-annually.

**Depth** Depth is the quantity the dealer is willing to sell at the bid or offer.

**Electronic communications networks (ECN)** The Securities and Exchange Commission defines electronic communications networks (ECNs) as “electronic trading systems that automatically match buy and sell orders at specified prices”.

**Market microstructure** Market microstructure is a field of economics that studies the price formation process and trading procedures in security markets.

**On-the-run** On-the-run refers to the most recently auctioned Treasury security of a particular maturity. After the next auction, the security goes *off-the-run*.

**Price discovery** The process by which prices adapt to new information.

**Primary dealers** Primary dealers are large brokerage firms and investment banks that are permitted to trade directly with the Federal Reserve in exchange for making markets in Treasuries. They provide the majority of liquidity in the Treasury market, participate in Treasury auctions, and provide information to assist the Fed in implementing open market operations.

**Secondary market** After the initial auction of Treasury instruments, trading in on-the-run and off-the-run securities makes up the *secondary* Treasury market.

**When issued** When-issued bonds are those Treasuries whose auctions have been announced but have not yet settled.

### Definition of the Subject

This article discusses the microstructure of the *US Treasury securities market*.

US Treasury securities are default risk free debt instruments issued by the US government. These securities play an important, even unique, role in international financial markets because of their safety, liquidity, and low transac-



tions costs. Treasury instruments are often the preferred safe haven during financial crises, a process often referred to as a “flight to quality”.

According to the US Treasury, there was more than \$9 trillion in US government debt outstanding as of August 31, 2007. Of this quantity, the public holds more than \$5 trillion and \$4.5 trillion is tradable on financial markets. Foreigners hold approximately \$2.4 trillion of the marketable supply, with Japan and China together holding more than \$1 trillion. According to the Securities Industry and Financial Markets Association (SIFMA), average daily trading volume in the US Treasury market in 2007 was \$524.7 billion.

*Microstructure* is the study of the institutional details of markets and trading behavior. Microstructural analysis takes three ideas seriously that are often overlooked: the institutional features of the trading process influence how private information is impounded into prices; agents are heterogeneous; and information is asymmetric. Empirical microstructure research studies topics such as the causes and effects of market structure, how market structure influences price discovery, how trading and order flow reveal private information, how quickly public information is impounded into prices, the volatility-volume relation, and the determinants of transactions costs (i.e., the components of bid-ask spreads). The relatively recent availability of tick-by-tick financial data and limit order book data, as well as the computer resources to manipulate them, have been a great boon to financial market microstructure research.

## Introduction

We begin by describing the types of Treasury issues and the major Treasury market participants, including the Federal Reserve, primary dealers and the major electronic brokers. We then outline the stages of the Treasury market, from auction announcements to the secondary market. Next, we examine several closely related areas of the literature: Seasonality in the Treasury market and the reactions of the Treasury market to macro and monetary announcements; discontinuities in Treasury prices; and the effect of order flow in Treasury markets. We then discuss modeling and other academic questions about the Treasury market.

## Types of Treasury Issues

As of October 2007, the US Treasury issued four types of debt instruments. The shortest-maturity instruments are known as Treasury *bills*. 22.6% of the marketable US debt is in bills, securities with maturities of 1 year or less. Bills

are sold at a discount and redeemed at their face value at maturity. They do not pay any coupons prior to maturity and currently have maturities up to 26 weeks. Treasury bill prices are usually quoted in “discount rate” terms, which are calculated with an actual/360 day count convention,

$$\text{T-bill discount rate} = [\text{face value} - \text{bill price}] \times (360/\text{number of days until maturity}) .$$

Thus, a bill with a face value of \$100,000, a cash price of \$97,500 and 90 days to maturity will have a discount rate of  $10\% = [100 - 97.5] \times (360/90)$  in a newspaper. Treasury bill yields are often quoted as “bond equivalent yields”, which are defined as,

$$\text{T-bill yield} = \left[ \frac{\text{face value} - \text{bill price}}{\text{bill price}} \right] \times (365/\text{number of days until maturity}) .$$

Treasury instruments with intermediate maturities (2-, 5- and 10-year) are known as *Treasury notes*. *Notes* pay semi-annual coupons, and make up 54.7% of the debt. In February 2006, the US Treasury also resumed issuing 30-year instruments, known as *Treasury bonds*. *Bonds* also pay semi-annual coupons, and make up 12.5% of the US debt.

The price of both notes and bonds are quoted as a percentage of their face value in thirty-seconds of a point. A quoted price of 98-08 means that the quoted price of the note (or bond) is  $(98 + 8/32) = \$98.25$  for each \$100 of face value. The cash price of bonds and notes is equal to the quoted price plus accrued interest since the last coupon payment, calculated with an actual/actual day count convention. Quoted prices are sometimes called “clean” prices, while cash prices are said to be “dirty”.

The US Treasury also issues 5-, 10-, and 20-year Treasury Inflation-Protected Securities (“TIPS”), whose pay-off is linked to changes in the US Consumer Price Index (CPI). These make up about 10.2% of the total value of Treasuries outstanding. The principal value of TIPS is adjusted daily and the semi-annual coupon payments and principal payment are then based on the adjusted principal amount. Economists extract inflation forecasts by comparing the TIPS yields to those on similar nominal instruments. The Federal Reserve Bank of Saint Louis provides “TIPS spreads” through its publication, *Monetary Trends*.

There is also an active market in STRIPS (Separate Trading of Registered Interest and Principal of Securities) which are popularly known as “zero coupon” bonds. These instruments are created by the Treasury through an accounting system which separates coupon interest payments and principal. Finally, the US Treasury also is-

sues savings bonds, low denomination securities for retail investors.

## Treasury Market Participants

### The Federal Reserve in the Treasury Market

The Federal Reserve Bank of New York, under the guidance of the Federal Open Market Committee (FOMC), is a uniquely important player in the Treasury market. The FOMC meets approximately every six weeks to review economic conditions and determine a target for the federal funds rate, the rate at which US banks borrow/lend reserve balances from/to each other. The manager of the Open Market Desk (a.k.a., “the Desk”) at the Federal Reserve Bank of New York is responsible for ensuring that the average federal funds transaction is close to the target by buying and selling Treasury instruments (primarily short-term). In practice, the Desk accomplishes this in two ways. First the Desk buys sufficient Treasuries to satisfy most but not all the markets’ demand for deposits at the Fed. Secondly, the Desk buys Treasuries via repurchase (repos) agreements (overnight and for terms of several days) to achieve a desired repo rate that influences the federal funds rate and other short-term interest rates through arbitrage.

To determine day-to-day actions, every morning, staff at both the Division of Monetary Affairs of the Board of Governors of the Federal Reserve System and the Desk forecast that day’s demand for reserve balances. The Desk staff also consults market participants to get their views on financial conditions. The relevant Desk and Board staffs then exchange views in a 9 am conference call. Finally, the relevant Desk staff, the Board staff, and at least one of the voting Reserve Bank Presidents then confer during a second conference call at about 9:20 am. The Desk staff summarizes market conditions, projects actions for the day and asks the voting Reserve Bank President(s) for comments. Open market operations commence shortly after the conclusion of this call.

When the Desk buys Treasuries, it increases available liquidity (reserves) in debt markets and tends to lower interest rates. Selling Treasuries has the opposite effect, lowering reserves and raising interest rates. If the intention is to make a permanent change in reserves, then outright purchases or sales are undertaken. In contrast, if the Desk anticipates that only temporary changes in reserves are necessary, it uses repos (for purchases) or reverse repos (for sales). Bernanke [7] notes that actual open market sales of debt instruments are rare; it is more common for the Federal Reserve to allow such securities to expire without replacing them. Both open market sales and allowing the Fed’s securities to expire have the same balance sheet

effects: The Fed holds fewer bonds and more cash, while the public will hold more bonds and less cash.

The Federal Reserve provides several valuable references on its operating procedures. The Annual Report of the Markets Group of the Federal Reserve Bank of New York describes open market operations and current procedures (Federal Reserve Bank of New York, Markets Group [26]). Meulendyke [57] provides a comprehensive view of Federal Reserve monetary policy operations with a historical perspective. Akhtar [1] explains how monetary policy is decided and how such policies affect the economy. Finally, Harvey and Huang [43] gives some historical perspective on operating procedures in the 1980s.

### Primary Dealers

Among the most important private sector players in the Treasury markets are the 21 *primary dealers*. The Federal Reserve Bank of New York explains that primary dealers must “participate meaningfully in both the Fed’s open market operations and Treasury auctions and ... provide the Fed’s trading desk with market information and analysis that are helpful in the formulation and implementation of monetary policy”. The Federal Reserve does not regulate primary dealers, but does subject them to capital requirements. The Federal Reserve can withdraw a firm’s primary dealer designation if it fails to participate in auctions or open market operations or if its capital reserves fall below desired levels.

### Interdealer Brokers

Prior to 2000, voice-assisted brokers dominated secondary market trading in Treasuries. Except for Cantor–Fitzgerald, all these brokers reported their trading activity to GovPX, a consortium. In the face of demands by the Securities and Exchange Commission and bond market dealers for greater transparency, five IDBs formed GovPX as a joint venture in 1991. In March 1999, Cantor–Fitzgerald opened up its internal electronic trading platform, eSpeed, to clients. The eSpeed system quickly grabbed a dominant market share, and Cantor Fitzgerald spun off eSpeed as a public company in December 1999. In 2000, a competing electronic brokerage, BrokerTec, joined the market. As in foreign exchange and equity markets, most interdealer and institutional trading in Treasuries quickly migrated from voice networks to these electronic communications networks (ECNs), which have dominated trading in Treasury instruments since 2001. Mizrach and Neely [58] describe the transition from voice assisted trading, largely through the primary dealers, to electronic trading in the Treasury market.

As of November 2007, the two dominant ECNs are eSpeed and BrokerTec. London-based ICAP, PLC, owns BrokerTec while eSpeed merged in the summer of 2007 with BGC, another London based interdealer brokerage. eSpeed and ICAP compete for both on- and off-the-run liquidity. Hilliard Farber and Tullett–Prebon hold the largest brokerage share outside of the dominant two platforms.

### Stages of the Treasury Bond Market

The sale of Treasuries undergoes four distinct phases: when issued, primary, on-the-run and off-the-run. Each of these stages has a distinct market structure.

#### The Primary Market

In the *primary* market, the US Treasury sells debt to the public via auction. The US Treasury usually publishes a calendar of upcoming tentative auction dates on the first Wednesday of February, May, August, and November and bids may be submitted up to 30 days in advance of the auction. In practice, however, the Treasury only announces firm auction information several days in advance and most bids are submitted at that time. Since August 8, 2002, the Treasury has made auction announcements (for all new securities) at 11:00 am Eastern Time (ET). 13- and 26-week bills are auctioned weekly; 2- and 5-year notes are auctioned monthly; 10-year notes are auctioned eight times a year. 30-year bonds, which were reintroduced on February 9, 2006 after a five year hiatus, are auctioned four times a year.

The US Treasury has used a single price auction exclusively since November 1998. Garbade and Ingber [35] discuss the transition from multiple price auctions to the current format single price auctions. All securities are allocated to bidders at the price that, in the aggregate, will result in the sale of the entire issue. This mitigates the risk of a “buyer’s curse” – the highest bidder paying more than other auction participants. To prevent a single large buyer from manipulating the auction, the Treasury restricts anyone from buying more than 35% of any single issue. Bids may be submitted up to thirty days prior to the auction, and large institutions make use of the Treasury Automated Auction Processing System (TAAPS). Retail investors can participate through the Treasury Direct program. The Treasury allocates a portion of nearly every auction to small investors at the same price as the large institutions. These are called *non-competitive bids*, and they are quantity only orders that are filled at the market clearing price.

Primary dealers dominate the auction process. In 2003, they submitted 86% of auction bids, totalling more than \$6 trillion. They were awarded \$2.4 trillion, or 78% of the total auction supply.

#### The Secondary Market

The secondary market is composed of the when-issued, on-the-run and off-the-run issues.

**When-Issued** Even prior to the primary auction, there is an active forward market in Treasury securities (apart from TIPS) that are about to be issued. Trading in the *when-issued* security market typically begins several days prior to an auction and continues until settlement of auction purchases. Nyborg and Sundaresan [61] document that when-issued trading provides important information about auction prices prior to the auction and also permits market participants to reduce the risk they take in bidding. Fabozzi and Fleming [25] estimate that 6% of total interdealer trading is in the when-issued market. Just prior to auctions though, these markets become substantially more active. In the bill market, when-issued trading volume exceeds the volume for the bills from the previous auction.

**On-the-Run** Upon completion of the auction, the most recently issued bill, note or bond becomes *on-the-run* and the previous on-the-run issue goes *off-the-run*. Overall Treasury trading volume is concentrated in a small number of on-the-run issues. Trading in these benchmark on-the-run issues, which Fabozzi and Fleming [25] say constitutes approximately 70% of total trading volume, has migrated almost completely to the electronic networks. Mizrach and Neely [58] estimate a 61% market share for the BrokerTec platform and a 39% share for eSpeed in 2005, which is consistent with industry estimates.

**Off-the-Run** With more than 200 off-the-run issues trading in October 2007 – 44 bills, 116 notes, and 45 bonds – most off-the-run volume takes place in voice and electronic interdealer networks. Barclay, Hendershott and Kotz [5] document the fall in ECN market share when issues go off the run. They also report that transaction volume falls by more than 90%, on average, once a bond goes off-the-run. The ECN market share falls from 75.2% to 9.9% for the 2-year notes, from 83.5% to 8.5% for the 5-year notes, and from 84.5% to 8.9% for the 10-year notes. Several IDBs handle most off-the-run securities trading.

**On- Versus Off-the-Run Liquidity and Prices** Off-the-run securities trade at a higher yield (lower price) than on-

the-run securities of similar maturity. Many researchers have attempted to explain the yield differential with relative liquidity. Vayanos and Weill [68] utilize a search theoretic model that is motivated by the fact that bonds may be difficult to locate once they go off-the-run. Goldreich, Hanke, and Nath [36] compare on-the-run and off-the-run Treasuries and show that the liquidity premium depends primarily on the amount of remaining future liquidity, which is highly predictable. The study exploits the fact that the liquidity of a Treasury is predictable. Duffie [18] argues that legal or institutional restrictions on supplying collateral induces “special” repo rates that are much less than market riskless interest rates. The price of the underlying instrument is increased by the present value of the savings in borrowing costs.

**Supply Variation and Prices** Although it is generally accepted that the on-the-run premium is due to greater liquidity, the theoretical relation between the supply of a given bond issue and prices is not clear. Do issue sizes produce lower yields (higher prices) through their liquidity effects or does downward-sloping demand for individual securities produce higher prices (lower yields) for larger issues? Empirically, the evidence is mixed. Simon [65,66], Duffie [18], Seligman [64] and Fleming [29] find that the larger issues lead to lower prices (higher yields), while Amihud and Mendelson [2], Kamara [51], Warga [69], and Elton and Green [23] find the opposite: The liquidity effect predominates, resulting in higher prices (lower yields) for larger issues. There might be

a nonlinear relationship. Liquidity may increase prices up to a certain point, but then finite demand for any individual security reduces the attractiveness of additional supply.

### The Treasury Futures Market

Spot markets are not the only markets for US Treasuries. The Chicago Board of Trade (CBOT) has active futures markets for 2-, 5-, 10- and 30-year US Treasuries. Table 1 briefly describes the CBOT contracts and pricing conventions.

Like other exchange-traded derivatives, Treasury futures have two advantages: trading is highly liquid and marking-to-market minimizes counterparty risk. The CBOT open auction trading hours are 7:20 am to 2:00 pm, Central Time, Monday through Friday; the CBOT electronic market functions from 6:00 pm to 4:00 pm, Central Time, Sunday through Friday. All Treasury contracts have a March–June–September–December cycle.

A variety of Treasury instruments meet the criteria to be deliverable issues. Table 1 describes the pricing conventions and the characteristics of the assets that may be delivered to satisfy the contracts. The CBOT defines “conversion factors” that adjust the quoted futures prices for the asset that is actually delivered. Despite these conversion factors, one issue will be the “cheapest to deliver”. Cash prices at delivery depend on both the conversion factor for a particular bond and the interest accrued on that bond since the last coupon payment.

**Treasury Market, Microstructure of the U.S., Table 1**  
**Contract Details from the CBOT Treasury Market**

Contract	Quote convention	Pricing example	Deliverable asset characteristics
2-year	1/32 and quarters of 32nds	$95 - 060 = 95 + 6/32$ $95 - 062 = 95 + 6.25/32$ $95 - 065 = 95 + 6.5/32$ $95 - 067 = 95 + 6.75/32$	US Treasury notes with a face value $\geq$ \$200,000 and original maturity $\leq$ 5 years and 3 months and remaining maturity $\geq$ 1 year and 9 months from the first day of the delivery month and and remaining maturity $\leq$ than 2 years from the last day of the delivery month.
5-year	1/32 and halves of 32nds	$90 - 170 = 90 + 17/32$ $90 - 175 = 90 + 17.5/32$	US Treasury notes with a face value $\geq$ \$100,000 and original maturity $\leq$ 5 years and 3 months and remaining maturity $\geq$ 4 year and 2 months from the first day of the delivery month
10-year	1/32 and halves of 32nds	$90 - 170 = 90 + 17/32$ $90 - 175 = 90 + 17.5/32$	US Treasury notes with a face value $\geq$ \$100,000 and remaining maturity $\leq$ 10 years remaining maturity $\geq$ 6 year and 6 months from the first day of the delivery month
30-year	1/32nds	$85 - 12 = 85 + 12/32$	US Treasury bonds with a face value $\geq$ \$100,000 and if callable: Not callable for at least 15 years from the first day of the delivery month; if not callable: Remaining maturity $\geq$ 15 years from the first day of the delivery month.

Although agents frequently use the futures markets for hedging or taking positions on future price movements, only a modest amount of microstructure research has focused on futures markets. Brandt, Kavajecz, and Underwood [11] show that futures and spot market order flow are useful in predicting daily returns in each market and that the type of trader influences the effect of order flow. Mizrahi and Neely [59] show that futures markets contribute a substantial amount of price discovery to US Treasury markets. Campbell and Hendry [12] compare price discovery in the 10-year bond and futures contracts in both the United States and Canada.

### Seasonality and Announcement Effects

Seasonality and announcement effects are intimately related to the microstructure literature in that the latter seeks to explain how markets with heterogeneous agents react to the release of information.

### Seasonality and Macroeconomic Announcements

The earliest studies considered the issue of daily seasonality in Treasuries. Flannery and Protopapadakis [27] document differing day-of-the-week patterns in Treasuries and stock indices. The patterns in the prices of Treasuries securities vary by maturity and differ from those found in stock indices. They conclude that no single factor explains seasonal patterns across asset classes. In contrast to this day-of-the-week effect in spot T-bills, Johnston et al. [50] find day-of-the-week effects in government national mortgage association (GNMA) securities, T-note, and T-bond futures, but not in T-bill futures. The fact that day-of-the-week effects exist in spot T-bills but not in T-bill futures points up the importance of futures settlement rules.

Later studies began to consider the effects of macro announcements on price changes, volatility, volume and spreads. Macroeconomic announcements have been an especially popular subject of study because they occur at regular intervals that can be anticipated by market participants. The existence of survey expectations about upcoming macro announcements permits researchers to identify the “shock” component of the announcement, which allows them to investigate the differential effects of anticipated and unanticipated news releases of different magnitudes.

Ederington and Lee [20,21] did the seminal modern work with intraday data on macro announcement effects in bond markets. They found that volatility increases before the announcement and remains elevated for some time afterwards. The employment, PPI, CPI and durable goods orders releases produce the greatest im-

pact of the 9 significant announcements, out of 16 studied. Ederington and Lee [22] follow up on their earlier studies by linking the literatures on seasonality and announcements in the bond market. Comparing the contributions of past volatility, seasonality and announcements in predicting intraday volatility bond futures data and exchange rates, these authors argue that announcements account for much of the apparent seasonality in interest rate volatility.

One of the earliest important results was that bond market prices react more strongly to macro announcements than do equity markets. Fleming and Remolona [32, 34] examined the 25 largest price changes in the GovPX data and related them all to macroeconomic announcements. Fleming and Remolona [34] note: “In contrast to stock prices, US Treasury security prices largely react to the arrival of public information on the economy”. Fleming and Remolona [32,33] attribute the relative sensitivity of bond markets to the fact that bond prices depend only on expected discount rates while stock prices are also determined by future expected dividends. Macro announcements can have little or no effect on stock prices if their effects on expected dividends and discount rates offset each other.

Several studies used more sophisticated econometric procedures to evaluate the impact of announcements on persistence in volatility in a full model. Jones, Lamont and Lumsdaine [49] examine volatility patterns in the 5-year Treasury market around US announcements. Daily volatility from an ARCH-M does not persist for days after announcements and the authors interpret this as indicating that agents rapidly incorporate announcement information into prices. Weekly volatility displays a U-shaped pattern; the largest price changes occur on Mondays and Fridays. Further, Jones, Lamont and Lumsdaine [49] find a risk premium in returns on days of announcements. Bollerslev, Cai, and Song [8] also consider the interaction of announcements and persistence in volatility with 5-minute US Treasury bond data. Modeling the intraday volatility patterns and accounting for announcements reveals long-memory in bond market volatility.

An important issue in microstructure is the determination of bid-ask spreads. Balduzzi, Elton, and Green [4] use intraday GovPX data to look at the effects of macro announcements on volume, prices and spreads. Confirming previous findings, prices adjust to news within one minute while increases in volatility and volume persist for up to 60 minutes. Spreads initially widen but then return to normal after 5 to 15 minutes. News releases explain a substantial amount of bond market volatility. Importantly, Balduzzi, Elton, and Green [4] argue that the dif-



ferential impact of news on long and short bond prices indicates that at least two factors will be needed for models of the yield curve. They also present evidence that discontinuities (jumps) will be important in modeling bond prices.

Some recent papers have relaxed the restrictive assumption that announcements influence Treasury market variables in a linear, symmetric fashion. For example, Christie–David, Chaudhry, and Lindley [15] allow the effects of announcement shocks to depend on the size and sign of the shock. They measure these nonlinear effects on the intraday 10- and 30-year Treasury futures from 1992 to 1996.

Most studies of the effects of volatility have measured such variation with some function of squared returns. One can use the volatility implied by options prices, however, to measure expected volatility over longer horizons. Heuson and Su [45], for example, show that implied volatilities from options on Treasuries rise prior to macro announcements and that volatilities quickly return to normal levels after announcements. Beber and Brandt [6] use intraday, tick data from 1995 to 1999 to determine that macro announcements reduce the variance of the option-implied distribution of US Treasury bond prices. The content of the news and economic conditions explain these changes in higher-order moments. The study attributes the results to time-varying risk premia rather than relative mispricing or changing beliefs.

In a comprehensive study of the impact of US macroeconomic announcements across asset markets, Andersen, Bollerslev, Diebold and Vega [3] study the reaction of international equity, bond and foreign exchange markets. They confirm that US macroeconomic news drives bond prices, as well as those of the other assets.

### Monetary Policy Announcements

Researchers have carefully investigated the effects of the Federal Reserve's actions on the Treasury market. While the literature has examined the effect of a wide variety of monetary policy behavior and communications – e.g., open market operations, FOMC news releases, speeches, etc. – on many aspects of Treasury market behavior, a large subset of these papers deal with one specific topic: The effect of federal funds target changes on the Treasury yield curve.

**Federal Funds Target Changes and the Treasury Yield Curve** The “expectations hypothesis of the term structure” motivates research on how the short- and long-end of the Treasury yield curve react to unexpected changes in

the federal funds target rate. That is, if the FOMC increases overnight interest rates, how does this change short- and long-term rates?

Using data on 75 changes in the federal funds target from September 1974 through September 1979, Cook and Hahn [16] find that these target changes caused larger movements in short-term rates than in intermediate- and long-term Treasury rates. A difficulty with interpreting the Cook and Hahn [16] results is that efficient markets presumably can often anticipate most or all of a target change and such expectations are already incorporated into the yield curve. To confront this problem, Kuttner [53] decomposes target changes into anticipated and unanticipated components, finding – unsurprisingly – that Treasury rates respond much more strongly to unanticipated changes and that the results are consistent with the expectations hypothesis of the term structure. That is, the anticipated component of an interest rate change does not affect expectations. Hamilton [41] carefully reexamines the work of Kuttner [53], showing that it is robust to uncertainty about the dates of target changes and the effect of learning by market participants.

Poole and Rasche [62] also decompose federal funds target changes into expected and unexpected components – but use a later contract month than Kuttner [53] to avoid problems associated with computation of the contract payoff. They find that interest rates across the maturity spectrum fail to respond to the anticipated components of the changes in the intended funds rate.

Poole, Rasche and Thornton [63] consider how changes in FOMC procedures affect the impact of target changes on interest rates. This study first succinctly describes the changes in FOMC procedures in the 1990s. The FOMC began to contemporaneously announce policy actions in 1994 and adopted this as formal policy in 1995. Starting in August 1997, each policy directive has included the quantitative value of the “intended federal funds rate”. And since 1999, the FOMC has issued a press release after each meeting with the value for the “intended federal funds rate” and, in most cases, an assessment of the balance of risks. After describing such procedural changes, Poole, Rasche and Thornton [63] go on to consider the response of the Treasury yield curve to funds rate target changes both before and after the FOMC began contemporaneously announcing target changes in 1994. In doing so, these authors account for measurement error in expectations and uncertainty about the dates of target changes and even whether market participants understood that the Federal Reserve was targeting the funds rate prior to 1994. They assess the market's knowledge of targeting by examining news reports. While short-rates respond similarly in

both subperiods, long rates do not respond as strongly to funds rate target changes after 1994. The authors interpret their results as being consistent with the Fed's greater transparency about long-run policy in the second subsample. With long-run expectations more firmly anchored, unexpected changes in the funds target have smaller effects on long rates.

One puzzle that has emerged from this literature is that the average effect of changes in the federal funds target on the yield curve is modest, despite the facts that such changes should be an important determinant of the yield curve and that yields are highly volatile around FOMC announcements. Fleming and Piazzesi [31] claim to partially resolve this puzzle by illustrating that such yield changes depend on the shape of the yield curve.

This literature on the reaction of the Treasury market to monetary policy has become progressively more sophisticated in assessing market expectations of Fed policy and modeling institutional features of the futures market and Fed operations. Nevertheless, the underlying conclusion that unanticipated target changes lead to large price increases on short-term Treasuries and smaller changes on the prices of long-term Treasuries has been remarkably robust.

**Other Federal Reserve Behavior and the Treasury Market** There has been a substantial literature analyzing how other types of Federal Reserve behavior have influenced the Treasury market. The literature has considered open market operations, FOMC statements, Congressional testimonies, and FOMC member speeches.

Open market operations are similar to macroeconomic announcements in that they are potentially important bond market events, occurring at regularly scheduled times. Harvey and Huang [43] used intraday data from 1982 to 1988 to examine how Federal Reserve open market operations influenced foreign exchange and bond markets. The paper finds that Treasury market volatility increases during open market operations, irrespective of whether they add or drain reserves. Oddly, volatility increases even more during the usual time for open market operations if there are no such transactions. The authors interpret this finding as indicating that open market operations actually smooth volatility.

Early studies made the simplifying assumption that the effect of macro announcements on the Treasury market was constant over time. This is not necessarily the case, of course. For example, the effect of macro announcements on the Treasury market might depend on monetary policy priorities. Kearney [52] characterizes the changing response of daily 3-month Treasury futures to the em-

ployment report over 1977 to 1997 and relates it to the changing importance of employment in the Fed's reaction function.

de Goeij and Marquering [17] also considers how both macro announcements and monetary policy events affect the US Treasury market. Using daily data from 1982 to 2004 de Goeij and Marquering [17] find that macro news announcements strongly affect the daily volatility of longer-term Treasury instruments while FOMC events affect the volatility of shorter-term instruments.

Some studies have explored more esoteric components of information about monetary policy. Boukus and Rosenberg [9], for example, use Latent Semantic Analysis to decompose the information content of FOMC minutes from 1987 to 2005. They then relate the information content to current and future economic conditions. Chirinko and Curran [13] argue that Federal Reserve speeches, testimonies, and meetings increase price and trading volatility on the 30-year bond market. FOMC meetings are the most important of the events considered. They go on to consider whether these Federal Reserve events merely create noise or transmit information about the future policy decisions or the state of the economy. They conclude that such events may reduce welfare by "overwhelming private information", creating herding behavior.

### Announcements and Liquidity Variation

The literature on variation in liquidity and price effects overlaps with the literature on macroeconomic announcements. The seminal work of Amihud and Mendelson [2] showed that yields on short-time-to-maturity Treasuries vary inversely with liquidity. That is, more liquid assets have lower yields/higher prices. Harvey and Huang [43] discovered elevated volatility in interest rate (and foreign exchange) futures markets, in the first 60–70 minutes of trading on Thursdays and Fridays. Ederington and Lee [20] confirmed Harvey and Huang [43]'s speculation that major macroeconomic announcements – especially the employment report, the PPI, the CPI, and durable goods orders – create the intraday and intraweek patterns in the volatility of Treasury bond futures. Volatility is very high after announcements and remains elevated for hours. Fleming and Remolona [32] extend this work to show that the 25 greatest surges in activity in the 5-year on-the-run bond market came on macroeconomic announcement days, within 70 minutes of the announcement. The most important announcements for trading surges were employment reports, fed funds targets, 30-year auctions, 10-year auctions, the CPI, NAPM surveys, GDP, retail sales, and 3-year auctions. Releases that affect prices also

matter for trading activity. Fleming and Remolona [32] observe that timeliness, the degree of surprise in the announcement and market uncertainty also increase announcements' impact on trading.

Researchers continued to explore the impact of variation in liquidity caused by other events. For example, Fleming [28] exploits exogenous variation in Treasury issuance to show that securities that are "reopened" – the Treasury sells additional quantities of existing securities – have greater liquidity, lower spreads, than comparable assets. Paradoxically, this higher liquidity does not produce lower yields for the reopened securities.

More recent papers have explored variation in liquidity and volatility across markets. Chordia, Sarkar and Subrahmanyam [14] estimate a vector autoregression (VAR) in liquidity and volatility variables in stock and bond markets. They find that common factors make the variables' innovations highly correlated. Volatility shocks predict liquidity variables.

### End-of-the-Year Patterns in One-Month Treasury Bills

The previous sets of papers studied daily and intraday seasonality, often as caused by macroeconomic or Federal Reserve announcements. Short-term Treasury bills also exhibit year-end seasonality, however. Market participants consider Treasury market instruments of 30 days or less to be highly liquid, close – but not perfect – substitutes for cash. The fact that short-term Treasuries are not perfect substitutes for cash is presumably what allows the New York Desk to use open market operations to manipulate short-term interest rates through a liquidity effect. A peculiar year-end pattern in one-month Treasury yields reinforces this evidence that such Treasuries are not perfect substitutes for cash.

Following on related work of Griffiths and Winters [40] in repos, Griffiths and Winters [39] find that yields on one month T-Bills (and other one-month securities) increase significantly at the beginning of December, remain high during December, and return to normal a few days before the year-end. This pattern does not exist in three-month T-bills. Neely and Winters [60] find similar patterns in the one-month LIBOR futures market.

Griffiths and Winters [38,39,40] explain this December effect by asserting that a year-end preference for liquidity drives the year-end surge in short-term interest rates. Debt holder (lenders in the money markets) start to liquidate their one-month securities in the last few days of November to meet cash obligations at the end of December. This preference for liquidity drives up one-month interest rates for most of December. Liquidity demand re-

turns to normal at the end of December as investors repurchase short-term instruments, and interest rates return to normal levels.

### Discontinuities in the US Treasury Market

The literature on discontinuities (or jumps) in Treasury prices is closely related to the literature on announcements, as announcements are obvious candidates to explain jumps. Three recent papers have looked at discontinuities in US Treasury prices. Huang [47] estimates daily jumps with bi-power variation on 10 years of 5-minute data on S&P 500 and US T-bond futures to measure the response of volatility and jumps to macro news. He identifies a major role for payroll news in bond market jumps by analyzing their conditional distributions and regressing continuous and jump components on measures of disagreement and uncertainty concerning future macroeconomic states. Huang [47] also finds that the bond market is relatively more responsive than the equity market.

Dungey, McKenzie, and Smith [19] estimate jumps and cojumps (simultaneous discontinuities in multiple markets) in the term structure of US Treasury rates. They find that the middle of the yield curve often cojumps with one of the ends, while the ends of the curve exhibit a greater tendency for idiosyncratic jumps. Macro news is strongly associated with cojumps in the term structure. Using BrokerTec data from 2003–2005, Jiang, Lo, and Verdelhan [48] extend this work by focusing on the role of liquidity shocks – estimated from the limit order book – in jumps and the relation of jumps to order flow and price discovery.

Lahaye, Laurent and Neely [54] examine jumps and cojumps across foreign exchange, stock, gold and 30-year Treasury futures. Discontinuities in bond futures prices were larger but less frequent than those in foreign exchange rates and smaller and about as frequent as those in equity markets. News announcements appear to cause many cojumps of bond prices with prices of other types of assets.

### Order Flow in the US Treasury Market

The effect of order flow on prices has been a popular recent topic in microstructure. Several papers have explored the impact of order flow on prices and the ways in which macro/monetary announcements influence these impacts.

Huang, Cai, and Wang [46] use intraday 1998 GovPX spot data on the 5-year Treasury note to characterize trading patterns of primary dealers, announcement effects and

volatility-volume relations. The paper finds that both public information (i.e., announcements) and dealer inventory/order flow affect trading frequency.

Green [37] uses the Madhavan, Richardson, and Roomans [55] model to study the impact of GovPX trading in 5-year around announcements. Order flow has its largest price impact after large macro surprises, times of greater uncertainty about the announcement, and times of high liquidity. Green [37] concludes that order flow does reveal information about riskless rates.

Brandt and Kavajecz [10] find that order flow imbalances can explain up to 26% of the day-to-day variation in yields on non-announcement days. In contrast to Green [37], they find that order flow has its strongest impact at times of low liquidity. Brandt, Kavajecz, and Underwood [11] extend the work of Brandt and Kavajecz [10] to control for trader type and macroeconomic announcements in explaining the impact of bond market order flow on futures prices.

Menkveld, Sarkar, and Van der Wel [56] confirm earlier conclusions that announcements have significant effects on 30-year Treasury yields and they also find that customer order flow is much more informative on announcement days than on non-announcement days. They go on to investigate the profits that different types of traders make on announcement and non-announcement days.

At high frequencies, order flow is highly autocorrelated. A dynamic analysis of the market resilience requires modeling this formally. We turn to empirical modeling of the Treasury market order book in the next section.

### Modeling the Limit Order Book

A purchase or a sale of a Treasury bond influences prices directly as trades work their way up the supply or demand curves. We would like to know whether these effects are large and long-lasting. To address this question, we must introduce a dynamic model of the limit order book.

Hasbrouck [44] proposed to study intra-day price formation with a standard bivariate vector autoregressive (VAR) model. Time  $t$  here is measured in 1-minute intervals. Let  $r_t$  be the percentage change in the transaction price and  $x_t^0$  be the sum of signed trade indicators (+1 for buyer initiated, -1 for seller initiated) over minute  $t$ . Treasury market data sets typically indicate trade initiation as a “hit” -1 or a “take” +1.

The bivariate vector autoregression assumes that causality flows from trade initiation to returns by permitting  $r_t$  to depend on the contemporaneous value for  $x_t^0$ , but not allowing  $x_t^0$  to depend on contemporaneous  $r_t$ . The

model for returns is specified as follows

$$\begin{bmatrix} r_t \\ x_t^0 \end{bmatrix} = \sum_{i=1}^5 \begin{bmatrix} a_{r,i} \\ a_{x,i} \end{bmatrix} r_{t-i} + \begin{bmatrix} \sum_{i=0}^{15} b_{r,i} \\ \sum_{i=1}^{15} b_{x,i} \end{bmatrix} x_{t-i}^0 + \begin{bmatrix} u_{r,t} \\ u_{x,t} \end{bmatrix}. \quad (1)$$

Mizrach and Neely [58] use 5 lags of the return series and 15 lags of the signed trades. The market impact is then defined as the dynamic effect of a buy shock to the return series,

$$\frac{\partial r_{t+n}}{\partial x_t}. \quad (2)$$

Mizrach and Neely [58] provide 15 minute market impact estimates from the GovPX market in 1999. The 2-year note is most resilient with prices only 0.0042% higher following a buyer initiated trade. The 30-year bond is the least liquid, with prices rising 0.0229% following a buy order. Mizrach and Neely also report 2004 estimates for the Cantor electronic limit order book. Market impacts range from 45 to 88% lower in the more liquid eSpeed ECN market. Fleming and Mizrach [30] find further reductions in market impacts on the BrokerTec ECN for 2005 and 2006.

### Price Discovery

A crucial issue in the market microstructure literature is *price discovery*. This is the process by which prices embed new information. In the Treasury market, price discovery occurs in both the secondary spot market and in the futures markets at the Chicago Board of Trade (CBOT). The degree to which each market contributes to price discovery is a natural issue to address.

To investigate relative price discovery in these two Treasury markets, Mizrach and Neely [59] follow Hasbrouck [44] and assume that the price series have a unit root, are cointegrated, and have an  $r^{\text{th}}$  order VAR representation,

$$p_t = \Phi_1 p_{t-1} + \Phi_2 p_{t-2} + \cdots + \Phi_r p_{t-r} + u_t.$$

It follows that the  $N$  returns,

$$r_t = \begin{bmatrix} p_{1,t} - p_{1,t-1} \\ \vdots \\ p_{N,t} - p_{N,t-1} \end{bmatrix} = \Delta p_t, \quad (3)$$

have the convenient Engle-Granger [24] error-correction

representation,

$$\Delta p_t = \alpha z_{t-1} + A_1 \Delta p_{t-1} + \dots + A_r \Delta p_{t-r-1} + u_t, \quad (4)$$

where  $z_t$  is an error-correction term of rank  $N - 1$ .

We analyze price discovery using the moving average representation of our return process (3),

$$\Delta p_t = \Theta(L)\varepsilon_t. \quad (5)$$

The disturbances are mean zero and serially uncorrelated,  $E[\varepsilon_{i,t}] = 0$  and  $\text{cov}[\varepsilon_{i,t}, \varepsilon_{i,t-r}] = 0$ , but they may be contemporaneously correlated,  $\text{cov}[\varepsilon_{i,t}, \varepsilon_{j,t}] \neq 0$ .

The information share is related to the long run impulse responses,  $\Theta(1) = \sum_{j=0}^{\infty} \Theta(L^j)$ , the permanent effect of the shock vector on the Treasury prices. Cointegration makes the long run multipliers common across all markets,

$$\Theta(1) = \begin{bmatrix} \theta_1 & \dots & \theta_N \\ \vdots & & \vdots \\ \theta_1 & \dots & \theta_N \end{bmatrix}. \quad (6)$$

To eliminate contemporaneous correlation among the error terms in (5), we decompose  $\Omega = E[\varepsilon_t \varepsilon_t']$ , the  $N \times N$  covariance matrix, to find a lower triangular matrix  $M$ , whose  $i, j$ th element we denote  $m_{ij}$ , such that  $MM' = \Omega$ . The Hasbrouck [44] information share for market  $j$  is defined as

$$H_j = \frac{\left[ \sum_{i=j}^n \theta_i m_{ij} \right]^2}{\left[ \sum_{i=1}^n \theta_i m_{i1} \right]^2 + \left[ \sum_{i=2}^n \theta_i m_{i2} \right]^2 + \dots + (\theta_n m_{nn})^2}, \quad (7)$$

where the  $\theta_i$ s are the elements of row  $i$  of the long-run multipliers in (6). Because the Choleski decomposition is not unique, the information share will vary with the order of the equations in the VAR.

Mizrach and Neely [59] pair spot and maturity matched futures for the 2-year, 5-year and 10-year on-the-run spot notes. This calculation requires us to adjust futures prices according to the on-the-run spot instruments with which we compare them. The CBOT provides adjustment factors for each instrument. These adjustments typically make a single bond the cheapest to deliver (CTD), but the CTD is typically off-the-run. Nevertheless, the CTD off-the-run bonds and the most liquid on-the-run bonds are very close substitutes – their daily returns are highly correlated – so it is reasonable to examine price discovery between futures prices and on-the-run bonds, despite the fact that they are not identical.

Mizrach and Neely [59] find that information shares rise with the growth of the GovPX market, but fall as the

ECNs take market share from GovPX voice markets. The spot market share is highest for the 2-year note, reaching 86%, while the 10-year spot market share never exceeds 50%. In addition, relative market liquidity measures like spreads, trades and volatility each strongly explain daily relative price discovery shares. Mizrach and Neely [59] compute both upper and lower bound estimates of the information shares. They also report estimates based on the Harris, McNish and Wood [42] methodology.

Campbell and Hendry [12] find similar results for the Canadian government bond market. They find that the information share in the 10-year spot note is below 50% in nearly all their sample of several months between 2002 and 2004. Upper and Werner [67] find that price discovery in the German Bund is dominated by the futures market, and in times of stress, like the 1998 Long Term Capital Management Crisis, the spot market information share falls to essentially zero. Upper and Werner [67], however, compare the futures market to the relatively illiquid, CTD bonds. This might explain their finding that the spot market does very little price discovery.

## Future Directions

This article has reviewed the microstructure of the US Treasury market. The Open Market Desk at the Federal Reserve Bank of New York plays a uniquely important role in the Treasury market by using transactions in those securities to adjust the level of bank reserves. Primary dealers are key players in both Treasury auctions and the Fed's open market operations. The Treasury market consists of several phases: when-issued, primary, on-the-run and off-the-run. Two ECNs, eSpeed and BrokerTec, intermediate the most active trading, during the on-the-run phase. The Treasury futures market at the CBOT complements trading in the spot market.

Treasury markets exhibit end-of-year, daily and intraday seasonality. Macro and Federal Reserve announcements are responsible for a substantial part of the daily and intraday seasonality. The literature studying the impact of order flows on Treasury prices has also considered how macro news and Federal Reserve actions influence such impact.

The futures markets in Chicago play an important role in price discovery, and a discussion of Treasury microstructure needs to take this into account. Both spot and futures markets are quite resilient and recent research on the Treasury ECNs suggest that the market continues to become more liquid. Fleming and Mizrach [30] report that volume has increased almost 5 times since 2001. This increase in trading volume accompanies a decline



in the importance of the primary dealers. The Financial Times reported in March 2007 that hedge funds accounted for 80% of trading activity in the Treasury market with only a 20% share for the primary dealers. One large fund alone, Citadel, accounts for 10% of the trading volume on eSpeed and BrokerTec. It was perhaps inevitable that trading by the millisecond would come to the Treasury market as it did to equities and foreign exchange. Perhaps we should only be surprised that it took so long.

The Treasury market plays a central role in the credit market. Times of financial crisis highlight the Treasury market's role as a safe haven for investors both in the US and overseas. Treasury securities also serve as benchmarks for complex derivatives like mortgage backed securities and structured loans like collateralized debt obligations. The microstructure of the US Treasury market is fundamental to our understanding of the global financial markets.

## Bibliography

- Akhtar MA (1997) Understanding open market operations. Public Information Department, Federal Reserve Bank of New York, New York
- Amihud Y, Mendelson H (1991) Liquidity, maturity, and the yields on US treasury securities. *J Finance* 46:1411–25
- Andersen TG, Bollerslev T, Diebold FX, Vega C (2007) Real-time price discovery in stock, bond and foreign exchange markets. *J Int Econ* 73:251–77
- Balduzzi P, Elton EJ, Green TC (2001) Economic news and bond prices: Evidence from the US treasury market. *J Financial Quant Anal* 36:523–43
- Barclay MJ, Hendershott T, Kotz K (2006) Automation versus intermediation: Evidence from treasuries going off the run. *J Finance* 61:2395–2414
- Beber A, Brandt MW (2006) The effect of macroeconomic news on beliefs and preferences: Evidence from the options market. *J Monetary Econ* 53:1997–2039
- Bernanke BS (2005) Implementing monetary policy, remarks at the redefining investment strategy education symposium, Dayton. This is a public speech by Bernanke <http://www.federalreserve.gov/boarddocs/speeches/2005/20050330/default.htm>
- Bollerslev T, Cai J, Song FM (2000) Intraday periodicity, long memory volatility, and macroeconomic announcement effects in the us treasury bond market. *J Empir Finance* 7:37–55
- Boukus E, Rosenberg JV (2006) The information content of FOMC minutes. Working paper, Federal Reserve Bank of New York, New York
- Brandt MW, Kavajecz KA (2004) Price discovery in the us treasury market: The impact of orderflow and liquidity on the yield curve. *J Finance* 59:2623–2654
- Brandt MW, Kavajecz KA, Underwood SE (2007) Price discovery in the treasury futures market. *J Futur Mark* 27:1021–1051
- Campbell B, Hendry S (2007) Price discovery in canadian and US 10-year government bond markets. Working Paper 07–43, Bank of Canada, Ottawa
- Chirinko RS, Curran C (2006) Greenspan shrugs: Formal pronouncements, bond market volatility, and central bank communication. Presented at: The American Economic Association Meetings
- Chordia T, Sarkar A, Subrahmanyam A (2005) An empirical analysis of stock and bond market liquidity. *Rev Financial Stud* 18:85–129
- Christie-David R, Chaudhry M, Lindley JT (2003) The effects of unanticipated macroeconomic news on debt markets. *J Financial Res* 26:319–39
- Cook T, Hahn T (1989) The effect of changes in the federal funds rate target on market interest rates in the 1970s. *J Monetary Econ* 24:331–351
- de Goeij P, Marquering W (2006) Macroeconomic announcements and asymmetric volatility in bond returns. *J Bank Finance* 30:2659–2680
- Duffie D (1996) Special repo rates. *J Finance* 51:493–526
- Dungey M, McKenzie M, Smith V (2007) News, no-news and jumps in the us treasury market. Cambridge University (unpubl)
- Ederington LH, Lee JH (1993) How markets process information: News releases and volatility. *J Finance* 48:1161–91
- Ederington LH, Lee JH (1995) The short-run dynamics of the price adjustment to new information. *J Financial Quant Anal* 30:117–34
- Ederington LH, Lee JH (2001) Intraday volatility in interest-rate and foreign-exchange markets: ARCH, announcement, and seasonality effects. *J Futur Mark* 21:517–52
- Elton EJ, Green TC (1998) Tax and liquidity effects in pricing government bonds. *J Finance* 53:1533–1562
- Engle R, Granger C (1987) Co-integration and error correction representation, estimation and testing. *Econometrica* 55: 251–276
- Fabozzi FJ, Fleming MJ (2005) US treasury and agency securities. In: Fabozzi FJ (ed) *The Handbook of Fixed Income Securities*, 7th edn. McGraw Hill, New York, pp 229–250
- Federal Reserve Bank of New York, Markets Group (2007) Domestic open market operations during 2006. Federal Reserve Bank of New York. <http://www.ny.frb.org/markets/omo/omo2006.pdf>
- Flannery M, Protopapadakis A (1988) From T-bills to common stocks: Investigating the generality of intra-week return seasonalities. *J Finance* 43:431–450
- Fleming M (2002) Are larger treasury issues more liquid? Evidence from bill reopenings. *J Money Credit Bank* 34: 707–735
- Fleming M (2003) Measuring treasury market liquidity. *Fed Reserv Bank New York Econ Policy Rev* 9:83–108
- Fleming M, Mizrahi B (2008) The microstructure of a US treasury ECN: The BrokerTec platform. Working paper. Federal Reserve Bank of New York, New York
- Fleming M, Piazzesi M (2005) Monetary policy tick-by-tick. Working paper No ID. Federal Reserve Bank of New York, New York
- Fleming M, Remolona EM (1997) What moves the bond market? *Fed Reserv Bank New York Econ Policy Rev* 3:31–50
- Fleming M, Remolona EM (1999) Price formation and liquidity in the US treasury market: The response to public information. *J Finance* 54:1901–1915
- Fleming M, Remolona EM (1999) What moves bond prices? *J Portf Manag* 25:28–38

35. Garbade KD, Ingber JF (2005) The treasury auction process: Objectives, structure, and recent adaptations. *Fed Reserv Bank New York Curr Issues Econ Finance* 11:1–11
36. Goldreich D, Hanke B, Nath P (2005) The price of future liquidity: Time-varying liquidity in the US treasury market. *Rev Finance* 9:1–32
37. Green TC (2004) Economic news and the impact of trading on bond prices. *J Finance* 59:1201–1233
38. Griffiths M, Winters D (1997) On a preferred habitat for liquidity at the turn-of-the-year: Evidence from the term-repo market. *J Financial Serv Res* 12:21–38
39. Griffiths M, Winters D (2005) The turn-of-the-year in money markets: Tests of risk-shifting window dressing and preferred habitat hypotheses. *J Bus* 78:1337–1364
40. Griffiths M, Winters D (2005) The year-end price of risk in a market for liquidity. *J Invest Manag* 3:99–109
41. Hamilton JD (2008) Assessing monetary policy effects using daily fed funds futures contracts. *Fed Reserv Bank St Louis Rev* 90:377–393
42. Harris F, McInish T, Wood R (2002) Security price adjustment across exchanges: An investigation of common factor components for dow stocks. *J Financial Mark* 5:277–308
43. Harvey CR, Huang RD (2002) The impact of the federal reserve bank's open market operations. *J Financial Mark* 5:223–57
44. Hasbrouck J (1991) Measuring the information content of stock trades. *J Finance* 46:179–207
45. Heuson AJ, Su T (2003) Intra-day behavior of treasury sector index option implied volatilities around macroeconomic announcements. *Financial Rev* 38:161–77
46. Huang RD, Cai J, Wang X (2002) Information-based trading in the treasury note interdealer broker market. *J Financial Intermed* 11:269–296
47. Huang X (2006) Macroeconomic news announcements, financial market volatility and jumps. Duke University (unpubl)
48. Jiang GJ, Lo I, Verdelhan A (2007) Why do bond prices jump? A study of the US treasury market. Eller College of Management, University of Arizona (unpubl)
49. Jones CM, Lamont O, Lumsdaine RL (1998) Macroeconomic news and bond market volatility. *J Financial Econ* 47:315–337
50. Johnston E, Kracaw W, McConnell J (1991) Day-of-the-week effects in financial futures: An analysis of GNMA, T-bond, T-note, and T-bill contracts. *J Financial Quant Anal* 26:23–44
51. Kamara A (1994) Liquidity, taxes, and short-term treasury yields. *J Financial Quant Anal* 29:403–417
52. Kearney AA (2004) The changing impact of employment announcements on interest rates. *J Econ Bus* 54:415–429
53. Kuttner KN (2001) Monetary policy surprises and interest rates: Evidence from the fed funds futures market. *J Monetary Econ* 47:523–544
54. Lahaye J, Laurent S, Neely CJ (2007) Jumps, cojumps and macro announcements. Working Paper 2007–032A, Federal Reserve Bank of St. Louis, St. Louis
55. Madhavan A, Richardson M, Roomans M (1997) Why do securities prices change? A transaction-level analysis of NYSE stocks. *Rev Financial Stud* 10:1035–1064
56. Menkveld AJ, Sarkar A, Van der Wel M (2006) Customer flow, intermediaries, and the discovery of the equilibrium riskfree rate. Working paper. Federal Reserve Bank of New York, New York
57. Meulendyke AM (1998) US monetary policy & financial markets. Federal Reserve Bank of New York, New York
58. Mizrach B, Neely CJ (2006) The transition to electronic communication networks in the secondary treasury market. *Fed Reserv Bank St. Louis Rev* 88:527–541
59. Mizrach B, Neely CJ (2008) Information shares in the US treasury market. *J Bank Finance* 32:1221–1233
60. Neely CJ, Winters DB (2006) Year-end seasonality in one-month LIBOR derivatives. *J Derivatives* 13:47–65
61. Nyborg KG, Sundaresan S (1986) Discriminatory versus uniform treasury auctions: Evidence from when-issued transactions. *J Financial Econ* 42:63–104
62. Poole W, Rasche RH (2000) Perfecting the market's knowledge of monetary policy. *J Financial Serv Res* 18:255–298
63. Poole W, Rasche RH, Thornton DL (2002) Market anticipations of monetary policy actions. *Fed Reserv Bank St. Louis Rev* 84:65–94
64. Seligman J (2006) Does urgency affect price at market? An analysis of US treasury short term finance. *J Money Credit Bank* 38:989–1012
65. Simon DP (1991) Segmentation in the treasury bill market: Evidence from cash management bills. *J Financial Quant Anal* 26:97–108
66. Simon DP (1994) Further evidence on segmentation in the treasury bill market. *J Bank Finance* 18:139–151
67. Upper C, Werner T (2002) Tail wags dog? Time-varying information shares in the bund market. Working Paper 24/02, Bundesbank, Frankfurt
68. Vayanos D, Weill P (2008) A search-based theory of the on-the-run phenomenon. *J Finance* 63:1361–1398
69. Warga A (1992) Bond returns, liquidity, and missing data. *J Financial Quant Anal* 27:605–17

## Tsunami Earthquakes

JASCHA POLET<sup>1</sup>, H. KANAMORI<sup>2</sup>

<sup>1</sup> Geological Sciences Department, California State Polytechnic University, Pomona, USA

<sup>2</sup> Seismological Laboratory, Caltech, Pasadena, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Characteristics of Tsunami Earthquakes](#)

[Factors Involved in the Seismogenesis and](#)

[Tsunamigenesis of Tsunami Earthquakes](#)

[A Model for Tsunami Earthquakes](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

***m<sub>b</sub>*** body wave magnitude, based on the amplitude of the direct P wave, period of the measurement: 1.0–5.0 s. Also see: Seismic Magnitude.

**$M_S$**  surface wave magnitude, based on the amplitude of surface waves, period of the measurement: 20 s. Also see: Seismic Magnitude.

**$M_w$**  moment magnitude, determined from the seismic moment of an earthquake, typical period of the measurement: > 200 s. Also see: Seismic Magnitude.

**Magnitude saturation** due to the shape of the seismic source spectrum, relatively short period measurements of seismic magnitude will produce similar magnitudes for all earthquakes above a certain size. The value of this threshold earthquake size depends on the period of the measurement: magnitude measurements using shorter period waves will saturate at lower values than magnitude measurements using longer period waves.  $M_w$  will not saturate.

**Run-up height** difference between the elevation of maximum tsunami penetration (inundation line) and the sea level at the time of the tsunami.

**Tsunami earthquake** an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude.

**Tsunami magnitude** a scale for the relative size of tsunamis generated by different earthquakes,  $M_t$  in particular is calculated from the logarithm of the maximum amplitude of the tsunami wave measured by a tide gauge distant from the tsunami source, corrected for the distance to the source (also see: Satake, this volume).

**Seismic magnitude** a scale for the relative size of earthquakes. Many different scales have been developed, almost all based on the logarithmic amplitude of a particular seismic wave on a particular type of seismometer, with corrections for the distance between source and receiver. These measurements are made for different wave types at different frequencies, and thus may lead to different values for magnitude for any one earthquake.

**Seismic moment** the product of the fault surface area of the earthquake, the rigidity of the rock surrounding the fault and the average slip on the fault.

## Definition of the Subject

The original definition of “tsunami earthquake” was given by Kanamori [37] as “an earthquake that produces a large size tsunami relative to the value of its surface wave magnitude ( $M_S$ )”. Therefore, the true damage potential that a tsunami earthquake represents may not be recognized by conventional near real-time seismic analysis methods and may only become apparent upon the arrival of the

tsunami waves on the local coastline. Although tsunami earthquakes occur relatively infrequently, the effect on the local population can be devastating, as was most recently illustrated by the July 2006 Java tsunami earthquake. This event (moment magnitude  $M_w = 7.8$ ) was quickly followed by tsunami waves two to seven meters high, traveling as far as two kilometers inland and killing at least 668 people ([http://www.searo.who.int/en/Section23/Section1108/Section2077\\_11956.htm](http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm)).

It is important to note that the definition of “tsunami earthquake” is distinct from that of “tsunamigenic earthquake”. A tsunamigenic earthquake is any earthquake that excites a tsunami. Tsunami earthquakes are a specific subset of tsunamigenic earthquakes, which we will later in this article more precisely define as earthquakes that directly cause a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from their seismic moment magnitude.

## Introduction

Shallow oceanic earthquakes may excite destructive tsunamis. Truly devastating tsunamis occur only infrequently, but as the natural disaster of the tsunami following the 2004 Sumatra–Andaman Islands earthquake has shown, may cause widespread damage (in this case in the region of the Indian Ocean) and lead to hundreds of thousands of casualties. In general, tsunamis are caused by shallow earthquakes beneath the ocean floor displacing large volumes of water. Thus, the magnitude of the earthquake plays an important role in determining its tsunamigenic potential. However, a particular subclass of shallow subduction zone earthquakes: “tsunami earthquakes”, poses a special problem.

For the purpose of this article, we will define the term “tsunami earthquake” as follows: an earthquake that directly causes a regional and/or teleseismic tsunami that is greater in amplitude than would be expected from its seismic moment magnitude. With this definition we specifically exclude seismic events that were followed by tsunamis directly caused by slides or slumps resulting from the original earthquake (as was the case for the 1992 Flores [28,32] and the 1998 Papua New Guinea earthquakes [22,73] for example). We further exclude events that only very locally caused large tsunamis as a result of, for example, focusing effects due to features of the ocean floor bathymetry (e.g. [63]) or directivity effects combined with the shape of the coastline, as was the case for the tsunamis that hit the harbor of Crescent City after the 1964 Aleutian Islands earthquake [12] and the November 15, 2006 Kurile Island event (<http://www.usc>).

[edu/dept/tsunamis/california/Kuril\\_2006/](http://edu/dept/tsunamis/california/Kuril_2006/)). Furthermore, this definition compares the size of the tsunami with the moment magnitude of the earthquake and not its surface wave magnitude, slightly modifying the definition given by Kanamori [37], in order to exclude great events for which the surface wave magnitude saturates.

Our primary objective in this article is to describe the characteristics of tsunami earthquakes and the possible factors involved in the anomalously strong excitation of tsunamis by these events. We will also discuss a possible model for these infrequent, but potentially very hazardous, events. The earthquakes listed in Table 1 and plotted in Fig. 1 are considered tsunami earthquakes, according to our modified definition presented in the previous paragraph, by the majority of the community of earthquake and tsunami researchers. However, we note that the interpretation of the 1994 Java and the 1946 Aleutian Islands earthquakes varies with investigators. The 1994 Java earthquake occurred off the southeastern coast of this island, near the east end of the Java Trench in the Indian Ocean, at 1:18 am local time. It generated a devastating tsunami that took the lives of more than 200 East Java coastal residents. Run-up measured along the southeastern Java coast ranged from 1 to 14 m, while run-up measured along the southwestern coast of Bali ranged from 1–5 m [74,82]. Although the anomalously high tsunami excitation of the 1994 event is not in doubt, its earthquake source characteristics have been debated [3,59]. The 1946 Aleutian Islands earthquake off Unimak Island produced one of the largest trans-Pacific tsunamis and had a tsunami magnitude of 9.3 [1], but its moment magnitude is only  $M_w = 8.2$ , making it a tsunami earthquake [36]. Some of the great tsunami heights measured (exceeding 30 m in height on Unimak Island and 16 m in run-up at the Hawaiian islands; [78]) can be attributed to slumping [19]. However, its anomalously high tsunamis are probably primarily due to the seismic source directly [47].

The other, less controversial, tsunami earthquakes listed in our table are: the 1896 Sanriku event near the coast of Japan, two events near the Kurile Islands: one in 1963 and the other in 1975, the 1992 Nicaragua earthquake, the Peru earthquake 4 years later, as well as an earlier event in this region in 1960 and most recently the 2006 Java earthquake. The June 15, 1896 Sanriku earthquake generated devastating tsunamis with a maximum run-up of 25 m and caused the worst tsunami disaster in the history of Japan with over 20,000 deaths, despite its moderate surface wave magnitude ( $M_S = 7.2$ ) and weak seismic intensity [2,26,76]. The November 20, 1960 Peru earthquake excited a tsunami that was anomalously large for an earthquake of moderate magnitude [57], resulting in 66 fatalities

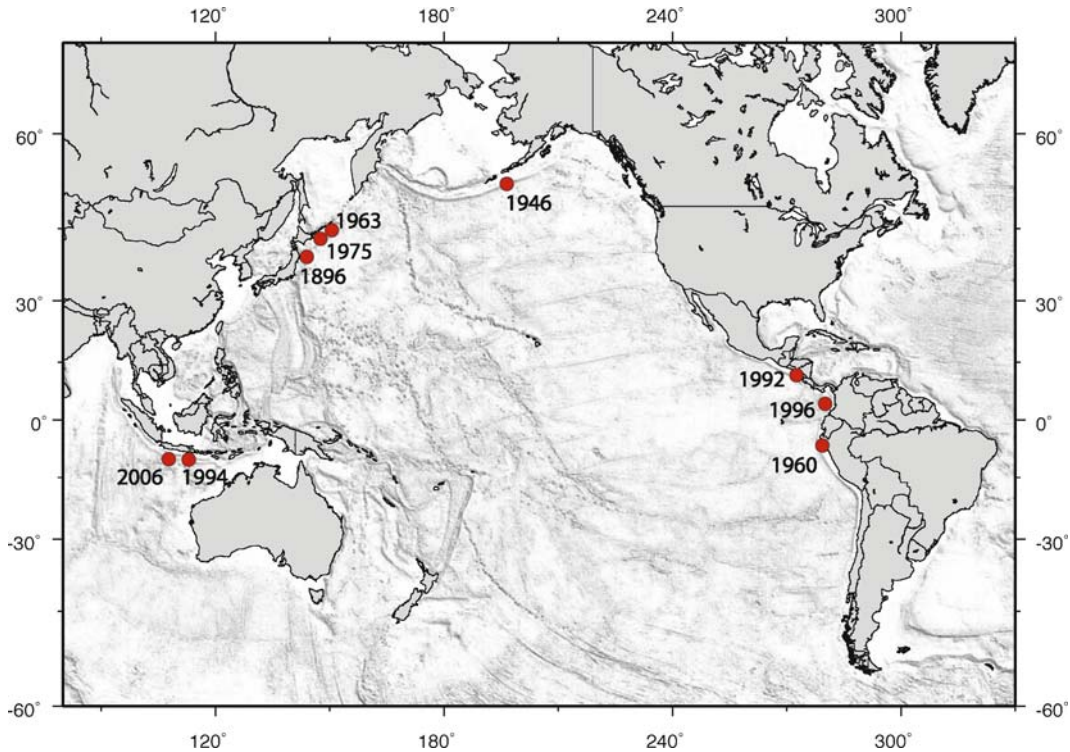
#### Tsunami Earthquakes, Table 1

Tsunami Earthquakes (references for most entries are listed in main text, others are from the National Geophysical Data Center Global Tsunami Database ([http://www.ngdc.noaa.gov/seg/hazard/tsu\\_db.shtml](http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml)) and the Centennial Earthquake Catalog [18])

Date	Geographical Region	$M_w$	$m_b$	$M_S$	$M_t$	Deaths
1896/06/15	Japan			7.2	8.0	26360
1946/04/01	Aleutian Islands	8.2		7.3	9.3	165
1960/11/20	Peru	7.6	7.0	7.0		66
1963/10/20	Kurile Islands	7.8	7.1	7.2		
1975/06/10	Kurile Islands	7.5	5.6	7.0		
1992/09/02	Nicaragua	7.7	5.4	7.2		179
1994/06/02	Java	7.8	5.7	7.1		250
1996/02/21	Peru	7.5	5.8	6.6		12
2006/07/17	Java	7.7	6.2	7.2		668

ties (from the tsunami event database of the National Geophysical Data Center, <http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7>). The October 20, 1963 Kurile earthquake was an aftershock to the great Kurile Islands underthrusting earthquake ( $M_w = 8.5$ ) of October 13, 1963 and produced a maximum run-up height of 10–15 m at Urup Island, much larger than the height of the main shock tsunami of 5 m [2]. The 1975 earthquake occurred south of the Kurile Islands and was weakly felt along the entire southern part of the Kurile Islands. Like the 1963 tsunami earthquake, this event can be considered an aftershock [21] of a larger event ( $M_S = 7.7$ ) that occurred essentially at the same location on June 17, 1973. The maximum run-up height was 5 m on Shikotan Island, while the main shock had a run-up height measured at 4.5 m. A fairly strong tsunami was also recorded on tide gauges in Alaska and Hawaii (from the tsunami event database of the National Geophysical Data Center <http://www.ngdc.noaa.gov/nndc/struts/form?t=101650&s=70&d=7>). After a time period of almost two decades without a significant tsunami earthquake, the 1992 Nicaragua earthquake was the first tsunami earthquake to be captured by modern broadband seismic networks. This tsunami caused 179 deaths (from the Emergency Disasters Database, <http://www.em-dat.net/disasters/list.php>) and significant damage to the coastal areas of Nicaragua and Costa Rica, reaching heights of up to 8 m. The 1996 Peru earthquake struck at 7:51 am local time, approximately 130 km off the northern coastal region of Peru. Approximately one hour after the main shock, a damaging tsunami reached the Peruvian coast, with run-up heights of 1 to 5 meters along a coastline of 400 km [27], resulting in twelve deaths [11]. Finally,





**Tsunami Earthquakes, Figure 1**

Map of tsunami earthquakes (listed in Table 1). Location for 1896 earthquake from [2] and for 2006 earthquake from the Global CMT catalog. All other earthquake locations from the Centennial Earthquake Catalog [18]

the 2006 Java earthquake was located only about 600 km west-northwest of the tsunami earthquake that occurred 12 years earlier in the same subduction zone. The Ministry of Health reported that approximately 668 people died and 65 are missing ([http://www.searo.who.int/en/Section23/Section1108/Section2077\\_11956.htm](http://www.searo.who.int/en/Section23/Section1108/Section2077_11956.htm)) due to a tsunami that had a maximum run-up height of 15.7 m along the coast of central Java [45].

Several other earthquakes in the last few years have produced damaging tsunamis and have been mentioned as possible tsunami earthquakes. Seno and Hirata [69] suggest that the great 2004 Sumatra–Andaman earthquake also likely involved a component of tsunami earthquakes, because tsunamis larger than expected from seismic slip occurred, possibly due to slow slip in the shallow subduction boundary. It has also been proposed that the Kurile Islands earthquake of November 15, 2006 [35] may have exhibited some characteristics of tsunami earthquakes and, even more recently, the Solomon Island earthquake of April 1, 2007 excited large tsunamis, at least locally (<http://soundwaves.usgs.gov/2007/04/>). However, the disparity between seismic and tsunami excitation by these events is

not nearly as large as for the events in Table 1, and we do not list these events as tsunami earthquakes.

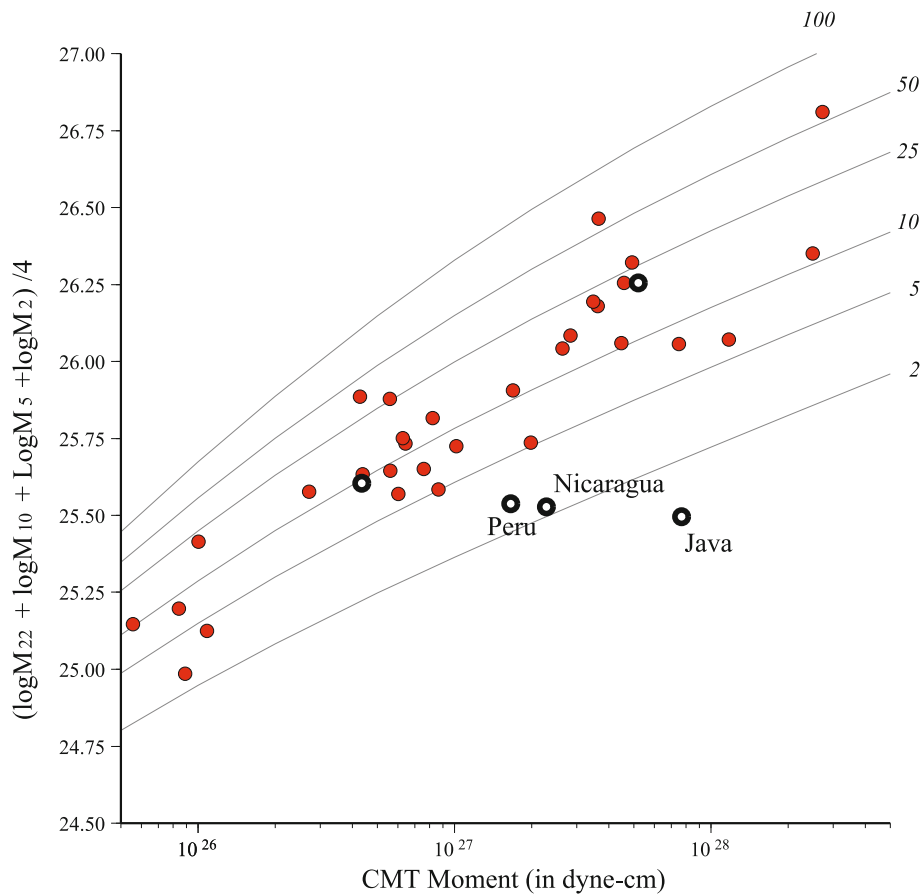
### Characteristics of Tsunami Earthquakes

Fortunately, tsunami earthquakes occur only infrequently. Fewer than ten of these events have occurred in the last three decades since the global installation of seismic broadband instruments and tide gauges and easy availability of their data were established. However, from the detailed investigations of the most recent events and comparisons with the limited data available for the older earthquakes in Table 1, several characteristics of these earthquakes clearly emerge.

#### Slow Character

The slow character of tsunami earthquakes manifests itself in several different, yet related, ways. One well-established characteristic of tsunami earthquakes is the discrepancy between the determined values of the different seismic magnitude types, calculated from various kinds of seismic waves or waves of different frequency ranges.



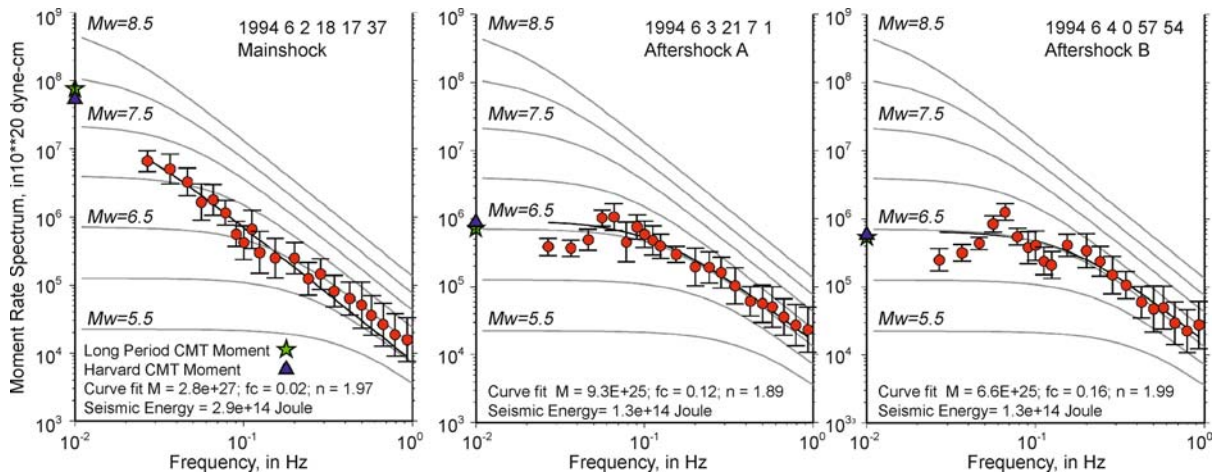


**Tsunami Earthquakes, Figure 2**

Average of log of moment rate spectrum at four periods (2, 5, 10, 22 s) as a function of the seismic moment as determined by CMT inversion of long-period surface waves. Reference curves were calculated for an  $\omega^2$  model [13]. Values next to the grey curves indicate the stress drop used to calculate the reference curve. Events shown are all shallow subduction zone earthquakes from 1992 to 1996 with moment magnitude 7.0 or greater. Earthquakes followed by anomalously large tsunamis are indicated with open circles. Of these events, only the 1992 Nicaragua, 1994 Java and 1996 Peru earthquakes are slow tsunami earthquakes, as is shown in this figure by their relatively low moment rate spectrum at shorter periods. Adapted from [59]

A typical comparison is that of the body wave magnitude,  $m_b$ , or the surface wave magnitude,  $M_S$ , with the moment magnitude of the earthquake,  $M_w$ . For tsunami earthquakes  $m_b$  is typically much smaller than the other two magnitudes,  $M_S$  and  $M_w$ , and  $M_w$  typically exceeds  $M_S$ . The discrepancy between these different magnitudes is more pronounced than for regular subduction zone earthquakes with similar moment magnitudes. For example: for the 1992 Nicaragua earthquake:  $m_b = 5.4$ ;  $M_S = 7.2$ ,  $M_w = 7.7$  [18]; for the 1994 Java tsunami earthquake:  $m_b = 5.7$ ;  $M_S = 7.1$ ,  $M_w = 7.8$  [59]; for the 2006 Java earthquake:  $m_b = 6.2$ ;  $M_S = 7.2$ ,  $M_w = 7.7$  [5]. Since the body wave magnitude is calculated from short period P-waves, the surface wave magnitude is determined by the amplitude of surface waves with a period of 20

seconds and the moment magnitude is generally based on longer periods for big events, this consistent discrepancy is an indication of the relatively greater seismic energy release at longer periods (or the “slow” character) of these tsunami events. Similarly, investigations of teleseismic P-waves [5,59] have shown that their source spectra are depleted in high frequency energy at periods shorter than 20 seconds as compared to other shallow subduction zone earthquakes (see Fig. 2, from [59]), as well as their own aftershocks (see Fig. 3, adapted from [60]). Modeling of the rupture processes shows that the rupture velocities for tsunami earthquakes are slower than for most other subduction zone earthquakes (for several events: [58] and [33]; for Aleutians 1946: [47]; for Nicaragua 1992: [40] and [30]; for Peru 1996: [31]; for



**Tsunami Earthquakes, Figure 3**

Moment rate spectra for the Java thrust mainshock (left) and two of its largest aftershocks, tensional events in the outer rise (right panels). The stars indicate the Harvard CMT moment, the triangles the moment determined using very long period surface waves. Grey reference curves were calculated for an  $\omega^2$  model [13] with a stress drop of 30 bars and an S-wave velocity of 3.75 km/s. The moment rate for the main shock is similar to that of its much smaller (in terms of  $M_w$ ) aftershocks for periods shorter than 10 seconds. Adapted from [60]

Java 2006: [5]). Correspondingly, the centroid times and source durations or rise times determined for these events are also relatively large with respect to other large subduction zone earthquakes ([59]; for Kurile Islands 1975: [71]; for Peru 1960: [57]; for Java 2006: [5] and [25]), although they may not be anomalous relative to other, smaller, subduction zone earthquakes at very shallow depth [9]. The energy that is radiated by these slow rupture processes is also anomalously low, as is shown by analyses of the radiated energy to moment ratio (for several recent tsunami earthquakes: [51]; for the 1946 Aleutian Islands earthquake: [47]) and radiation efficiency [83].

Unfortunately, the slow character of tsunami earthquakes also means that local residents are not warned by strong ground shaking of the possibility of an impending tsunami. Field surveys and first-person accounts describe the motion of tsunami earthquakes more as a weak “rolling motion” than the usual impulsive character of local events. In the case of the Nicaragua earthquake, some felt a very feeble shock before the tsunami, but most did not feel the earthquake at all [30]. For the 1994 Java event earthquake-induced ground shaking was not noticed by the coastal residents interviewed in Bali and Java [74]. Interviews with local residents carried out for the 2006 Java earthquake [50] also indicate that they felt little or no shaking.

Most designs for tsunami earthquake discriminators and early warning systems make use of a number of the manifestations of the unusually slow character of tsunami

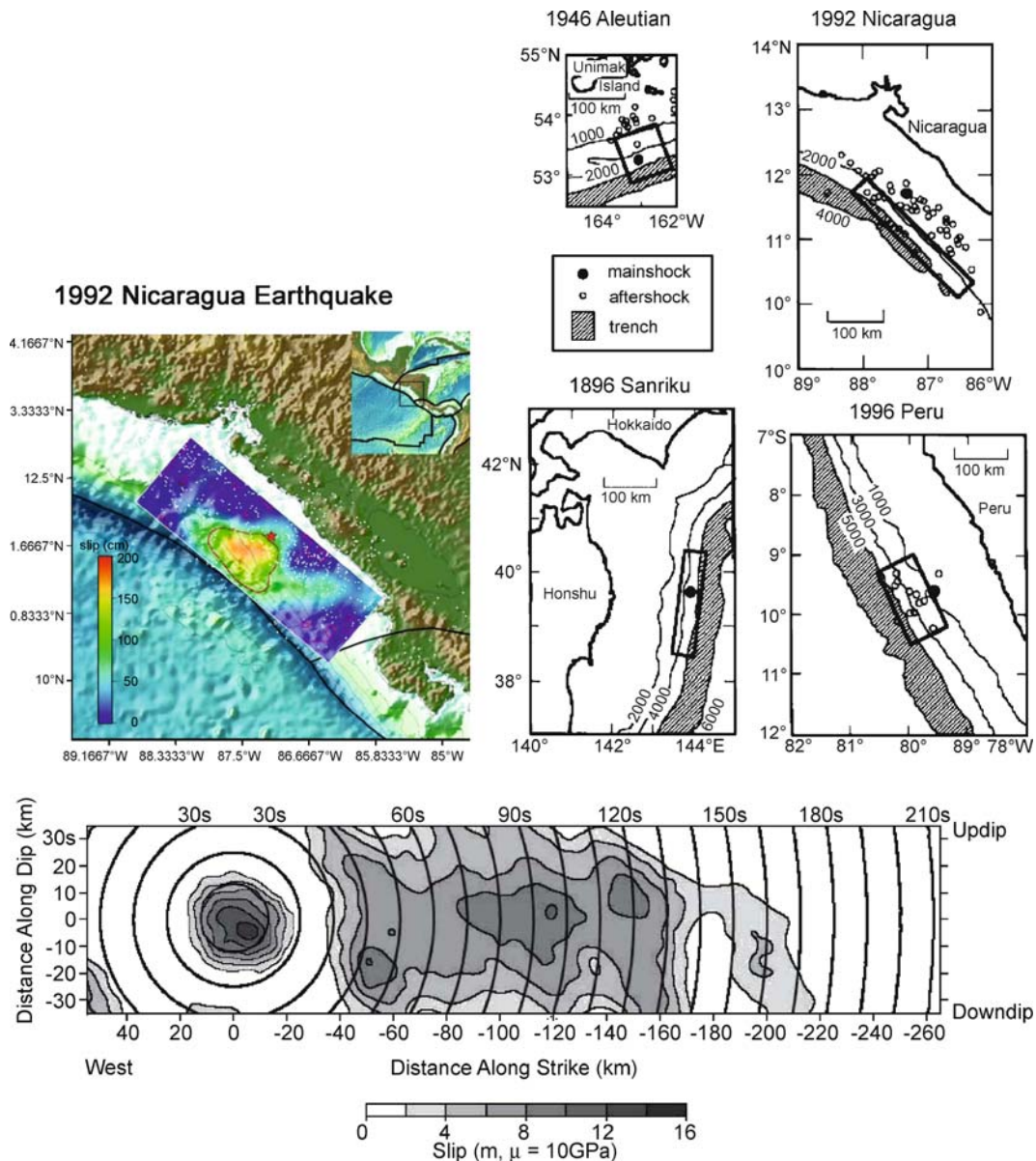
earthquakes listed above and many incorporate the use of long period seismic waves for robust estimation of the size of the event. For example, the pulse width of the P-wave, used to calculate moment magnitude  $M_{wp}$ , can give an accurate estimate of source duration time. The combination of  $M_{wp}$  and the source duration can provide an effective tool to issue early tsunami warnings [81]. A slightly later arrival on a seismogram, the W-phase, is a distinct ramp-like long-period (up to 1000 s) phase that begins between P and S waves on displacement seismograms and is particularly pronounced for slow earthquakes; thus it can be used for identification of these types of events [38]. Another method for fast regional tsunami warning uses the ratio of the total seismic energy to the high-frequency energy (between 1 and 5 Hz), computed from the seismograms [70]. Similarly, the detection of deficient values of seismic energy-to-seismic moment ratio can be accomplished in automated, real-time mode [51].

### Location: Close to Trench

The hypocenters of the recent tsunami earthquakes are located relatively close to the trench, as compared to regular subduction zone earthquakes. The Global CMT and other [59] centroid locations for several of these events are located even on the seawards side of the trench (also see Fig. 4). It may be possible that the inversion process mislocates the centroid of the event due to the unusually long duration of the seismic source for its moment

and thus its unusually late centroid time. Inversions using seismic and/or tsunami waveforms and other waveform investigations for the 1896 Sanriku event [76], the 1946 Aleutian Islands earthquake [36,78], the 1960 Peru earthquake [58], the 1963 Kurile earthquake [7,58,86], the 1975 Kurile earthquake [58,86], the 2006 Java earthquake [5,20], the Nicaragua earthquake [40,41,62] and the

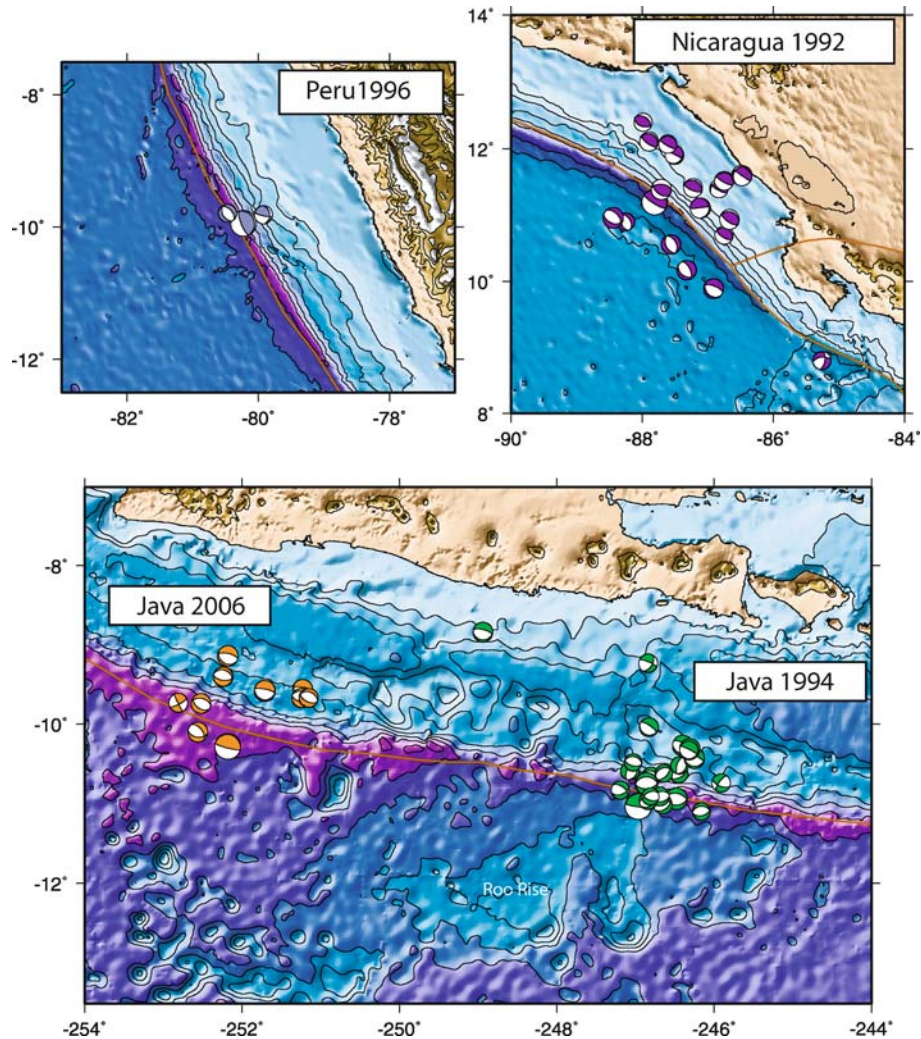
1996 Peru event [31,64] indicate the presence of concentrated slip in a narrow region near the trench (see Fig. 4). Although in many of these inversions only 1-D Green's functions were used, preliminary research results using a more realistic velocity model for the shallow subduction zone [54] shows a similar picture for the 2006 Java tsunami earthquake.



**Tsunami Earthquakes, Figure 4**

Slip or fault models determined for various tsunami earthquakes. The models shown are for: Nicaragua 1992 (top left, from [33]) Java 2006 (bottom, from [5]) and four tsunami earthquakes (top right, from [64]; the main shock and aftershock epicenters are shown and the hatched area indicates the trench). These models all show the presence of substantial slip close to the trench





**Tsunami Earthquakes, Figure 5**

Aftershock sequences (seismicity in the region of the main shock within a period of 4 years after its occurrence) for several tsunami earthquakes, from the Global CMT catalog. Relatively few interplate aftershocks occur on the fault plane after a tsunami earthquake occurs, but a preponderance of normal-fault, probably intraplate, aftershocks is apparent

### Aftershocks

The aftershock sequences of (recent) tsunami earthquakes are unusual in the preponderance of events not located on the interface between overriding and subducting plates [59]. Some of these aftershocks are located in the outer rise according to their Global CMT centroid locations and, in the case of the Java 2006 aftershocks, relocations using a 2.5-D model of the subduction zone [54] confirm this location. Others are located in the overriding plate ([8] for the 2006 Java earthquake), with some deeper within the accretionary prism (for the Java 2006

earthquake: [54]). The low number, or non-existence, of large (greater than magnitude 5.5) interplate earthquakes suggests that the main shock almost completely relieved the stress on the interface or may be related to the frictional properties of the fault. Several explanations have been proposed for the anomalously high number of intraplate earthquakes following tsunami earthquakes. Because of the proximity of the areas of high slip to the trench, the stress-change in the outer rise and trench area due to a tsunami earthquake are greater than for the “standard” subduction zone earthquake. Several modeling studies of stress alterations caused by large subduction earthquakes

suggest that the subduction slip will act to increase the tensional stress and favor normal events in zones towards the ocean from the upper limit of the rupture [16,80]. The subducting plate also seems to have a highly-broken up or rough character in many areas in which tsunami earthquakes have occurred ([49,75]; for Java: [48]; for Peru: [29] and [44]), which suggests the presence of more pervasive pre-existing weak zones, due to, for example, seafloor spreading related fabric. These weak zones may be re-activated in outer rise, or deeper intraplate, earthquakes following a tsunami earthquake.

### Factors Involved in the Seismogenesis and Tsunamigenesis of Tsunami Earthquakes

Based on the consistent characteristics of tsunami earthquakes, as described in the previous section, and observations of their tectonic environments, hypotheses have been developed as to the cause of their extraordinary tsunami excitation and unusual seismic source process. In this section, we will document the factors that have been proposed to affect the seismo- and tsunami-genesis of tsunami earthquakes. Some are associated with the numerical prediction of tsunami wave-heights based on observed seismic waveforms, and others with possible unusual conditions of the tectonic environment in which these events occur; many of these factors are closely or at least somewhat related.

#### Slow Character May Lead to Underestimation of Earthquake Size

Prior to the installation of the broadband Global Seismic Network, the magnitudes of earthquakes were often determined from the amplitude of their teleseismic P-waves only. In the case of tsunami earthquakes, using this technique to determine their magnitude would lead to an initial underestimation of their true size, since their source spectra are depleted in the relatively high frequency energy that usually dominates the direct P-wave signals of regular earthquakes [59]. A similar issue would occur, although to a lesser degree, when using surface waves of periods of 20 seconds to determine the surface wave magnitude of these slow events [58]. With the advent of broadband sensors in the past several decades, it has now however become possible to investigate seismic waves to very long periods, hundreds or even thousands of seconds. Using more sophisticated techniques and the waveforms from technologically advanced sensors, the source spectrum of the recent tsunami earthquakes can now be modeled to these very long periods. Thus, no long period seismic energy that would excite tsunami waves should be “hid-

den” from the view of seismologists in the computation of moment magnitude or full rupture models using long period surface and body waves. However, for the earthquakes discussed in this paper, the observed tsunami are still larger than would be expected, even for their moment magnitude.

#### Effect of the Presence of Weak Materials with Low Shear Modulus

Most earthquake source inversions (either for full rupture or Centroid Moment Tensor parameters) implement a simple one-dimensional velocity and rigidity model to compute synthetic seismograms and model the recorded waveforms. If tsunami earthquakes are unusually shallow and/or involve sediments of low seismic wave speed, this is probably a poor approximation of the actual structure near the source region [23,64]. Some authors have attempted to rectify this error by using moment or slip distributions determined by seismic inversions in a structural model with significantly reduced shear modulus, more appropriate for the shallow trench region, and forward modeling the tsunami waves (e.g. [23]). This approach, however, is not satisfactory because a seismic inversion for moment or slip using this more appropriate rigidity model would produce a different distribution of moment or slip. Thus, a simple “correction” for the use of an inappropriate value of rigidity cannot be carried out after the rupture model has already been computed. If available, a correct rigidity model should be part of the seismic inversion process itself. To do this correctly, Green’s functions should be computed for a three-dimensional (or possibly two dimensional, if the velocity structure is relatively uniform in the trench-parallel direction) velocity structure of the shallow subduction zone and incorporated in the modeling of the seismic waveforms. However, such sophisticated models are currently only available for very few subduction zones and the computational power required for these calculations (for body wave frequencies) is substantial. Unfortunately, whichever approach is chosen to go from recordings of seismic waveforms of tsunami earthquakes to the prediction of tsunami wave heights, a good model of the velocity and elastic properties of the shallow subduction zone is an unavoidable requirement.

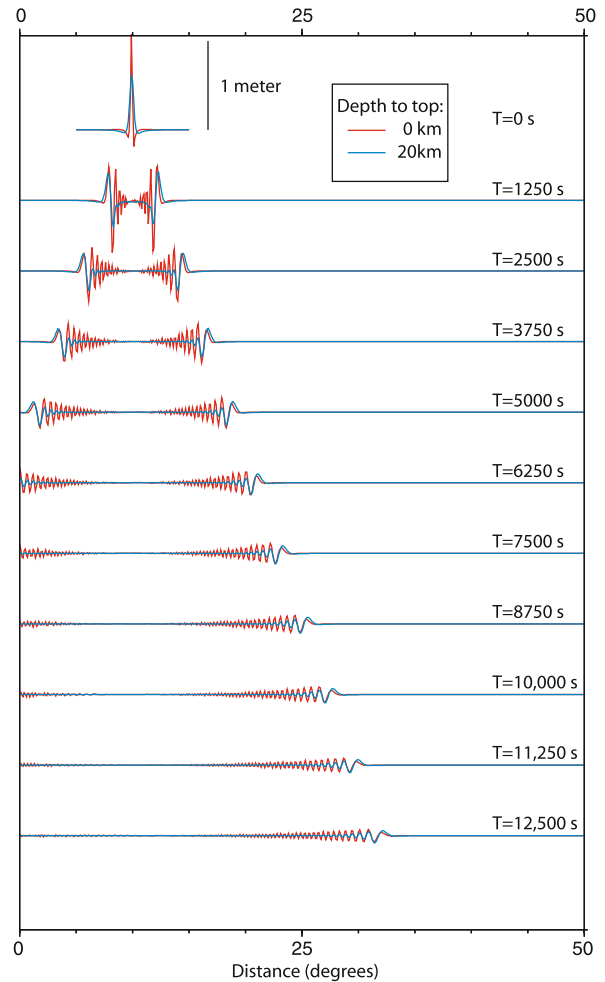
Even when the moment distribution or moment magnitude of a tsunami earthquake has been determined using an appropriate model for the velocity and elastic parameters, there exists also the issue of enhanced tsunami excitation in material with weaker elastic properties, such as sedimentary layers. Modeling suggests that an event for which 10% of the moment is in sediments generates



a tsunami 10 times larger than its seismic moment would suggest [52] mainly because the slip in this material would be much greater than that for the same seismic moment in a stronger material (moment is the product of slip, area and rigidity after all). Therefore, the moment of a tsunami earthquake, even if determined correctly, may not directly reflect its tsunamigenic potential when low velocity sediments are present in the rupture zone. Since tsunami wave heights are mainly determined by the vertical displacement of the ocean floor, which in turn is primarily controlled by the slip on the fault plane, the slip (distribution) is more directly indicative of tsunamigenic potential. Since variations in shear modulus of a factor of five are not uncommon in shallow subduction zones [23], earthquakes with similar moments can result in substantially different slip models and tsunami excitation.

### Shallow Depth of Slip Causes Relatively Great Displacement of Ocean Floor

Shallower earthquakes produce greater and shorter wavelength vertical displacement of the ocean floor, and thus greater and shorter wavelength tsunami waves right above the source region. However, higher frequency waves travel more slowly than longer period waves and, after a few hundred or thousand kilometers of travel, they drift to the back of the wave train and do not contribute to the maximum amplitude. Beyond about 2000 km distance, any earthquake at a depth less than 30 km appears to be equally efficient in tsunamigenesis ([85] and Fig. 6). Therefore, the exact depth of the slip in a shallow earthquake is not a key factor in determining its teleseismic tsunami wave-heights. Although the teleseismic tsunami wave-heights for a shallow slip event may not be significantly greater in amplitude than those for a somewhat deeper slip event [85], at local and regional distances the depth of the slip is an important factor. Therefore, for tsunami earthquakes, which have anomalously great tsunami height at mainly local and regional distances close to the rupture, the depth of the rupture should be a significant factor in their tsunamigenesis (see Fig. 6). Furthermore, modeling of shallow subduction zone earthquakes using a specific crack model [24] indicate that a rupture intersecting the free surface results in approximately twice the average slip. However, under the assumption of other specific frictional and crack models, the modeling of subduction zone earthquakes by Wang and He [84] produces slip models that have less vertical displacement of the ocean floor when the slip reaches the surface, due to a different slip distribution and the curvature of the subduction interface.



**Tsunami Earthquakes, Figure 6**

Cross sections of an expanding tsunami from a M7.5 thrust earthquake. The fault strikes north south (into the page) and the sections are taken east west. Elapsed time in seconds is given at the left and right sides. Red lines are for a fault that breaks the surface and blue lines for a fault with its top at a depth of 20 km. Deeper earthquakes make smaller and longer wavelength tsunamis at relatively short distances

### Shallow Fault Dip May Lead to Underestimation of Slip from Seismic Waves

It is notoriously difficult to resolve moment,  $M_0$ , and dip,  $\delta$ , independently for shallow thrust earthquakes. The excitation functions of Rayleigh waves for shallow thrust earthquakes show that only the product of dip and moment (more precisely  $M_0 \sin(2\delta)$ ) can be resolved [39] when using these surface waves in inversions for a source mechanism. If Love wave data are also included in the inversion, it may be possible to add constraints by concentrating on fitting the amplitudes of those Love waves

recorded at azimuths corresponding to the along strike direction [34], however, this approach may be complicated by directivity effects. The polarity of body waves can only be used to constrain the focal mechanism of an earthquake at a limited range of incidence angles and thus also cannot provide any additional constraints on the dip of the shallowly dipping plane, unless assumptions are made about the rake angle. It is thus possible to (severely) underestimate the amount of slip in the earthquake, if the dip of the mechanism is poorly constrained and the inversion leads to a value for dip that is too high. This could be particularly important for very shallow subduction earthquakes, since the dip is expected to be small for these events. Thus, a difference between a dip of 3 or 6 degrees in a CMT solution may not seem significant, but it could lead to a difference in moment (and thus slip) of a factor of two. To illustrate the importance of this issue for very shallowly dipping thrust events, we show in Fig. 7 three different homogeneous slip models (only slip was changed, fault surface area and rigidity were kept constant) that will produce similar surface waves because the product of their slip and dip is identical. However, the vertical deformation of the ocean floor and the tsunami waveforms resulting from these three different earthquakes would be significantly different in amplitude.

#### Horizontal Deformation of the Ocean Floor May Lead to Great Displacement of Water, Yet Is Neglected in Tsunami Modeling

Most tsunami modelers only consider a water surface displacement identical to the vertical deformation of the ocean bottom due to faulting when computing the tsunami height resulting from an earthquake and neglect the effect of horizontal deformation. However, when the tsunami source is located close to a steep slope and the horizontal displacement is large relative to the vertical displacement, which is generally the case for tsunami earthquakes due to their mechanism and shallow depth, the effect of horizontal deformation may become significant [77].

Furthermore, it has been suggested that the lateral collision force of a continental slope into the ocean due to faulting could also play an important part in the tsunami genesis of these events [72]. This type of dynamic excitation of tsunami waves would be particularly important for very shallowly dipping, shallow, thrust events, which would have a large component of horizontal motion. However, this concept has still been debated. This problem will be resolved if tsunami excitation is computed using the three-dimensional bathymetry in the source re-



**Tsunami Earthquakes, Figure 7**

Tsunami wave heights as a function of time for three different slip models for a shallow thrust earthquake, which will produce similar surface wave recordings (because the product of dip and moment is held constant). Thus, it is difficult to resolve between these different models using an inversion of surface waves, yet they produce very different vertical displacement of the ocean floor and thus very different tsunamis

gion with the displacement of the sea-floor as the initial condition.

#### Subduction of Bathymetric Features May Enable Seismic Slip in Usually Aseismic Region and Be Related to Unusual Aftershock Sequences

Sandbox experiments show the pervasive influence on the geomorphology of the shallow subduction zone margin when a seamount on the subducting plate is being subducted [17]. Subduction and underplating of relatively undeformed and water-laden sediments beneath the rear

part of the margin could, together with the dense fracture network generated by seamount subduction, modify the fluid pressure and introduce significant variations of the effective basal friction and thus the local mechanical plate coupling. More directly, subduction of a seamount may increase the normal stress across the subduction interface and hence enhance seismic coupling [15,67]. Unusual earthquakes have been documented in regions where ridges, seamounts or other bathymetric features are being subducted (e.g. [14,43]) and investigations of rupture characteristics of large underthrusting earthquakes provide evidence that seamounts can be subducted to seismogenic depths and that variations in seafloor bathymetry of the subducting plate may strongly influence the earthquake rupture process [10,61]. In the case of the Java tsunami earthquake of 1994, the bathymetry of the area landwards of the trench suggests that a local high is in the process of being subducted close to the area of maximum slip ([3], Fig. 4). A bathymetric high in the form of the Roo Rise can also be found just seaward of the trench region. In the bathymetry of the area around the 2006 Java event no such pronounced local feature can be found (Fig. 4), but the regional bathymetry south of the Java trench region is distinguished by an overall rough character. Similarly, the Nicaragua subduction zone is characterized by a highly developed horst and graben structure in the subducting plate, but no large-scale features, like a subducting seamount, are obvious. However, in case of the Peru earthquakes, both the 1996 and 1960 tsunami earthquakes occur at the intersection of the trench with major topographic features on the Nazca plate: the Mendaña fracture zone and the Trujillo trough, respectively [53].

Thus, subduction of either pronounced local bathymetric features or more regional seafloor roughness or horst-and-graben structures, which may modify the local coupling between subducting and overriding plates, has been documented in or near the rupture zone of many tsunami earthquakes.

### **Accretionary Prism: Uplift, Slides or Splay Faulting May Displace a Relatively Great Volume of Water**

Tsunami earthquakes may involve seismic slip along the normally aseismic basal decollement of the accretionary prism [58,76]. Sediments near the toe of an inner trench slope may be scraped off by a large horizontal movement over the decollement due to an earthquake and thus cause an additional inelastic uplift, which could have a large effect on tsunami generation (for the 1896 Sanriku earthquake: [79]; for the 1946 Aleutian Islands earthquake: [78]).

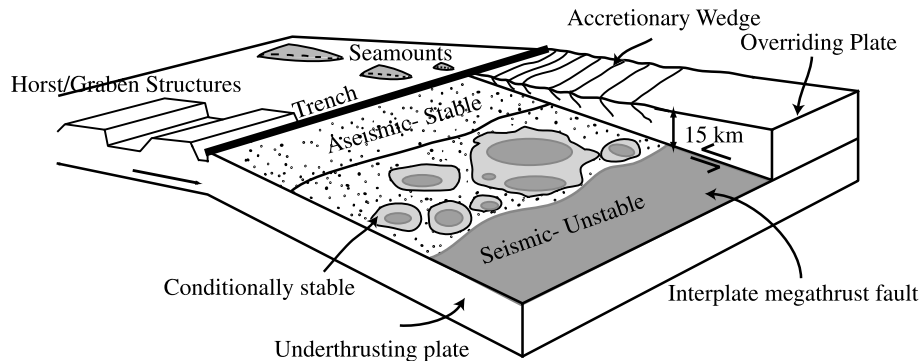
The existence of splay faulting, which would be more effective in exciting tsunamis due to their steeper dip, within the accretionary prism itself has been suggested to be a cause of the large tsunami excitation for the 1994 Java earthquake [3], the 2004 Sumatra mega-thrust event [46] and the 1963/1975 Kurile earthquakes [21]. Splay faulting can further promote extensive vertical deformation of the ocean floor, and hence large tsunamis, through partitioning or branching of a rupture upwards from the interface along multiple splay faults leading up to the surface [21,56].

### **Presence of Fluids Influences Seismic Behavior**

In subduction zones, fluids expelled from the subducting plate play an important role in many different subduction related phenomena such as volcanism, metamorphism and seismogenesis. Zones of high pore fluid pressure in the shallow subduction zone would change the effective normal stress significantly, possibly extending the region in which seismic slip is possible to shallower depths and generate slip of a slow nature. The presence of such zones has been suggested to be related to the occurrence of silent slip in the Nankai trough region [42]. In the case of tsunami earthquakes, the presence of large zones of elevated fluid pressure has been proposed to cause fairly rapid seismic slip close to the trench axis, following the breakage of asperities [68].

### **A Model for Tsunami Earthquakes**

Most explanations for the slowness of tsunami earthquakes involve the presence of low velocity, low strength and low rigidity sediments in the accretionary prism and between the overriding and subducting plate in the shallow subduction zone. Rupture through these slow sediments is thought to promote a slow rupture velocity. In Scholz's [65,66] model of the typical subduction zone (see Fig. 8 for an interpretation of this model from [9]), three possible stability regimes exist. In the stable zone, seismic slip cannot be supported and aseismic creep releases all strain. In the unstable zone episodic seismic slip occurs. In the conditionally stable zone, in between these two zones, slip can be abrupt if it experiences loading from a nearby slip patch. This conditionally stable zone may be very heterogeneous due to roughness of the thrust fault [75] creating isolated asperities or due to permeability changes [55] from the subduction of low permeability materials or the presence of fluids [68]. Tsunami earthquakes may represent a rupture of one or several large "unstable" asperities, which then propagates for a significant distance in the conditionally stable sedimentary materials [9,59].



### Tsunami Earthquakes, Figure 8

Cartoon illustrating frictional conditions of the subduction interface between subducting and overriding plate. Individual unstable sliding contact areas (dark gray) can provide the nucleation sites for rupture in the shallow subduction zone environment, which is typically a stable (stippled) or conditionally stable (light gray) frictional region. From [9]

Alternatively, but consistently, we could interpret tsunami earthquakes in the context of fracture mechanics as displaying a lack of radiated energy and a low rupture speed in a high  $G_C$  (fracture energy) environment [5,83]. As stated above, if these ruptures also involve localized patches of relatively strong unstable friction (that would be associated with high rupture speed and low  $G_C$ ) it would allow the rupture to propagate seismically, instead of as a continuous creep process. From this point of view, tsunami earthquakes dissipate a large amount of energy during the fracture process and are left with little energy to radiate. It is possible that the highly faulted trench and deformed sediments result in larger energy dissipation during failure due to an excessive amount of branching and bifurcation of cracks which gives rise to inelastic behavior and hence a large dissipation of energy [6,83], possibly involving the branching of the rupture into multiple splay faults in the accretionary prism.

Tsunami earthquakes therefore would represent slip at unusually shallow depths that would typically be dominated by creep-processes. Their nucleation would be made possible through the existence of localized asperities or patches of unstable friction in a typically stable or conditionally stable region. The presence of compartments of elevated fluid pressure may aid in the propagation of the seismic slip by creating zones of nearly zero friction surrounding the asperities [68]. These asperities may be created by the subduction of bathymetric features like seamounts or ridges or by the broken up nature of the subducting plate itself, creating a horst-and-graben system, which would act as buckets for sediment subduction [59,75]. The stress-release on these asperities would be near complete, and any additional unloading of stress on the plate interface due to the rupture may occur mostly

through creep. This would result in a relatively low number of aftershocks occurring on the interface between overriding and subducting plate. However, the static stress change in the outer rise area would be significant due to the shallow nature of most of the slip and thus normal faulting outer rise earthquakes would be more likely to be triggered [16,29,80]. The subduction of a bathymetric feature would also likely result in fracturization of the margin [17] in the overriding plate, thus further promoting the occurrence of intraplate aftershocks in this area of the shallow subduction zone. If the subducting plate itself is highly broken-up and thus contains pre-existing weak zones, this may facilitate further faulting within the subducting plate, in particular close to the lower edge of the rupture where the stress change due to the main shock is relatively large.

In this model for tsunami earthquakes discussed above, the unusually high effectiveness in the excitation of tsunami waves can be attributed to several factors, with the shallowness of the slip as the main underlying cause. Other important concerns coming into play are the possible failure of seismological techniques to provide an accurate estimate of the slip due to complexities associated with a very shallowly propagating rupture in a subduction zone, the possible failure of tsunami modeling to determine accurate wave heights due to similar complexities and the possible involvement of splay faulting or uplift of sediments near the trench in the accretionary prism.

### Future Directions

The Sumatra–Andaman earthquake and tsunami of 2004 renewed interest in the development of near real-time methods to estimate the true size of large earthquakes and

the tsunamis that might follow them, and the installation of instrumentation that will facilitate these measurements. Because the time between the earthquake and the arrival of the first tsunami waves at the local coastline is short, it is still unclear how effective these types of early warning systems are for saving lives at short distances from the tsunami source, but they will be useful at large distances.

New technologies and surveys will enhance our knowledge of the geomorphology and velocity structure of the shallow subduction zone. Ocean bottom seismometers, tide gauges, buoys and other seafloor monitoring devices will provide high quality data, which will enable us to place better constraints on where exactly the slip in shallow earthquakes occurs and in what tectonic environment. Although tsunami earthquakes occur relatively infrequently and thus may be difficult to capture, comprehensive characterizations of their rupture processes placed in the context of detailed three-dimensional models of the shallow subduction zones they occurred in will be an important next step in understanding their unusual seismo- and tsunami-genic processes.

## Bibliography

### Primary Literature

- Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. *J Geophys Res* 84:1561–1568
- Abe K (1989) Quantification of tsunamigenic earthquakes by the  $M_t$  scale. *Tectonophysics* 166:21–34
- Abercrombie RE, Antolik M, Felzer K, Ekstrom G (2001) The 1994 Java tsunami earthquake—Slip over a subducting seamount. *J Geophys Res* 106:6595–6608
- Ammon CJ, Ji C, Thio HK, Robinson D, Ni S, Hjorleifsdottir V, Kanamori H, Lay T, Das S, Helmberg D, Ichinose G, Polet J, Wald D (2005) Rupture Process of the 2004 Sumatra–Andaman Earthquake. *Science* 308:1133. doi:10.1126/science.1112260
- Ammon CJ, Kanamori H, Lay T, Velasco AA (2006) The 17 July 2006 Java tsunami earthquake. *Geophys Res Lett* 33:24. doi:10.1029/2006GL028005
- Barragan BE, Giaccio GM, Zerbino RL (2001) Fracture and failure of thermally damaged concrete under tensile loading. *Mater Struct* 34:312–319
- Beck SL, Ruff LJ (1987) Rupture process of the great 1963 Kuril Islands earthquake sequence: asperity interaction and multiple event rupture. *J Geophys Res* 92:14123–14138
- Bilek SL, Engdahl ER (2007) Rupture characterization and relocation aftershocks of for the 1994 and 2006 tsunami earthquakes in the Java subduction zone. *Geophys Res Lett* 34:L20311. doi:10.1029/2007GL031357
- Bilek SL, Lay T (2002) Tsunami earthquakes possibly widespread manifestations of frictional conditional stability. *Geophys Res Lett* 29:18-1. doi:10.1029/2002GL01521
- Bilek SL, Schwartz SY, Deshon HR (2003) Control of seafloor roughness on earthquake rupture behavior. *Geology* 31:455–458. doi:10.1130/0091-7613(2003)031
- Bourgeois J, Petroff C, Yeh H, Titov V, Synolakis CE, Benson B, Kuroiwa J, Lander J, Norabuena E (1999) Geologic Setting, Field Survey and Modeling of the Chimbote, Northern Peru, Tsunami of 21 February 1996. *Pure Appl Geophys* 154: 513–540
- Brown DL (1964) Tsunami activity accompanying the alaskan earthquake of 27 March 1964. US Army Engr Dist, Alaska, 20 pp
- Brune JN (1970) Tectonic stress and spectra of seismic shear waves from earthquakes. *J Geophys Res* 75:4997–5009
- Chung WY, Kanamori H (1978) Subduction process of a fracture zone and aseismic ridges – the focal mechanism and source characteristics of the New Hebrides earthquake of 1969 January 19 and some related events. *Geophys J Int* 54(1):221–240. doi:10.1111/j.1365-246X.1978.tb06764.x
- Cloos M (1992) Thrust-type subduction-zone earthquakes and seamount asperities; a physical model for seismic rupture. *Geology* 20:601–604
- Dmowska R, Zheng G, Rice JR (1996) Seismicity and deformation at convergent margins due to heterogeneous coupling. *J Geophys Res* 101:3015–3029
- Dominguez S, Malavieille J, Lallemand SE (2000) Deformation of accretionary wedges in response to seamount subduction: insight from sandbox experiments. *Tectonics* 19:182–196
- Engdahl ER, Villaseñor A (2002) Global seismicity: 1900–1999. In: Lee WHK, Kanamori H, Jennings PC, Kisslinger C (eds) *International Handbook of Earthquake and Engineering Seismology*. Academic Press, Amsterdam, Part A, chapt 41, pp 665–690
- Fryer GJ, Watts P, Pratson LF (2004) Source of the great tsunami of 1 April 1946: a landslide in the upper Aleutian forearc. *Mar Geol* 203:201–218
- Fujii Y, Satake K (2006) Source of the July 2006 West Java tsunami estimated from tide gauge records. *Geophys Res Lett* 33:L24317.1–L24317.5. doi:10.1029/2006GL028049
- Fukao Y (1979) Tsunami earthquakes and subduction processes near deep-sea trenches. *J Geophys Res* 84:2303–2314
- Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea tsunami: Earthquake or landslide? *Seism Res Lett* 71:344–351
- Geist EL, Bilek SL (2001) Effect of depth-dependent shear modulus on tsunami generation along subduction zones. *Geophys Res Lett* 28:1315–1318
- Geist EL, Dmowska R (1999) Local tsunamis and distributed slip at the source. *Pure Appl Geophys* 154:485–512
- Hara T (2006) Determination of earthquake magnitudes using duration of high-frequency energy radiation and maximum displacement amplitudes: application to the July 17, 2006 Java earthquake and other tsunami earthquakes. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract S21A-0132
- Hatori T (1967) The generating area of the Sanriku earthquake of 1896 and its comparison with the tsunami of 1933. *J Seismol Soc Jap Ser 2*, 20:164–170
- Heinrich P, Schindele F, Guibourg S, Ihmlé PF (1998) Modeling of the February 1996 Peruvian tsunami. *Geophys Res Lett* 25:2687–2690
- Hidayat D, Barker JS, Satake K (1995) Modeling the seismic source and tsunami generation of the December 12, 1992 Flores island, Indonesia, earthquake. *Pure Appl Geophys* 144: 537–554
- Hilde TWC (1983) Sediment subduction versus accretion around the Pacific. *Tectonophysics* 99:381–397
- Ide S, Imamura F, Yoshida Y, Abe K (1993) Source character-



- istics of the Nicaraguan tsunami earthquake of September 2, 1992. *Geophys Res Lett* 20:863–866
31. Ihmlé PF, Gomez JM, Heinrich P, Guibourg S (1998) The 1996 Peru tsunamigenic earthquake: Broadband source process. *Geophys Res Lett* 25:2691–2694
  32. Imamura F, Gica E, Takahashi T, Shuto N (1995) Numerical simulation of the 1992 Flores tsunami: Interpretation of tsunami phenomena in northeastern Flores Island and damage at Babi Island. *Pure Appl Geophys* 144:555–568
  33. Ji C (2006) A comparison study of 2006 Java earthquake and other Tsunami earthquakes. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract
  34. Ji C (2006) Resolving the trade-off between the seismic moment and fault dip of large subduction earthquakes and its impact on tsunami excitation. *Tsunami Sources Workshop*. Menlo Park
  35. Ji C, Zeng Y, Song AT (2007) Rupture process of the 2006 Mw 8.3 Kuril Island Earthquake inferred from joint inversion of teleseismic body and surface waves. *SSA meeting*. Kona
  36. Johnson JM, Satake K (1997) Estimation of seismic moment and slip distribution of the April 1, 1946, Aleutian tsunami earthquake. *J Geophys Res* 102:11765–11774
  37. Kanamori H (1972) Mechanism of tsunami earthquakes. *Phys Earth Planet Inter* 6:346–359
  38. Kanamori H (1993) W phase. *Geophys Res Lett* 20:1691–1694
  39. Kanamori H, Given JW (1981) Use of long-period surface waves for rapid determination of earthquake-source parameters. *Phys Earth Planet Int* 27:8–31
  40. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake – A slow tsunami earthquake associated with subducted sediments. *Nature* 361:714–716
  41. Kikuchi M, Kanamori H (1995) Source characteristics of the 1992 Nicaragua tsunami earthquake inferred from teleseismic body waves. *Pure Appl Geophys* 144:441–453
  42. Kodaira S, Iidaka T, Kato A, Park JO, Iwasaki T, Kaneda Y (2004) High pore fluid pressure may cause silent slip in the Nankai trough. *Science* 304:1295–1298. doi:10.1126/science.1096535
  43. Kodaira S, Takahashi N, Nakanishi A, Miura S, Kaneda Y (2000) Subducted seamount imaged in the rupture zone of the 1946 Nankaido earthquake. *Science* 289:104–106. doi:10.1126/science.289.5476.104
  44. Kulm LD, Prince RA, French W, Johnson S, Masias A (1981) Crustal structure and tectonics of the central Peru continental margin and trench. In: Kulm LD, Dymond J, Dasch EJ, Hussong DM (eds) *Nazca Plate: Crustal formation and Andean Convergence*. *Geol Soc Am Mem* 154:445–468
  45. Lavigne F, Gomes C, Giffo M, Wassmer P, Hoebreck C, Mardiatno D, Priyono J, Paris R (2007) Field observations of the 17 July 2006 Tsunami in Java. *Nat Hazards Earth Syst Sci* 7: 177–183
  46. Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RA, Beck SL, Bilek BL, Brudzinski MR, Butler R, DeShon HR, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. *Science* 308:1127–1133. doi:10.1126/science.1112250
  47. Lopez AM, Okal EA (2006) A seismological reassessment of the source of the 1946 Aleutian ‘tsunami’ earthquake. *Geophys J Int* 165(3):835–849. doi:10.1111/j.1365-246X.2006.02899.x
  48. Masson DG, Parson LM, Milsom J, Nichols G, Sikumbang N, Dwiyanto B, Kallagher H (1990) Subduction of seamounts at the Java trench – a view with long-range sidescan sonar. *Tectonophysics* 185:51–65
  49. McAdoo BG, Capone MK, Minder J (2004) Seafloor geomorphology of convergent margins: Implications for Cascadia seismic hazard. *Tectonics* 23:TC6008. doi:10.1029/2003TC001570
  50. Mori J, Mooney WD, Afnimar Kurniawan S, Anaya AI, Widiyan-toro S (2007) The 17 July 2006 tsunami earthquake in west Java, Indonesia. *Seismol Res Lett* 78:291
  51. Newman AV, Okal EA (1998) Teleseismic estimates of radiated seismic energy: The  $E/M_0$  discriminant for tsunami earthquakes. *J Geophys Res* 103:26885–26898
  52. Okal EA (1988) Seismic parameters controlling far-field tsunami amplitudes: A review. *Nat Haz* 1:67–96
  53. Okal EA, Newman AV (2001) Tsunami earthquakes: The quest for a regional signal. *Phys Earth Planet Inter* 124:45–70
  54. Okamoto T, Takenaka H (2006) Source process of the July 17, 2006 off Java island earthquake by using a fine crustal structure model of the Java trench and a 2.5D FDM computations. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract
  55. Pacheco JF, Sykes LR, Scholz CH (1993) Nature of seismic coupling along simple plate boundaries of the subduction type. *J Geophys Res* 98:14133–14159
  56. Park J-O, Tsuru T, Kodaira S, Cummins PR, Kaneda Y (2002) Splay Fault branching along the Nankai subduction zone. *Science* 297:1157–1160
  57. Pelayo AM, Wiens DA (1990) The November 20, 1960 Peru tsunami earthquake: Source mechanism of a slow event. *Geophys Res Lett* 17:661–664
  58. Pelayo AM, Wiens DA (1992) Tsunami earthquakes – Slow thrust-faulting events in the accretionary wedge. *J Geophys Res* 97:15321–15337
  59. Polet J, Kanamori H (2000) Shallow subduction zone earthquakes and their tsunamigenic potential. *Geophys J Int* 142:684–702. doi:10.1046/j.1365-246x.2000.00205.x
  60. Polet J, Thio HK (2003) The 1994 Java Tsunami earthquake and its “Normal” Aftershocks. *Geophys Res Lett* 30:27–1. doi:10.1029/2002GL016806
  61. Robinson DP, Das S, Watts AB (2006) Earthquake rupture stalled by a subducting fracture zone. *Science* 312:1203–1205. doi:10.1126/science.1125771
  62. Satake K (1994) Mechanics of the 1992 Nicaragua tsunami earthquake. *Geophys Res Lett* 21:2519–2522
  63. Satake K, Kanamori H (1991) Abnormal tsunamis caused by the June 13, 1984, Torishima, Japan, earthquake. *J Geophys Res* 96:19933–19939
  64. Satake K, Tanioka Y (1999) Sources of tsunami and tsunamigenic earthquakes in subduction zones. *Pure Appl Geophys* 154:467–483. doi:10.1007/s000240050240
  65. Scholz CH (1990) *The mechanics of earthquakes and faulting*. Cambridge Univ Press, New York
  66. Scholz CH (1998) Earthquakes and friction laws. *Nature* 391: 37–42
  67. Scholz CH, Small C (1997) The effect of seamount subduction on seismic coupling. *Geol* 25:487–490
  68. Seno T (2002) Tsunami earthquakes as transient phenomena. *Geophys Res Lett* 29(10):58.1–58.4. doi:10.1029/2002GL014868
  69. Seno T, Hirata K (2007) Did the 2004 Sumatra–Andaman earthquake involve a component of tsunami earthquakes? *Bull Seismol Soc Am* 97:S296–S306. doi:10.1785/0120050615
  70. Shapiro NM, Singh SK, Pacheco J (1998) A fast and simple diag-

nostic method for identifying tsunamigenic earthquakes. *Geophys Res Lett* 25:3911–3914

71. Shimazaki K, Geller RJ (1977) Source process of the Kurile Islands tsunami earthquake of June 10, 1975. *Eos Trans Am Geophys Union* 58:446
72. Song Y, Fu L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum C, Yi Y (2006) Horizontal motions of faulting dictate the 26 December 2004 tsunami genesis. *Eos Trans AGU* 87(52):Fall Meet Suppl, Abstract U53C-02
73. Synolakis CE, Bardet JP, Borrero JC, Davies HL, Okal EA, Silver EA, Sweet S, Tappin DR (2002) The slump origin of the 1998 Papua New Guinea Tsunami. *Proc Royal Soc A Math Phys Eng Sci* 458(2020):763–789. doi:10.1098/rspa.2001.0915
74. Synolakis CE, Imamura F, Tsuji Y, Matsutomi H, Tinti S, Cook B, Chandra YP, Usman M (1995) Damage, conditions of East Java tsunamis of 1994 analyzed. *EOS* 76:26
75. Tanioka Y, Ruff L, Satake K (1997) What controls the lateral variation of large earthquake occurrence along the Japan trench. *Isl Arc* 6:261–266
76. Tanioka Y, Satake K (1996) Fault parameters of the 1896 Sanriku tsunami earthquake estimated from tsunami numerical modeling. *Geophys Res Lett* 23:1549–1552
77. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
78. Tanioka Y, Seno T (2001) Detailed analysis of tsunami waveforms generated by the 1946 Aleutian tsunami earthquake. *Nat Haz Earth Syst Sci* 1:171–175
79. Tanioka Y, Seno T (2001) Sediment effect on tsunami generation of the 1896 Sanriku tsunami earthquake. *Geophys Res Lett* 28:3389–3392
80. Taylor MAJ, Zheng G, Rice JR, Stuart WD, Dmowska R (1996) Cyclic stressing and seismicity at strong coupled subduction zones. *J Geophys Res* 101:8363–8381
81. Tsuboi S (2000) Application of  $M_{wp}$  to tsunami earthquake. *Geophys Res Lett* 27:3105–3108
82. Tsuji Y, Imamura F, Matsutomi H, Synolakis CE, Nanang PT, Jumadi, Harada S, Han SS, Arai K, Cook B (1995) Field survey of the east Java earthquake and tsunami of June 3, 1994. *Pure Appl Geophys* 144(3–4):839–854
83. Venkataraman A, Kanamori H (2004) Observational constraints on the fracture energy of subduction zone earthquakes. *J Geophys Res* 109:B05302.1–05302.20. doi:10.1029/2003JB002549
84. Wang K, He J (2008) Effects of frictional behavior and geometry of subduction fault on coseismic seafloor deformation. *Bull Seismol Soc Am* 98(2):571–579
85. Ward SN (2002) Tsunamis. In: Meyers RA (ed) *The Encyclopedia of Physical Science and Technology*, vol 17. Academic Press, San Diego, pp 175–191
86. Wiens D (1989) Bathymetric effects on body waveforms from shallow subduction zone earthquakes and application to seismic processes in the Kurile Trench. *J Geophys Res* 94:2955–2972

## Books and Reviews

- Bebout G, Kirby S, Scholl D, Platt J (eds) (1996) *Subduction from Top to Bottom*. American Geophysical Union Monograph, no 96. American Geophysical Union, Washington DC
- Satake K, Imamura F (1995) Tsunamis 1992–1994. Special Issue of *Pure Appl Geophys* 144(3–4):373–890

## Tsunami Forecasting and Warning

OSAMU KAMIGAICHI

Japan Meteorological Agency, Tokyo, Japan

### Article Outline

Glossary

Definition of the Subject

Introduction

Complexity Problem in Tsunami Forecasting

Components of a Tsunami Early Warning System (TEWS)

Tsunami Early Warning System in Japan

Future Outlook

Acknowledgments

Bibliography

### Glossary

**Tsunami early warning system** It consists of four components, namely, 1) Seismic network, 2) Seismic data processing system, 3) Tsunami forecast system, and 4) Sea level data monitoring system. In a broader sense, warning transmission system (downlink to disaster management organizations and public) is also included.

**Tsunami amplitude** Amplitude is measured from undisturbed sea level to peak or trough of the wave. By definition, it can be positive or negative. Tsunami amplitude can be measured in real-time by instruments like tide gauge, pressure sensor, etc., and can be reproduced by numerical tsunami propagation simulation. One needs not to be confused with the term ‘run-up height’. It is the maximum height of inundation on land, and measured in post-tsunami field surveys from traces of tsunami (i. e. damage of constructions, vegetative markers, etc.).

**Simulation point** In this article, it is defined as the surface projection location of the hypothetical earthquake fault center. The vertical component of ocean bottom deformation due to earthquake fault dislocation calculated by elastic theory gives the initial tsunami waveform for the numerical tsunami propagation simulation.

**Forecast point** In this article, it is defined as the location of the offshore point where tsunami amplitude is evaluated by using numerical tsunami propagation simulation.

**Intergovernmental coordination group** The group established under UNESCO/IOC to facilitate interna-

tional cooperation for the tsunami disaster mitigation. There are four ICGs as of now (Pacific Ocean, Indian Ocean, Caribbean Sea, North Eastern Atlantic Ocean and Mediterranean Sea). One of the most important characteristics of tsunami is that it can cause huge disaster even after long distance propagation due to amplification near the coast. Therefore, international cooperation, especially the prompt data and information exchange, is essential for the disaster mitigation.

**Centroid moment tensor solution** One of the representation of seismic source process. It is represented by the moment tensor, which is a combination of six independent equivalent force couples, and is a weighted average of the source process in time and space. Since it represents an overall image of the source process, it is suitable to evaluate tsunamigenic potential of the earthquake.

**Earthquake early warning** Earthquake Early Warning is to enable countermeasures in advance for strong motion disaster by detecting seismic *P* wave at stations near the epicenter, quickly estimate seismic intensity and arrival time of *S* wave, and transmit these estimation before the *S* wave arrival. The Japan Meteorological Agency (JMA) started to provide EEW to the general public in October, 2007. This technique is applicable to quicken tsunami warning dissemination.

## Definition of the Subject

A tsunami is, along with strong motion, one of the two major disasters caused by earthquake. To mitigate tsunami disaster, it is important to integrate software countermeasures like tsunami forecast to enable timely evacuation from area at risk before tsunami strikes the coast, as well as to intensify hardware countermeasures particularly in vulnerable coastal areas like building banks and water-gates. Tsunami disaster mitigation can be achieved effectively by the appropriate combination of the software and hardware countermeasures. Also, improving people's awareness on the tsunami disaster, necessity of spontaneous evacuation when they notice an imminent threat of tsunami on their own (feeling strong shaking near the coast, seeing abnormal sea level change, etc.) and how to respond to a tsunami forecast, and conducting tsunami evacuation drill are very important issues for disaster mitigation.

In this article, a tsunami forecast, as the most typical software countermeasure that a national organization can provide, is mainly described. Recent progress in science has deepened our understanding of the tsunami-generating source mechanisms, tsunami propagation and inundation process and enabled the development of sophis-

ticated numerical simulation programs. But at the same time, complexities of focal process and tsunami behavior, especially near the coast, have also been revealed. Therefore, careful consideration of the complicated nature of tsunami phenomena is essential in order to make tsunami forecast contents and the dissemination of warnings effective for disaster mitigation.

## Introduction

A tsunami is an oceanic gravity wave generated by submarine fault dislocation or other origins such as mud or rock slumps on steep continental margin slopes, marine volcanic eruptions and others. A large tsunami may cause disasters along densely populated or built-up coasts and sometimes also by the inundation of low land areas, up to several km inland. Most tsunamis are generated by large earthquakes occurring in oceanic areas, and it is possible to estimate tsunami generation to a certain extent from seismic wave analysis. By taking advantage of the propagation velocity difference between the much faster seismic and the slower tsunami waves, it is possible to mitigate tsunami disasters by issuing tsunami forecast before the tsunami arrives at the coast, thus enabling evacuation and other countermeasures. For other origins of tsunami, it is still difficult to quantitatively forecast tsunami generation until it is observed actually by the sea level change sensors.

In earlier years, an empirical method had been used to estimate tsunami amplitude via a regression formula that relates tsunami amplitude to the magnitude of the earthquake and its epicentral distance from the coast of interest. Recently, significant progress has been made in the understanding of the tsunami characteristics. This permits numerical simulation of the tsunami propagation once the initial tsunami wave distribution in the source area is correctly known, and bathymetry data are available with sufficient spatial resolution. In Japan, numerical simulation technique has been used in the operational tsunami forecasting procedure in the Japan Meteorological Agency (JMA) since 1999. Efforts have been made also in other countries to incorporate numerical simulation technique in tsunami forecasting. In this article, mainly based on JMA experience, procedures are introduced how to conduct tsunami forecast service by using numerical simulation techniques and giving due consideration to the complexity of this problem.

## Complexity Problem in Tsunami Forecasting

The complexity of tsunami forecasting with numerical methods is twofold and mainly due to the following:

- The uncertainty of the initial tsunami wave distribution in the source area
- The complexity of bathymetry and coastal topography

### Uncertainty of the Initial Tsunami Wave Distribution

For a local tsunami, only limited time is left from the generation of tsunami to its arrival at the coast. In order to assure maximum lead time for evacuation, this necessitates the tsunami forecast to be based on the earliest available data from seismic wave analysis. On the other hand, in the case of a distant tsunami sources, a relatively long lead time is left until the tsunami strikes the coast, and data of sea level change can be recorded in near real-time by tidal stations on the way from the source to the coast. This allows the tsunami forecast to be based on the actual observation of generated tsunami waves and thus to reduce the uncertainty in the initial spatial distribution of the tsunami wave in the source area. Also for local events it is possible to improve the accuracy and reliability of the tsunami forecast in a step-by-step manner, if more detailed data and reliable analysis results become available after the first forecast has been issued. Therefore, the most practical approach in view of the disaster mitigation is to assure the rapid issuance of the first forecast based on the seismic wave analysis, and then to update it with improved data.

There are several uncertainties when the initial tsunami wave distribution is inferred from seismological data analysis alone.

### Uncertainty of the Relative Location of the Hypocenter in the Rupture Area

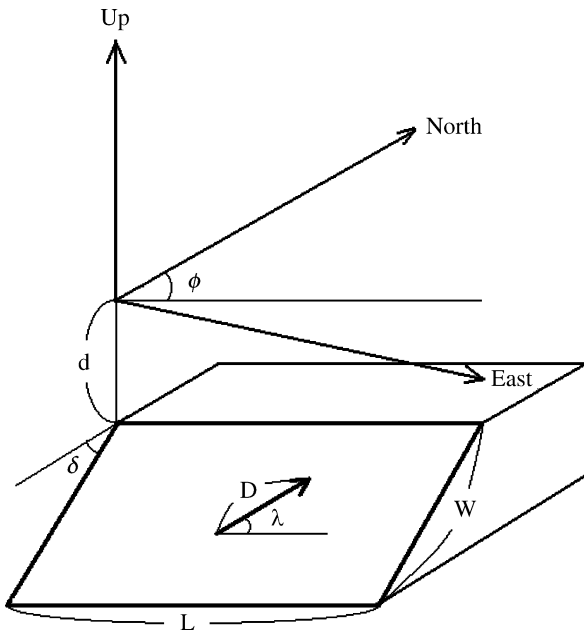
The hypocenter of an earthquake is merely the location from where the rupture starts. It is calculated from first arriving seismic waves. Depending on the direction of rupture propagation (e.g., unilateral, bilateral, radial) and the complexity/irregularity of the actual fault rupture, the seismologically determined hypocenter may neither lie in the center of the rupture area nor coincide with the area of maximum fault displacement. The bigger the earthquake and thus the larger the spatial extent of the rupture area, the less representative the seismologically derived hypocenter as a parameter for characterizing the earthquake rupture is. To reduce this uncertainty, some methods have been developed, like a rupture process inversion by using teleseismic broadband seismogram [20], or near-field strong motion data [11]. A fault model inversion by using a co-seismic step estimated by real-time GPS data analysis is under development [5,22,25]. And the spatial extent of the tsunami source area can also be constrained by using inverse refraction diagrams from tidal stations [1] or deep ocean wa-

ter pressure sensor (DART (Deep-ocean Assessment and Reporting of Tsunamis))-type stations [6] that recorded tsunami arrivals as well as by satellite sea surface altimetry data taken over the source area [7,10].

In general, the rupture process is complicated, and the slip distribution on the fault is not uniform. Non-uniformity of the slip distribution can make the initial tsunami wave distribution much more complicated, especially in its short wavelength component, and may strongly affect the tsunami behavior on the coast. To estimate the slip distribution, in addition to the seismological and geodetic methods mentioned above, fault slip distribution inversion techniques have been developed which compare actually observed tsunami waveforms with pre-calculated theoretical tsunami waveforms from unit fault dislocation segments as Green functions [26] ► [Tsunamis, Inverse Problem of](#).

**Uncertainty of the Magnitude** The maximum amplitude of a seismic body or surface wave group is usually used for estimating the earthquake magnitude (see this volume) as a measure of earthquake size in a certain period range, which is often limited by the bandwidth of the seismometer response. Therefore, the tsunamigenic potential may be underestimated, especially for gigantic earthquakes (such as the Sumatra-Andaman Mw9.3 earthquake of 26 December 2004) and “tsunami earthquakes” that generate bigger tsunami than expected from its shaking potential and magnitude values in the short-period range. To overcome this difficulty, some fast body-wave methods have been developed that consider besides single maximum amplitudes also multiple rupture amplitudes over the rupture duration [2], ► [Earthquake Magnitude](#), integrated seismic energy and/or total rupture duration time [8,21], while others use successfully long-period spectral mantle surface-wave amplitudes in real-time applications in the context of tsunami warning [38], and *W*-phase for quick estimation of CMT solution for a practical tsunami warning service [17]. Besides these seismological methods for estimating earthquake magnitude, geodetic methods that use the co-seismic step or ultra-long period component observed by crustal deformation sensors [5,13,18] play an increasing role for estimating the size of great earthquakes.

**Uncertainty of the Fault Parameters** It is difficult to estimate fault geometry promptly after the earthquake occurrence. Sea floor deformation due to fault motion is usually assumed to be identical to initial tsunami wave distribution. Accordingly, the latter is significantly affected by the specifics of fault geometry, fault and slip orienta-



**Tsunami Forecasting and Warning, Figure 1**

Definitions of fault parameters.  $L$ ,  $W$ ,  $D$  and  $d$  denote the fault length, width, average slip amount and depth of the top margin of the fault respectively. The fault dip  $\delta$  is measured down from horizontal, strike  $\phi$  clockwise round from North and slip (rake)  $\lambda$  anti-clockwise round from strike direction along the fault plane (motion of the hanging wall relative to the footwall)

tion in space as well as rupture complexity. These complexities are usually neglected and the real fault rupture is roughly approximated by a dislocation model that is represent by a rectangular fault that is described by 6 parameters, namely the fault length  $L$ , width  $W$ , average slip amount  $D$ , and the three angles of fault dip  $\delta$ , strike  $\phi$  and slip (rake)  $\lambda$  (Fig. 1).

**Fault Length, Width and Average Slip Amount** Different scaling laws exist between  $L$ ,  $W$  and  $D$  with earthquake magnitude (e. g., [3,4,16,37,39]). They allow us to roughly estimate these values from determined magnitudes. Such scaling relations are based on the assumption of more or less constant stress drop (i. e. High stress drop earthquake has larger  $D$  for the same  $L$  and  $W$  of low stress drop earthquake). But the stress drop is different between inter-plate thrust events and intra-plate events. In general, intra-plate event has higher stress drop than inter-plate event. Such differences may cause significant errors when estimating these fault parameters via scaling relations with magnitudes. Best estimates of these parameters, that characterize the earthquake size, are possible via the determination of the seismic moment  $M_0 = \mu DA$  with  $\mu$  being the rigid-

ity of the crustal/lithosphere material in the rupture area  $A = L \times W$ , and  $L$  and  $W$  estimation by rupture process analysis described in Sect. “Uncertainty of the Relative Location of the Hypocenter in the Rupture Area”.

**Fault Dip, Strike and Slip Angle** Representative values for dip, strike and slip angles can be set based on the analysis of past great earthquakes having occurred in each region. The fault plane of a newly occurred earthquake is likely to be close to either one of the two nodal planes derived from centroid moment tensor (CMT) solutions. But to select from these two candidates the real acting fault plane, precise aftershock location or rupture process analysis as described in Subsect. “Uncertainty of the Relative Location of the Hypocenter in the Rupture Area” are necessary.

### Complexity due to Complicated Bathymetry

If the spatial distribution of the initial tsunami wave is given correctly, the tsunami propagation can be forecast in a deterministic manner by using numerical simulation technique. In a long-wave approximation, which considers only wavelengths of the tsunami that are much larger than the sea depth, the propagation velocity  $v$  of the tsunami is represented by  $v = \sqrt{gd}$ , where  $g = 9.81 \dots \text{m/s}^2$  is the Earth’s gravity acceleration and  $d$  the sea depth in m. This approximation is valid for most tsunami cases. As wavelength is the product of velocity times period, tsunami wavelength is proportional to the square root of sea depth. Accordingly, the wavelength of the tsunami become shorter as it approaches the coast because the sea depth becomes shallower, and the wave becomes much more sensitive to finer bathymetry changes and coastal feature. Therefore, a finer mesh of bathymetry and coastal features is necessary near the coast in order to represent the tsunami behavior correctly. But, even if very fine mesh bathymetry data are available, it is not appropriate to use it in an early stage of tsunami propagation simulation when the initial tsunami wave distribution is still uncertain as mentioned in Subsect. “Uncertainty of the Initial Tsunami Wave Distribution”.

Further, very fine-mesh simulations require a substantially long time for the conduct of simulation. Time constraints are crucial when incorporating numerical simulation techniques in an operational tsunami forecast procedure, except for distant events. Therefore, especially for local events, it is most practical to conduct tsunami simulations for a variety of scenarios in advance, to store these results in a database, and to conduct tsunami forecast by retrieving the most appropriate case for the determined hypocenter. As described later, the JMA has adopted this



way. Even in that case, the mesh size to be used in the simulations must be carefully examined taking into account the required accuracy and spatial resolution in the tsunami forecast for guiding disaster mitigation efforts and the need to complete the simulations for all scenarios in a realistic time span.

When very fine mesh is used, depending on the complexity of the bathymetry and topography near the coast, tsunami amplitude distribution along the coast shows significant scatter, and sometimes, extremely large tsunami amplitude will be estimated very locally. Therefore, it must be clearly defined in advance as to what kind of statistical average value, using a finite number of estimated tsunami amplitudes for the area of interest, should be adopted in the tsunami forecast.

### Components of a Tsunami Early Warning System (TEWS)

In general, a tsunami early warning system consists of the following constituents:

1. Seismic network (seismometers and real-time data transmission link)
2. Real-time seismic data processing system for hypocenter and magnitude determination
3. Tsunami forecast system (including warning criteria, assembling of the text, and dissemination)
4. Sea level data monitoring system (tide gauge/tsunami-meter and real-time/near real-time data transmission link)

Such a composition is adopted in Japan, USA, Russia, Chile, Australia, French Polynesia, New Caledonia and in other countries around the Pacific Ocean. The up-to-date status of the TEWS of these countries can be consulted on the UNESCO/Intergovernmental Oceanographic Commission's (IOC) website as based on the national reports submitted to the latest meeting of the Intergovernmental Coordination Group/Pacific Tsunami Warning System (ICG/PTWS). Also in Indian Ocean countries, after the tremendous tsunami disaster brought by the Great Sumatra Earthquake in 2004, efforts have been made to establish tsunami early warning systems of the similar composition in different countries (e. g. GITEWS: German Indonesian Tsunami Early Warning System, see <http://www.gitews.org/>). Similar efforts are under way in the Caribbean, North Eastern Atlantic and Mediterranean regions too, inspired by the Sumatra event and taking into account potential tsunami risk assessments for these areas based on historical records.

To accomplish tsunami warning for local events, local seismic networks (see Subject. "[Seismic Network](#)") and real-time seismic data processing systems (see Subject. "[Real-Time Seismic Data Processing System](#)") are indispensable. For a distant event, one can utilize the hypocenter parameters contained in the international tsunami watch information provided by the Pacific Tsunami Warning Center (PTWC), West Coast/Alaska Tsunami Warning Center (WC/ATWC) of US and North West Pacific Tsunami Advisory Center (NWPTAC) of the JMA.

As a typical example, technical details of Japan's Tsunami Early Warning System, in which all four constituents are fully deployed, is explained in the following section.

As for the item 4 (sea level data monitoring system), the US and also GITEWS are now deploying DART type of buoys [6] system. This is to measure a sea surface vertical displacement by observing the related pressure change at the ocean bottom. DART buoys are placed in far offshore regions where bathymetry is simple, and very simple tsunami waveforms can be observed without the influence of complicated bathymetry near a coast. Such measured data are preferable for a comparison with simulated waveforms. And by placing such buoys with proper spacing at a certain distance from the tsunamigenic zone over the deep ocean basin, tsunami waves, that are sufficiently separated in time from seismic waves, can be observed early enough after the earthquake occurrence to be useful for tsunami EW from distant sources.

The Pacific Marine Environmental Laboratory (PMEL) of the National Oceanic and Atmospheric Administration (NOAA) is developing a tsunami forecasting system named SIFT (Short-term Inundation Forecasting) [33] based on the DART buoy data. Like in the JMA's tsunami simulation database, described in detail below, tsunami simulations are conducted for many different unit faults located along the tsunamigenic zone, and the calculated tsunami waveforms at DART buoy locations are stored in a database as Green functions, together with simulated waveforms at coastal points. In the case that a tsunami wave is observed by DART buoys, the observed waveforms are represented by a linear combination of several Green functions. Then, tsunami amplitudes at coastal points can be estimated from the linear combination coefficients and the simulated waveforms at coastal points (see web-page of NOAA/PMEL (<http://nctr.pmel.noaa.gov/>) for applications of their method to recent actual tsunami events). This system is planned to be introduced in the actual tsunami forecasting procedures of the US tsunami warning centers.

In Japan, a similar study is in progress at Tohoku-University [35]. They conduct tsunami simulations originat-

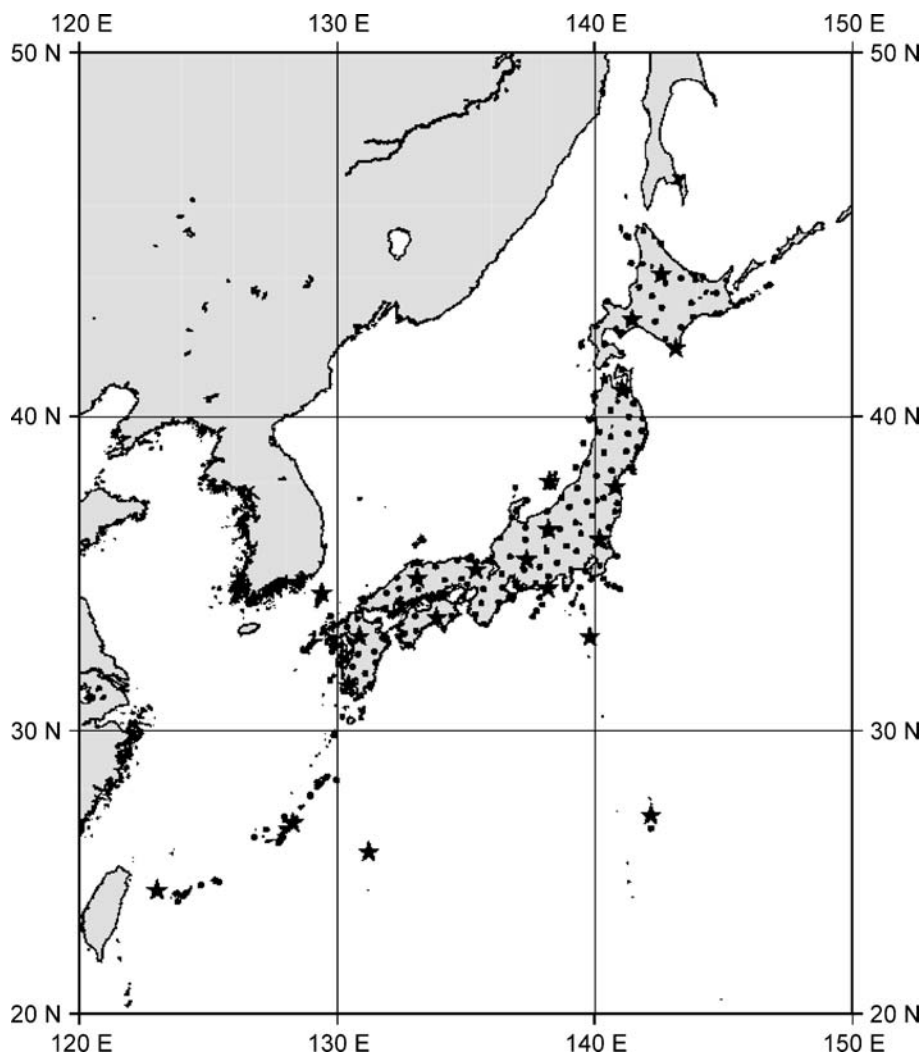
ing from unit vertical sea surface displacements in a unit sea area, and the resulting simulated waveforms at offshore tsunami meters are stored as Green functions in a database. They do not adhere to a specific fault model. The procedure for the estimation of tsunami waveform at coastal points is similar to that of SIFT.

### Tsunami Early Warning System in Japan

Before describing the present status of the JMA's TEWS, a historical review is briefly given.

The JMA started tsunami warning service in 1952. Since then, the JMA has been making an effort to integrate its TEWS, and this is exactly a history of 'fight against time'.

At the commencement of the tsunami warning service, the JMA had 46 seismometers at meteorological observatories (basically near populated area on a sedimental layer). In case of earthquake occurrence, seismograms were read at observatories, and their results ( $P$  and  $S$  arrival times and maximum amplitudes) were transmitted to the headquarters in a telegram format. At the headquar-



**Tsunami Forecasting and Warning, Figure 2**

Seismic network of the JMA used for tsunami forecasting. *Solid circles* and *stars* denote the locations of seismic stations operated by the JMA. The average spacing between the about 180 stations is 50 to 60 km. *Stars* denote the location of stations where STS-2 velocity broadband seismometers have been installed in addition to the Japanese short-period velocity-type seismometers and accelerometers

ters, hypocenter and magnitude were determined manually, and the tsunami warning grade was determined by using an empirical chart based on the relation between tsunami amplitude, earthquake magnitude and epicentral distance to a coast of the past events. It took about 15 to 20 minutes to disseminate a tsunami warning.

From the late 1960s to early 1980s, telemetry technology to transmit seismic waveform data from observatories to the headquarters and a processing computer were introduced. Detection capability of the earthquake was improved by monitoring collected seismic waveforms at one place, and the *P* and *S* picking precision was also improved by introducing a digitizer. By these, dissemination time was reduced to 12 to 13 minutes.

In 1983, the Mid-Japan-Sea earthquake (Mjma7.7) occurred. The JMA disseminated a tsunami warning in 14 minutes, but in about 7 minutes, the tsunami struck the nearest coast, killing 104 people. After this event, the JMA deployed a more sophisticated computer system for the seismic waveform processing. A graphical man-machine interface was introduced for more accurate and quicker phase reading and hypocenter and magnitude calculation. Still, the empirical method was used for the tsunami amplitude estimation. Dissemination time was reduced to between 7 and 8 minutes.

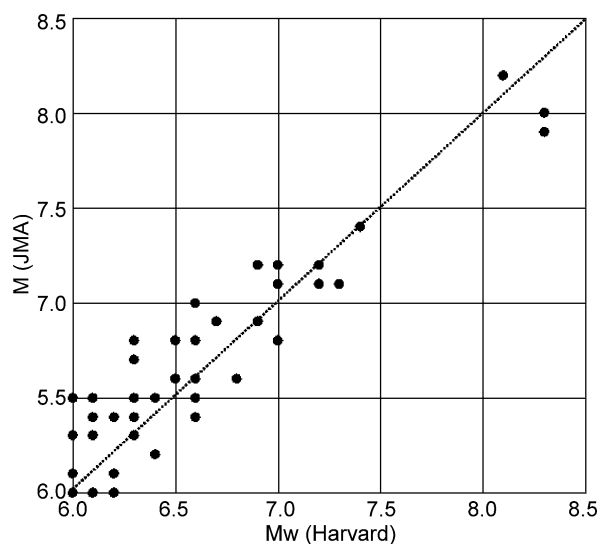
In 1993, the Southwest off Hokkaido earthquake (Mjma7.8) occurred. The JMA disseminated a tsunami warning in 5 minutes, but the tsunami struck Okushiri Island in a few minutes, and 230 people were killed or lost by tsunami. After this event, the JMA totally replaced the seismic network. All seismometers were installed in remote un-manned sites on hard rock, and the total number of the site increased to about 180. Dissemination time was reduced to 3 to 5 minutes.

The present status of the JMA's TEWS is as follows.

### Seismic Network

The JMA operates about 180 seismometers installed in Japan (Fig. 2). The seismic waveform data are sent to JMA continuously on a real-time basis through dedicated telephone lines.

There are two kinds of sensors deployed at each station: A short-period velocity sensor and an accelerometer. The records of the short-period sensor are mainly used for precise picking of the onset times of *P* and *S* phases which are required for an accurate hypocenter determination. Accelerometer records are used for the calculation of magnitude of large earthquakes in the case that the ground motion amplitude exceeds the dynamic range of the short-period sensor.



**Tsunami Forecasting and Warning, Figure 3**

Relation between *M<sub>w</sub>* (Global CMT) and *M<sub>jma</sub>*, based on events that occurred after September 2003 in and around Japan. Both magnitudes agree on average well up to around magnitude 8.0

STS-2 broad-band seismometers have additionally been installed at 20 stations. These broadband velocity seismic records are mainly used for obtaining CMT solutions and moment magnitudes *M<sub>w</sub>*.

### Real-Time Seismic Data Processing System

The JMA applies the short term average/long term average (STA/LTA) method as trigger criteria. It is most widely adopted in the seismological community and used to limit seismic recordings to relevant seismic event waveforms. When the ratio of STA/LTA exceeds the trigger threshold, it is supposed that a significant signal, possibly an earthquake signal, has been detected by the considered station.

STA/LTA trigger may occur due to some troubles or local noise at one station. To avoid "false trigger", the JMA also applies other criteria including the group trigger criterion. A group is set to include several neighboring stations. When the STA/LTA trigger turns on at a certain number of stations in one group, the system considers it to be caused by an earthquake.

For the precise picking of the *P* onset, an autoregressive (AR) model is applied to represent seismogram time series. Akaike's information criterion (AIC) is used to search the optimal time to separate the seismogram into two sub-stationary time series (noise part and signal part), which are represented by AR models, respectively, in a certain time window around trigger time [32,40].

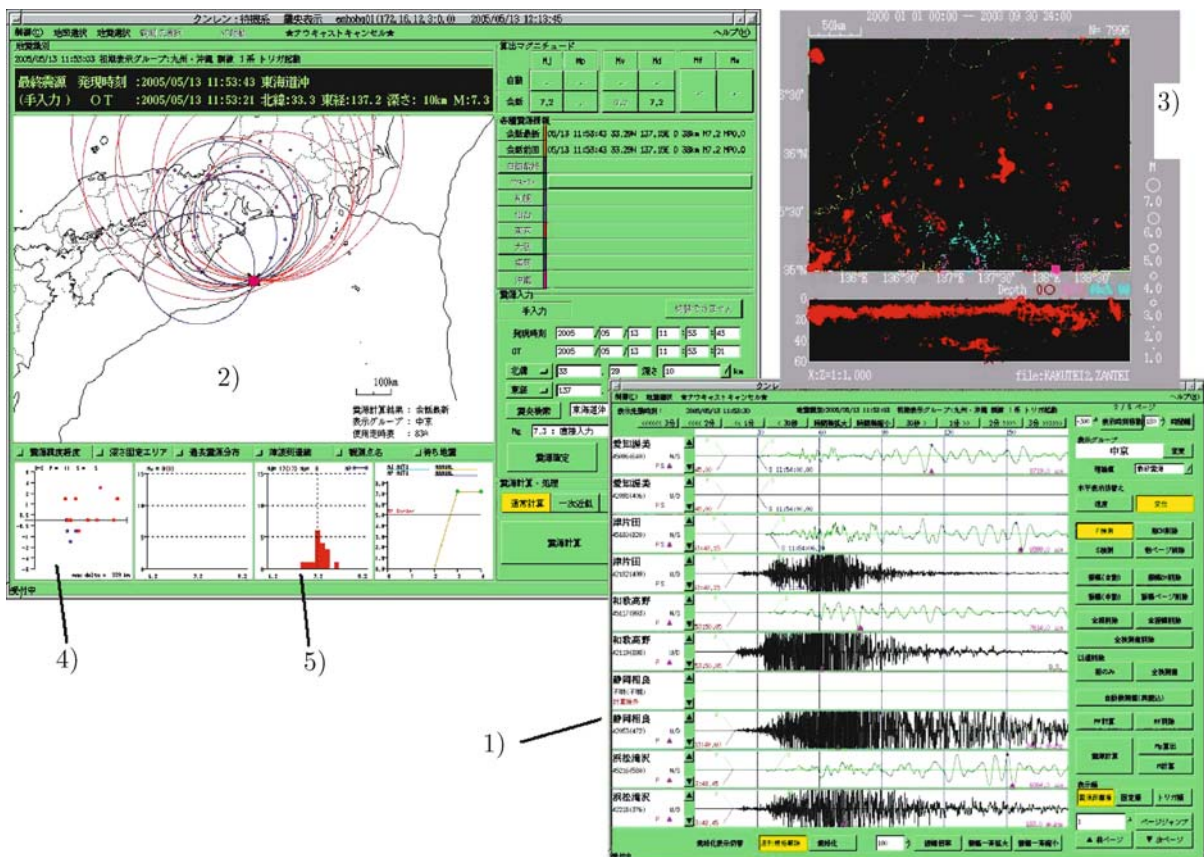
The usual least square method is used in the hypocenter calculation. As for the velocity structure, the Japan local standard model JMA2001 [36] is applied. To better locate the hypocenter,  $S$  arrival times are used in addition to  $P$  arrival times. First, a preliminary hypocenter is calculated by using only  $P$  arrival times picked by an automatic picker. Then,  $S$  arrival times are picked automatically by applying the AR model and AIC in a time window centered at the theoretical  $S$  arrival times estimated from the preliminary hypocenter location and origin time. The final hypocenter is then calculated by using both the  $P$  and  $S$  arrival times.

For magnitude calculation, one of the local magnitude definitions in Japan, the JMA magnitude ( $M_{jma}$ ) according to Katsumata [19] is used, in which average amplitude decay characteristic in Japan is taken as a decay correction term of the formula. Maximum amplitude of the horizontal ground displacement is the measured input value in the formula. The displacement waveform is calculated by applying real-time recursive filter that inte-

grates twice the accelerometer data and applies a high-pass filter to avoid computational instability. Due to historical reasons, the long-period cut-off of the high-pass filter is set at 5 seconds. Since unsaturated strong-motion data are used, magnitude calculation can be started promptly after the earthquake occurrence by using data from stations very close to the epicenter. Therefore, such magnitude data are suitable for the tsunami warning purposes of local events.

The comparison between  $M_{jma}$  and the moment magnitude  $M_w$  of the Global CMT Project's solution is shown in Fig. 3.  $M_{jma}$  values are on average comparable with  $M_w$  for earthquakes with magnitudes between  $M_{6.0}$  and about 8.0.

All procedures mentioned above are run automatically by the computer. Considering the possible impact of a tsunami warning on the potentially affected communities, a human check of the adequacy of the calculated results is indispensable for minimizing false alarms. The fol-



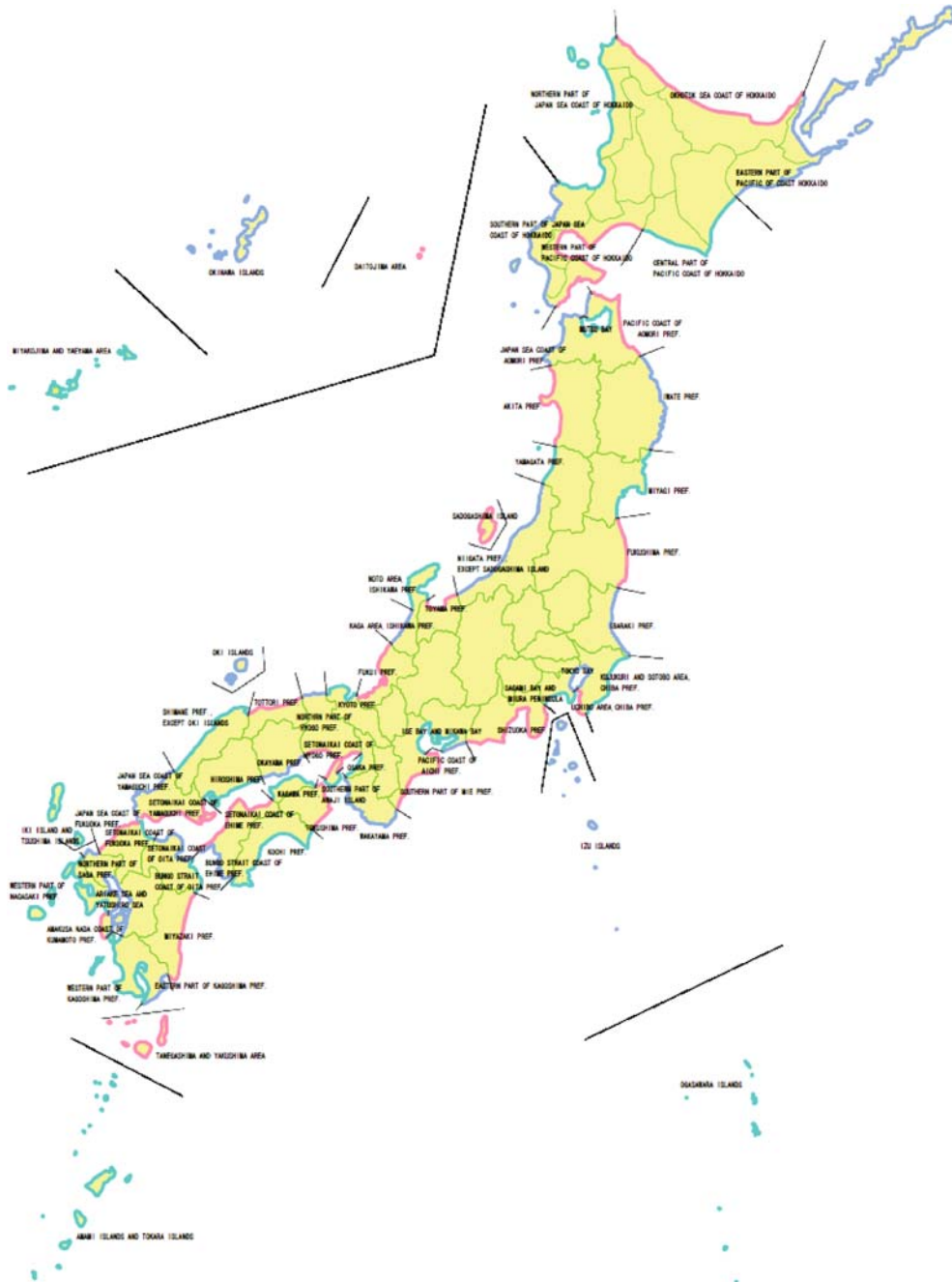
Tsunami Forecasting and Warning, Figure 4

Examples of the JMA's man-machine interface screen image in the seismic data processing system. Numbers in the figure are related to the respective numbers of items explained in the main text



lowing items are checked by the operators, using a man-machine interface:

1. Adequacy of the phase pickings and maximum amplitude readings: The seismic waveforms are plotted together with picked and theoretical arrival times for  $P$
2. Adequacy (mathematical) of the hypocenter location: A hypocenter plot map is used in which the locations of used seismic stations and the circles denoting the calcu-



Tsunami Forecasting and Warning, Figure 5

Coastal block partitioning for tsunami forecast in Japan. *Straight lines* denote the borders between the 66 coastal blocks



**Tsunami Forecasting and Warning, Table 1**

Tsunami forecast grades and corresponding levels of expected maximum tsunami amplitudes as used in tsunami information issued by the JMA for each coastal block after a tsunami forecast has been made

Forecast Grade		Levels of Estimated Tsunami Amplitude
Warning	Major Tsunami	3 m, 4 m, 6 m, 8 m, 10 m or greater
	Tsunami	1 m, 2 m
Advisory		0.5 m

- lated epicentral distance from each station are shown. If the circles intersect densely at one point, and if the used stations assure a wide azimuthal coverage around the source, then the calculation is judged as fine.
3. Adequacy (seismological) of the hypocenter location: The hypocenter plot on the background seismicity map is used together with the vertical cross section. If the hypocenter is located in a region where no background seismicity is found, it should be re-examined carefully.
  4. Adequacy of the depth estimation: Travel time residuals are plotted as a function of epicentral distance. If the residuals depend on epicentral distance, the depth estimation should be revised.
  5. Adequacy of the magnitude calculation: A histogram of station magnitudes is used. Outliers are checked and excluded in the re-calculation.

Sample images of the man-machine interface are shown in Fig. 4.

The JMA's man-machine interface is designed so that re-picking and recalculation of hypocenter and magnitude can be done within about 30 seconds.

Calculated hypocenter parameters (latitude, longitude, depth and magnitude) are transferred to the following tsunami forecast system as its input.

### Tsunami Forecast System

**Tsunami Forecast in Japan** For a tsunami forecast to be useful, a mere general description of the tsunamigenic potential estimated from calculated hypocentral parameters is not sufficient. From the issued message, the recipient disaster management organization should be able to clearly grasp the severity of the expected disaster in its own jurisdiction. In Japan, a tsunami forecast is disseminated to each coastal block (the coast line of Japan is separated into certain number of blocks), giving three kinds of grade corresponding to the anticipated severity of the disaster.

**Tsunami Forecast to Coastal Block** The coastal blocks divide the Japanese coasts into 66 regions (Fig. 5). These blocks are defined by taking into account:

- a) The administrative districts of local governments, which are the disaster management units to take actions in emergency situations
- b) The sea areas to which each coast is facing
- c) The uniformity in the behavior of tsunami
- d) The precision of tsunami forecast technique

Basically, one coastal block corresponds to one prefecture, or finer.

Before the introduction of numerical tsunami simulation technique, the total number of coastal blocks was 18, due to a low precision of empirical tsunami amplitude estimation method.

**Tsunami Forecast Criterion and Category** The JMA categorizes tsunami forecast into "Tsunami Warning" and "Tsunami Advisory". Further, "Tsunami Warning" is divided into two grades, "Tsunami" and "Major Tsunami".

In Japan, there exist ample number of materials on the relation between tsunami disaster and observed tsunami amplitude along the coast. From the materials, high correlation between these two can be seen. Therefore, estimated maximum tsunami amplitude at the coast is used as a criterion of the tsunami forecast in Japan. The speed of water current in the sea near the coast could be another good index for assessing the tsunami disaster potential. But, up to now, only a few examples of the current observation are available. Therefore, the estimated tsunami amplitude at the coast is still the most appropriate criterion for tsunami forecast.

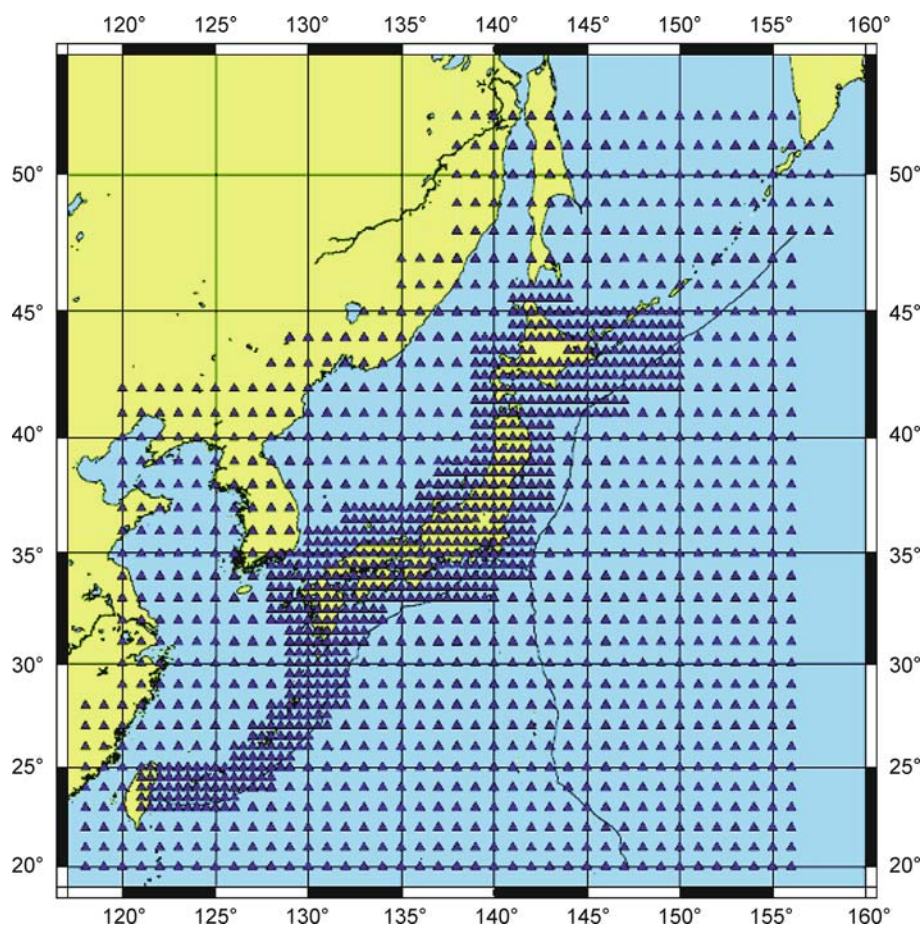
In Japan, tsunami disaster in the land area occurs when the tsunami amplitude exceeds 1 meter [30], so the warning criterion is set at 1 meter. When tsunami amplitude exceeds 3 meters, the proportion of damaged wooden houses and fishing boats increases remarkably [9,28,29]. Therefore, the JMA issues the highest warning grade "Major Tsunami" when the estimated maximum tsunami amplitude is 3 meters or higher. When the estimated maximum tsunami amplitudes range between 1 meter to 3 meters, the grade name is "Tsunami". If a "Tsunami Warning" is issued, the head of municipality must order evacuation to the residents in the "Tsunami Evacuation Zone" assigned by each municipality in advance.

Even when the tsunami amplitude is less than 1 meter, it is occasionally the case that bathing people are affected and aquaculture rafts for marine products industry are damaged. Considering safety criteria in sea bathing places (beaches, bays), and the relation between the tsunami amplitude near the coast and damage on aquaculture rafts in the past, the JMA's criterion for "Tsunami Advisory" is set at 20 centimeters. In the case that a "Tsunami Advisory" is issued, there is no need of evacuation in the land areas, except vulnerable very low-land areas, but people on beaches or swimming should go to higher places.

Table 1 shows the tsunami forecast grades and levels of the estimated maximum tsunami amplitude that are used by the JMA when issuing tsunami information for each coastal block after a tsunami forecast has been made.

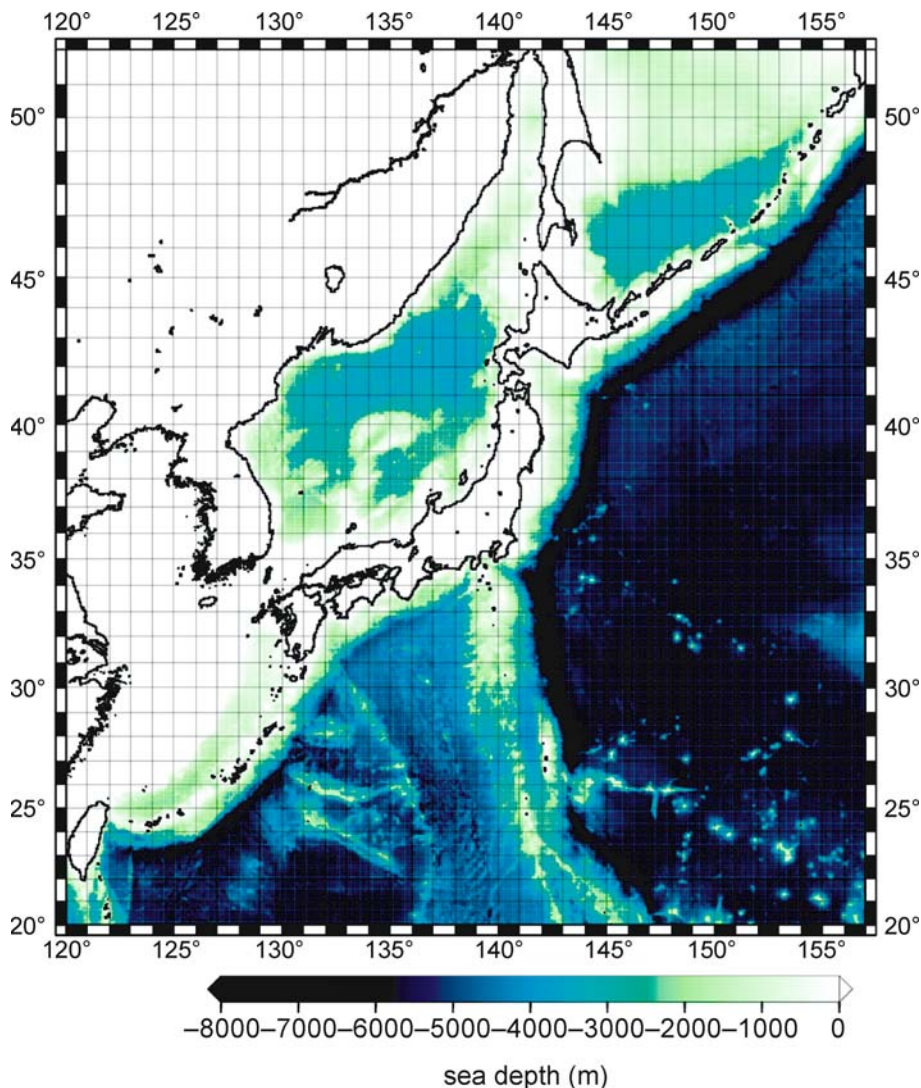
When the estimated tsunami amplitude is less than 20 cm, no warnings or advisories are issued, but a message "No threat of tsunami disaster" is promptly provided to the public in order to prevent unnecessary spontaneous evacuations in accordance with the widely spread recommendation in Japan: "When you feel a strong shaking near a coast, run to a high place without waiting for tsunami forecasts from the JMA".

**Tsunami Database Creation** The JMA introduced in 1999 a numerical simulation technique to implement accurate tsunami warning. However, even most-advanced computers require substantial time for the completion of the calculations. Therefore it is impossible to issue prompt tsunami warning based on numerical simulation tech-



**Tsunami Forecasting and Warning, Figure 6**

Distribution of simulation points in and around Japan. *Solid triangles* denote locations of simulation points to which likely hypothetical fault models are attributed. *Spacing* between the simulation points is 0.5 degrees in near coastal areas, and 1.0 degree in off-shore areas



**Tsunami Forecasting and Warning, Figure 7**

Computational area for creating the Japanese tsunami simulation database. Bathymetry is given by a *color grade scale*. One arc minute mesh bathymetry data are used throughout this area for the tsunami simulation

nique even if the calculations start simultaneously with the occurrence of an earthquake.

Alternatively, the JMA employs a database method to achieve a breakthrough on this issue. Tsunami propagation originating from various locations, fault sizes and likely rupture mechanisms are calculated in advance and the results, namely estimates of maximum tsunami amplitudes and arrival times, are stored in a database together with the associated hypocenter location and magnitude.

If an earthquake occurs, the most appropriate case for the actual hypocenter location and magnitude is retrieved

from the database and the tsunami forecast is issued accordingly.

**Hypocentral Location** The first step to create a tsunami database is to set fault models. ‘Simulation points’ are defined as surface projection locations of the center of possible causing faults, and are placed in the areas where tsunamigenic earthquakes are likely to occur. These points are selected on the basis of background seismicity and the records of the past earthquakes that generated tsunami.

Figure 6 shows the locations of simulation points in Japan and surrounding areas.

The total number of scenarios calculated for these points is about 64,000 (lat.  $\times$  lon.  $\times$  dep.  $\times$  mag.). Earthquakes with hypocenter depth of more than 100 km or magnitude less than 6.2 are not considered, because they do not cause disastrous tsunami [23].

**Fault Parameters Setting** Figure 1 shows fault parameters used in the numerical simulation. Length  $L$ , width  $W$ , and slip amount  $D$  of the fault are represented in terms of empirical formulas as functions of magnitude. The following are common relations used in Japan [37] based on  $M = M_j$ ma:

$$\begin{aligned} \log L &= 0.5M - 1.8 \quad \text{with } L \text{ in km,} \\ W &= L/2 \quad \text{or } \log W = 0.5M - 2.1 \\ &\quad \text{with } W \text{ in km,} \\ \log D &= 0.5M - 3.3 \quad \text{with } D \text{ in m.} \end{aligned}$$

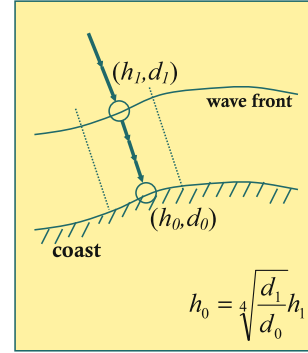
These formulas are consistent with the next two formulas, assuming a common rigidity  $\mu = 2.0 \times 10^{10}$  Newton/m<sup>2</sup> of the rock-material in the source area

$$\begin{aligned} Mo &= \mu \times L \times W \times D \\ \log Mo &= 1.5 Mw + 9.1 \\ &\quad \text{(in international standard units, i. e., Mo} \\ &\quad \text{in Nm = Newton meter).} \end{aligned}$$

The dip ( $\delta$ ), strike ( $\phi$ ) and slip ( $\lambda$ ) angles of the fault at each simulation point are based on average values of past earthquakes at or near these locations. But if they are uncertain, they are assumed to be those of a pure reverse fault ( $\lambda = 90$  degrees) whose strike is parallel to the trench or nearby coast, and with a dip angle of 45 degrees, corresponding to an efficient tsunami generation in view of disaster management.

**Initial Value for Numerical Simulation of Tsunami Propagation** The vertical deformation of sea floor due to fault motion is calculated by the elastic theory [24], using fault parameters as set above. The initial deformation of the sea surface is assumed to be identical to the vertical deformation of the sea floor and used as input value for the numerical simulation, because in most cases, rupture propagation velocity is much higher than the tsunami propagation velocity, and the ocean bottom deformation can be treated to occur instantaneously [14].

**Numerical Simulation of Tsunami Propagation** Non-linear long-wave approximation equation with advection term and ocean-bottom friction term is adopted as the equation of motion, and solved together with the equation of continuity as shown below by finite difference method



**Tsunami Forecasting and Warning, Figure 8**  
Schematic sketch to illustrate the application of Green's Law for tsunami amplitude calculation in near coastal areas. Ray convergence and divergence are ignored

with staggered leap frog scheme [27].

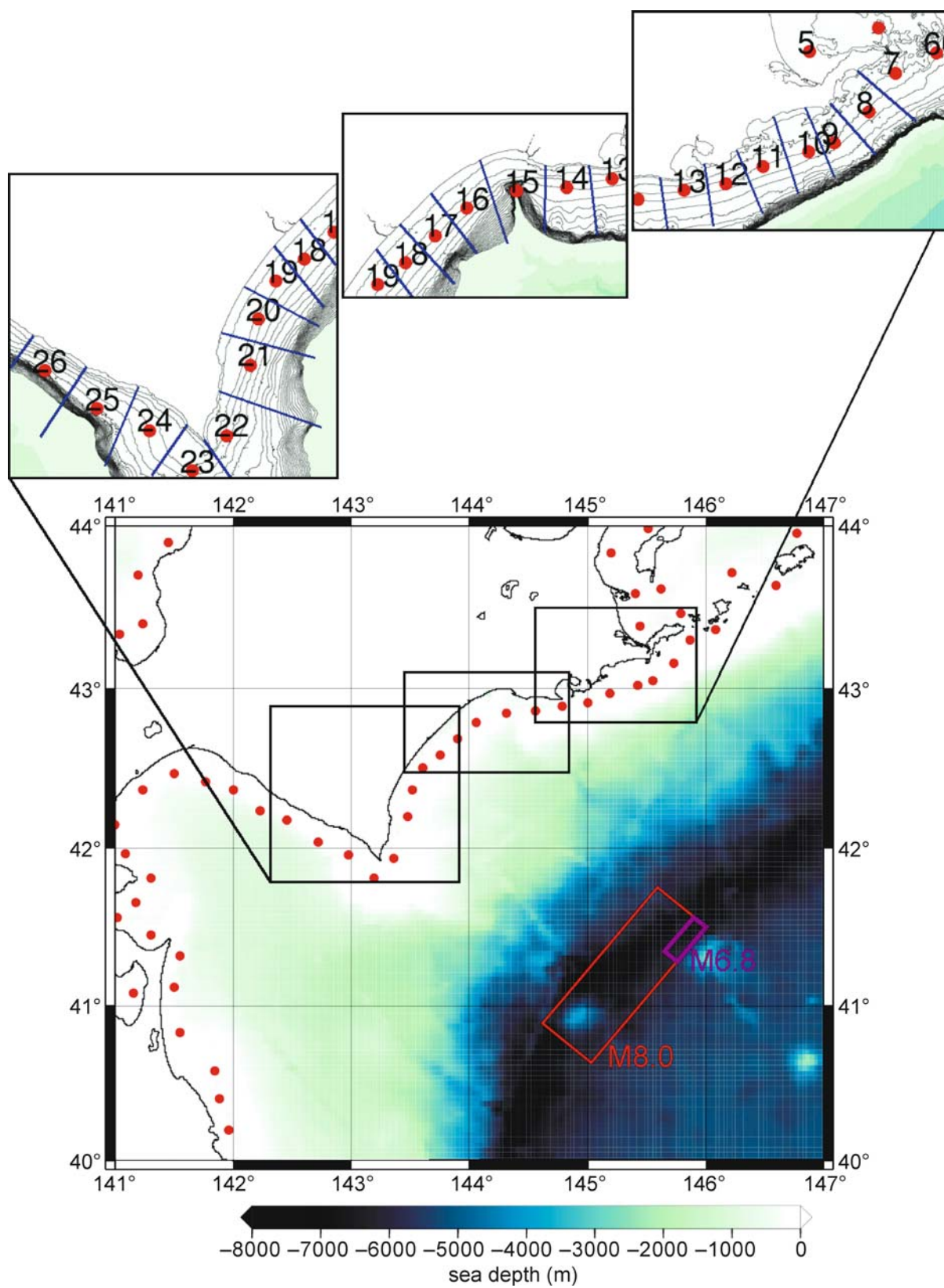
$$\begin{aligned} \frac{\partial V_x}{\partial t} + V_x \frac{\partial V_x}{\partial x} + V_y \frac{\partial V_x}{\partial y} &= -f V_y - g \frac{\partial h}{\partial x} - C_f \frac{V_x \sqrt{V_x^2 + V_y^2}}{d + h} \\ \frac{\partial V_y}{\partial t} + V_x \frac{\partial V_y}{\partial x} + V_y \frac{\partial V_y}{\partial y} &= f V_x - g \frac{\partial h}{\partial y} - C_f \frac{V_y \sqrt{V_x^2 + V_y^2}}{d + h} \\ \frac{\partial h}{\partial t} + \frac{\partial}{\partial x} \{V_x(h + d)\} + \frac{\partial}{\partial y} \{V_y(h + d)\} &= 0, \end{aligned}$$

where  $V_x$  and  $V_y$  are the  $x$  (east) and  $y$  (south) components of the average water particle motion velocity in the depth direction, and  $h$  and  $d$  are sea surface displacement and sea depth, respectively.  $f$  is the Coriolis parameter ( $= 2\Omega \cos \theta$ ,  $\Omega$  is the angular frequency of Earth's self rotation,  $\theta$  is the co-latitude), and  $C_f$  is the sea bottom friction coefficient.

#### ► Tsunami Forecasting and Warning, Figure 9

The map shows the distribution of forecast points (red dots) along the southern and eastern shore of Hokkaido and part of NE Honshu together with the surface projections of two hypothetical faults representing earthquakes with magnitude 8.0 (red rectangle) and 6.8 (violet rectangle), respectively. The depth of the fault top is 1 km from the sea bottom for both cases. Pure reverse fault with dip angle 45 degrees is assumed. In the upper half of the figure the coastal blocks are enlarged, showing the locations of forecast points denoted by numerals and the blue separation lines between sub-sections of the coast. The numbers correspond to the 'forecast point number' given on the abscissa of Fig. 10. Also depicted are bathymetry contour lines near the coast at 20 m depth intervals







Advection and ocean bottom friction are considered only in sea areas with sea depth less than 100 m. Coriolis force is considered only in the distant tsunami case (see Sect. “[Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory \(NWPTA\)](#)”).

Figure 7 shows the computational area. Mesh size is 1 arc minute throughout the computational area. Eight hours of propagation simulation is conducted for all scenarios with a fixed time interval three seconds for the integration. The Coriolis force is not considered for local and regional tsunami simulations. Total reflection boundary condition is used at the land-ocean boundary, and open boundary condition is used outside of the computational area.

To represent the tsunami waveform correctly in a shallow sea area, very fine bathymetry data mesh is necessary (in a strict sense, 20 or more grid points are necessary within one wave-length [31]), and a vast time is required for the completion of such detailed calculations. To overcome this difficulty, the numerical simulation with the long-wave approximation is applied only to points which are a few to a few ten kilometers seaward from the coast (“forecast points”) where sea depth is about 50 m. Then, tsunami amplitude at the coast is calculated by using Green’s law described in the next section.

**Derivation of Tsunami Amplitude at Coast** To estimate tsunami amplitude at the coast, Green’s law (= energy conservation law) is applied to the estimated amplitude at the forecast points just seaward of the coast. As shown in Fig. 8, Green’s law says that the tsunami amplitude at the coast is represented by fourth root of the ratio of sea depth at the forecast point and at the coast:

$$h_0 = \sqrt[4]{\frac{d_1}{d_0}} h_1 .$$

$h$  and  $d$  denote tsunami amplitude and sea depth, and suffix ‘0’ and ‘1’ denotes the value at the coast and forecast point.

As the forecast points are set close enough to the coast, horizontal convergence and divergence of the rays can be ignored, i. e., the wave front is regarded as almost parallel to the coast due to a refraction towards the direction of the largest sea-depth gradient. As for the “ $d_0$ ” value, the sea depth at the coast, 1 m is assumed, but actually measured value would be more appropriate, if available.

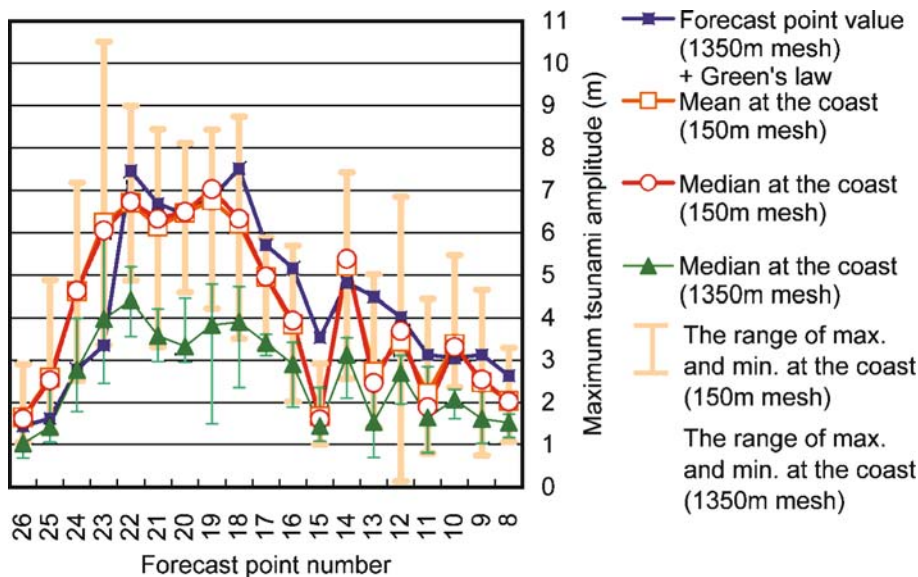
**Adequacy of the Green’s Law Application** The adequacy of Green’s Law application to estimate tsunami amplitude at the coast is examined by using the TSUNAMI-

N2 [12] program. It is equivalent to the program used at the JMA [27] to create tsunami simulation database. Up to 150 m mesh at the finest was used by adopting nesting technique, and the tsunami amplitudes estimated at the coastal mesh are considered “true” ones. We compared the case using 1,350 m mesh throughout the computational area (which is equivalent to 1 arc minute mesh) and then estimating the tsunami amplitude at the coast by applying Green’s Law to the value at the offshore forecast point, with the expected “true” amplitude at the coastal point using finer mesh. Forecast points actually used for the database creation are shown in Fig. 9. Hypothetical fault models corresponding to two different magnitudes (8.0 and 6.8) have been depicted in this Figure as red and violet rectangles, respectively, south off the Pacific coast of Hokkaido Island. The results are shown in Fig. 10a,b. For both events the estimated tsunami amplitudes at the coast by using a 1,350 m mesh and applying Green’s Law to the forecast points (solid squares) agree well with the means (open squares) and medians (open circles) of tsunami amplitudes calculated by using a finer mesh in each coastal subsection. The coastal subsections are separated by middle points between the forecast points (see upper part of Fig. 9). Thus, it was confirmed that the application of Green’s Law to tsunami amplitudes at the offshore forecast points yields amplitude estimates at the coast that are comparable with those calculated with much more time consuming fine mesh simulations. Therefore, we consider the former as representative estimates of the tsunami amplitude in the coastal subsection centered at the coastal projection of the forecast point.

It could also be shown that direct estimates of tsunami amplitudes at the coast based on coarse mesh simulations will underestimate values. This trend is even clearer for smaller event with shorter typical tsunami wavelengths (Fig. 10b). Therefore, it is not proper to estimate the tsunami amplitude directly at the coastal mesh when the mesh size is coarse relative to the typical tsunami wavelength.

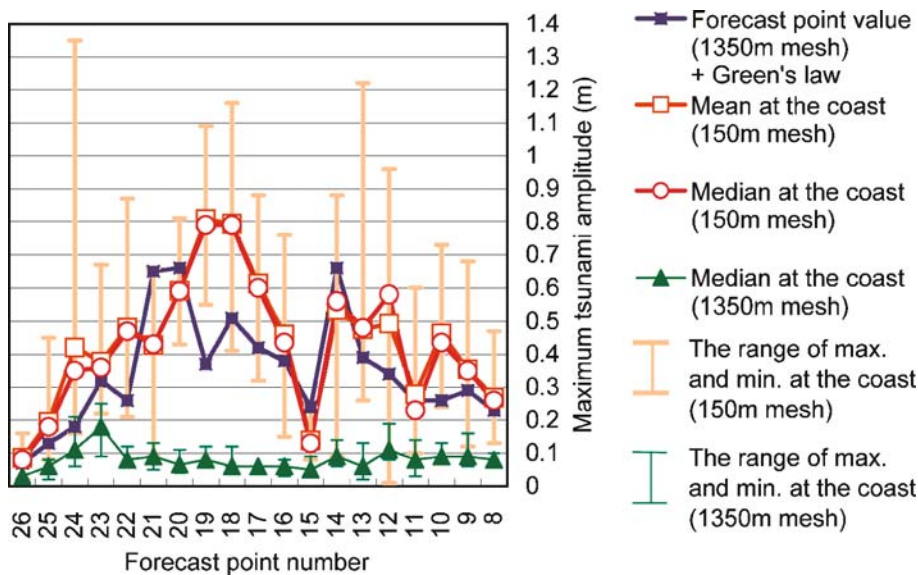
However, there will be some issues to be considered in future in this method as follows.

- (a) Strictly speaking, the Green’s Law should only be applied to a direct wave. Therefore, in case the maximum amplitude is created by the superposition of multiply reflected waves, then this method can give overestimated tsunami amplitude.
- (b) Scatter of estimated tsunami amplitude in one coastal subsection differs from section to section depending on the complexity of the coast line feature. It would be very difficult to precisely represent tsunami ampli-

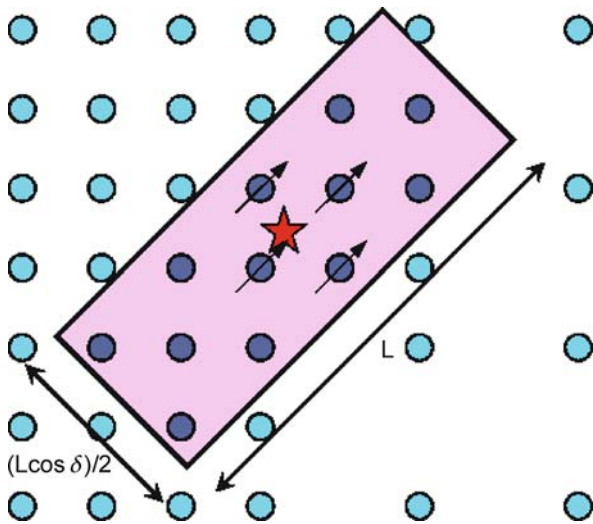


Tsunami Forecasting and Warning, Figure 10a

The presented diagram illustrates the adequacy of the Green's law application. Shown are the tsunami forecast results for the hypothetical earthquakes of magnitude 8.0 (*above*) and 6.8 (*below*). On the abscissa the numbers of the 'forecast points' depicted in Fig. 9 are given. On the ordinate the maximum tsunami amplitudes in meter are given, as they result from the application of different modeling procedures (fine mesh as well as course mesh with and without applying Green's law; see legend). *Vertical thick bars* denote the range between maximum and minimum values of estimated tsunami amplitudes in each sub-section using an up to 150 m mesh. *Vertical thin bars* denote the range of directly estimated maximum and minimum tsunami amplitudes at the coast in each sub-section using a 1,350 m mesh



Tsunami Forecasting and Warning, Figure 10b  
(continued)



**Tsunami Forecasting and Warning, Figure 11**  
Maximum Risk Method: The star denotes the epicenter and the circles the locations of simulation points. The rectangle delimitates the searching area from which the simulation points have to be retrieved. This rectangle is centered at the epicenter, and its length and width are  $L$  and  $0.5L \cos \delta$ , respectively, where  $L$  is given by the scaling law (see Subsect. "Fault Parameters Setting"). The strike of the length is the same as that given to the nearest simulation point to the epicenter. All simulation points inside of the rectangle (solid circles) are selected. Small arrows point into the strike direction assigned to simulation points

tudes for individual coastal points even if very fine bathymetry data are used in a simulation, because a short wavelength tsunami is significantly affected by small scale bathymetry and coastal feature or by a small change in an initial tsunami waveform. Therefore, incorporating a standard deviation of estimated tsunami amplitude to represent a degree of scatter in one coastal subsection, as well as a median or mean, in the tsunami warning/advisory grade determination would be effective for disaster management.

**Retrieval of the most Appropriate Case from the Database** In general, no identical case for actually determined latitude, longitude, hypocenter depth, magnitude exists in the database. Thus, some kind of selection rule is necessary to choose the most appropriate one.

**Hypocenter Depth and Magnitude** The logarithm of the estimated tsunami amplitude changes approximately positively linear with magnitude and negatively linear with hypocenter depth. Therefore, first, one has to calculate the logarithm of the estimated tsunami amplitudes for

**Tsunami Forecasting and Warning, Table 2a**  
Tsunami forecast/information examples of the JMA. Original texts are in Japanese and these are English translations: a Tsunami forecast of the first issuance, b Tsunami information (estimated tsunami arrival time and amplitude) that follows a, c Tsunami information (estimated high tide and tsunami arrival times) that follows a, d Tsunami information (Observed results), e Tsunami forecast revision, f Tsunami forecast cancellation

Tsunami Warning
Issued by the Japan Meteorological Agency (JMA)
Issued at 2029JST 15 Nov 2006
"Tsunami Warning"
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory"
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.

four cases (i.e., for two different depths and magnitudes each) which include the calculated depth and magnitude in between, then apply a two-dimensional linear interpolation and finally convert the log values back to a tsunami amplitudes.

**Latitude and Longitude** As for the location of the epicenter, either one of the two methods is used depending on the hypocentral area.

**Interpolation method** One selects the nearest four locations of simulation points in the data base which enclose the epicenter determined for the real event and performs a two-dimensional linear interpolation. This method is used far off the coast areas where horizontal shifts of the source location cause relatively smooth changes in the estimated tsunami amplitude at the coast. This is applicable in areas with a simulation point spacing of 1 degree (see Fig. 6).

**Maximum risk method** In the case that the assumed fault locations in the data base differ from the real location, tsunami amplitude estimates near the epicenter may differ substantially from observed ones for events oc-

**Tsunami Forecasting and Warning, Table 2b**  
(continued)

Tsunami Information (Estimated Tsunami Arrival Time and Amplitude) Issued by the Japan Meteorological Agency (JMA) Issued at 2030JST 15 Nov 2006		
— Estimated Tsunami Arrival Time and Amplitude —		
Coastal Block	Arrival Time	Amplitude
"Tsunami Warning"		
Eastern Part of Pacific Coast of Hokkaido	2110 15 Nov	1 m
Okhotsk Sea Coast of Hokkaido	2120 15 Nov	2 m
"Tsunami Advisory"		
Central Part of Pacific Coast of Hokkaido	2130 15 Nov	0.5 m
Western Part of Pacific Coast of Hokkaido	2150 15 Nov	0.5 m
Northern Part of Japan Sea Coast of Hokkaido	2250 15 Nov	0.5 m
Pacific Coast of Aomori Pref.	2140 15 Nov	0.5 m
Iwate Pref.	2140 15 Nov	0.5 m
Miyagi Pref.	2140 15 Nov	0.5 m
Fukushima Pref.	2210 15 Nov	0.5 m
Ibaraki Pref.	2210 15 Nov	0.5 m
Kujukuri and Sotobo Area of Chiba Pref.	2210 15 Nov	0.5 m
Uchibo Area of Chiba Pref.	2210 15 Nov	0.5 m
Izu Islands	2210 15 Nov	0.5 m
Sagami Bay and Miura Peninsula	2220 15 Nov	0.5 m
Shizuoka Pref.	2220 15 Nov	0.5 m
Tsunami may be higher than the estimation in some places.		
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

curing near to the coast. The uncertainty in the relative location of the epicenter on the surface projection of the seismic fault has to be considered at an early stage of the seismic data analysis, and all possible cases should be taken into account. Four cases with the epicenter located on one of the four corners of the fault are the most extreme cases, and the truth lies in between. Therefore, in the case of near coastal sources and with a view to disaster management, cases corresponding to every simulation point inside the rectangular area shown in Fig. 11 (solid circles) have to be considered. This rectangular area is centered at the epicenter (star mark). Its length and width are determined via the empirical formulas given in Subsect. "Fault Parameters Setting". Note, however, that the projected width in Fig. 11 is only  $0.5L \cos \delta$  and depends on the source dip. Also one has to remember that each

simulation point is located in the center of the respective scenario fault in the data base. Therefore, the length and width of the rectangle to select all possible 'simulation points' are,  $L$  (not  $2L$ ) and  $0.5L \cos \delta$  respectively, and the strike of the length is that given to the simulation point (denoted by an arrow), which is closest to the determined epicenter. And the worst case is chosen amongst them for each of the forecast points.

When readings of the tsunami arrivals at tidal gauges or tsunami meter become available, improbable simulation points can be excluded by applying inverse refraction diagrams. Thus the uncertainty in the spatial extent of the tsunami source area can be reduced.

If the number of the simulation points inside of the rectangular area is less than 4, the nearest 4 simulation points from the epicenter are retrieved.

**Tsunami Forecasting and Warning, Table 2c**  
(continued)

Tsunami Information (Estimated High Tide and Tsunami Arrival Time)		
Issued by the Japan Meteorological Agency (JMA)		
Issued at 2030JST 15 Nov 2006		
— Estimated High Tide and Tsunami Arrival Time —		
If tsunami arrives at coasts at around high tide time, tsunami becomes higher.		
Coastal Block	High Tide Time	Tsunami Arrival
“Tsunami Warning”		
Eastern Part of Pacific Coast of Hokkaido		2110 15 Nov
Kushiro	2335 15 Nov	2130 15 Nov
Hanasaki	2338 15 Nov	2120 15 Nov
Okhotsk Sea Coast of Hokkaido		2120 15 Nov
Abashiri	2226 15 Nov	2140 15 Nov
Monbetsu	2206 15 Nov	2200 15 Nov
Esashi-ko	2228 15 Nov	2220 15 Nov
“Tsunami Advisory”		
Central Part of Pacific Coast of Hokkaido		2130 15 Nov
Urakawa	2353 15 Nov	2140 15 Nov
Tokachi-ko	2359 15 Nov	2140 15 Nov
Western Part of Pacific Coast of Hokkaido		2150 15 Nov
Muroran	2337 15 Nov	2210 15 Nov
Hakodate	0008 16 Nov	2220 15 Nov
Tomakomai-nishi-ko	0007 16 Nov	2210 15 Nov
Yoshioka	0026 16 Nov	2230 15 Nov
Northern Part of Japan Sea Coast of Hokkaido		2250 15 Nov
Wakkanai	0229 16 Nov	2310 15 Nov
Rumoi	0114 16 Nov	2250 15 Nov
Otaru	0105 16 Nov	2250 15 Nov
Pacific Coast of Aomori Pref.		2140 15 Nov
Hachinohe	0000 16 Nov	2200 15 Nov
Sekinehama	0005 16 Nov	2200 15 Nov
Iwate Pref.		2140 15 Nov
Miyako	0006 16 Nov	2150 15 Nov
Ofunato	0017 16 Nov	2150 15 Nov
Kamaishi	0011 16 Nov	2150 15 Nov
(abbreviated)	(abbreviated)	(abbreviated)
Sea level may slightly fluctuate in other coastal areas but no danger.		
— Earthquake Information —		
Origin Time: 2015JST 15 Nov 2006		
Epicenter: 46.6 North, 153.6 East		
Depth: 30 km		
Magnitude: 8.1		

*Tsunami Arrival Time Estimation* Forecast points are located not along the coast but offshore. Travel time estimation by adding the travel time from the source area to the forecast point to that from the forecast point to the coastal point may lead to substantial errors, especially when the

tsunami source area is close to the coast. Therefore, for the tsunami arrival time estimation at the coastal point, an inverse refraction diagram from the coastal point to the source area is used to reduce such errors. Finite spatial extent of the source area, depending on the calculated mag-



**Tsunami Forecasting and Warning, Table 2d**  
(continued)

Tsunami Information(Observation Results)				
Issued by the Japan Meteorological Agency (JMA)				
Issued at 2207JST 15 Nov 2006				
— Tsunami Observation at Sea Level Stations as of 2205 15 Nov —				
Higher tsunami may have arrived at some places other than the sea level stations.				
Tsunami may glow higher later.				
Kushiro	First Wave	2143 15 Nov	(+)	0.2 m
	Maximum Wave	2155 15 Nov		0.2 m
Hanasaki	First Wave	2129 15 Nov	(+)	0.4 m
	Maximum Wave	2143 15 Nov		0.4 m
Tokachi-ko	First Wave	2149 15 Nov		unclear
	Maximum Wave	(arriving)		
“Tsunami Warning” has been in effect for the following coastal blocks.				
Eastern Part of Pacific Coast of Hokkaido				
Okhotsk Sea Coast of Hokkaido				
“Tsunami Advisory” has been in effect for the following coastal blocks.				
Central Part of Pacific Coast of Hokkaido				
Western Part of Pacific Coast of Hokkaido				
Northern Part of Japan Sea Coast of Hokkaido				
Pacific Coast of Aomori Pref.				
Iwate Pref.				
Miyagi Pref.				
Fukushima Pref.				
Ibaraki Pref.				
Kujukuri and Sotobo Area of Chiba Pref.				
Uchibo Area of Chiba Pref.				
Izu Islands				
Sagami Bay and Miura Peninsula				
Shizuoka Pref.				
Sea level may slightly fluctuate in other coastal areas but no danger.				
— Earthquake Information —				
Origin Time: 2015JST 15 Nov 2006				
Epicenter: 46.6 North, 153.6 East				
Depth: 30 km				
Magnitude: 8.1				

nitude value, is considered, and the earliest arrival time corresponding to the crossing point of the inverse refraction diagram and the outer rim of the source area is used with a view to disaster management.

**Tsunami Forecast Assembling** The tsunami forecast grade for a considered coastal block depends on the maximum of the expected tsunami amplitudes at the forecast points located in that block which are converted to amplitudes at the coast by applying Green’s law. Forecast points

are placed off-shore parallel to the coast with spacing of about 20 km. The average number of forecast points in one coastal block is 9. As mentioned above, expected tsunami amplitude at a forecast point, including Green’s law application when using 1 arc minute bathymetry mesh, agrees reasonably well with expected mean or median tsunami amplitudes on the coast that have been calculated by using very fine bathymetry mesh in a coastal subsection separated in the middle between forecast points. This means that the expected tsunami amplitude at the forecast point

**Tsunami Forecasting and Warning, Table 2e**  
(continued)

<b>Tsunami Warning (Revision)</b>
<b>Issued by the Japan Meteorological Agency (JMA)</b>
<b>Issued at 2330JST 15 Nov 2006</b>
"Tsunami Warning" has been revised into "Tsunami Advisory"
for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
"Tsunami Advisory" has been in effect for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Okhotsk Sea Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Northern Part of Japan Sea Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Shizuoka Pref.
Ogasawara Islands

can be regarded as being representative for the respective coastal subsection. We think this approach is reasonable with a view to disaster management requirements because the forecast grades for a coastal block are based on the maximum of these representative values in each coastal subsection. This prevents that the forecast grade is affected by abnormally large local maxima as they may result from very fine-mesh simulations (see Fig. 10). However, the JMA informs in its public information releases that tsunami amplitude might very locally be much higher than in the forecast.

The content of tsunami warning/advisory is very simple, namely, tsunami grades and corresponding coastal block names. This is to enable recipient organizations to understand the most important information easily, namely the necessity of evacuation. Tsunami information that follows just after the warning/advisory provides then more detailed estimates of tsunami amplitudes and arrival times for each coastal block, as well as hypocentral parameters. Examples of tsunami forecast/information are given in Table 2a–f.

**Tsunami Forecasting and Warning, Table 2f**  
(continued)

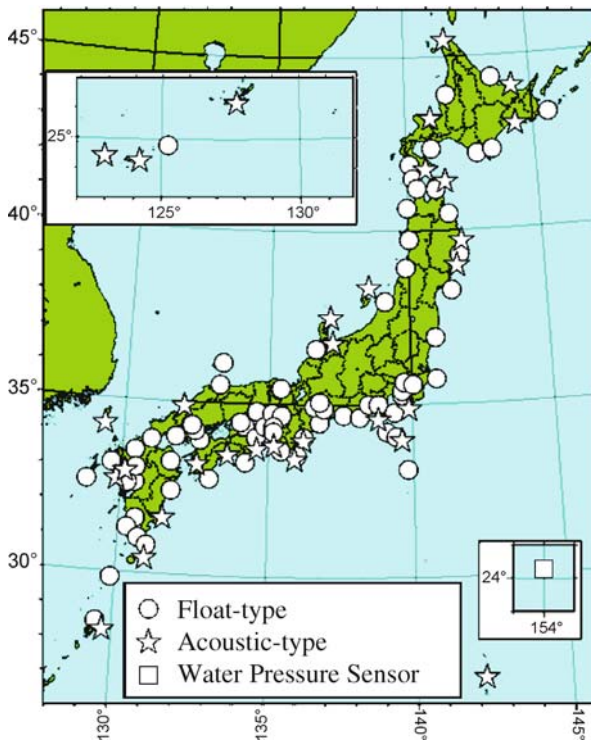
<b>Cancellation of Tsunami Warning</b>
<b>Issued by the Japan Meteorological Agency (JMA)</b>
<b>Issued at 0130JST 16 Nov 2006</b>
Tsunami warning has been all canceled.
"Tsunami Advisory" has been canceled for the following coastal blocks.
Eastern Part of Pacific Coast of Hokkaido
Central Part of Pacific Coast of Hokkaido
Western Part of Pacific Coast of Hokkaido
Pacific Coast of Aomori Pref.
Iwate Pref.
Miyagi Pref.
Fukushima Pref.
Ibaraki Pref.
Kujukuri and Sotobo Area of Chiba Pref.
Uchibo Area of Chiba Pref.
Izu Islands
Sagami Bay and Miura Peninsula
Ogasawara Islands
However, sea level is expected to fluctuate in above coastal areas.
Be careful in sea bathing and fishing.

**Tsunami Forecast Dissemination** Like other forecast/information disseminated from the JMA, tsunami forecast/information is transmitted to relevant organizations by fully utilizing existing online facilities. In case of land line failure, forecast/information is also transmitted through satellite link. Loud speaker/sires operated by municipalities/prefectures are used in 'the last mile' to reach the public. Police and fire departments have their own robust transmission routes, which are also used. Additionally, the public broadcasting company plays a very important role because of its wide outreach and timeliness.

### Tsunami Monitoring System

Data from more than 100 tidal observation stations (including those operated by other institutes, like the Japan Coast Guard) are collected on-line at the JMA (Fig. 12). Data sampling and transmission rate are every one second.

Arrival time, initial amplitude, polarity, maximum tsunami amplitude and corresponding time are measured on the man-machine interface depicted in Fig. 13, after prior removal of the ocean tidal component. Tsunami observation results are issued in tsunami information.



**Tsunami Forecasting and Warning, Figure 12**

Tidal stations network used for tsunami monitoring by the JMA. More than 100 tidal stations are monitored in real time. Symbols denote instrumentation type. Circle, star and square denote float-type, acoustic-type and pressure sensor type, respectively

Sea level data are used for the following: 1) Deciding about the time of tsunami warning/advisory cancellation; 2) Revision of the grade of warning/advisory in the case that estimated and observed tsunami amplitudes differ significantly and 3) Informing the public on the up-to-date status of tsunami observation.

#### **Tsunami Forecast for Distant Event and Northwest Pacific Tsunami Advisory (NWPTA)**

For a distant event, a similar method to a local event as mentioned above is used, but the seismic waveform data from global network which are available in real-time through the Internet is used for hypocenter and magnitude calculation, instead of those from domestic seismic network. The PTWC's bulletins are also incorporated in the framework of ICG/PTWS. For magnitude estimation, *M<sub>w</sub>* [34] is used commonly with PTWC and WC/ATWC. Tsunami simulation database has been created also for distant events, taking the Coriolis force effect into account considering long distance propagation.

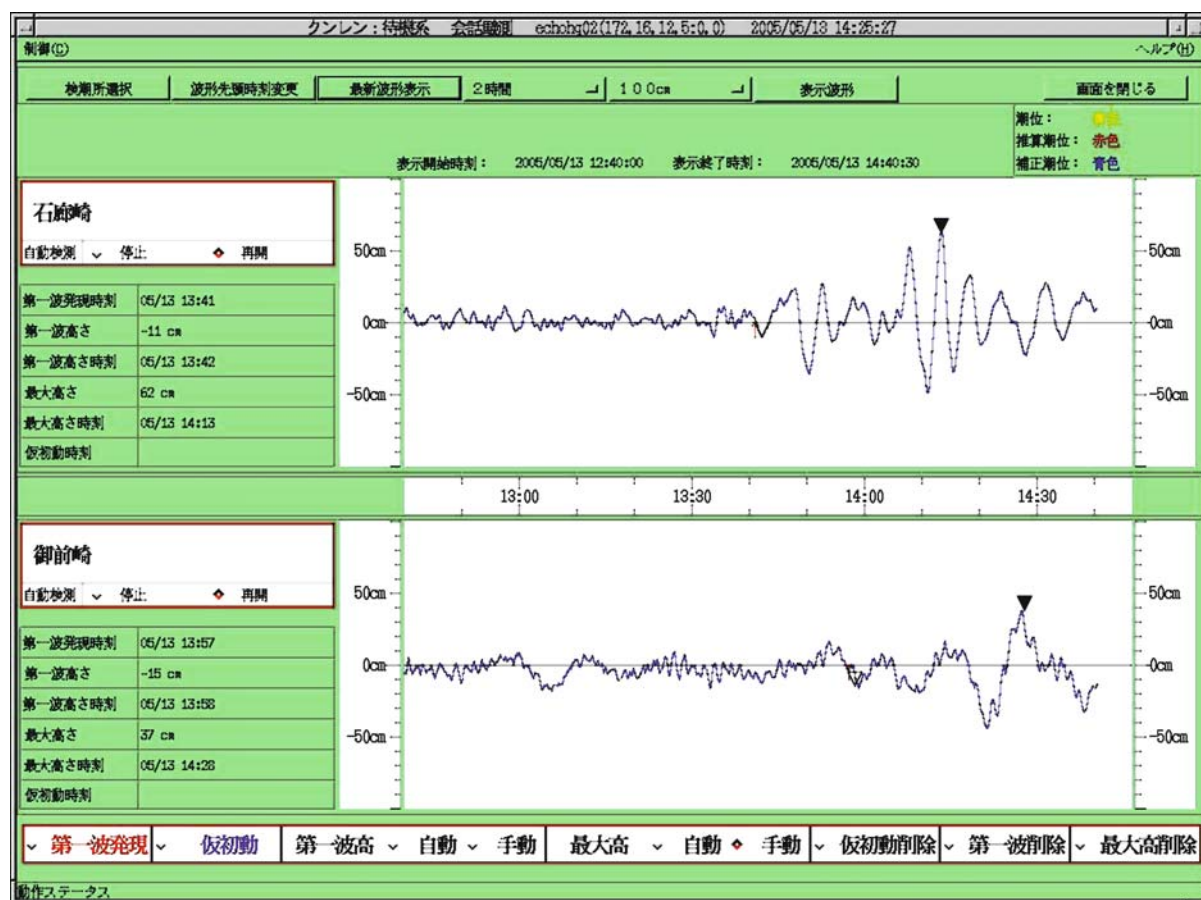
In case enough time is left until the earliest estimated tsunami arrival time at Japan coast, the JMA monitors sea level data collected via the data collection platform function of geostationary meteorological satellites, and circulated through GTS (global telecommunication network operated by the World Meteorological Organization of UN) circuit in near-real-time. After confirming the generation of tsunami at sea level observation sites near the source, tsunami warnings/advisories are disseminated for the same coastal blocks as for local tsunami if necessary. Estimated tsunami amplitudes in the database can be calibrated by the actually observed tsunami amplitudes before the dissemination. Furthermore, if a reliable causing fault model is available, the JMA conducts tsunami propagation simulation for a specific setting of fault location, depth, magnitude and mechanism so that tsunami warnings/advisories can be based on more reliable tsunami estimation.

By using the same method as described above in this section, the JMA, as a regional center of the ICG/PTWS, is providing international tsunami information (NWPTA: Northwest Pacific Tsunami Advisory) to relevant countries when an earthquake with magnitude 6.5 or larger occurs in the northwest Pacific area since March 2005. NWPTA contains estimated tsunami amplitudes and arrival times, as well as estimated hypocenter parameters.

#### **Some Lessons Learnt from a Recent Event**

At 11:14 (UTC) on Nov. 15, 2006, a large earthquake with *M<sub>w</sub>*8.3 occurred in the Kuril Island Region which generated a Pacific-wide tsunami. Tsunami records at tidal gauges in Japan are shown in Fig. 14. The JMA disseminated tsunami forecasts at 11:29 (UTC) (corresponding to 20:29 JST in the related case example given in Table 2) based on the estimated tsunami amplitudes in the database. About 4 hours after the first detection of the tsunami on the Japan coast, the JMA cancelled all the forecasts at 16:30 (UTC) judging from the amplitude decay in the sea level record of Chichijima station (cf. Fig. 14) that no higher tsunami will be observed. But, after the cancellation, maximum amplitudes were recorded at many of the tidal gauges along the Pacific coast of Japan up to 5 hours later. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota).

The JMA conducted a close examination of this case and found that these large amplitudes were tsunami waves reflected from the Emperor Sea Mount Chain in the middle of the Pacific (Fig. 15). Since the computational area for the tsunami simulation database of the JMA was limited around Japan (cf. Fig. 7), such reflected wave could



Tsunami Forecasting and Warning, Figure 13

Example of the JMA's man-machine interface screen image in the Sea Level Data Monitoring System. Shown are two sea level recordings by different tidal gauge stations. The ocean tide component has been removed for more precise reading of the relevant tsunami wave parameters. Arrival time, initial wave amplitude, polarity, maximum amplitude and corresponding time are picked. The reversed solid triangles denote the maximum amplitude within the analyzed time window

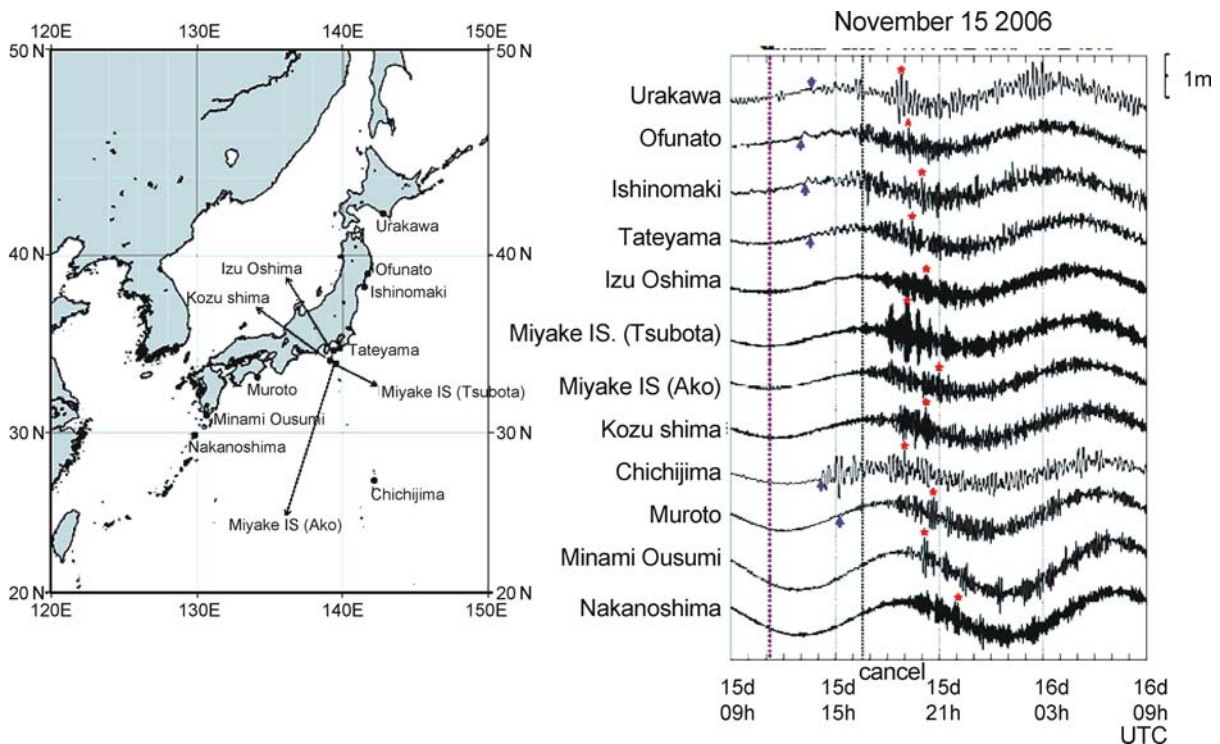
not be represented by the simulation. Therefore, the JMA examines now the scenarios for which the computational area should be expanded.

### Recent Improvements

**Application of Earthquake Early Warning (EEW) Technique to Quicken Tsunami Forecast** In the near-coastal area (i. e., roughly within 100 km from the coast) EEW hypocenter and magnitude estimates [15] are equivalent to those resulting from the common procedures described above. Therefore, EEW results can be used as input data for tsunami forecast. The JMA started this incorporation in October 2006. In the case of the Mjma6.8 Niigataken Chuetsu-oki earthquake of 16 July 2007, the JMA disseminated tsunami advisory one minute after the first detection of the seismic wave.

**Quicker Revision/Cancellation of Tsunami Forecasts by Utilizing CMT Solutions** In JMA, automatic CMT solutions are available in 10 to 20 minutes after the earthquake occurrence. They yield more realistic magnitudes than Mjma for really great earthquakes ( $M_w > 8$ ) and reliable fault plane solutions. On the other hand, it is not proper to wait for the CMT solution for issuing the first tsunami warning, and it is not realistic neither to prepare the tsunami simulation database for so many different dip, strike and slip angle settings with fine increments and choose the best fitting one, because it requires huge storage and computational time for the conduct of simulations for all parameter settings. Therefore, we use the CMT solution for the revision/cancellation of the first tsunami warning, and we limit the number of different fault parameter settings to 4, namely pure reverse fault with dip angles 10, 45 and 80 degrees and pure strike fault. Normal fault with





**Tsunami Forecasting and Warning, Figure 14**

Tidal station records of regular ocean tides superimposed by tsunami waves generated by the 15 Nov. 2006 Kuril earthquake ( $M_w = 8.3$ ). *Left*: Locations of tidal gauges whose data are shown in the right figure. *Right*: Tidal records of the stations shown in the left figure. Time is in UTC. The *thick blue dotted line* denotes the origin time of the earthquake and the blue arrows denote the arrival time of tsunami at each site. The *thin dotted line* denotes the time of tsunami forecast cancellation. *Red stars* mark the time of maximum amplitude at each site. At all sites maximum amplitudes were recorded about 2 to 5 hours after the forecast had been canceled. The largest tsunami amplitude recorded in Japan for this event was 84 cm at Miyake Is. (Tsubota)

the opposite slip angle ( $\lambda$ ) direction to the reverse fault with the same values for other fault parameters has almost identical tsunamigenic potential, and the difference is just a polarity of initial tsunami wave. Therefore, normal fault is treated as reverse fault with the same dip angle for this purpose. Strike angles have been fixed to the representative values of the considered regions, or set parallel to the nearby trench axis or coast line. The case with the highest resemblance to the actually calculated CMT solution with respect to its tsunamigenic potential is then retrieved.

The JMA incorporates CMT solutions in the revision/cancellation of tsunami forecast as follows:

- In the case of reverse or normal faults, and if the centroid depth is less than the depth determined by  $P$  and  $S$  arrival times  $+30$  km, and  $M_w$  0.5 or more larger than  $M_{jma}$ , the tsunami forecast is then upgraded in accordance with respective database cases.
- In the case of strike faults, taking the database into account, and after confirmation by low or not observed

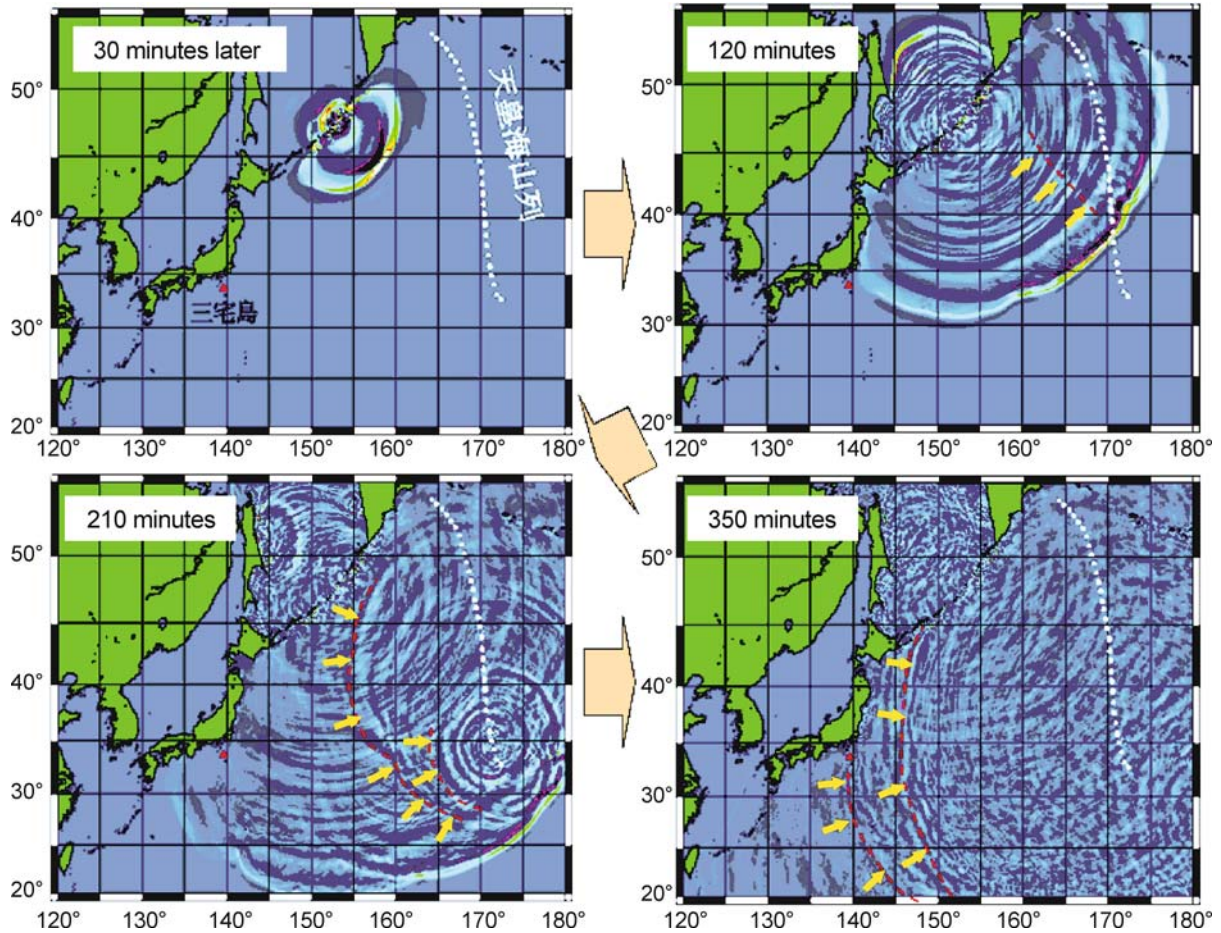
tsunami amplitudes at tide gauges, the tsunami forecast is then downgraded or canceled.

The JMA started this incorporation in July, 2007 for the sea area where additional database creation has been completed.

### Future Outlook

In tsunami forecasting, trade-off exists between promptness and accuracy/reliance. The JMA's tsunami forecasting strategy, especially for local events, is to satisfy both by using state-of-the-art technologies. They assure promptness of the first issuance of tsunami forecast, based on preliminary seismological results including that of EEW, and its subsequent revision, if required, as soon as more accurate and reliable data such as sea level change, and the results of a more detailed complex data analysis have become available. And as for accuracy/reliability there are two ways to be taken. One is quicken the process of reduc-





**Tsunami Forecasting and Warning, Figure 15**

Snap shots of numerical tsunami simulation for the 2006 Nov. 15 Kuril event (Mw8.3). Coriolis force has been considered for this case. Four snap shots (30, 120, 210 and 350 minutes after the earthquake occurrence) are shown. The *white dotted line* marks the position of the Emperor Sea Mount Chain. *Broken red curves with arrows* denote reflected tsunami waves from the Emperor Sea Mount Chain. The estimated arrival times of the reflected waves at the coast of Japan are consistent with the actual tidal records

ing the uncertainty of initial tsunami wave distribution assessments. Quick focal process inversion technique, new magnitude definitions applicable for gigantic or tsunami earthquakes and tsunami source inversion techniques using sea level data (also from satellite altimetry) will become effective for this purpose soon. The other way is very detailed numerical simulation, after reduction of uncertainties in the initial tsunami wave distribution, using fine bathymetry mesh data. Along with the development of integrated calculation algorithms, the improvement of the computer performance might solve this problem in future. But even then, due to the stochastic nature of tsunami behavior near the coast, one has carefully to examine on what statistical quantity of simulated results the tsunami forecast criterion has to be based.

### Acknowledgments

We thank Dr. Peter Bormann and Dr. Kenji Satake for reviewing the manuscript, their comments and suggestions greatly improved it.

### Bibliography

#### Primary Literature

1. Abe K (1973) Tsunami and mechanism of great earthquakes. *Phys Earth Planet Inter* 7:143–153
2. Bormann P, Wylegalla K (2005) Quick estimator of the size of great earthquakes. *Eos* 86(46):464
3. Bormann P, Baumbach M, Bock G, Grosser H, Choy GL, Boatwright J (2002) Seismic sources and source parameters. In: Bormann P (ed) *IASPEI new manual seismological observatory*

- practice, vol 1, Chap 3. GeoForschungsZentrum Potsdam, Potsdam, pp 1–94
4. Geller RJ (1976) Scaling relations for earthquake source parameters and magnitudes. *Bull Seism Soc Am* 66:1501–1523
  5. Geographical Survey Institute of Japan (2006) Real-time collection and analysis of crustal deformation data, Report on technical development and promotion plan concerning prompt disaster mitigation countermeasures based on disaster information, Chap 2. Ministry of Land, Infrastructure and Transport, Tokyo
  6. González FI, Bernard EN, Meinig C, Eble M, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. *Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program)* 35(1):25–39
  7. Jim Gower J (2005) Jason 1 detects the 26 December 2004 tsunami. *EOS Trans Am Geophys Union* 86(4):37–38
  8. Hara T (2007) Measurement of duration of high-frequency energy radiation and its application to determination of magnitudes of large shallow earthquakes. *Earth Planets Space* 59:227–231
  9. Hatori T (1984) On the damage to houses due to tsunamis. *Bull Earthq Res Inst* 59:433–439 (in Japanese)
  10. Hayashi Y (2008) Extracting the 2004 Indian Ocean tsunami signals from sea surface height data observed by satellite altimetry. *J Geophys Res* 113:C01001
  11. Ide S, Takeo M, Yoshida Y (1996) Source process of the 1995 Kobe earthquake: Determination of spatio-temporal slip distribution by Bayesian modeling. *Bull Seism Soc Am* 87: 547–566
  12. Imamura F (1997) IUGG/IOC TIME PROJECT Numerical method of tsunami simulation with the leap-frog scheme, part 3 (Programme lists for near field tsunami), vol 35. IOC Manuals and Guides, Paris
  13. Japan Meteorological Agency (2005) A magnitude estimation using borehole volume strainmeters for earthquake events near the coast of Sumatra, Indonesia. *Rep Coord Comm Earthq Predict* 74:575–577 (in Japanese)
  14. Kajiura K (1970) Tsunami source, energy and the directivity of wave radiation. *Bull Earthq Res Inst (Univ. of Tokyo)* 48: 835–869
  15. Kamigaichi O (2004) JMA earthquake early warning. *J Japan Assoc Earthq Eng (Special Issue)* 4:134–137
  16. Kanamori H, Anderson DL (1975) Theoretical basis of some empirical relations in seismology. *Bull Seism Soc Am* 65: 1073–1095
  17. Kanamori H, Rivera L (2007) Speeding up seismic tsunami warning using *W* phase. In: *Abstracts of AGU Fall Meeting 2007*, S43C-06
  18. Kasahara M, Sasatani T (1986) Body wave analyses of strain seismograms observed at Erimo, Hokkaido, Japan. *J Fac Sc Hokkaido Univ Ser. VII (Geophysics)* 8:83–108
  19. Katsumata A (2004) Revision of the JMA displacement magnitude. *Q J Seismol* 67:1–10 (in Japanese)
  20. Kikuchi M, Kanamori H (1991) Inversion of complex body waves, III. *Bull Seism Soc Am* 81:2335–2350
  21. Lomax A, Michelini A, Piatanesi A (2007) An energy-duration procedure for rapid determination of earthquake magnitude and tsunamigenic potential. *Geophys J Int* 170:1195–1209
  22. Matsu'ura M, Hasegawa Y (1987) A maximum likelihood approach to nonlinear inversion under constraints. *Phys Earth Planet Inter* 47:179–187
  23. Okada M, Tanioka Y (1998) Relation of tsunami generation ratio with earthquake magnitude and hypocentral depth. *Mon Kaiyo (Special issue)* 15:18–22 (in Japanese)
  24. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seism Soc Am* 75:1135–1154
  25. Ozawa S (1996) Geodetic inversion for the fault model of the 1994 Shikotan Earthquake. *Geophys Res Lett* 23(16): 2009–2012
  26. Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – the 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. *J Geophys Res* 94:5627–5636
  27. Satake K (1995) Linear and nonlinear computations of the 1992 Nicaragua earthquake tsunami. *PAGEOPH* 144:455–470
  28. Shuto N (1991) Historical changes in characteristics of tsunami disasters. In: *Proc of international symposium on natural disaster reduction and civil engineering*. Japan Society of Civil Engineering, Tokyo, pp 77–86
  29. Shuto N (1992) Tsunami intensity and damage, Tsunami engineering technical report. *Tohoku Univ* 9:101–136 (in Japanese)
  30. Shuto N (1998) Present state of tsunami research and defense works. *Bull Coastal Oceanogr* 35(2):147–157 (in Japanese)
  31. Shuto N et al (1986) A study of numerical techniques on the tsunami propagation and run-up. *Sci Tsunami Hazard* 4: 111–124
  32. Takanami T, Kitagawa G (eds) (2002) Methods and application of signal processing in seismic network operations. *Lecture Notes in Earth Science* vol 98. Springer, Berlin
  33. Titov VV, González FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. *Nat Hazards (Special Issue, US National Tsunami Hazard Mitigation Program)* 35(1):41–58
  34. Tsuboi S, Abe K, Takano K, Yamanaka Y (1995) Rapid determination of Mw from broadband *P* waveforms. *Bull Seism Soc Am* 83:606–613
  35. Tsushima H, Hino R, Fujimoto H, Tanioka Y (2007) Application of cabled offshore ocean bottom tsunami gauge data for real-time tsunami forecasting. In: *Proc symposium on underwater technology 2007/Workshop on scientific use of submarine cables and related technologies 2007*. The University of Tokyo, Tokyo, pp 631–639
  36. Ueno H, Hatakeyama S, Aketagawa T, Funasaki J, Hamada N (2002) Improvement of hypocenter determination procedures in the Japan Meteorological Agency. *Q J Seism* 65:123–134 (in Japanese)
  37. Utsu T, Shima E, Yoshii T, Yamashina K (2001) *Encyclopedia of Earthquakes*, 2nd edn. Asakura, Tokyo, pp 657
  38. Weinstein S, Okal E (2005) The mantle magnitude *M<sub>m</sub>* and the slowness parameter *theta*: Five years of real-time use in the context of tsunami warning. *Bull Seism Soc Am* 85: 779–799
  39. Wells DL, Coppersmith KJ (1994) New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement. *Bull Seism Soc Am* 84(4): 974–1002
  40. Yokota T, Zhou S, Mizoue M, Nakamura I (1981) An automatic measurement of arrival time of seismic waves and its application to an on-line processing system. *Bull Earthq Res Inst* 56:449–484 (in Japanese)

## Books and Reviews

- Bormann P (ed) (2002) IASPEI new manual of seismological observational practice, vol 1 and 2. GeoForschungsZentrum Potsdam, Potsdam
- Satake K (2007) Tsunamis, chap 4, 17. Treatise on geophysics, vol 4. Elsevier, Amsterdam, pp 483–511

## Tsunami Inundation, Modeling of

PATRICK J. LYNETT

Texas A&M University, College Station, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Brief Review of Tsunami Generation  
 and Open Ocean Propagation  
 Physics of Nearshore Tsunami Evolution  
 Effects of Bathymetric and Topographical Features  
 on Inundation  
 Hydrodynamic Modeling of Tsunami Evolution  
 Moving Shoreline Algorithms  
 Future Directions  
 Bibliography

### Glossary

- Beach profile** A cross-shore, or normal to the beach, survey of the seafloor and dry ground elevation (bathymetry and topography); a series of spatial location and bottom elevation data pairs.
- Bore** A steep hydraulic front which transitions between areas of different water level. Tsunamis can approach land as a turbulent, breaking bore if the incident tsunami is of sufficiently large height.
- Boussinesq equations** An approximate equation model, used for waves with wave length of at least two times the local water depth; a long-wave-based model, but includes some frequency dispersion
- Dispersion, amplitude** The separation of wave components due to a wave-height related difference in wave speed; all else being equal, a wave with a large height will travel faster than one with a small height.
- Dispersion, frequency** The separation of wave components due to a frequency related difference in wave speed; all else being equal, a wave with a longer period will travel faster than one with a short period.
- Navier–Stokes equations** The full equations of fluid motion, including dissipation through the fluid molecular

viscosity only. Other models discussed here, namely the Shallow Water Wave and Boussinesq equations, are approximations to these equations.

**Runup, or runup height** The ground elevation (a vertical measure) at the furthest point of landward inundation.

**Shallow water wave equations** An approximate equation model, used for waves with wave length many times larger than the water depth; a non-dispersive, long-wave model; there is no frequency dispersion in this model.

**Tsunami inundation** The spatial area flooded as a tsunami rushes inland.

### Definition of the Subject

Tsunami inundation is the one of the final stages of tsunami evolution, when the wave encroaches upon and floods dry land. It is during this stage that a tsunami takes the vast majority of its victims. Depending on the properties of the tsunami (e. g. wave height and period) and the beach profile (e. g. beach slope, roughness), the tsunami may approach as a relatively calm, gradual rise of the ocean surface or as an extremely turbulent and powerful bore – a wall of white water. The characteristics of this approach determine the magnitude and type of damage to coastal infrastructure and, more importantly, the actions required of coastal residents to find a safe retreat or shelter.

To gage the nearshore impact of tsunami inundation, engineers and scientists rely primarily on three different methods: 1) Field survey of past events, 2) Physical experimentation in a laboratory, and 3) Numerical modeling. It is the last of these methods – numerical simulation of tsunami inundation – that will be the focus of this article. With numerical simulation, it is possible to predict the consequence of future tsunamis on existing coastal towns and cities. This information allows for the establishment of optimum evacuation routes, identification of high-risk and/or unprepared areas, re-assessment of building codes to withstand wave impact, and placement of tsunami shelters, for example. It is the hope that, through accurate prediction of tsunami effects, in conjunction with policy makers willing to implement recommended changes and a strong public education program, communities will show resiliency to tsunami impact, with minimal loss of life and damage to critical infrastructure.

### Introduction

On December 26, 2004, the boundary between the Indo-Australian and Eurasian plates off the coast of north-



ern Sumatra ruptured in a great (Mw 9.3) earthquake at 00:58:53 universal time (U.T.). Up to 15m of thrust on the plate interface [31] displaced tens of cubic kilometers of seawater and propagated a tsunami across the Indian Ocean. The earthquake was widely felt throughout South Asia and was locally destructive in Sumatra and the Andaman and Nicobar islands, but it was the tsunami that caused widespread damage to densely populated coastal communities both nearby and thousands of kilometers away.

Due to the extensive damage left behind by large tsunamis such as the Indian Ocean tsunami, it is difficult if not impossible to put together a complete picture of the event with field observations alone. Additionally, for some parts of the world that have not seen a tsunami in recent times, there are no field observations on which to develop safety procedures and protect residences from future tsunamis. It is for these purposes – understanding the detail of tsunami inundation and to estimate tsunami hazard – that we must rely on modeling of tsunamis. There are two primary modeling approaches - physical and numerical. The physical, or experimental, approach uses scaled-down models to look at a particular aspect of a phenomenon. While this approach is integral to the fundamental understanding of waves, because of the huge wavelengths of tsunamis, experiments are limited. For example, a tsunami might have a wavelength of 100 km in a deep ocean depth of 1 km, with a wave height of 1 m. Note that the above values represent approximate order of magnitudes for a large subduction zone tsunami. Now, to scale this down for the laboratory with a wave tank depth of 1 m – the tank would have to be 100 m long and the created lab-tsunami would have a hardly measurable height of 1 mm. Numerical modeling, while not “real” in the sense that modeling is done on a computer chip with approximated equations of motion rather than in the laboratory, does not suffer from this scaling problem, and can generally accommodate any type of arbitrary wave and ocean depth profile.

Numerical simulations of tsunami propagation have been greatly improved in the last 30 years. In the United States, several computational models are being used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration (NOAA), to produce tsunami inundation maps and predict tsunami runup in real-time for the warning system. In addition, there are numerous other models used by researchers and engineering companies in an attempt to better understand tsunami impact. In this article, an overview of these models, as well as how they are validated and utilized, is provided.

## Brief Review of Tsunami Generation and Open Ocean Propagation

Before introducing the physics behind propagating a tsunami across oceans and overland, we must first discuss how a tsunami is created. For earthquake generated mega-tsunamis, such as the Indian Ocean event, a huge undersea earthquake along a great fault length of a subduction zone must occur. These earthquakes create large vertical motions of the seafloor. This vertical motion of the seafloor pushes the water above it, essentially creating a small displacement of water above the earthquake. This displacement of water will immediately try to spread out and reach a gravitational equilibrium, and it does so as waves propagating away from the earthquake zone – this is the tsunami.

To represent the tsunami in numerical models, we use an initial condition. Simply put, there is placed some irregular ocean surface profile at the instant after the earthquake, when the numerical simulations will start. Then, based on physics – Newton’s Laws written for fluid – the initial condition evolves and transits oceans. As a tsunami travels unhindered across ocean basins, it does so quickly and with little noticeable change. In the deep ocean, even the largest tsunamis have heights only near 1 m and currents of 10 cm/s, and are not likely to be identified by ships or surface buoys in the presence of wind waves.

## Physics of Nearshore Tsunami Evolution

A tsunami in the deep ocean is long and travels extremely fast. As the wave reaches shallow water, near the coastline, the tsunami begins the shoaling process. The speed at which long wave such as a tsunami moves, or celerity, is a function of the local water depth. The less the depth, the slower the wave moves. A tsunami, with its very long length, experiences different water depths at any given instant as it travels up a slope; the depth at the front of the wave, the portion of the tsunami closest to the shoreline, will generally be in the shallowest water and thus is moving the slowest. The back of the tsunami, on the other hand, will be in deeper water and will be moving faster than the front. This leads to a situation where the back part of the wave is moving faster than the front, causing the wavelength to shorten. With a shortening tsunami, the wave energy is in essence squeezed into a smaller region, forcing the height to grow. It is for this reason that, despite having a height of only a meter in the deep ocean, the tsunami elevation over land can easily exceed 10 m. With this great increase in wave height comes a more dynamic and complex phenomenon.

Presented in a more technical manner, a tsunami in the open ocean is generally a linear, non-dispersive wave. First, what is meant by a non-dispersive tsunami will be discussed. Also, the discussion here will be in terms of a large earthquake generation tsunami, such as the 2004 Indian Ocean event. Other impulsive waves, such as landslide or asteroid impact generated waves, are more difficult to generalize and will be introduced separately at the end of this section.

Any wave condition, whether it is a tsunami or a typical wind wave in the ocean, can be mathematically described as a superposition, or summation, of a series of separate sine (or cosine) waves, each with independent amplitude and speed. For example, with the right choice of individual sine waves, it is possible to construct even the idealized tsunami: a single soliton. If a wave is considered a dispersive wave, then the various sine wave components will have different wave speeds, and the wave will disperse as the faster moving components move away from the slower ones. If a wave is non-dispersive, then all the components move at the same speed, and there is no lengthwise dispersal, or spreading, of the tsunami wave energy. It is for this reason that tsunamis can be devastating across such a large spatial region; the tsunami wave energy will not disperse but will remain in a focused pulse.

The dispersion described above is generally what scientists are referring to during a discussion of dispersive vs non-dispersive waves. However, it is more precisely called “frequency dispersion” as it is dominantly dependent on the period of the component. There is another type of dispersion, called “amplitude dispersion.” This second type of dispersion is a function of the nonlinearity of the wave, and is usually discussed under the framework of linear vs nonlinear waves. For tsunamis, the nonlinearity of the waves is given by the ratio of the tsunami height to the water depth. When this ratio is small, such as in the open ocean, the wave is linear; on the other hand, in shallow water the ratio is order unity and thus the wave is no longer linear. The linear/nonlinear nomenclature is not an intuitive physical description of the waves, but comes from the equations describing the tsunami motion, described later in this section. When this nonlinear effect is taken into account, it is found that the wave speed is no longer just a function of the local depth, but of the wave height as well. More specifically, looking at two components of the same period but with different amplitudes, the component with the larger amplitude will have a slightly larger wave speed. Except for the interesting cases of wave fission, discussed later in this section, the nonlinear effect of amplitude dispersion does not spread tsunami energy with an end result of lessening nearshore impact; in fact it will act to focus

wave energy at the front, often leading to a powerful breaking bore.

Thus, open ocean propagation of a conventional tsunami is a relatively uncomplicated process which translates wave energy across basins, subject to wave speed changes that are a function of the local depth. As a tsunami enters the nearshore region, roughly characterized by water depths of 100m and less, the wave can undergo a major physical transformation. The properties of this transformation depend heavily on the characteristics of the beach profile and the wave itself. In the simplest inundation case, the beach profile is relatively steep (footnote: here “steep” should be thought of in terms of the tsunami wavelength. If the horizontal distance along the slope connecting deep water to the shoreline is small compared to the tsunami wavelength, the beach would be considered steep) and the tsunami wave height is small, then the runup process closely resembles that of a wave hitting a vertical wall, and the runup height will be approximately twice the offshore tsunami height. In these special cases, a breaking bore front would not be expected; in fact horizontal fluid velocities near the shoreline would be very small. Here, the tsunami inundation would closely resemble that of a quickly rising tide with only very minor turbulent, dynamic impacts. However, even in these cases, overland flow constrictions and other features can create localized energetic inundation.

If the beach profile slope is mild, typical of continental margins, and/or the tsunami wave height is large, then the shallow water evolution process becomes highly nonlinear. However, while the nonlinear effect becomes very important, in the large majority of cases, frequency dispersion is still very small and can be neglected. Nearshore nonlinear evolution is characterized by a strong steepening, and possible breaking, of the wave front with associated large horizontal velocities. In these cases, turbulent dissipation can play a major role.

While it may be intuitive to postulate that wave breaking dissipation at the tsunami front plays a significant role in the tsunami inundation, this may not be altogether correct. This breaking dissipation, while extraordinarily intense, is fairly localized at the front which, both spatial and temporal, often represents only a small fraction of the tsunami. So, for tsunamis such as the 2004 Indian Ocean event, the related dissipation likely had only minor impact on leading-importance quantities such as the maximum runup and inland (off-beachfront) flow velocities. However, the properties of breaking are of great importance to other aspects of tsunami inundation. The maximum forces on beachfront infrastructure, such as ports, terminals, piers, boardwalks, and houses, should include



the bore impact force as well as the drag force associated with the following quasi-steady flow [54,68]. If one was interested in understanding how bottom sediments are suspended, transported, and deposited by a tsunami, the bore turbulence again may play an important role. Thus, understanding the dynamics of a breaking tsunami front is not of particular importance for near real-time or operational tsunami forecast models. This information is of great use for engineers and planners, who can utilize it to design tsunami-resistant structures, for example.

A second energy dissipation mechanism, one that does play a major role in determining maximum runup, is bottom friction. On a fundamental level, this dissipation is caused by the flow interaction with the bottom, where bottom irregularities lead to flow separations and the resulting turbulence. All natural bottoms result in some bottom friction; a smooth, sandy beach may generate only minor dissipation, while a coral reef or a mangrove forest can play a huge role in reducing tsunami energy [12]. Such features will be discussed in additional detail in the next section. Other means of energy dissipation will be largely local, and may include enhanced mixing due to sediment or debris entrainment, large shallow-flow vortex generation by headlands or other natural or artificial obstacles and the resulting dissipation, and flow through/around buildings and other infrastructure, sometimes termed macro-roughness and grouped with bottom friction.

Up to this point, we have only discussed the “typical” nearshore tsunami evolution which is portrayed as a wave without frequency dispersion, and may be called a linear or nonlinear tsunami, depending on a number of physical properties. The rest of this section will be devoted to those situations where the above characterization may no longer be adequate. Looking first to the tsunami source, waves that are generated by underwater landslides, underwater explosions, or asteroid impacts will often not behave as non-dispersive waves in the open ocean [44,72]. These source regions tend to be at least an order of magnitude smaller in spatial extent compared to their subduction zone counterparts. Physically, this implies that the generated waves will be of shorter wavelength. As a rule of thumb, if these generated waves have length scales of less than 10 times the local depth, then it should be anticipated that frequency dispersion will play a role [42]. Under this constraint, individual component wave speeds near the dominant period become frequency dependent.

Understanding that an impulsively generated wave can be dispersive has serious implications. Take, for example, a hypothetical landslide located in the Atlantic Ocean which generates a dispersive tsunami (e. g. Ward and Day, 2001). As this tsunami travels across the Atlantic, to ei-

ther the USA east coast or the European west coast, frequency dispersion effects will spread the wave energy in the direction of propagation. This will convert the initial short-period pulse into a long train of waves. By spreading this energy out, the inundation impact will be greatly reduced. First, by taking a high-density energy pulse and stretching it into a longer, lower-density train, the maximum energy flux, and thus intensity, hitting the shoreline will decrease. Second, by increasing the duration of the time series, and creating many individual crests, energy dissipation can play a bigger role. Using simple energy arguments, it can be shown that, comparing a high-density, short-period pulse to a low-density, long-period train, more energy will be removed through bottom friction and breaking. This increase will be related to the ratio of the period of the entire dispersive wave train to the period of the pulse. Numerically studies have shown that for such cases, the individual wave crests are largely dissipated, and runup is dominated by the carrier wave, or in other terms it becomes a time-dependent, wave setup problem [26].

While a topic of current research in the tsunami community, frequency dispersion may occasionally play a non-negligible role in even the long wavelength, subduction zone tsunamis. To date, there have been two categories of argument that dispersion is important for these tsunamis: 1) short-period energy generated at the source is significant and leads to different patterns of runup if included (e. g. [21,28]), and 2) shallow-water nonlinear interactions can generate short-period components which can become decoupled (or un-locked) from the primary wave, and will change the incident tsunami properties [47].

Thinking of an arbitrary and complex initial free surface displacement generated by an undersea earthquake, there does exist the possibility that dispersive wave energy can be initially generated here. This irregular wave condition can be constructed as a continuous wave energy spectrum, and by definition there will be finite (albeit small) energy at all frequencies. The obvious question in this case is: what length scale characteristics of the initial free surface displacement, or the preceding earthquake, will lead to a significant measure of dispersive wave energy? To provide an answer to this question, a simple order-of-magnitude scaling argument is presented here; see Hammack and Segur [19], for example, for a mathematically rigorous attempt at insight. Let us define a characteristic change in vertical free surface elevation,  $\Delta\eta$ , and a horizontal length scale across which this vertical change occurs,  $\Delta L$ . Reducing to the simplest case, a regular wave with wave height equal to  $\Delta\eta$ , then the wavelength would be  $2\Delta L$ . Follow-

ing this analogy, which will hold in a proportional sense for a Fourier series of wave components, for any  $\Delta\eta$  measured along a tsunami initial condition, there exists a wave component with wave length equal to  $2\Delta L$ . For that component to be significant to the tsunami evolution, the local vertical change,  $\Delta\eta$ , must be some non-negligible fraction of the maximum tsunami height,  $H$ . Note that the  $H$  discussed here is a global property of the entire initial tsunami wave condition. For the individual wave component, with length  $2\Delta L$ , to be dispersive, its length should be less than roughly 10 times the local water depth,  $h_0$ . Additionally, the difference between dispersive wave propagation and a non-dispersive propagation is cumulative. For example, if the full linear wave theory predicts, or a specific wave component, a wave speed that is 10% less than the long wave speed, then the predicted arrival time difference will grow by 0.1 times the period for each wavelength of propagation. Thus, the impact of dispersion is related to the distance of propagation, and is proportional to  $D/\lambda$ , where  $D$  is the total distance traveled by the wave, and  $\lambda$  is the average wavelength of the wave across  $D$  which can be expected to be proportional to  $2\Delta L$ . Assuming that  $\lambda$  is approximately equal to  $2\Delta L$ , it can be said that in order for frequency dispersion effects to play a role in tsunami evolution,

$$\max \left| \frac{h_0 \Delta\eta}{\Delta L^2} \right| \frac{D}{H} > \delta, \quad (1)$$

where  $\delta$  is some minimum threshold for importance. What this value should be is an open question, although it is likely to be near 0.1. From this term, it is clear the impact of dispersion is a function of a number of the properties of the initial tsunami condition, and should be taken into consideration when creating tsunami initial conditions. For example, use of discontinuous block-type segments (e.g. [30]), with sharp edges (very small  $\Delta L$ ) may lead to the conclusion that frequency dispersion is important, while it could be a direct result of a coarsely approximated initial tsunami condition. Also note that this exercise does not include the effects of radial spreading, which could very likely be important for small-scale irregularities in the initial condition. Wave height decrease by radial spreading is proportional to the horizontal curvature of the initial condition and to  $(\lambda/D)^n$ , and decreases faster for dispersive waves ( $n \sim 1$ ) as compared to non-dispersive waves ( $n = 0.5$ ) (e.g. [72]). Thus, for this case of significant radial spreading, it would be very difficult for source-based dispersion effects to play a meaningful role in the far field.

Under certain conditions, namely a nonlinear tsunami propagation across a wide shallow shelf, a process called

fission may occur. Wave fission is a separation process where wave energy, initially part of a primary wave or pulse, attains certain properties, such as higher or lower phase speed, that allow it to disconnect from the primary wave and propagate as an independent wave. In the context of nearshore tsunami evolution, there is a standard mechanism which is the cause of this fission. First, it is necessary to describe what a nonlinear, phase-locked wave is. To do this, we will examine the acceleration terms of the 1D conservation of momentum equation for the velocity component  $u$ :

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = -\frac{\partial p}{\partial x} + \mu \frac{\partial^2 u}{\partial x^2}. \quad (2)$$

Now assume that there is a single wave component, under which the velocity oscillates as

$$\cos(kx - \omega t) = \cos \theta, \quad (3)$$

where  $k$  is the wavenumber,  $\omega$  the frequency, and the speed of the wave is given by  $\omega/k$ . If the wave is nonlinear, which is to say that the convective acceleration term in the above momentum equation is not negligible, the convective term will include the product of

$$\cos \theta * \cos \theta = \cos 2\theta = \cos(2kx - 2\omega t). \quad (4)$$

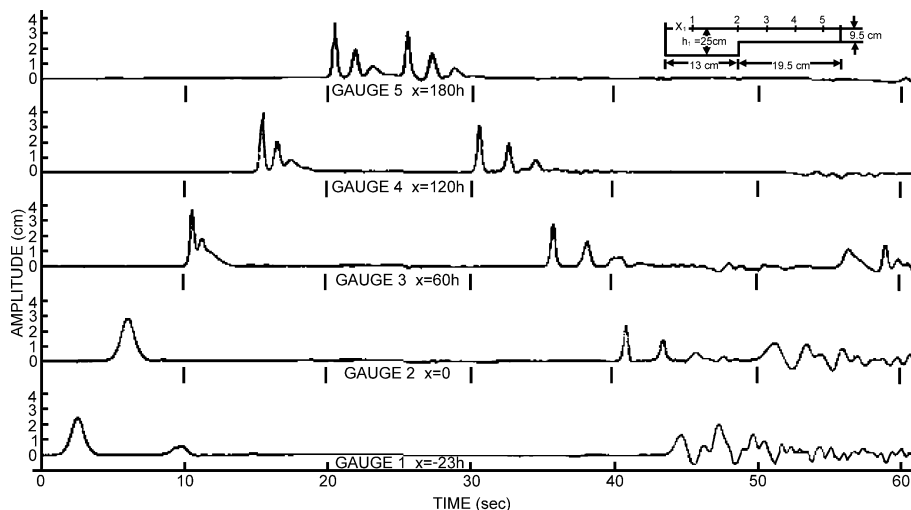
Thus, through this nonlinear term, a new wave component, with twice the wavenumber and frequency (or half the wavelength and period) has been generated. From linear wave theory, it is expected that this new wave, with a shorter period, will have a different wave speed than the original, primary wave. However, from the phase function of this new wave, there is the speed  $2\omega/2k = \omega/k$ , which is the identical speed of the primary wave. Thus this new wave is *locked* to the *phase* of the original wave. This connection can be rather delicate, and any disruptions to the primary wave, such as a varying seafloor, dissipation, or interactions with other free waves in the train or wave pulse, can cause the new waves to become unlocked. When this occurs, the now free waves retain their frequency  $2\omega$ , but take a wavenumber as given by linear wave theory. Since these freed waves will be of a shorter period than the primary wave, they will travel at a slower speed and generally trail the main wave front.

Long wave fission is most commonly discussed in the literature via a solitary wave propagating over an abrupt change in depth, such as a step (e.g. [17,24,33,37,45,55]). In these cases, there is a deep water segment of the seafloor profile, where a solitary wave initially exists. In this depth, the solitary wave is of permanent form. As the solitary wave passes over the change in depth, into shallower water,

the leading wave energy will try to re-discover a balance between nonlinearity and dispersion; the solitary wave. Since this new solitary wave will be a different shape and contain a lower level of mass, by conservation there must be some trailing disturbance to account for the deficient. This trailing disturbance will take the form of a rank-ordered train of solitons. Figure 1 depicts this process. The solitons in the trailing train, while smaller in height than the leading solitary wave, tend to have a similar wavelength; this has been shown both analytically, numerically, and experimentally. Note, however, that discussion of fission in this sense is not particularly relevant to “real” tsunami modeling, where the offshore wave approaching the shelf break rarely resembles a solitary wave solution [62]. However, the offshore wave does not need to specifically be a solitary wave for this process to occur.

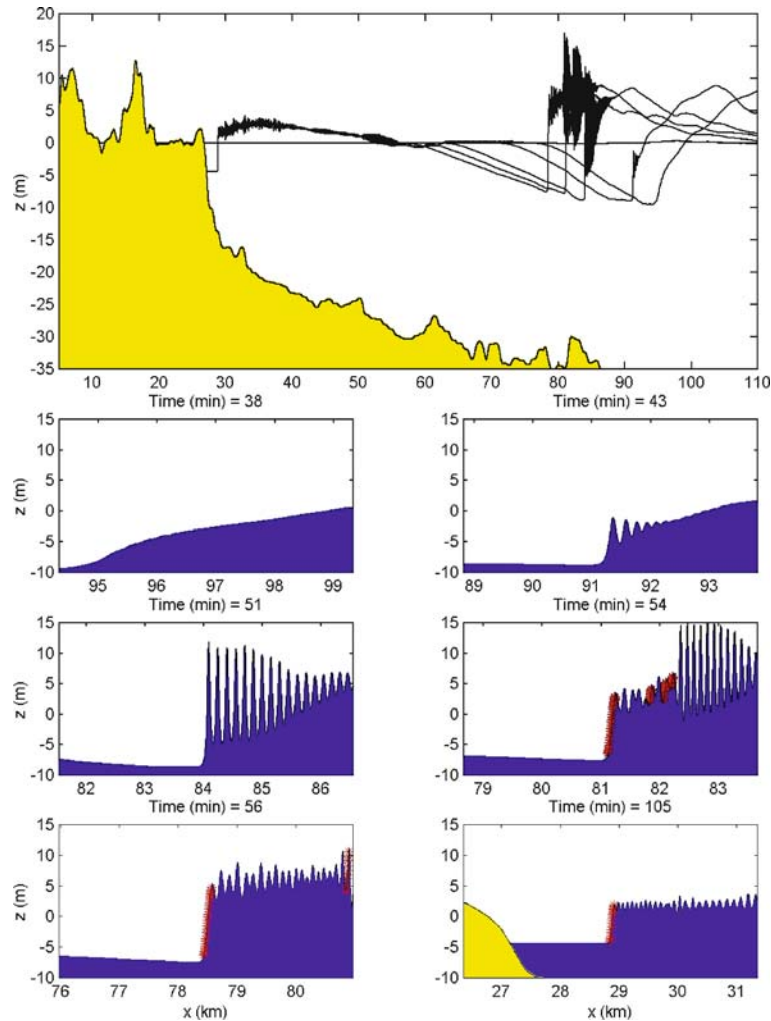
In numerous eyewitness accounts and videos recorded of the 2004 Indian Ocean tsunami, there is evidence of the tsunami approaching the coastline as a series of short period (on the order of 1 min and less) breaking fronts, or bores (e.g. [30]). These short period waves may be the result of fission processes of a steep tsunami front propagating across a wide shelf of shallow depth. Along the steep front of a very long period wave, nonlinearity will be very important. There will be a large amount of energy in high-frequency components with wavelengths similar the horizontal length of the tsunami front (on the order of 1 km). As the wave continues to shoal, the high-frequency locked waves may eventually become free waves, and will

take the form of very short waves “riding” the main wave pulse. This situation is akin to an undular bore in a moving reference frame. This process is, in fact, identical to that described in the above paragraph, it simply takes place over a much longer distance. The newly freed waves, in the nonlinear and shallow environment, will attempt to reach an equilibrium state, where frequency dispersion and nonlinearity are balanced. Thus, the fission waves will appear as solitary waves, or more generally, cnoidal waves. This fact provides some guidance as to the wavelength of these fission waves; they can be approximately calculated via solitary wave theory using the tsunami height and depth of the shelf. For example, on a shelf with depth of 30 m and an incident tsunami height of 5 m, fission waves with a wavelength of approximately 240 m and period of 13 s would be generated. In recent work looking at tsunamis along the eastern USA coast, where there exists a wide shallow shelf, this fission process has been investigated [14]. Figure 2 gives a few numerical simulation snapshots, and shows where the fission occurs, and the eventually impact on the waveform. This simulation, run with the dispersive equations, generated fission waves with lengths in the range of 100–200 m, and required a grid size of 5 m to attain numerically convergent results. In this example, the steep fission waves break offshore, and have little impact on the maximum runup. A conclusion of this fission issue is that, if one attempts to simulate tsunami propagation with dispersive equations, and if the grid is not chosen to be fine enough to resolve the short fission waves,



**Tsunami Inundation, Modeling of, Figure 1**

Example of experimental data looking at solitary wave fission by propagation onto a shelf from Goring and Raichlen [17]. Note that the flume layout and measurement location is given up in the upper right. The initial solitary wave undergoes the fission process and results in three distinct solitary waves



**Tsunami Inundation, Modeling of, Figure 2**

Example of tsunami fission. Simulation results are from Geist et al. [14] for a landslide-generated tsunami off the east coast of the USA. The *top plot* shows the beach profile and six free surface profiles at different times. The *lower subplots* are zoom-in's of those six profiles, with the times given in the individual plot titles. The *red marks* visible in the *lowest plots* indicate regions where the wave is breaking

the justification to use the dispersive model is greatly degraded.

### Effects of Bathymetric and Topographical Features on Inundation

It is well established that large-scale coastal features, such as small islands, large shoals, canyons, and shelves, can play an important role in tsunami inundation due to conventional shallow water effects such as shoaling and refraction (e.g. [4,5,16,36,70]). On the other hand, understanding of the impact of smaller scale features is just now being developed. This work was largely initiated by field

observations. Synolakis et al. [60], surveying the coast of Nicaragua for information about the 1992 tsunami in the region, noted that the highest levels of damage along a particular stretch of beach were located directly landward of a reef opening used for boat traffic. It was postulated that the reef gap acted as a lower resistance conduit for tsunami energy, behaving like a funnel and focusing the tsunami. Along neighboring beaches with intact reefs, the tsunami did not have the intensity to remove even beach umbrellas. Investigating impacts from the same tsunami, Borrero et al. [3], discussed how small scale bathymetry variations affected coastal inundation. One of the conclusions of this work was that bathymetry features with





**Tsunami Inundation, Modeling of, Figure 3**

Example of foundation scour. This image was taken by the International Tsunami Survey Team to Sri Lanka in January 2005

length scales 50 m and less had leading order impact on the runup. Looking to the recent Indian Ocean tsunami, a survey team in Sri Lanka inferred from observations that reef and dune breaks lead to locally increased tsunami impact [38,39]. Also in Sri Lanka, Fernando et al. [12] performed a more thorough survey along the southeastern coastline, and concluded that there was a compelling correlation between coral mining and locally severe tsunami damage. While additional research is needed to quantify the effects of small scale features, the observations hint that defense measures such as seawalls, once thought to be inconsequential to tsunami inundation, may provide some protection.

Onshore, tsunami propagation is effected by the general topography (ground slope), ground roughness, and obstacles (e.g. [41,59,61,67]). The composition of the ground, be it sand, grass, mangroves, or pavement, controls the roughness and the subsequent bottom friction damping. To predict tsunami inundation with high confidence, the ground type must be well mapped and the hydrodynamic interaction with that type must be well under-



**Tsunami Inundation, Modeling of, Figure 4**

Example of damage to the backside of a coastal residence. These images were taken by the International Tsunami Survey Team to Sri Lanka in January 2005. The *top photograph* shows the front side of the structure, facing the ocean; there is damage but the main structure is intact. The *lower photo* shows the backside of the same building, showing the walls blown out, away from the center of the structure

stood. If the tsunami approaches the shoreline as a bore, the process of “bore collapse”, or the conversion of potential to kinetic energy, will cause the fluid to rapidly accelerate [56,69]. This fast flow equates to high fluid forces on obstacles such as buildings. Tsunami interaction with these obstacles can lead to a highly variable local flow pattern (e.g. [10,67]). As the flow accelerates around the corners of a building, for example, the scour potential of that flow increases greatly, and foundation undermining is a concern (e.g. see Fig. 3). As with any fluid flow past an obstacle, the backface of the obstacle is characterized by a low-pressure wake. Combined with the interior flooding of a building, this low pressure wake may lead to an outward “pull” force on the back wall, causing it to fail by falling away from the center of the building. Such failures were observed during field surveys of the 2004 event, as shown in Fig. 4. Increasing the topographical complexity, in built coastal environments, structures are located within



close enough proximity to each other such that their disturbances to the flow may interact. This can lead to irregular and unexpected loadings, where for example a 2nd row building experiences a larger force than beach front buildings due to a funneling effect. These types of interactions are very poorly understood, and require additional research.

### Hydrodynamic Modeling of Tsunami Evolution

Numerical simulations of tsunami propagation have made great progress in the last thirty years. Several tsunami computational models are currently used in the National Tsunami Hazard Mitigation Program, sponsored by the National Oceanic and Atmospheric Administration, to produce tsunami inundation and evacuation maps for the states of Alaska, California, Hawaii, Oregon, and Washington. The computational models include MOST (Method Of Splitting Tsunami), developed originally by researchers at the University of Southern California [66]; COMCOT (Cornell Multi-grid Coupled Tsunami Model), developed at Cornell University [35]; and TUNAMI-N2, developed at Tohoku University in Japan [29]. All three models solve the same depth-integrated and 2D horizontal (2DH) nonlinear shallow-water (NSW) equations with different finite-difference algorithms. There are a number of other tsunami models as well, including the finite element model ADCIRC (ADvanced CIRCulation Model For Oceanic, Coastal And Estuarine Waters; e.g., [53]). For a given source region condition, existing models can simulate propagation of a tsunami over a long distance with sufficient accuracy, provided that accurate bathymetry data exist.

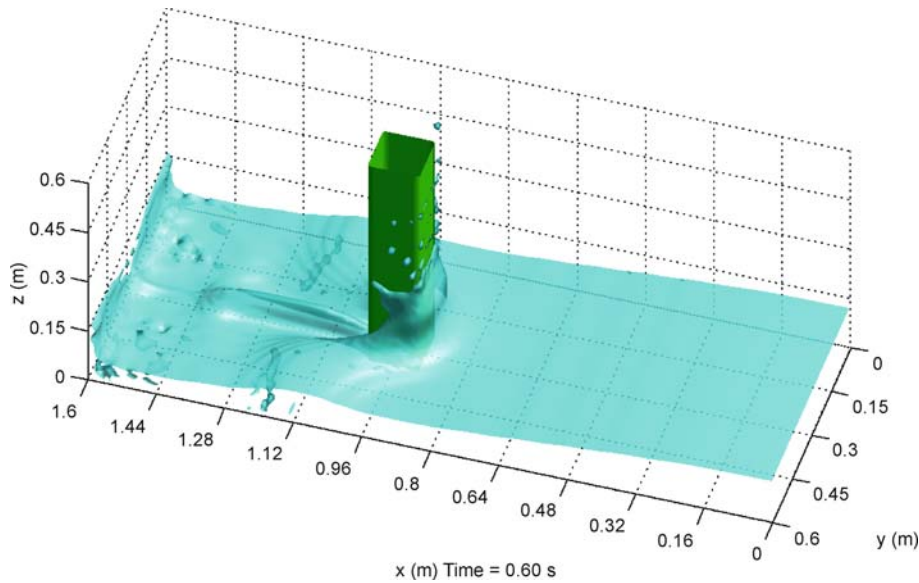
The shallow-water equation models commonly lack the capability of simulating dispersive waves, which, however, could well be the dominating features in landslide-generated tsunamis and for the fission processes described previously. Several high-order depth-integrated wave hydrodynamics models (Boussinesq models) are now available for simulating nonlinear and weakly dispersive waves, such as COULWAVE (Cornell University Long and Intermediate Wave Modeling Package; [42]) and FUNWAVE [25]. The major difference between the two is their treatment of moving shoreline boundaries. Lynett et al. [44] applied COULWAVE to the 1998 PNG tsunami with the landslide source; the results agreed with field survey data well. Recently, several finite element models have also been developed based on Boussinesq-type equations (e.g., [73]). Boussinesq models require higher spatial and temporal resolutions, and therefore are more computationally intensive. Moreover, most of model validation

work was performed for open-ocean or open-coast problems. In other words, the models have not been carefully tested for wave propagation and oscillations in semi-enclosed regions – such as a harbor or bay – especially under resonant conditions.

Being depth-integrated and horizontally 2D, NSW and Boussinesq models lack the capability of simulating the details of many coastal effects, such as wave overturning and the interaction between tsunamis and coastal structures, which could be either stationary or movable. At present, stationary coastal structures are parametrized as bottom roughness and contribute to frictional effects in these 2DH models. Although by adjusting the roughness and friction parameter satisfactory results can be achieved for maximum runup and delineation of the inundation zone (e.g., [35]), these models cannot provide adequate information for wave forces acting on coastal structures.

As a tsunami propagates into the nearshore region, the wave front undergoes a nonlinear transformation while it steepens through shoaling. It is in this nearshore region that dissipative effects can be important. Bottom friction can play a major role in the maximum runup and area of inundation (e.g. [67]). In depth-integrated models, bottom friction is typically approximated through a quadratic (drag) friction term, where the friction factor is calculated often through a Manning's coefficient or a Darcy-Wiesbach type friction factor (e.g. [25,35]). The validity of these steady-flow based coefficients has yet to be rigorously validated for use with tsunamis. If the tsunami is large enough, it can break at some offshore depth and approach land as a bore – the white wall of water commonly referenced by survivors of the Indian Ocean tsunami. Wave breaking in traditional NSW tsunami models has not been handled in a satisfactory manner. Numerical dissipation is commonly used to mimic breaking (e.g. [36]), and thus results become grid dependant. In Boussinesq models, this breaking is still handled in an approximate manner due to the fact that the depth-integrated derivation does not allow for an overturning wave; however these breaking schemes have been validated for a wide range of nearshore conditions (e.g. [40]).

Being depth-integrated, NSW and Boussinesq models lack the capability of simulating the vertical details of many coastal effects, such as strong wave breaking/ overturning and the interaction between tsunamis and irregularly shaped coastal structures. To address this deficiency, several 2D and 3D computational models based on Navier–Stokes equations have been developed, with varying degrees of success. An example is COBRAS (Cornell Breaking waves and Structures model Lin and Liu 1998a,b, Lin et al. 1999), which is capable of describing the



#### Tsunami Inundation, Modeling of, Figure 5

An simulation snapshot taken from a 3D Navier–Stokes solver with an LES turbulence closure. This setup is looking at a bore impacting a column; the wake and vertical splash are clearly visible (Image provided by Philip L.-F. Liu, Cornell University)

interactions between breaking waves and structures that are either surface piercing or submerged [6,22]. COBRAS adopted the Volume of Fluid (VOF) method to track free surface movement along with a Large Eddy Simulation (LES) turbulence closure model; several other computational models using different free surface tracking methods are also in use, such as the micro surface cell technique developed by Johnson et al. [23]. This 3D Navier–Stokes equation model has been tested by two tsunami related experiments. The first is 3D landslide experiments [38,39], while the second involves measurements of solitary wave forces on vertical cylinders. Both experiments were conducted in the NEES tsunami basin at Oregon State. An example of a LES numerical solution of a solitary wave impinging on a circular is shown in Fig. 5.

Due to their high computational costs, full 3D models would best be used in conjunction with a depth-integrated 2DH model (i. e., NSW or Boussinesq). While the 2DH model provides incident far-field tsunami information, the 3D model computes local wave-structure interactions. The results from 3D models could also provide a better parametrization of small-scale features (3D), which could then be embedded in a large-scale 2DH model. One-way coupling (e. g. using a NSW-generated time series to drive a 3D model, but not permitting feedback from the 3D model back into the NSW) is fairly straightforward to construct (e. g. [18]). Two-way coupling, however, is difficult and requires consistent matching of physics and nu-

merical schemes across model interfaces. Previous work in this area of two-way coupling of hydrodynamic models is limited. Fujima et al. [13] two-way coupled a NLSW model with a fully 3D model. While the results appear promising, the approach used by Fujima et al. requires ad-hoc and unphysical boundary conditions at the model matching locations, in the form of spatial gradients forced to zero, to ensure numerical stability. Even with these ad-hoc treatments, their hybrid model compares very well with the completely-3D-domain simulation, requiring roughly 1/5 of the total 3D CPU time to achieve similar levels of accuracy. Sittanggang et al. [58] presented work on two-way coupling of a Boussinesq model and 2D Navier–Stokes model. These results indicate that there is large potential for hybrid modeling, in terms of more rapid simulation as well as the ability to approach a new class of problems.

#### Moving Shoreline Algorithms

In order to simulate the flooding of dry land by a tsunami, a numerical model must be capable of allowing the shoreline to move in time. Here, the shoreline is defined as the spatial location where the solid bottom transitions from submerged to dry, and is a function of the two horizontal spatial coordinates and time. Numerical models generally require some type of special consideration and treatment to accurately include these moving boundaries; the logic

and implementation behind this treatment is called a moving shoreline, or runup, algorithm.

For typical tsunami propagation models, it is possible to divide runup algorithms into two main approaches: those on a fixed grid and those on a Lagrangian or transformed domain. Both approaches have their advantages and disadvantages; currently fixed grid methods are found more commonly in operational-level models (e.g. [66]), likely due in large part to their conceptual simplicity. A review of these two classes of models will be given in this section, followed by a review of the standard analytical, experimental, and field benchmarks used to validate the runup models. For additional information, the reader is directed to the comprehensive review given in Pedersen [49].

With a fixed grid method, the spatial locations of the numerical grid points or control volumes are determined at the start of a simulation, and do not change shape or location throughout the simulation duration. These methods can be classified into extrapolation, stairstep, auxiliary shoreline point, and permeable beach techniques. The extrapolation method has its roots in Sielecki and Wurtele [57], with extensions by Hibberd and Peregrine [20], Kowalik and Murty [27], and Lynett et al. [43]. The basic idea behind this method is that the shoreline location can be extrapolated using the nearest wet points, such that its position is not required to be locked onto a fixed grid point; it can move freely to any location. Theoretically, the extrapolation can be of any order; however, from stability constraints a linear extrapolation is generally found. Hidden in the extrapolation, the method is roughly equivalent to the use of low-order, diffusive directional differences taken from the last wet point into the fluid domain [43]. Additionally, there are no explicit conservation constraints or physical boundary conditions prescribed at the shoreline, indicating that large local errors may result if the flow in the extrapolated region cannot be approximately as linear in slope. The extrapolation approach can be found in both NLSW and Boussinesq models with finite difference, finite volume, and finite element solution schemes, and has shown to be accurate for a wide range of non-breaking, breaking, two horizontal dimension, and irregular topography problems (e.g. [8,9,26,44,51]).

Stairstep moving shoreline methods, one of the more common approaches found in tsunami models (e.g. [35]), reconstruct the naturally continuous beach profile into a series of constant elevation segments connected through vertical transitions. In essence, across a single cell width, the bottom elevation is taken as the average value. A cell transitions from a dry cell to a wet cell when the water elevation in a neighboring cell exceeds the bottom elevation,

and transitions from wet to dry when the local total water depth falls below some small threshold value. These methods are particularly useful in finite volume and C-grid [1] type approaches (e.g. [32,36]), but can be difficult to implement in centered difference models, particularly high-order models or those sensitive to fluid discontinuities, where the “shock” of opening and closing entire cells can lead to numerical noise.

Auxiliary shoreline point methods require dynamic re-gridding very near the shoreline, such that the last wet point is always located immediately at the shoreline. Obviously, this method requires a numerical scheme that can readily accommodate non-uniform and changing node locations. There is some relation to the extrapolation methods discussed above; the moving shoreline point must be assigned some velocity, and it is extrapolated from the neighboring wet points. However, it is fundamentally different in that the shoreline point is explicitly included in the fluid domain. Thus, it would be expected that the governing conservation equations near the shoreline are more precisely satisfied here, although still dependent on the appropriateness of the extrapolation. One such method can be found in Titov and Synolakis [65], and has been successfully applied in NSLW equation models.

Another fixed grid treatment of moving boundary problems is employing a slot or permeable-seabed technique [63,64]. Conceptually, this method creates porous slots, or conduits, through the dry beach, such that there is always some fluid in a “dry” beach cell, although it may exist below the dry beach surface. These porous, “dry” nodes use a modified form of the NLSW; it is noted here that although in concept this approach is modeling a porous beach, it is not attempting to simulate the groundwater flow under a real, sandy beach, for example. The equations governing the “dry” domain contain a number of empirical parameters that are tuned to provide reasonable runup agreement with benchmark datasets. The advantage of this approach is that it allows the entire domain; including the fluid and “dry” nodes, to be determined via a somewhat consistent set of governing equations, without requiring a direct search routine to determine the shoreline location. The method has gained some popularity in wind wave models (e.g. [25,46]) when a highly accurate estimate of the shoreline location is not the highest priority. However, the approach has been used with some success in tsunami studies (e.g. [30]) despite the fact that the empirical coefficients that govern the model accuracy cannot be universally determined for a wide range of problems [7].

Alternative to fixed grid methods is the Lagrangian approach. Here, the fluid domain is discretized into particles,

or columns of fluid in depth-integrated models, that are transported following the total fluid derivative. There are no fixed spatial grid locations; the columns move freely in space and time and thus these techniques require numerical flexibility, in terms of utilizing constantly changing space and time steps. The Lagrangian approach can be described as both the more physically consistent and mathematical elegant method of describing shoreline motion. The shoreline “particle” is included in the physical formulation just as any other point in the domain (i. e. no extrapolations are necessary), and thus the shoreline position accuracy will be compromised only by the overarching physical approximation (e. g. long wave approximation) and the numerical solution scheme (e. g. second-order time integration). The cost for this accuracy is a mathematical system that can be more difficult and tedious to solve numerically, typically requiring domain transformations, mappings, and/or re-griddings. Lagrangian methods have been used successfully in finite difference and finite element nonlinear shallow water (NLSW) and Boussinesq equation models (e. g., [2,15,48,50,52,74]).

### Future Directions

Towards a more robust simulation of tsunami inundation, there are two major issues which require additional fundamental investigations: dissipation mechanisms and interaction with infrastructure. Bottom friction, known to play an important role in inundation, needs to be re-examined starting from its basic formulation. Can a steady-flow based Mannings-type expression for bottom friction be used for tsunami? Does the unsteady nature of the tsunami flow make use of these existing formulations invalid? The answer to these questions may be different depending on what part of the wave is investigated (e. g. front). In addition, the hydrodynamic effect of common coastal vegetation, such as mangroves, needs to be quantified. There is current discussion of the use of such natural roughness as a tsunami defense (e. g. [11]); confidence cannot be put in such measures until it is understood how they behave. In addition to bottom friction, which exists at all locations and times under an inundating tsunami, wave breaking can increase the total energy dissipation. While breaking is generally confined to the leading front of a tsunami, the characteristics of this front are important for hydrodynamic loadings on beachfront structures, and may be significant to the net sediment and debris transport of a tsunami. Three-dimensional tsunami breaking is poorly understood and has received little attention.

Wave loadings and interactions with infrastructure are not well understood. To tackle this problem, tsunami hy-

drodynamic models need to be coupled with structural and geotechnical models. Ideally, these models should all be two-way coupled, such that the displacement of a structure, be it a single collapsed wall, will change the flow pattern, and scour underneath the foundation will change the structure stability. Additionally, impacts of flow-transported debris (e. g. cars) should be included in this framework. If such a modeling capacity existed, engineering design of coastal structures could be undertaken in a very efficient manner.

### Bibliography

1. Arakawa A, Lamb VR (1977) Computational design of the basic dynamical processes of the UCLA general circulation model. In: Chang J (ed) *Methods in computational physics*. Academic Press, New York, pp 174–267
2. Birknes J, Pedersen G (2006) A particle finite element method applied to long wave run-up. *Int J Numer Methods Fluids* 52(3):237–261
3. Borrero JC, Bourgeois J, Harkins G, Synolakis CE (1997) How small-scale bathymetry affected coastal inundation in the 1992 Nicaraguan tsunami. Fall AGU Meeting, San Francisco
4. Briggs MJ, Synolakis CE, Harkins GS, Green D (1994) Laboratory experiments of tsunami runup on a circular island. *PAGEOPH* 144(3/4):569–593
5. Carrier GF (1966) Gravity waves on water of variable depth. *Fluid J Mech* 24:641–659
6. Chang K-A, Hsu T-J, Liu PL-F (2001) Vortex generation and evolution in water waves propagating over a submerged rectangular obstacle. Part I solitary waves. *Coast Eng* 44:13–36
7. Chen Q, Kirby JT, Dalrymple RA, Kennedy AB, Chawla A (2000) Boussinesq modeling of wave transformation, breaking, and runup: Part I 2D. *J Waterw Port Coast Ocean Eng* 126(1): 57–62
8. Cheung K, Phadke A, Wei Y, Rojas R, Douyere Y, Martino C, Houston S, Liu PL-F, Lynett P, Dodd N, Liao S, Nakazaki E (2003) Modeling of storm-induced coastal flooding for emergency management. *Ocean Eng* 30:1353–1386
9. Cienfuegos R, Barthelemy E, Bonneton P (2007) A fourth-order compact finite volume scheme for fully nonlinear and weakly dispersive Boussinesq-type equations. Part II: Boundary conditions and model validation. *Int J Numer Meth Fluids* 53(9):1423–1455
10. Cross RH (1967) Tsunami surge forces, ASCE JI Waterways & Harbors Division WW4:201–231
11. Danielsen F, Sørensen MK, Olwig MF, Selvam V et al (2005) The Asian tsunami: A protective role for coastal vegetation. *Science* 310:643
12. Fernando HJS, McCulley JL, Mendis SG, Perera K (2005) Coral poaching worsens tsunami destruction in Sri Lanka. *Eos Trans AGU* 86(33):301
13. Fujima K, Masamura K, Goto C (2002) Development of the 2d/3d hybrid model for tsunami numerical simulation. *Coastal Eng J* 44(4):373–397
14. Geist E, Lynett P, Chaytor J (2008) Hydrodynamic modeling of tsunamis from the currituck landslide, *Marine Geology* (in press)



15. Gopalakrishnan TC, Tung CC (1983) Numerical analysis of a moving boundary problem in coastal hydrodynamics. *Intl J Numer Meth Fluids* 3:179–200
16. González FI, Satake K, Boss EF, Mofjeld HO (1995) Edge wave and non-trapped modes of the 25 April 1992 Cape Mendocino tsunami. *Pure Appl Geophys PAGEOPH* 144(3–4): 409–426
17. Goring DG, Raichlen F (1992) Propagation of long waves onto shelf. *J Waterw Port Coast Ocean Eng* 118(1):43–61
18. Guignard S, Grilli ST, Marcer R, Rey V (1999) Computation of shoaling and breaking waves in nearshore areas by the coupling of BEM and VOF methods. *Proc. 9th Offshore and Polar Engng. Conf.*, vol III, ISOPE99, Brest, France, pp 304–309
19. Hammack J, Segur H (1978) Modelling criteria for long water waves. *J Fluid Mech* 84(2):359–373
20. Hibberd S, Peregrine DH (1979) Surf and run-up on a beach. *J Fluid Mech* 95:323–345
21. Horrillo, Juan, Kowalik, Zygmunt, Shigihara, Yoshinori (2006) Wave dispersion study in the Indian Ocean–Tsunami of December 26, 2004. *Marine Geodesy* 29(3):149–166(18)
22. Hsu T-J, Sakakiyama T, Liu PL-F (2002) A numerical model for waves and turbulence flow in front of a composite breakwater. *Coast Eng* 46:25–50
23. Johnson DB, Raad PE, Chen S (1994) Simulation of impacts of fluid free surfaces with solid boundaries. *Int J Num Methods Fluids* 19:153–176
24. Johson RS (1972) Some numerical solutions of a variable-coefficient Korteweg–de Vries equation (with application to solitary wave development on a shelf). *J Fluid Mech* 54:81
25. Kennedy AB, Chen Q, Kirby JT, Dalrymple RA (2000) Boussinesq modeling of wave transformation, breaking, and runup. 1: 1D. *J Waterway Port Coastal Ocean Eng ASCE* 126:39–47
26. Korycansky DG, Lynett P (2007) Runup from impact tsunami. *Geophys J Int* 170:1076–1088
27. Kowalik Z, Murty TS (1993) Numerical simulation of two-dimensional tsunami runup. *Marine Geodesy* 16:87–100
28. Kulikov E (2005) Dispersion of the Sumatra tsunami waves in the Indian Ocean detected by satellite altimetry. Report from P.P. Shirshov Institute of Oceanology, Russian Academy of Sciences, Moscow
29. Imamura F (1995) Review of tsunami simulation with a finite difference method, long-wave runup models. *World Scientific*, Singapore, pp 25–42, [http://en.wikipedia.org/wiki/2004\\_Indian\\_Ocean\\_earthquake](http://en.wikipedia.org/wiki/2004_Indian_Ocean_earthquake)
30. Ioualalen M, Asavanant JA, Kaewbanjak N, Grilli ST, Kirby JT, Watts P (2007) Modeling of the 26th December 2004 Indian Ocean tsunami: Case study of impact in Thailand. *J Geophys Res* 112:C07024, doi:10.1029/2006JC003850
31. Lay T, Kanamori H, Ammon C, Nettles M, Ward S, Aster R, Beck S, Bilek S, Brudzinski M, Butler R, DeShon H, Ekström G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of December 26, 2004. *Science* 308:1127–1133, doi:10.1126/science.1112250
32. LeVeque RJ, George DL (2004) High-resolution finite volume methods for the shallow water equations with bathymetry and dry states. In: Liu PL, Yeh H, Synolakis C (eds) *Advanced numerical models for simulating tsunami waves and runup*. vol 10 of *Advances in Coastal and Ocean Engineering*. World Scientific, Singapore
33. Losada MA, Vidal V, Medina R (1989) Experimental study of the evolution of a solitary wave at an abrupt junction. *J Geophys Res* 94:14557
34. Liu PL-F, Synolakis CE, Yeh H (1991) Report on the international workshop on long-wave run-up. *J Fluid Mech* 229:675–688
35. Liu PL-F, Cho Y-S, Yoon SB, Seo SN (1994) Numerical simulations of the 1960 Chilean tsunami propagation and inundation at Hilo, Hawaii. in: El-Sabh MI (ed) *Recent development in tsunami research*. Kluwer, Dordrecht, pp 99–115
36. Liu PL-F, Cho Y-S, Briggs MJ, Kanoglu U, Synolakis CE (1995) Runup of solitary waves on a circular island. *J Fluid Mech* 320:259–285
37. Liu PL-F, Cheng Y (2001) A numerical study of the evolution of a solitary wave over a shelf. *Phys Fluids* 13(6):1660–1666
38. Liu PL-F, Lynett P, Fernando H, Jaffe BE, Fritz H, Higman B, Morton R, Goff J, Synolakis C (2005) Observations by the International Tsunami Survey Team in Sri Lanka. *Science* 308: 1595
39. Liu PL-F, Wu T-R, Raichlen F, Synolakis CE, Borrero JC (2005) Runup and rundown generated by three-dimensional sliding masses. *J Fluid Mech* 536:107–144
40. Lynett P (2006) Nearshore modeling using high-order Boussinesq equations. *J Waterway Port Coastal Ocean Eng (ASCE)* 132(5):348–357
41. Lynett P (2007) The effect of a shallow water obstruction on long wave runup and overland flow velocity. *J Waterway Port Coastal Ocean Eng (ASCE)* 133(6):455–462
42. Lynett P, Liu PL-F (2002) A numerical study of submarine landslide generated waves and runup. *Proc R Soc London A* 458:2885–2910
43. Lynett P, Wu T-R, Liu PL-F (2002) Modeling wave runup with depth-integrated equations. *Coast Eng* 46(2):89–107
44. Lynett P, Borrero J, Liu PL-F, Synolakis CE (2003) Field survey and numerical simulations: A review of the 1998 Papua New Guinea tsunami. *Pure Appl Geophys* 160:2119–2146
45. Madsen OS, Mei CC (1969) The transformation of a solitary wave over an uneven bottom. *J Fluid Mech* 39:781
46. Madsen PA, Sorensen OR, Schaffer HA (1997) Surf zone dynamics simulated by a Boussinesq-type model: Part I. Model description and cross-shore motion of regular waves. *Coast Eng* 32:255–287
47. Matsuyama M, Ikeno M, Sakakiyama T, Takeda T (2007) A study on tsunami wave fission in an undistorted experiment. *Pure Appl Geophys* 164:617–631
48. Özkan-Haller HT, Kirby JTA (1997) Fourier-Chebyshev collocation method for the shallow water equations including shoreline run-up. *Appl Ocean Res* 19:21–34
49. Pedersen G (2006) On long wave runup models. In: *Proceedings of the 3rd International Workshop on Long-Wave Runup Models*, in Catalina Island, California
50. Pedersen G, Gjevik B (1983) Runup of solitary waves. *J Fluid Mech* 142:283–299
51. Pedrozo-Acuna A, Simmonds DJ, Otta AK, Chadwick AJ (2006) On the cross-shore profile change of gravel beaches. *Coast Eng* 53(4):335–347
52. Petera J, Nassehi V (1996) A new two-dimensional finite element model for the shallow water equations using a Lagrangian framework constructed along fluid particle trajectories. *Int J Num Methods Eng* 39:4159–4182
53. Priest GR et al (1997) Cascadia subduction zone tsunamis: Haz-



ard mapping at Yaquina Bay, Oregon. Final Technical Report to the National Earthquake Hazard Reduction Program DOGAMI Open File Report 0-97-34, p 143

54. Ramsden JD (1996) Forces on a vertical wall due to long waves, bores, and dry-bed surges. *J Waterway Port Coast Ocean Eng* 122(3):134–141
55. Seabra-Santos FJ, Renouard DP, Temperville AM (1987) Numerical and experimental study of the transformation of a solitary wave over a shelf or isolated obstacles. *J Fluid Mech* 176:117
56. Shen M, Meyer R (1963) Climb of a bore on a beach Part 3. Runup. *J Fluid Mech* 16(1):113–125
57. Sielecki A, Wurtele MG (1970) The numerical integration of the nonlinear shallow-water equations with sloping boundaries. *J Comput Phys* 6:219–236
58. Sittanggang K, Lynett P, Liu P (2006) Development of a Boussinesq-RANSVOF Hybrid Wave Model. in *Proceedings of 30th ICCE, San Diego*, pp 24–25
59. Synolakis CE (1987) The runup of solitary waves. *J Fluid Mech* 185:523–545
60. Synolakis CE, Imamura F, Tsuji Y, Matsutomi S, Tinti B, Cook B, Ushman M (1995) Damage, Conditions of East Java tsunami of 1994 analyzed. *EOS, Transactions, American Geophysical Union* 76(26):257 and 261–262
61. Tadepalli S, Synolakis CE (1994) The runup of N-Waves on sloping beaches, *Proc R Soc London A* 445:99–112
62. Tadepalli S, Synolakis CE (1996) Model for the leading waves of tsunamis. *Phys Rev Lett* 77:2141–2144
63. Tao J (1983) Computation of wave runup and wave breaking. Internal Report. Danish Hydraulics Institute, Denmark
64. Tao J (1984) Numerical modeling of wave runup and breaking on the beach. *Acta Oceanol Sin* 6(5):692–700 (in chinese)
65. Titov VV, Synolakis CE (1995) Modeling of breaking and nonbreaking long wave evolution and runup using VTCS-2. *J Harbors Waterways Port Coast Ocean Eng* 121(6):308–316
66. Titov VV, Synolakis CE (1998) Numerical modeling of tidal wave runup. *J Waterway Port Coast Ocean Eng ASCE* 124(4):157–171
67. Tomita T, Honda K (2007) Tsunami estimation including effect of coastal structures and buildings by 3D model. *Coastal Structures '07, Venice*
68. Yeh H (2006) Maximum fluid forces in the tsunami runup zone. *J Waterway Port Coast Ocean Eng* 132(6):496–500
69. Yeh H, Ghazali A, Marton I (1989) Experimental study of bore runup. *J Fluid Mech* 206:563–578
70. Yeh H, Liu P, Briggs M, Synolakis C (1994) Propagation and amplification of tsunamis at coastal boundaries. *Nature* 372:353–355
71. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. *Geophys Res Lett* 28(17):3397–3400
72. Weiss R, Wunemann K, Bahlburg H (2006) Numerical modelling of generation, propagation, and run-up of tsunamis caused by ocean impacts: model strategy and technical solutions. *Geophys J Int* 67:77–88
73. Woo S-B, Liu PL-F (2004) A finite element model for modified Boussinesq equations. Part I: Model development. *J Waterway Port Coast Ocean Eng* 130(1):1–16
74. Zelt JA (1991) The run-up of nonbreaking and breaking solitary waves. *Coast Eng* 15(3):205–246

## Tsunamis, Inverse Problem of

KENJI SATAKE

Earthquake Research Institute, University of Tokyo,  
Tokyo, Japan

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Tsunami Generation by Earthquakes](#)

[Tsunami Propagation](#)

[Tsunami Observations](#)

[Estimation of Tsunami Source](#)

[Estimation of Earthquake Fault Parameters](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Inverse problem** Unlike a forward problem which starts from a tsunami source then computes propagation in the ocean and predicts travel times and/or water heights on coasts, an inverse problem starts from tsunami observations to study the generation process. While forward modeling is useful for tsunami warning or hazard assessments, inverse modeling is a typical approach for geophysical problems.

**Shallow water (long) waves** In hydrodynamics, water waves can be treated as shallow water, or long, waves when the wavelength is much larger than the water depth. In such a case, the entire water mass from water bottom to surface moves horizontally and the wave propagation speed is given as a square root of the product of the gravitational acceleration and the water depth.

**The 2004 Indian Ocean tsunami** On December 26, 2004, a gigantic earthquake, the largest in the last half century in the world, occurred off the west coast of Sumatra Island, Indonesia. With the source extending more than 1,000 km through Nicobar and Andaman Islands, the earthquake generated tsunami which attacked the coasts of Indian Ocean and caused the worst tsunami disaster in history. The total casualties were about 230,000 in many countries as far away as Africa.

**Fault parameters** Earthquake source is modeled as a fault motion, which can be described by nine static parameters. Once these fault parameters are specified, the

seafloor deformation due to faulting, or initial condition of tsunamis, can be calculated by using the elastic dislocation theory.

**Refraction and inverse refraction diagrams (travel time map)** Refraction diagram is a map showing isochrons or lines of equal tsunami travel times calculated from the source toward coasts. Inverse refraction diagram is a map showing arcs calculated backwards from observation points. The tsunami source can be estimated from the arcs corresponding to tsunami travel times.

### Definition of the Subject

Forward modeling of tsunami starts from given initial condition, computes its propagation in the ocean, and calculates tsunami arrival times and/or water heights on coasts. Once the initial condition is provided, the propagation and coastal behavior can be numerically computed on actual bathymetry (Fig. 1).

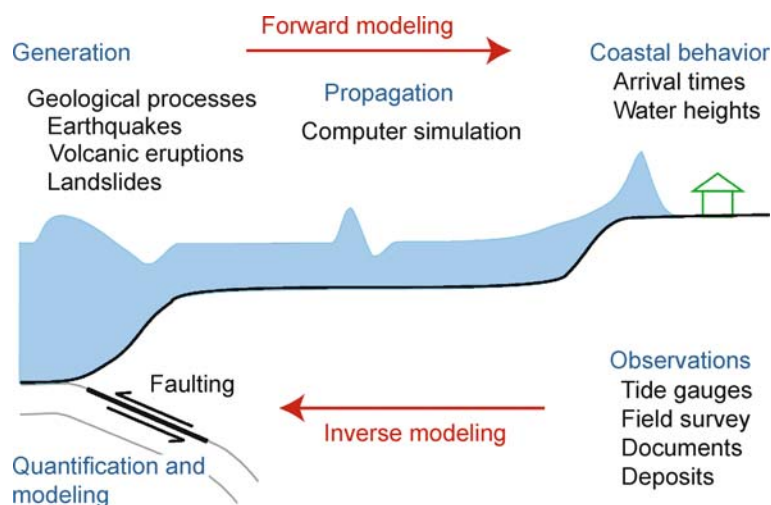
Recent technological developments make it possible to carry out tsunami forward modeling with speed and accuracy usable for the early tsunami warning or detailed hazard assessments. However, the initial condition, or the tsunami generation process, is still poorly known, because large tsunamis are rare and the tsunami generation in the ocean is not directly observable. Indirect estimation of tsunami source, mostly on the basis of seismological analyzes, is used as the initial condition of tsunami forward modeling. More direct estimation of tsunami source is essential to better understand the tsunami generation process and to more accurately forecast the tsunami on coasts.

Inverse modeling of tsunami starts from observed tsunami data, to study the tsunami source. The propagation process can be evaluated by using numerical simulation, as in the forward modeling. As the observed tsunami data, tsunami arrival times, heights or waveforms recorded on instruments are used. For historical tsunamis, tsunami heights can be estimated from description of damage on historical documents. For prehistoric tsunamis, geological studies of tsunami deposits can be used to estimate the coastal tsunami heights or flooding areas.

### Introduction

Tsunamis are oceanic gravity waves generated by seafloor deformation due to submarine earthquakes or other submarine geological processes such as volcanic eruptions, landslides, or asteroid impacts. While earthquake tsunamis, such as the 2004 Indian Ocean tsunami caused by the Sumatra–Andaman earthquake, are most frequent, large volcanic eruptions such as the 1883 Krakatau eruption off Sumatra Island also cause ocean-wide tsunamis. Landslides, often triggered by earthquakes, cause locally large tsunamis, but the effects are usually limited to the area around the source.

Most tsunamigenic geological processes produce seafloor deformation. When horizontal scale, or wavelength, of the seafloor deformation is much larger than the water depth, a similar disturbance appears on the water surface and becomes the source of tsunami. This is called shallow water, or long-wave, approximation. For large earthquakes, wavelength of seafloor deformation is



**Tsunamis, Inverse Problem of, Figure 1**

Schematic diagram showing tsunami generation, propagation and coastal behavior. Forward modeling starts from tsunami source and forecasts the coastal behavior, while inverse modeling starts from observed data to estimate the tsunami source

an order of several tens to hundreds of km, while the ocean depth is up to several km, hence the long-wave approximation is valid. For small scale disturbance relative to water depth, such as submarine landslides or volcanic eruptions in deep seas, the shallow water approximation may not be valid.

This paper reviews inverse methods to study tsunami sources from the observations. Section “[Tsunami Generation by Earthquakes](#)” describes the tsunami generation by earthquakes, with emphasis on the fault parameters and their effects on tsunamis. Section “[Tsunami Propagation](#)” describes tsunami propagation: shallow water theory and numerical computation. Section “[Tsunami Observations](#)” summarizes the tsunami observation: instrumental sea-level data, runup height estimates for modern, historical and prehistoric tsunamis. Section “[Estimation of Tsunami Source](#)” describes methods of modeling and quantifying tsunami source, and of analyzing tsunami travel times, amplitudes and waveforms, including some historical developments. Section “[Estimation of Earthquake Fault Parameters](#)” focuses on the estimation of earthquake fault parameters, including the waveform inversion of tsunami data to estimate heterogeneous fault motion and its application for tsunami warning.

## Tsunami Generation by Earthquakes

### Fault Parameters and Seafloor Deformation

The seafloor deformation due to earthquake faulting can be calculated by using the elastic theory of dislocation. The displacement,  $u_k$ , in an infinite homogeneous medium due to dislocation  $\Delta u_i$  across surface  $\Sigma$  is given by the Volterra's theorem as [1]

$$u_k = \frac{1}{8\pi\mu} \int_{\Sigma} \Delta u_i \{ \lambda \delta_{ij} u_k^{n,n} + \mu (u_k^{i,j} + u_k^{j,i}) \} v_j dS \quad (1)$$

where  $\lambda$  and  $\mu$  are Lamé constants,  $\delta_{ij}$  is Kronecker's delta,  $v$  is the unit normal to the surface. The expression  $u_i^j$  denotes the  $i$ th component of the displacement due to the  $j$ th component of a point force at the source whose magnitude is  $8\pi\mu$ , and  $u_i^{j,k}$  indicates its spatial derivative with respect to the  $k$ th coordinate. For a half-space with free surface, a mirror image can be used to cancel the stress components on the free surface. The explicit formulas are given by Mansinha and Smylie [2] or Okada [3].

The fault parameters needed to compute surface deformation are summarized in Fig. 2. They are: fault length ( $L$ ), width ( $W$ ), strike ( $\phi$ ), dip ( $\delta$ ), rake ( $\lambda$ ), slip amount ( $u$ ) and location ( $x, y, z$ ). The slip  $u$  can be decomposed into

strike-slip component  $u_s$  and dip-slip component  $u_d$ . The strike  $\phi$  is measured clockwise from North, dip angle  $\delta$  is downward from horizontal, and rake angle  $\lambda$  is a movement of hanging wall measured counter-clockwise from horizontal (see Fig. 2). Therefore, the fault motion is reverse if  $\lambda > 0^\circ$  and normal if  $\lambda < 0^\circ$ . The fault motion has left-lateral component if  $|\lambda| < 90^\circ$  and right-lateral component if  $|\lambda| > 90^\circ$ .

The physical parameter to quantify the fault motion is the seismic moment,  $M_0$ , defined as

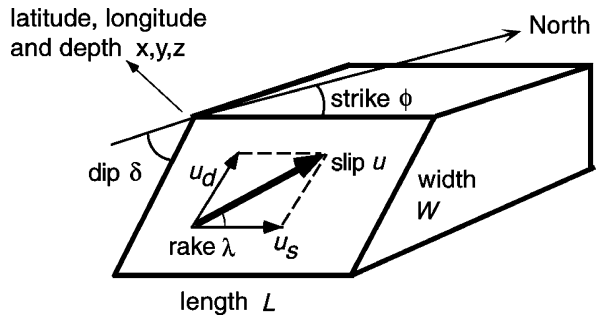
$$M_0 = \mu u S = \mu u L W. \quad (2)$$

More conventional parameter of earthquake size is a magnitude scale, which has been determined from amplitudes of seismograms. To relate the seismic moment and magnitude scales, the moment magnitude scale,  $M_w$ , is defined as [4,5]

$$M_w = \frac{2}{3} \log M_0 - 10.7 \quad (3)$$

where  $M_0$  is given in dyne.cm ( $10^{-7}$  Nm).

Most of the above fault parameters can be estimated from seismic wave analysis. The location and depth of fault ( $x, y, z$ ) correspond to hypocenter, which is estimated from arrival times of seismic waves. The fault geometry ( $\phi, \delta, \lambda$ ) is estimated from the polarity distribution of body wave first motions or azimuthal distribution of surface wave amplitudes. The seismic moment is estimated from waveform modeling of seismic waves. The fault size,  $L$  and  $W$ , are more difficult to estimate; they are usually estimated from aftershock distribution or detailed waveform modeling of seismic body waves. The slip amount,  $u$ , is indirectly estimated, from seismic moment  $M_0$  by assuming  $\mu$  and fault size ( $L$  and  $W$ ). All such estimates assume



**Tsunamis, Inverse Problem of, Figure 2**

Fault parameters. Seafloor deformation can be computed from these static parameters

that the faulting is planar and continuous, which most often is a simplification of real, more complex faulting.

The 2004 Sumatra–Andaman earthquake was the largest earthquake since the 1960 Chilean earthquake ( $M_w$  9.5) or 1964 Alaskan earthquake ( $M_w$  9.2). The seismic moment estimates range  $4 - 12 \times 10^{22}$  Nm, and the corresponding moment magnitude  $M_w$  ranges 9.0–9.3 from the seismological analyzes [6,7,8]. The aftershock area extended from off Sumatra through the Nicobar to Andaman Islands with the total fault length of 1,200 to 1,300 km [6]. For such a gigantic earthquake, multiple fault planes with different strike and slip amounts are needed to represent the fault motion, as shown later (Sect. “[Estimation of Earthquake Fault Parameters](#)”).

### Effect of Fault Parameters on Tsunami Generation

Among the above static fault parameters, the slip amount has the largest effect on the vertical seafloor deformation and the tsunami amplitude. The dip angle and fault depth are also important parameters to control tsunami amplitude [9,10]. Dynamic parameters such as rupture velocity are found to be insignificant for tsunami generation. However, for a gigantic earthquake such as the 2004 Sumatra–Andaman earthquake with the source length over 1,000 km, rupture propagation effect on tsunamis is not negligible [11].

Amplitude of far-field seismic waves, either body waves or surface waves, is controlled by seismic moment, while amplitude of tsunami is controlled by fault slip. Satake and Tanioka [12] found for the 1998 Papua New Guinea tsunami that the far-field tsunami amplitudes are proportional to the volume of displaced water, while the near-field tsunami amplitudes are controlled by the potential energy of the displacement.

Traditionally, only vertical component of seafloor deformation has been considered for tsunami generation (Fig. 3a). If an earthquake occurs on a steep ocean slope

such as trench slope, horizontal displacement due to faulting moves the slope and contributes the tsunami generation [13]. The effective vertical movement (positive upward) due to faulting can be written as follows (Fig. 3b),

$$u_z + u_x \frac{\partial H}{\partial x} + u_y \frac{\partial H}{\partial y} \quad (4)$$

where  $u_z$  is vertical component and  $u_x$  and  $u_y$  are horizontal components of seafloor deformation, and  $H$  is water depth (measured positive downward).

Recently, Song et al. [14] proposed that horizontal motion of slope also transfers kinetic energy from seafloor to water. They claimed that the kinetic energy thus transferred was about five times larger than the potential energy for the 2004 Indian Ocean tsunami, from comparisons of the observed tsunami (sea surface height) data with the computed ones by using an ocean-general-circulation-model.

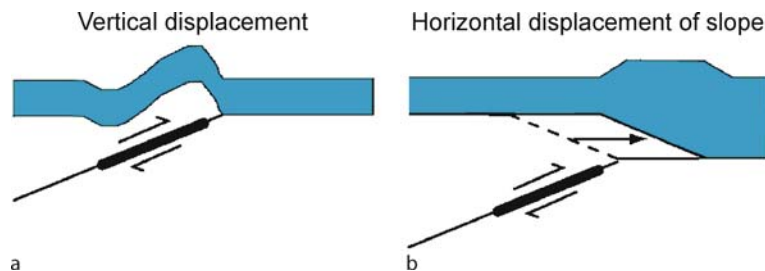
### Tsunami Propagation

#### Shallow Water Theory

The equation of motion, or conservation of momentum, for shallow water, or long-wave, theory is given as follows.

$$\frac{\partial \mathbf{V}}{\partial t} + (\mathbf{V} \cdot \nabla) \mathbf{V} = -g \nabla h + C_f \frac{\mathbf{V} |\mathbf{V}|}{d + h} \quad (5)$$

where  $\mathbf{V}$  is the depth-averaged horizontal velocity vector,  $h$  is the water height or tsunami amplitude,  $d$  is the water depth and  $g$  is the gravitational acceleration. On the left-hand side, the first term represents local acceleration and the second term represents nonlinear advection. On the right-hand side, the first term represents pressure gradient, or restoring force due to gravity, and the second term represents nonlinear bottom friction where  $C_f$  is the non-dimensional frictional coefficient.



**Tsunamis, Inverse Problem of, Figure 3**

Seafloor deformation and tsunami source. **a** Vertical seafloor deformation becomes the tsunami source. **b** When the seafloor is not flat, horizontal displacement of the slope also affects the tsunami generation [13]

The equation of continuity, or conservation of mass, can be written as

$$\frac{\partial(d+h)}{\partial t} = -\nabla \cdot \{(d+h) \mathbf{V}\}. \quad (6)$$

These equations can be linearized as

$$\frac{\partial \mathbf{V}}{\partial t} = -g \nabla h \quad (7)$$

$$\frac{\partial h}{\partial t} = -\nabla \cdot (d \mathbf{V}) \quad (8)$$

when the tsunami amplitude  $h$  is small compared to water depth  $d$ , and the bottom friction can be neglected. Such an assumption is valid for deep ocean, or most of the tsunami propagation path. Near the coasts, nonlinear terms play important roles, hence linearization may not be valid. The major advantage of the linear theory is the superposition principle; the computational results can be easily scaled to estimate with different initial water heights.

From Eqs. (7) and (8), the wave equation with wave velocity (celerity)  $\sqrt{gd}$  can be derived. This indicates that the tsunami speed is controlled by water depth. Once the water depth distribution, or ocean bottom bathymetry, is known, then the tsunami propagation can be computed numerically.

### Numerical Computations

Equations (5) and (6), or (7) and (8) for linearized case, can be directly solved by numerical methods, once the initial condition is given. The tsunami wave velocity distribution, which is given by the bathymetry, is much better known than the velocity distribution of seismic waves, hence actual values can be used. Finite-difference method with staggered grids is popularly used [15,16], while use of other methods such as finite-element methods have been also proposed [17]. Grid size for finite-difference computations is typically a few km for deep ocean, but grids as fine as several tens to hundreds of meters are used near coasts. The temporal changes in water height at grid points corresponding to the observation points are used as computed tsunami waveforms.

The database of global water depth or bathymetry data such as ETOPO2 (NOAA/NGDC) or GEBCO (British Oceanographic Data Centre) are popularly used. The ETOPO2 database is based on predicted bathymetry from satellite altimetry data [18] with interval of 2 minutes (about 3.5 km), while GEBCO data are digitized from nautical charts with grid interval of 1 minute. Higher resolution bathymetry data near coasts are open to public in some countries such as US (NOAA/NGDC) or Japan (JODC).

## Tsunami Observations

### Instrumental Data

Traditional instrumental data for tsunami observation are tide gauge records. Tide gauges are typically installed on ports or harbors to define datum or to monitor ocean tides. The temporal resolution is usually low with a sampling interval of several minutes or longer. For the tsunami monitoring, higher sampling rate, at least 1 min or shorter interval, is required. While the recorded tsunami waveforms contains coastal effects such as coastal reflections or resonance particularly for the later phase, the initial part of tsunami signals is more dominant by the tsunami source effect hence the source information can be retrieved. Currently, sea level measurement data from many tide gauge stations are transmitted through weather satellite and available in real time. Figure 4 (left) shows some of tide gauge records of the 2004 Indian Ocean tsunami.

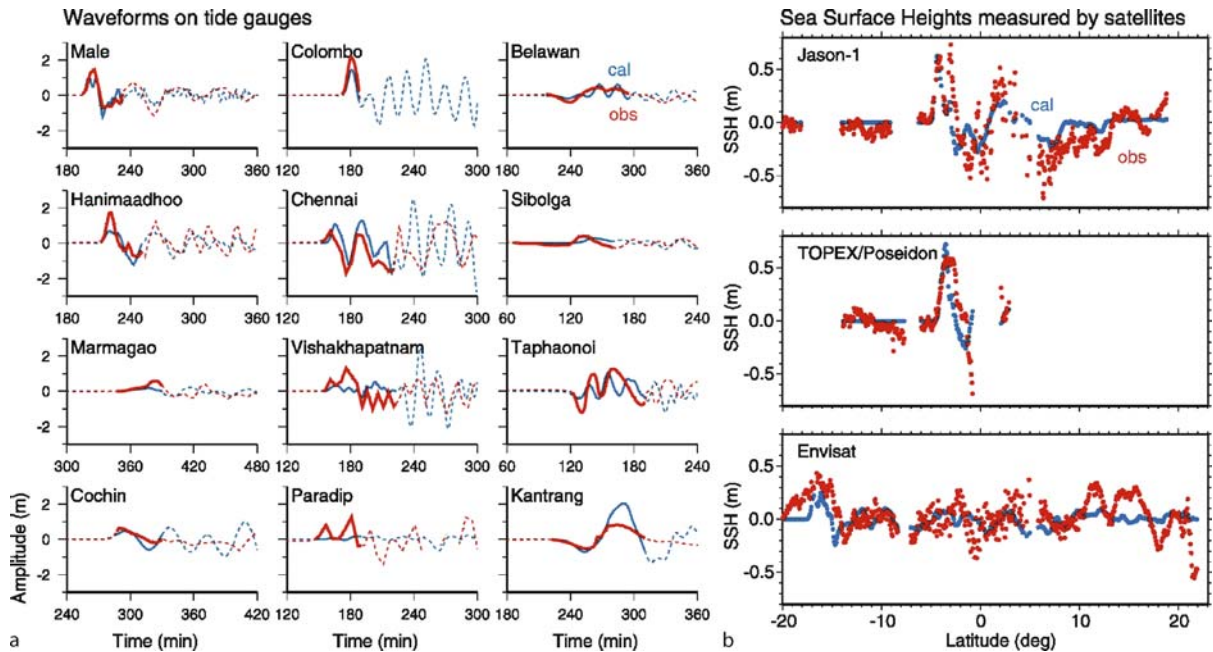
Tsunami waveforms are simpler offshore or in deep ocean, free from nonlinear coastal effects, though the signal is smaller. Offshore and deep ocean tsunami observation facilities have been significantly developed recently. Offshore tsunami gauges such as GPS tsunami gauge have been developed and deployed around Japan [19]. In 2003, cabled bottom pressure gauges have recorded the tsunami generation process in the source area [20]. The US NOAA deployed dozens of bottom pressure gauges, called Deep-ocean Assessment and Reporting of Tsunamis (DART) or simply Tsunameters [21]. The bottom pressure signals are sent to surface buoys via acoustic telemetry in the ocean, then to land station via satellite. As described later, the DART data are used for real-time data assimilation and tsunami warning purposes [22]. After the 2004 Indian Ocean tsunami, many DART-type bottom pressure gauges have been deployed in the Pacific and Indian Oceans.

Satellite altimeters captured the propagation of the 2004 Indian Ocean tsunami (Fig. 4). Three satellites flew over the Indian Ocean at a few hours after the earthquake and measured the sea surface height (SSH) of about 0.8 m in the middle of Indian Ocean. The tsunami amplitudes in deep ocean are much smaller than the maximum coastal heights of more than 10 m. The SSH data are used to study the tsunami source [11,23].

### Modern, Historical and Prehistoric Tsunami Heights

After damaging tsunamis, tsunami height distribution is often measured by survey teams [24]. Measurements are usually made for flow depth above ground, on the basis of





**Tsunamis, Inverse Problem of, Figure 4**

The sea level data from the 2004 Indian Ocean tsunami [11]. **a** Tsunami waveforms on tide gauges. Red curves indicate observed waveforms and blue ones are computed. Data shown in solid lines are used for the waveform inversion. **b** Sea surface heights measured by three different satellites (see Fig. 5 for the tracks). Red shows the observed data and blue is for computed surface heights

various watermarks, then converted to inundation height above sea level [25]. The tsunami inundation heights are usually not constant along a profile from beach, and the height at most inland point is called runup height.

For historical tsunamis, coastal tsunami heights can be estimated from descriptions of tsunami or its damage recorded in historical documents. Such estimates include various assumptions on sea levels and a relationship between tsunami damage and flow depth, but provide important tsunami data for historical tsunamis. For example, date and size of the last gigantic earthquake in the Cascadia subduction zone off North America were estimated as January 26, 1700 and  $M_w \sim 9.0$  from the Japanese tsunami records [26].

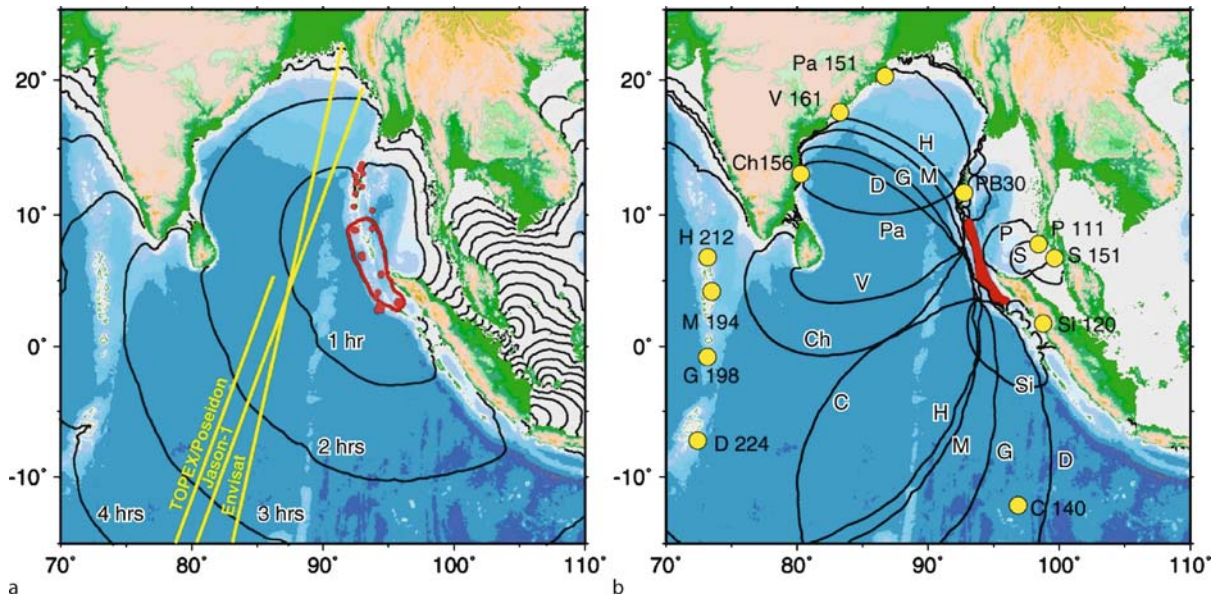
Geological traces such as tsunami deposits can also be used to estimate tsunami heights for prehistoric tsunamis. In the last few decades, many studies of tsunami deposits, combined with numerical computations, have been made to analyze prehistoric tsunamis [27,28]. For example, in Hokkaido, prehistoric tsunami deposits indicate past tsunamis with larger inundation area and longer recurrence interval than those from the recent plate-boundary earthquakes along the southern Kuril trench, which were attributed to multi-segment earthquakes with  $M_w \sim 8.4$  [29].

## Estimation of Tsunami Source

### Refraction Diagram

Tsunami propagation can be computed and described as a refraction diagram. When the tsunami wavelength is smaller than the scale length of velocity heterogeneity, or the water depth variation is smooth, then the geometrical ray theory of optics can be applied. The wavefronts of propagating tsunami can be drawn on the basis of Huygens' principle. Alternatively, propagation of rays, which is orthogonal to wavefronts, can be traced from an assumed source. While refraction diagram do not provide information on water height, the relative amplitudes can be estimated from density of rays [30].

Refraction diagrams can be prepared for major tsunami sources and used for tsunami warning; as soon as the epicenter is known, the tsunami arrival times can be readily calculated. The refraction diagrams are usually drawn from a point source, but it is possible to draw it from an extended source for a great or giant earthquake. Figure 5a shows the refraction diagram from the 2004 Sumatra-Andaman earthquake with wavefronts at each hour. To the east of the assumed source, the tsunami is expected to arrive at the Thai coast in about two hours through Andaman Sea. To the west, through deeper Bay



**Tsunamis, Inverse Problem of, Figure 5**

**a** Tsunami refraction diagram for the 2004 Sumatra–Andaman earthquake. Red dots indicate aftershocks within 1 day according to USGS. The red curve shows the assumed tsunami source. Tracks of three satellites with altimeters are shown by yellow lines. Black curves indicate tsunami wavefronts at each hour after the earthquake. **b** Tsunami inverse refraction diagram for the same event. Station code and tsunami arrival times (in min) are attached to tide gauge stations (yellow circles) where the tsunami was instrumentally recorded. Black curves are the travel-time arcs computed for each station. Red area indicates inferred tsunami source

of Bengal, the tsunami is expected to arrive at Sri Lanka also in two hours. The predicted tsunami arrival times are similar to the actually observed values [31].

### Inverse Refraction Diagram

Refraction diagram can be drawn backwards from coasts. Such a diagram is called inverse refraction diagram and is used to estimate the tsunami source area. When the tsunami travel time, that is tsunami arrival time minus earthquake origin time, is known, the corresponding wavefront, or travel-time arc, drawn from the tsunami observation point (typically tide gauge stations) would indicate the initial wavefront at the tsunami source. The tsunami inverse refraction diagram was first drawn for the 1933 Sanriku tsunami [32], although the estimated tsunami source was much larger than modern estimates, because both tsunami travel times and the bathymetry were poorly known.

The 2004 Indian Ocean tsunami was observed at many tide gauge stations in the Indian Ocean [33,34]. The tsunami arrival times were read from the tide gauge records and tsunami travel times were calculated from the earthquake origin time. The tsunami propagation was then computed from each tide gauge station, and wave-

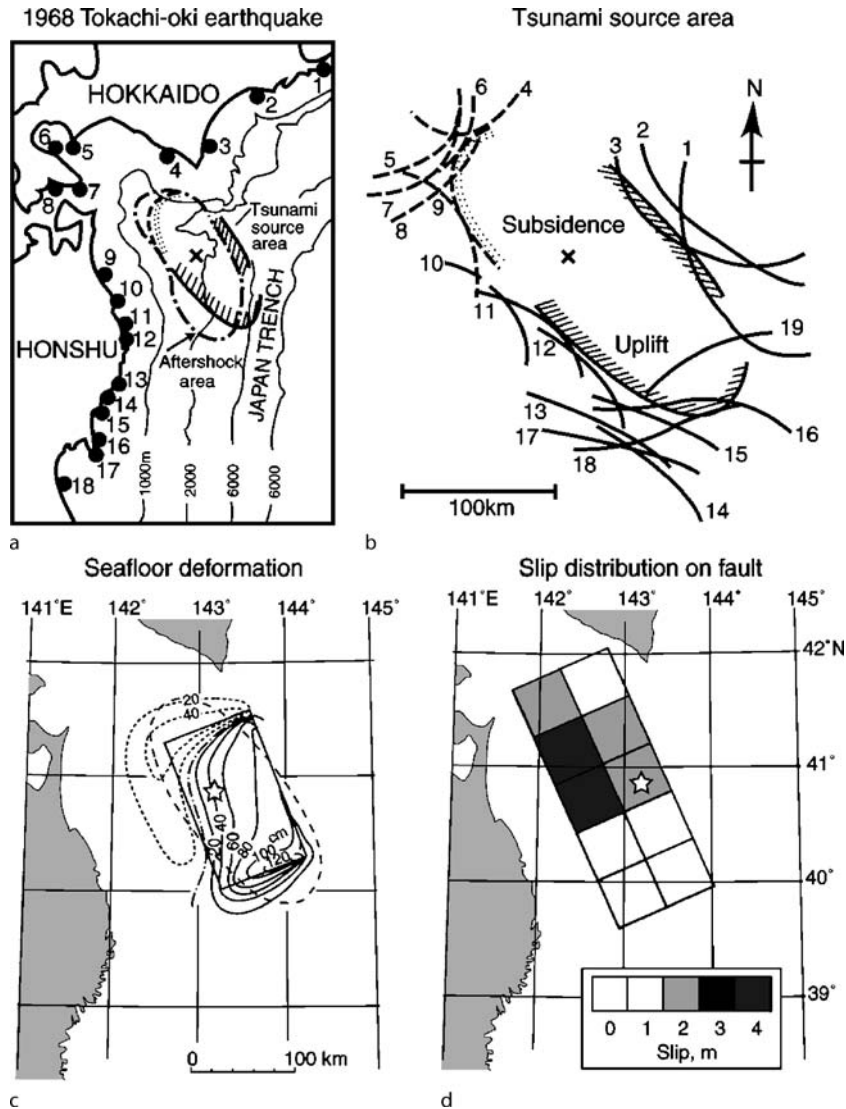
fronts corresponding to the travel time were drawn as travel-time arcs (Fig. 5b). These travel-time arcs surround the tsunami source, and the source area was estimated as about 900 km long [6,35].

### Estimation of Tsunami Source

Tsunami data can be used to study earthquake source processes in a similar way that seismic waves are used. This was first demonstrated for the 1968 Tokachi-oki earthquake ( $M_0 = 2.8 \times 10^{21}$  Nm or  $M_w = 8.3$ ) [36]. The tsunami source area estimated from an inverse refraction diagram agrees well with the aftershock area (Fig. 6a, b). In addition, the initial water surface disturbance was estimated as uplift at the southeastern edge and subsidence at the northwestern edge, from the first motion of recorded tsunami waveforms on tide gauges. This pattern is very similar to the vertical bottom deformation due to the faulting, which was independently estimated from seismological analysis (Fig. 6c).

### Green's Law and Tsunami Heights

The water height in the tsunami source area can be estimated from the observed tsunami heights along the coasts, by using the Green's law. The Green's law is derived from



**Tsunamis, Inverse Problem of, Figure 6**

The tsunami source area and seafloor deformation of the 1968 Tokachi-oki earthquake [36,37]. **a** Estimated tsunami source, aftershock area and distribution of tide gauge stations. **b** Travel-time arcs drawn by inverse refraction diagram. The numbers correspond to tide gauge stations in (a). Solid and dashed curves show uplift and subsidence, respectively. **c** Seafloor deformation computed from a seismological fault model. **d** Slip distribution on fault estimated by an inversion of tsunami waveforms

the conservation of potential energy along rays [38],

$$b_0 d_0^{1/2} h_0^2 = b_1 d_1^{1/2} h_1^2 \quad (9)$$

where  $d$  is the water depth,  $b$  is the distance between the neighboring rays,  $h$  is the tsunami amplitude, and the subscripts 0 and 1 indicate two different locations on the same ray. If the tsunami amplitude at location 0 (e.g., on the coast) is known, the tsunami amplitude at location 1 (e.g.,

at the source) can be estimated as

$$h_1 = \left( \frac{b_0}{b_1} \right)^{1/2} \left( \frac{d_0}{d_1} \right)^{1/4} h_0. \quad (10)$$

The ratio  $b_0/b_1$  represents the spreading of rays, which can be graphically obtained from refraction diagrams. For the Tokachi-oki earthquake, the average tsunami height at the source was estimated as 1.8 m using the Green's law, which is very similar to 1.6 m, the average vertical seafloor displacement computed from the fault model [36].

The Green's law is also used to estimate the shoaling effects. For plane waves approaching the coast, the spreading ratio is unity, hence the amplitude is proportional to a 1/4 power of water depth change. For example, when the water depth becomes a half, the amplitude becomes 1.18 times larger.

### Tsunami Magnitude

Tsunami magnitude scale,  $M_t$ , was introduced to quantify earthquake source that generated tsunamis [39]. The formulas were calibrated with the moment magnitude scale,  $M_w$ , of earthquakes. It is different from other tsunami magnitude or intensity scales that simply quantify the observed tsunamis. The definition of  $M_t$  for a trans-Pacific tsunami is [39]:

$$M_t = \log H + C + 9.1 \quad (11)$$

and for a regional ( $100 \text{ km} < \Delta < 3500 \text{ km}$ ) tsunami is [40]:

$$M_t = \log H + \log \Delta + 5.8 \quad (12)$$

where  $H$  is a maximum amplitude on tide gauges in meters,  $C$  is a distance factor depending on a combination of the source and the observation points, and  $\Delta$  is the nautical distance in km. The tsunami magnitude  $M_t$  was assigned as  $M_t = 8.2$  for the 1968 Tokachi-oki earthquake and  $M_t = 9.0$  for the 2004 Sumatra-Andaman earthquake.

Because the tsunami magnitude scale  $M_t$  is defined from tsunami amplitudes, it can be used to characterize "tsunami earthquakes" that produce much larger tsunamis than expected from seismic waves (see Polet and Kanamori: ► [Tsunami Earthquakes](#)). Abe [41] defined "tsunami earthquakes" for such events with tsunami magnitude  $M_t$  larger than surface wave magnitude  $M_s$  by more than 0.5. It should not be confused with "tsunamigenic earthquake" which refers to any earthquake that generates tsunami.

### Estimation of Earthquake Fault Parameters

For earthquake tsunamis, the fault parameters can be estimated by inverse modeling of tsunamis. Such attempts were first made by a trial and error approach. In order to estimate the heterogeneous fault parameters, inversion of tsunami waveforms or runup heights has been introduced.

### Trial and Error Approach

Numerical simulation of tsunami has been carried out for many tsunamigenic earthquakes around Japan [42].

For the 1968 Tokachi-oki earthquake, tsunami waveforms were computed from two models, one based on seismological analysis (Fig. 6c) and another horizontally shifted by 28 km, and were compared with the observed tsunami waveforms recorded on tide gauges. Comparison of waveforms indicates that the latter model, shifted from that based on seismological analysis, shows better match in terms of tsunami arrival times. The slip amount on the fault was estimated as 4 m.

The best fault models are judged by comparison of the observed and computed tsunami waveforms. A few statistical parameters used to quantify the comparison are geometric mean, logarithmic standard deviation and correlation coefficient. The geometric mean  $K$  of the amplitude ratio  $O_i/C_i$ , where  $O_i$  and  $C_i$  are the observed and computed tsunami amplitudes at station  $i$ , is given as

$$\log K = \frac{1}{n} \sum_i \log \frac{O_i}{C_i}. \quad (13)$$

The logarithmic standard deviation  $\kappa$  is defined as

$$\log \kappa = \left[ \frac{1}{n} \sum_i \left( \log \frac{O_i}{C_i} \right)^2 - (\log K)^2 \right]^{1/2}. \quad (14)$$

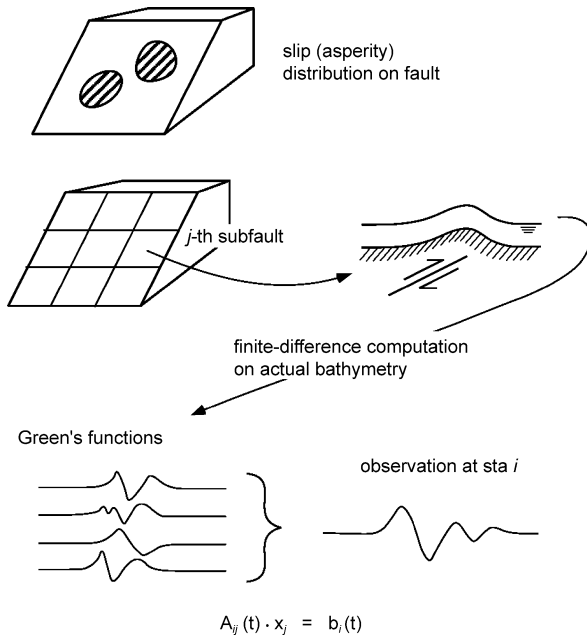
If the logarithmic amplitude ratios obey the normal distribution  $N(\log K, \log \kappa)$ , then parameter  $\kappa$  can be considered as an error factor, because its logarithm shows the standard deviation. The geometric mean  $K$  indicates the relative size of the observed and computed tsunami models. The logarithmic standard deviation  $\kappa$  indicates the goodness of the model; the smaller  $\kappa$  means the better model. The arrival times of observed and computed waveforms are also compared, as indicated in the above example. Another parameter is correlation coefficient of the observed and computed waveforms, which are also used for the comparison of models.

While the above parameters ( $K$  and  $\kappa$ ) were originally defined for maximum amplitudes of waveforms, they are also used for comparison of observed and computed runup heights. For tsunami hazard evaluation of nuclear power plants in Japan, tsunami source models need to satisfy  $0.95 < K < 1.05$  and  $\kappa < 1.45$  for the observed and computed coastal heights [43].

### Heterogeneous Fault Motion

Seismological studies of large earthquakes have indicated that the slip amount is not uniform but heterogeneous on faults. For the 1968 Tokachi-oki earthquake, inversion of far-field body waves or regional Rayleigh waves showed





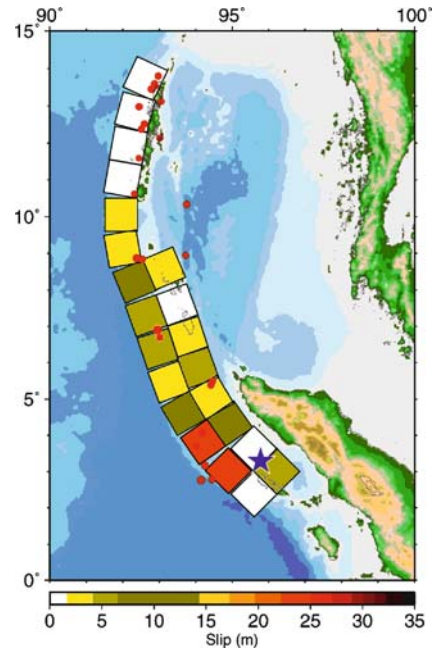
**Tsunamis, Inverse Problem of, Figure 7**

Schematic illustration of tsunami waveform inversion method

the slip distributions similar to that estimated from tsunami waves [44,45]. The large slip area, called asperity, produces high-frequency seismic waves thus important for strong-motion prediction for earthquake hazard assessments. The asperity produces large seafloor deformation, hence it is also important for tsunami generation and its hazard estimation. Lay and Kanamori [46] suggested that characteristic size of asperities differs from one subduction zone to another and is controlled by geological setting. Yamanaka and Kikuchi [47], from studies of recurrent earthquakes off northern Honshu, showed that the same asperity ruptures in repeated earthquakes. Their asperity map can be used for earthquake and tsunami hazard assessment.

### Waveform Inversion

The asperity distribution can be estimated by an inversion of tsunami waveforms. In this method (Fig. 7), the fault plane is first divided into several subfaults, and the seafloor deformation is computed for each subfault with a unit amount of slip. Using these as the initial conditions, tsunami propagation is numerically computed for actual bathymetry and the waveforms at tide gauge stations, called Green's functions, are computed. Assuming that the tsunami generation and propagation are linear process, the observed tsunami waveforms are expressed as



**Tsunamis, Inverse Problem of, Figure 8**

Slip distribution on 22 subfaults of the 2004 Sumatra-Andaman earthquake estimated from a joint inversion of tsunami waveforms on tide gauges and sea surface heights measured by satellite altimeters [11]

a linear superposition of Green's functions as follows,

$$A_{ij}(t) \cdot x_j = b_i(t) \quad (15)$$

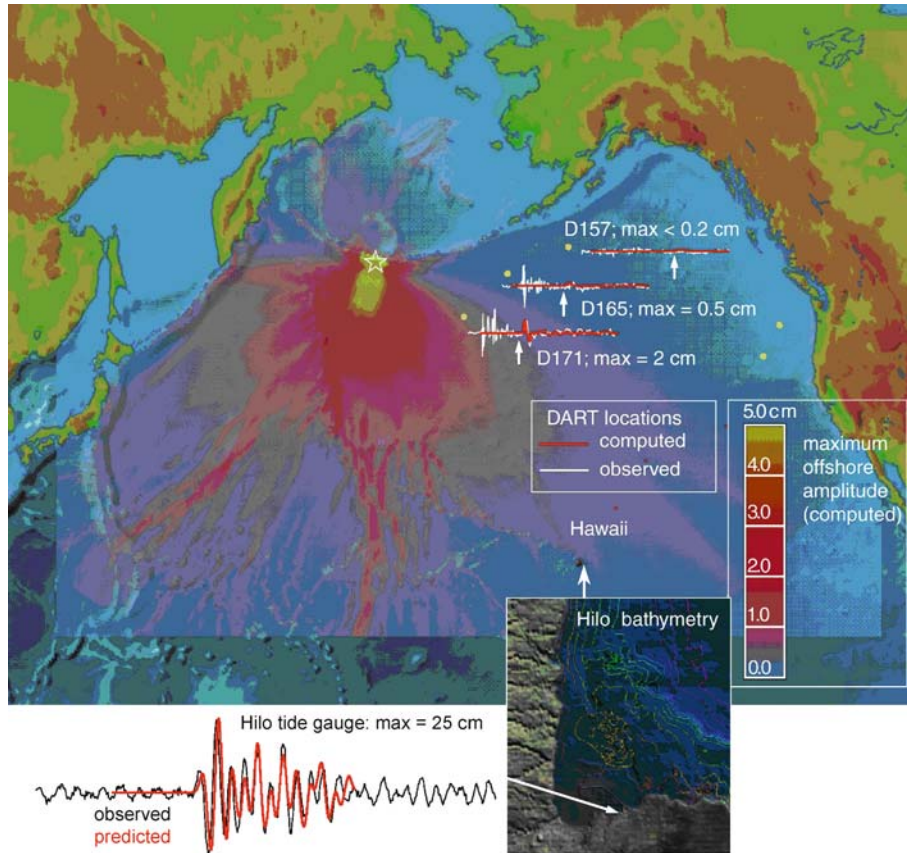
where  $A_{ij}(t)$  is the computed waveform as a function of time  $t$ , or Green's function, at the  $i$ th station from the  $j$ th subfault;  $x_j$  is the amount of slip on the  $j$ th subfault; and  $b_i(t)$  is the observed tsunami waveform at the  $i$ th station. The slip  $x_j$  on each subfault can be estimated by a least-square inversion of the above set of equations, by minimizing the  $l_2$  norm of the residuals  $J$

$$J = \|A \cdot \mathbf{x} - \mathbf{b}\| \rightarrow \min \quad (16)$$

where  $A$ ,  $\mathbf{x}$ , and  $\mathbf{b}$  indicate matrix representations of elements in Eq. (15). Figure 6d shows the slip distribution on the fault for the 1968 Tokachi-oki earthquake. The source fault was divided into 10 subfaults, and slip distribution on the subfaults were estimated. The largest slip, about 3.7 m, was estimated on the subfaults to the west of epicenter, but the average slip is 1.2 m, much smaller than that estimated by the trial and error method (4 m) which compared the maximum tsunami amplitudes [37].

The 2004 Indian Ocean tsunami, caused by the Sumatra-Andaman earthquake, was recorded by satellite altimeters, as well as tide gauges. A joint inversion of





**Tsunamis, Inverse Problem of, Figure 9**

Real-time data assimilation for the November 17, 2003 Rat Island Tsunami [22]. Star indicates the epicenter. Yellow dots are locations of DART systems. The tsunami waveforms recorded on DART (white curves) are compared with computed waveforms (red). Filled colors on ocean show computed maximum tsunami amplitudes of the updated source. The bottom plot shows location of Hilo tide gauge (right) and comparison of the predicted tsunami waveforms (red) with the recorded tide gauge signal (black)

tsunami waveforms recorded at 12 tide gauge stations and the sea surface heights measured by three satellites indicates that the tsunami source was about 900 km long [11]. The estimated slip distribution (Fig. 8) indicates that the largest slip, about 13 to 25 m, was located off Sumatra Island and the second largest slip, up to 7 m, near the Nicobar Islands. Inversion of satellite altimeter data alone supports a longer, about 1,400 km long, tsunami source [23], but such a model produces much larger tsunami waveforms than observed at Indian tide gauge stations. Inversion of tide gauge data alone does not support tsunami source beneath Andaman Islands [48]. The slip distribution estimated by the joint inversion is similar to those estimated from seismological analyses. The fault slip was the largest near off the northern Sumatra, followed by off Nicobar Islands [49]. Fault slip around Andaman Islands was estimated to be small from seismological analysis [50].

### Nonlinear Inversion Methods

In the above inversion method, both tsunami generation and propagation process are assumed to be linear. Because slip amount  $u$ , among the nine static fault parameters, is linearly related with the seafloor deformation and tsunami amplitudes, and it has the largest effect on tsunamis, other parameters are fixed in the above method. However, the tsunami propagation, particularly near coasts, might be affected by some nonlinear process such as advection or bottom friction.

Nonlinear inversion methods to overcome these limitations have been proposed. Pires and Miranda [51] proposed an adjoint method, which consists of four steps: source area delimitation by backward ray-tracing, the optimization of the initial sea state, nonlinear adjustments of the fault model, and final optimization of fault parameters.

The minimum of residual  $J$  is obtained iteratively through gradient descent method using the partial derivative or gradient of  $J$  with respect to parameters to be inverted.

### Inversion of Tsunami Heights

Maximum tsunami heights on coasts, rather than tsunami waveforms, have also been used for tsunami inversion. Distribution of maximum tsunami heights along the coasts is available from field surveys, historical or geological studies, and is valuable to study tsunami source. Piatanesi et al. [52] used coastal tsunami heights of the 1992 Nicaragua “tsunami earthquake” and estimated the slip distribution on the fault, as well as the mean amplification factor of computed coastal heights and measured runup heights.

Annaka et al. [53] proposed a method of joint inversion of tsunami waveforms and runup heights. As a residual to be minimized, they used a weighted sum of difference in waveforms (similar to Eq. 16) and logarithm of runup heights. They first tried linear inversion to estimate the initial value, then estimated the perturbation by the nonlinear inversion. Similar nonlinear inversion methods, based on the initial solution estimated by either linear inversion or other data such as geodetic data, have been proposed [54,55].

### Real-Time Data Assimilation

The tsunami waveform inversion can be done in real-time for the purpose of tsunami warning. The real-time data assimilation using the DART records and tsunami forecast have been made by NOAA [22]. They first use seismological information to determine the source location and parameters, then using the database of pre-computed simulation results, invert the DART data to estimate the tsunami source size (slip amounts). The tsunami forecast is made for farther locations where the tsunami has not arrived. For the 2003 Rat Island (Aleutians) earthquake, they successfully forecasted tsunami waveforms at Hilo, Hawaii, before the tsunami arrivals (Fig. 9).

### Future Directions

Inverse modeling methods of tsunami need to be further developed to better understand the tsunami generation process. Future developments are expected in each field of observation, propagation modeling and application to seismic and nonseismic tsunamis.

The tsunami observation system has been improved recently, particularly after the 2004 Indian Ocean tsunami.

Many instrumental data, both coastal and offshore, become available for the studies of tsunami generation process [56]. Maintenance of the systems, particularly for offshore systems, is sometimes costly, but essential to record infrequent tsunami. Open and real-time availability of such data is also important for tsunami studies as well as for tsunami warning purposes.

For the past tsunamis, more studies are needed to estimate tsunami heights from historical documents, as well as geological data such as distribution of tsunami deposits. Such historical tsunami database has been developed, e. g., at NOAA/NGDC ([http://www.ngdc.noaa.gov/seg/hazard/tsu\\_db.shtml](http://www.ngdc.noaa.gov/seg/hazard/tsu_db.shtml)).

For modeling tsunamis recorded on coastal tide gauges or runup heights, nonlinear computations with very fine bathymetry data are essential. While computational methods have been developed, fine bathymetry data are not always available. Developments of nonlinear inversion methods are also important.

Finally, inversion of tsunami data can be applied to tsunamis generated from submarine processes other than earthquakes, such as volcanic eruptions or landslides. For such nonseismic tsunamis, parametrization is essential to quantify the geological process and to solve inverse problems.

## Bibliography

### Primary Literature

1. Steketee JA (1958) On Volterra's dislocations in a semi-infinite elastic medium. *Can J Phys* 36:192–205
2. Mansinha L, Smylie DE (1971) The displacement fields of inclined faults. *Bull Seism Soc Am* 61:1433–1440
3. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seism Soc Am* 75:1135–1154
4. Hanks T, Kanamori H (1979) A moment magnitude scale. *J Geophys Res* 84:2348–2350
5. Kanamori H (1977) The energy release in great earthquakes. *J Geophys Res* 82:2981–2987
6. Lay T, Kanamori H, Ammon CJ, Nettles M, Ward SN, Aster RC, Beck SL, Bilek SL, Brudzinski MR, Butler R, DeShon HR, Ekstrom G, Satake K, Sipkin S (2005) The great Sumatra–Andaman earthquake of 26 December 2004. *Science* 308:1127–1133
7. Stein S, Okal EA (2005) Speed and size of the Sumatra earthquake. *Nature* 434:581–582
8. Tsai VC, Nettles M, Ekstrom G, Dziewonski AM (2005) Multiple CMT source analysis of the 2004 Sumatra earthquake. *Geophys Res Lett* 32. doi:10.1029/2005GL023813
9. Geist E (1998) Local tsunamis and earthquake source parameters. *Adv Geophys* 39:117–209
10. Yamashita T, Sato R (1974) Generation of tsunami by a fault model. *J Phys Earth* 22:415–440
11. Fujii Y, Satake K (2007) Tsunami source model of the 2004 Sumatra–Andaman earthquake inferred from tide gauge and satellite data. *Bull Seism Soc Am* 97:S192–S207

12. Satake K, Tanioka Y (2003) The July 1998 Papua New Guinea earthquake: Mechanism and quantification of unusual tsunami generation. *Pure Appl Geophys* 160: 2087–2118
13. Tanioka Y, Satake K (1996) Tsunami generation by horizontal displacement of ocean bottom. *Geophys Res Lett* 23:861–864
14. Song YT, Fu L-L, Zlotnicki V, Ji C, Hjorleifsdottir V, Shum CK, Yi Y (2008) The role of horizontal impulses of the faulting continental slope in generating the 26 December 2004 tsunami. *Ocean Modell* 20:362–379
15. Intergovernmental Oceanographic Commission (1997) IUGG/IOC TIME Project Numerical Method of Tsunami Simulation with the Leap-frog Scheme. UNESCO, Paris
16. Mader CL (1988) Numerical modeling of water waves. University of California Press, Berkeley
17. Yeh H, Liu P, Synolakis C (1996) Long-wave runup models. World Scientific, Singapore
18. Smith WHF, Scharroo R, Titov VV, Arcas D, Arbic BK (2005) Satellite altimeters measure tsunami, early model estimates confirmed. *Oceanography* 18:10–12
19. Kato T, Terada Y, Kinoshita M, Kakimoto H, Isshiki H, Matsui M, Yokoyama A, Tanno T (2000) Real-time observation of tsunami by RTK-GPS. *Earth Planet Space* 52:841–845
20. Mikada H, Mitsuzawa K, Matsumoto H, Watanabe T, Morita S, Otsuka R, Sugioka H, Baba T, Araki E, Suyehiro K (2006) New discoveries in dynamics of an M8 earthquake-phenomena and their implications from the 2003 Tokachi-oki earthquake using a long term monitoring cabled observatory. *Tectonophysics* 426:95–105
21. Gonzalez FI, Bernard EN, Meinig C, Eble MC, Mofjeld HO, Stalin S (2005) The NTHMP tsunameter network. *Nat Hazard* 35: 25–39
22. Titov VV, Gonzalez FI, Bernard EN, Eble MC, Mofjeld HO, Newman JC, Venturato AJ (2005) Real-time tsunami forecasting: Challenges and solutions. *Nat Hazard* 35:41–58
23. Hirata K, Satake K, Tanioka Y, Kuragano T, Hasegawa Y, Hayashi Y, Hamada N (2006) The 2004 Indian Ocean tsunami: Tsunami source model from satellite altimetry. *Earth Planet Space* 58:195–201
24. Synolakis CE, Okal EA (2005) 1992–2002: Perspective on a decade of post-tsunami surveys. In: Satake K (ed) *Tsunamis: Case studies and recent developments*. Springer, Dordrecht, pp 1–29
25. Intergovernmental Oceanographic Commission (1998) Post-tsunami survey field guide. UNESCO, Paris
26. Satake K, Wang KL, Atwater BF (2003) Fault slip and seismic moment of the 1700 Cascadia earthquake inferred from Japanese tsunami descriptions. *J Geophys Res* 108. doi:10.1029/2003JB002521
27. Atwater BF, Musumi-Rokkaku S, Satake K, Tsuji Y, Ueda K, Yamaguchi DK (2005) The orphan tsunami of 1700. *USGS Prof Paper* 1707:133
28. Dawson AG, Shi SZ (2000) Tsunami deposits. *Pure Appl Geophys* 157:875–897
29. Nanayama F, Satake K, Furukawa R, Shimokawa K, Atwater BF, Shigeno K, Yamaki S (2003) Unusually large earthquakes inferred from tsunami deposits along the Kuril trench. *Nature* 424:660–663
30. Satake K (1988) Effects of bathymetry on tsunami propagation – Application of ray tracing to tsunamis. *Pure Appl Geophys* 126:27–36
31. Rabinovich AB, Thomson RE (2007) The 26 December 2004 Sumatra tsunami: Analysis of tide gauge data from the world ocean Part 1, Indian Ocean and South Africa. *Pure Appl Geophys* 164:261–308
32. Miyabe N (1934) An investigation of the Sanriku tsunami based on mareogram data. *Bull Earthq Res Inst Univ Tokyo Suppl* 1:112–126
33. Merrifield MA, Firing YL, Aarup T, Agricole W, Brundrit G, Chang-Seng D, Farre R, Kilonsky B, Knight W, Kong L, Magori C, Manurung P, McCreery C, Mitchell W, Pillay S, Schindele F, Shillington F, Testut L, Wijeratne EMS, Caldwell P, Jardin J, Nakahara S, Porter FY, Turetsky N (2005) Tide gauge observations of the Indian Ocean tsunami, December 26, 2004. *Geophys Res Lett* 32. doi:10.1029/2005GL022610
34. Nagarajan B, Suresh I, Sundar D, Sharma R, Lal AK, Neetu S, Shenoi SSC, Shetye SR, Shankar D (2006) The great tsunami of 26 December 2004: A description based on tide-gauge data from the Indian subcontinent and surrounding areas. *Earth Planet Space* 58:211–215
35. Neetu S, Suresh I, Shankar R, Shankar D, Shenoi SSC, Shetye SR, Sundar D, Nagarajan B (2005) Comment on “The great Sumatra–Andaman earthquake of 26 December 2004”. *Science* 310:1431a
36. Abe K (1973) Tsunami and mechanism of great earthquakes. *Phys Earth Planet Inter* 7:143–153
37. Satake K (1989) Inversion of tsunami waveforms for the estimation of heterogeneous fault motion of large submarine earthquakes – The 1968 Tokachi-Oki and 1983 Japan Sea earthquakes. *J Geophys Res* 94:5627–5636
38. Mei CC (1989) The applied dynamics of ocean surface waves. World Scientific, Singapore
39. Abe K (1979) Size of great earthquakes of 1873–1974 inferred from tsunami data. *J Geophys Res* 84:1561–1568
40. Abe K (1981) Physical size of tsunamigenic earthquakes of the northwestern Pacific. *Phys Earth Planet Inter* 27:194–205
41. Abe K (1989) Quantification of tsunamigenic earthquakes by the Mt scale. *Tectonophysics* 166:27–34
42. Aida I (1978) Reliability of a tsunami source model derived from fault parameters. *J Phys Earth* 26:57–73
43. Yanagisawa K, Imamura F, Sakakiyama T, Annaka T, Takeda T, Shuto N (2007) Tsunami assessment for risk management at nuclear power facilities in Japan. *Pure Appl Geophys* 164: 565–576
44. Kikuchi M, Fukao Y (1985) Iterative deconvolution of complex body waves from great earthquakes – The Tokachi-oki earthquake of 1968. *Phys Earth Planet Inter* 37:235–248
45. Mori J, Shimazaki K (1985) Inversion of intermediate-period Rayleigh waves for source characteristics of the 1968 Tokachi-oki earthquake. *J Geophys Res* 90:11374–11382
46. Lay T, Kanamori H (1981) An asperity model of large earthquake sequences. In: Simpson DW, Richards PG (eds) *Earthquake prediction – An international review*. American Geophysical Union, Washington DC, pp 579–592
47. Yamanaka Y, Kikuchi M (2004) Asperity map along the subduction zone in northeastern Japan inferred from regional seismic data. *J Geophys Res* 109:B07307, doi:10.1029/2003JB002683
48. Tanioka Y, Yudhicara, Kusunose T, Kathirolu S, Nishimura Y, Iwasaki S-I, Satake K (2006) Rupture process of the 2004 great Sumatra–Andaman earthquake estimated from tsunami waveforms. *Earth Planet Space* 58:203–209
49. Ammon CJ, Ji C, Thio HK, Robinson D, Ni SD, Hjorleifsdottir V,

- Kanamori H, Lay T, Das S, Helmberger D, Ichinose G, Polet J, Wald D (2005) Rupture process of the 2004 Sumatra–Andaman earthquake. *Science* 308:1133–1139
50. Velasco AA, Ammon CJ, Lay T (2006) Search for seismic radiation from late slip for the December 26, 2004 Sumatra–Andaman (Mw=9.15) earthquake. *Geophys Res Lett* 33:L18305, doi:10.1029/2006GL027286
  51. Pires C, Miranda PMA (2001) Tsunami waveform inversion by adjoint methods. *J Geophys Res* 106:19773–19796
  52. Piatanesi A, Tinti S, Gavagni I (1996) The slip distribution of the 1992 Nicaragua earthquake from tsunami run-up data. *Geophys Res Lett* 23:37–40
  53. Annaka T, Ohta K, Motegi H, Yoshida I, Takao M, Soraoka H (1999) A study on the tsunami inversion method based on shallow water theory. *Proc Coastal Engin JSCE* 46:341–345
  54. Yokota T, Nemoro M, Masuda T (2004) Estimate of slip distribution by tsunami height data inversion. *Abst Jpn Earth Planet Sci Joint Meeting* S043-P0005
  55. Namegaya Y, Tsuji Y (2007) Distribution of asperities of the 1854 Ansei Nankai earthquake. *Abst Jpn Earth Planet Sci Joint Meeting* S142-009
  56. Satake K, Baba T, Hirata K, Iwasaki S, Kato T, Koshimura S, Takenaka J, Terada Y (2005) Tsunami source of the 2004 off the Kii peninsula earthquakes inferred from offshore tsunami and coastal tide gauges. *Earth Planet Space* 57:173–178

### Books and Reviews

- Lawson CL, Hanson RJ (1974) Solving least squares problems. Prentice-Hall, Englewood Cliffs. (Republished by Society for Industrial and Applied Mathematics, 1995)
- Lay T, Wallace TC (1995) Modern global seismology. Academic Press, San Diego
- Menke W (1989) Geophysical data analysis: Discrete inverse theory (revised edition). Academic Press, San Diego
- Satake K (2007) Tsunamis. In: Kanamori H (ed) *Treatise on Geophysics*, vol 4. Elsevier, Amsterdam

## Tunneling Through Quantum Dots with Discrete Symmetries

YSHAI AVISHAI<sup>1,2</sup>, KONSTANTIN KIKOIN<sup>3</sup>

<sup>1</sup> Physics Department and Ilse Katz Institute for Nanotechnology, Ben-Gurion University of the Negev, Beer-Sheva, Israel

<sup>2</sup> RTRA-Triangle de la Physique, LPS (Orsay) and CEA-SPHT (Saclay), Gif sur Yvette, France

<sup>3</sup> Department of Physics, Tel-Aviv University, Tel-Aviv, Israel

### Article Outline

Glossary

Definition of the Subject

Introduction

Complex Quantum Dots

Dynamical Symmetries and Kondo Effect in Tunneling Through Complex Quantum Dots

Discrete Degrees of Freedom in Kondo Tunneling Tunneling in External Fields

Connections with other Nano-objects and Future Prospects

Bibliography

### Glossary

**Quantum dot** Quantum dot (QD) is an element of artificial nanostructures. It arises in a situation when a finite number of electrons is confined in a puddle of  $10^0$ – $10^2$  nanometer size either by means of external electrical potential imposed on semiconductor heterostructure or in a process of formation of non-equilibrium self-assembled structures. The electron wavelength in a QD is comparable with its size. As a result all energy levels are spatially quantized like in atoms or molecules. Quantum systems with fully quantized energy levels are defined as zero-dimensional nano-objects. Complex quantum dots consist of several QDs organized in linear or ring structures.

**Discrete symmetry** Group theory classifies all physical objects in accordance with their symmetry properties relative to symmetry transformations (rotations and translations) in space and time. These operations may be continuous or discrete. In the latter case one speaks about discrete rotational or translational symmetry.

**Dynamical symmetry** Dynamical symmetry characterizes not only the eigenstates of a Hamiltonian, but also the symmetry of transition between the states belonging to different irreducible representation of the symmetry group of the Hamiltonian under external dynamical perturbation. Dynamical symmetry may be continuous and discrete or combine both types of symmetry operations.

**Kondo effect** Kondo effect is the many particle phenomenon which arises in magnetically doped metals due to exchange scattering of metallic electrons on localized magnetic moments of impurities. Multiple electron scattering processes result in dynamical screening of impurity magnetic moment. As a result, impurity-related electrical resistivity has a minimum at some temperature, and reaches the unitarity limit at  $T \rightarrow 0$ . This limit corresponds to maximum backward electron scattering. Kondo effect exists also in tunneling through small QDs. In this case the unitarity limit corresponds to maximum tunnel transparency of QD.

**Quantum tunneling** Quantum tunneling of electrons through potential barriers arises at low temperatures,



where all overbarrier activation processes are frozen. In nanostructures this type of transport is realized on the interfaces between different materials forming nanostructure or in especially designed tunneling channels.

### Definition of the Subject

Electrons may be confined within a nano-size quantum box by various methods. The first example of such confinement was demonstrated in the studies of optical properties of semiconductor precipitates in glasses [10]. Later on such confinement was realized in planar quantum dots. These dots are fabricated in semiconductor heterostructures, where the electrons already confined in a two-dimensional layer between two semiconductors (usually GaAs/GaAlAs) are locked in a nano-size puddle by electrostatic potential created by electrodes superimposed on the heterostructure (see [25,51] for a description of the early stage of the physics of quantum dots). QDs may be also prepared by means of colloidal synthesis [4], grown as self-assembled structures of semiconductor droplets on a strained surface of another semiconductor [36], etc. In particular, quantum dots may be fabricated in a form of vertical structures possessing cylindrical symmetry [32].

Discrete symmetries become relevant in tunneling through QD when this nanoobject consists of several individual dots. In such complex quantum dots (CQDs) electrons may occupy any of the individual quantum wells, having the same or nearly the same energy provided all wells are identical or nearly identical. Thus, additional degeneracy of the electron spectrum arises, which is characterized by the discrete permutation symmetry group. If CQD has the form of an “artificial molecule” or has a form of a regular polygon, this permutation symmetry transforms into point symmetry. Interplay between the discrete symmetries of CQD,  $SU(2)$  spin symmetry of the electron confined in the dot and  $U(1)$  symmetry of its charge results in new exciting properties of these artificial systems which are unimaginable in “natural” objects such as atoms, molecules or impurities in metals and semiconductors.

### Introduction

To study electron tunneling through QDs, one should build it into an electric circuit. In the case of the planar quantum dot the metallic electrodes (source and drain leads) are formed in the two-dimensional layer filled by electrons. The tunnel current emerges when the bias voltage  $V_b$ , is applied to the leads, and the number of electrons in QD is regulated by the gate voltage  $v_g$  applied to the puddle. The number of electrons in such QD may be var-

ied experimentally from 1–2 to several tens. The most remarkable property of QDs is a single electron tunneling, which is realized in small enough dots, provided the charging energy  $Q$  spent for injecting an electron from the dot exceeds both the inter-level spacing  $\delta\varepsilon$  and the tunneling rate  $\Gamma$ . In this regime the current-voltage characteristics  $I(V_b)$  acquires the step-wise form of the “Coulomb ladder” instead of the conventional linear Ohm’s law. Tunnel conductance  $G = \delta I / \delta V_b$  is nothing but a sequence of delta-function-like peaks. Each peak corresponds to injection of the next electron in the dot. This possibility of “counting electrons by number” opens the way to construction of single electron transistors and other nano-devices.

Semiconductor QDs may be used as single photon emitters (“single photon on demand”). Besides, an ensemble of QDs is proposed as a reliable carrier of spin qubits for quantum computers [37,53].

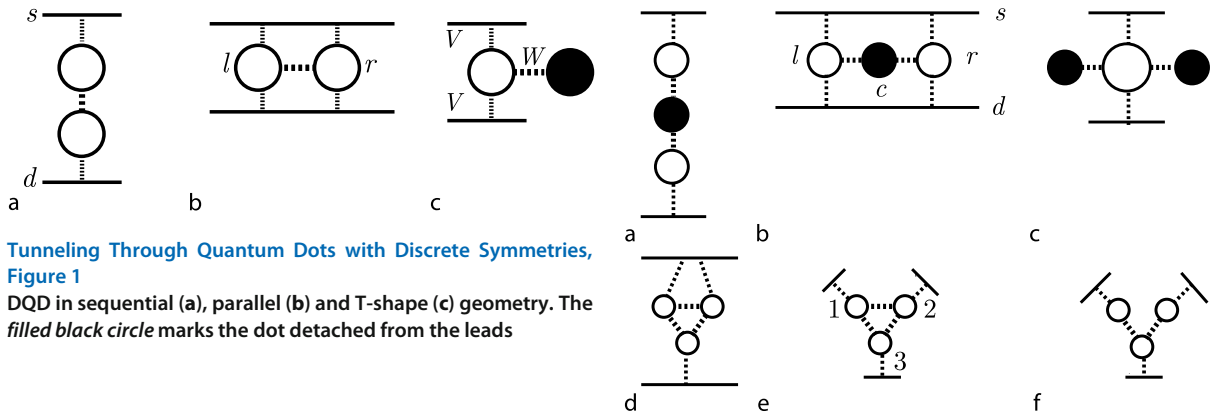
On the other hand, a quantum dot with strong Coulomb-exchange correlation in a tunnel contact with two electron reservoirs (Fermi-seas) is an excellent model system for studying fundamental quantum mechanical and quantum statistical phenomena, such as the Kondo effect [15,42], edge singularity phenomenon [39], Fano-resonances [24,29], etc. Quantum dots may be organized in complex self-assembled arrays, and these arrays possess features, which combine the properties of solid state and atomic physics [6,55].

### Complex Quantum Dots

In the case of a strong Coulomb blockade, the number of electrons in a QD is fixed, and one should discriminate between the dots with even and odd electron occupation. Electrons in a QD occupy discrete levels in accordance with the Pauli principle. Then the dot with odd occupation is characterized by spin  $1/2$ , whereas the dot with even occupation has zero spin. Disk-like planar QDs and vertical QDs possess cylindrical symmetry, so these nano-objects may be considered as few electron systems with shell structures (“artificial cylindrical atoms” [32]). The electron shells are occupied in accordance with Hund rules modified for cylindrical symmetry [23]. As a result the ground and low-lying excited states in these objects are manifolds consisting of states with spin  $S = 1/2, 3/2$  (odd occupation) and  $S = 1, 0$  (even occupation).

Complex quantum dots may be treated as “artificial molecules”. The simplest of such complex objects is the double quantum dot (DQD), first realized experimentally in planar geometry [18,41]. It consists of two islands with confined electrons. Both capacitive (electrostatic) and tunneling coupling may exist between these islands. If two is-





**Tunneling Through Quantum Dots with Discrete Symmetries, Figure 1**

DQD in sequential (a), parallel (b) and T-shape (c) geometry. The filled black circle marks the dot detached from the leads

lands are equivalent, then the structure of the electronic spectrum mimics that of elementary molecules: DQD with two electrons simulates  $H_2$ , the dots with one or three electrons look like positively and negatively charged H molecules, etc. If the two dots are not equivalent (e. g., two potential wells have different depth and/or width) then the electronic structure of DQD mimics that of a polar molecule, e. g. DQD with four electrons reminds us of the LiH molecule. One may tune the balance between ionic and covalent components of the inter-dot coupling by varying the gate voltage applied to two dots and the width of the tunnel channel between them.

DQD may be coupled to source and drain electrodes in several ways (Fig. 1). In cases (a) and (b) there is a one and two tunneling channel, respectively. In case (c) tunneling through the single channel is controlled by charge and spin states of the side dot.

In the case of even occupation and strong left-right ( $l$ - $r$ ) tunnel coupling, all electrons are shared between two islands, the dot is characterized only by its spin state (singlet or triplet), and the electron motion from source ( $s$ ) to drain ( $d$ ) is possible only in the co-tunneling regime (where the electron from the source tunnels into the DQD only when another electron leaves it on its way into the drain). In case of odd occupation one “unpaired” electron oscillates between the two islands. As a result, the DQD acquires an orbital degree of freedom, which is characterized by a discrete  $l$ - $r$  symmetry.

Like in other two-level systems, this symmetry may be characterized by a pseudo-spin operator  $\tau$  acting in ( $l$ - $r$ ) space.

More diverse discrete symmetries are realized in triple quantum dots (TQD). These dots may have linear structure [8,16] or form closed loops (triangles) [56]. Some interesting and experimentally accessible geometric structures of tunneling devices with TQD are shown in Fig. 2.

The linear TQDs shown in the upper panel are characterized by  $l$ - $r$  reflection symmetry, whereas the triangular

**Tunneling Through Quantum Dots with Discrete Symmetries, Figure 2**

TQD in numerous configurations: sequential (a), parallel (b), cross (c), two-terminal triangular (d), three-terminal triangular (e) and fork (f). The filled black circles mark the dots which are detached from the leads

TQD with equivalent constituent smaller dots (configurations (d, e) in the lower panel), have a discrete symmetry of a triangle, that is represented by the group  $C_{3v}$  which is equivalent to a permutation symmetry  $P_3$ . This symmetry is violated in case (d) by the tunnel contacts with source and drain. The QD with cross symmetry (c) is a generalization of the T-shape symmetric QD displayed in Fig. 1c.

### Dynamical Symmetries and Kondo Effect in Tunneling Through Complex Quantum Dots

Any characteristic (internal) symmetry of a QD may be (and generically is) violated due to an interaction with the environment (usually electrons in the leads forming a Fermi sea). All systems which are schematically depicted in Figs. 1, 2 can be described by the Hamiltonian

$$H = H_d + H_r + H_c \quad (1)$$

Here the indices  $d$ ,  $r$ ,  $c$  denote dot, electron reservoir and dot-reservoir coupling, respectively. In a simple special case the variables, which are conserved in isolated QD are charge and spin, so that

$$H_d = H_{d0} + QN^2 + \lambda S^2. \quad (2)$$

Here  $H_{d0}$  describes the discrete energy spectrum of the QD,  $Q$  is the electrostatic (capacitive) energy fixing the number of electrons  $N$  in the dot, and  $\lambda < 0$  is the exchange energy responsible for stabilization of the dot spin (described by the operator  $S$ ). The Hamiltonian (1) is a generic Hamiltonian for a class of strongly correlated

electron systems (SCES) [13,27]: it describes coupling between two subsystems, one of which  $H_T$  contains nearly free particles with weak interaction, whereas the low-energy excitations in the dot Hamiltonian  $H_d$  are fully determined by the strong correlation, which acquires the form of the constraint as in (2). As a rule, the coupling term  $H_c$  between the two subsystems results in a complete reconstruction of the low-energy part of the energy spectrum of SCES within some interval  $E_K$  due to various many-body effects characterized by infrared singularities. Since the symmetry of the Hamiltonian (2) is violated by  $H_c$ , this coupling initiates transitions between the states belonging to different irreducible representations (irreps) of the symmetry group  $G_H$  of  $H_d$  within the energy interval  $E_K$ . This means that the dynamical symmetry of CQD influences the physics of tunneling through this object.

In the context of a quantum tunneling problem, the dynamical symmetry group  $G_D$  is determined as the group generated by the collection of operators  $\mathbf{R}$  inducing transitions both within a given irrep and between different irreps of the symmetry group  $G_H$  [28]. This set of operators form a closed algebra with commutation relations

$$[R_i, R_j] = f_{ijk} R_k, \quad (3)$$

where  $f_{ijk}$  are the structure factors. The simplest example of such a group is the group  $SO(4)$  describing the dynamical symmetry of triplet-singlet manifold  $S = 1, 0$  of DQD with even occupation [26]. This group is formed by six generators. Three of them are the components of the operator  $\mathbf{S}$ , and the other three are the components of another vector  $\mathbf{P}$  with the components

$$\begin{aligned} P^+ &= \sqrt{2}(|T_\uparrow\rangle\langle S| - |S\rangle\langle T_\downarrow|), \\ P_z &= -(|T_0\rangle\langle S| + |S\rangle\langle T_0|) \end{aligned} \quad (4)$$

describing transitions between singlet ( $S$ ) and three components of spin triplet ( $T_\nu$ ). In the case of DQD with  $N = 2$ , where each dot in Fig. 1 is occupied by one electron with spin  $s$ , the two vectors may be represented as

$$\mathbf{S} = \mathbf{s}_1 + \mathbf{s}_2, \quad \mathbf{P} = \mathbf{s}_1 - \mathbf{s}_2. \quad (5)$$

Kinematics imposes two constraints on these vectors, namely

$$s_1^2 + s_2^2 = 3/2; \quad s_1^2 - s_2^2 = 0.$$

These two constraints are in fact the two Casimir operators  $\hat{C}_1, \hat{C}_2$  supplementing the  $o(4)$  algebra of the  $SO(4)$  group. In other words, these two constraints are  $\hat{C}_1 = \mathbf{S}^2 + \mathbf{P}^2 = 3$ ,  $\hat{C}_2 = \mathbf{S} \cdot \mathbf{P} = 0$ . The first of them arises in  $H_d$  instead

of the usual Casimir operator  $\mathbf{S}^2 = S(S+1)$ . This Hamiltonian then acquires the form

$$H_d = H_{d0} + QN^2 + (E_T \mathbf{S}^2 + E_S \mathbf{P}^2)/2, \quad (6)$$

where  $E_T, E_S$  are the energies of triplet and singlet states respectively, and  $E_T - E_S = -\lambda$ .

Among the many-body features, which accompany tunneling through QDs the most salient and universal is the Kondo effect. This effect, originally found in many-particle scattering of electrons by magnetic impurities in metals [31] is responsible also for the well pronounced temperature and magnetic field dependent zero bias anomalies in conductance through quantum dots [15,42]. The physical explanation of this mapping of the scattering problem in bulk metals on the tunneling problem in low-dimensional systems is that the strong Coulomb blockade allows electron propagation through the QD only in the co-tunneling regime (see above), which does not preserve the spin of the dot. The co-tunneling Hamiltonian which arises in second order in  $H_c$ , projects out charge excitations and takes into account spin reversal processes. It has the form of the effective exchange term,

$$H_{cot} = J \mathbf{S} \cdot \boldsymbol{\sigma}, \quad (7)$$

where the operator  $\boldsymbol{\sigma}$  describes spin excitations of the itinerant electrons in a metallic reservoir and  $J \sim V^2$  is the coupling constant quadratic in the tunneling amplitude  $V$  between the reservoir and the QD.

In complex quantum dots the Kondo effect should take into account dynamical symmetries of CQD, provided the characteristic energy scale  $E_K \sim \exp(-1/J)$  (Kondo energy) is comparable with the scale of low-lying states in the manifold of eigenstates of  $H_d$ . In DQD where this manifold includes the states  $E_T, E_S$ , dynamical symmetry  $SO(4)$  is involved in Kondo tunneling provided the singlet-triplet gap  $|\lambda| \sim E_K$  [11,26]. Then both vectors  $\mathbf{S}$  and  $\mathbf{P}$  are involved in Kondo cotunneling, and the effective Hamiltonian has the form

$$H_{cot} = J_1 \mathbf{S} \boldsymbol{\sigma} + J_2 \mathbf{P} \cdot \boldsymbol{\sigma}. \quad (8)$$

In this situation the Kondo energy scale is a function of the exchange gap  $E_K(\lambda)$ , and the zero bias anomaly in the tunnel conductance becomes sensitive to its value.

### Discrete Degrees of Freedom in Kondo Tunneling

The short chains of QDs represented in Fig. 1a-c and Fig. 2a-c, are among the most simple objects, where the interplay between discrete symmetries and the original Kondo physics may be demonstrated. An electron in

a DQD in the charge sector  $N = 1$  is represented not only by its spin  $\sigma$  with projections  $\pm$ , but also by the pseudo-spin vector operator  $\tau$  describing its position in  $l$ - $r$  space within the double well:

$$\begin{aligned}\tau_z &= \sum_{\sigma} \left( d_{l\sigma}^{\dagger} d_{l\sigma} - d_{r\sigma}^{\dagger} d_{r\sigma} \right), \quad \tau^+ = \sum_{\sigma} d_{l\sigma}^{\dagger} d_{r\sigma}, \\ \tau^- &= \sum_{\sigma} d_{r\sigma}^{\dagger} d_{l\sigma}.\end{aligned}\quad (9)$$

This vector, together with the four spin vectors  $S_{ij}$  representing spin in a double well ( $i, j = l, r$ ) form a set of 15 generators of the  $SU(4)$  group:

$$\{(\tau^+, \tau^-, \tau_z, I) \otimes (\sigma^+, \sigma^-, \sigma_z, I)\} - \{I \otimes I\}. \quad (10)$$

The effective co-tunneling Hamiltonian in this case acquires the form

$$H_{\text{cot}} = \sum_{ij} J_{ij} S_{ij} \cdot s_{ji} + K \tau \cdot t. \quad (11)$$

Here  $I$  is the unit  $2 \times 2$  matrix,  $s_{ij}$  is the spin operator for electrons in the leads ( $s, d$ ),  $t$  is the pseudospin operator for lead electrons similar to (9). In sequential tunneling geometry of Fig. 1a, only the coupling ( $ls$ ) and ( $rd$ ) is taken into account.

Electron levels in the double well are split due to the inter-well tunneling  $\tau^{\pm}$ . Both split levels are involved in Kondo cotunneling provided the splitting energy  $\Delta$  is comparable with the Kondo energy.

The Kondo energy itself is a function of this splitting,  $E_K(\Delta)$ . This energy splitting plays the same role in the discrete dynamical  $SU(4)$  symmetry as the exchange splitting  $\lambda$  does in the dynamical symmetry  $SO(4)$ . In both cases, involvement of excited states in the Kondo tunneling results in the increase of  $E_K$ , and thereby it causes the enhancement of the zero bias anomaly of the tunnel conductance [11,12,26,49]. This increment is encoded within the following asymptotic power law,

$$E_K(\delta) = E_K(0)[E_K(0)/\delta]^{\gamma} \quad (\delta = \lambda, \Delta). \quad (12)$$

Here the exponent  $\gamma$  is specific for a given dynamical symmetry.

In many cases, the exchange splitting energy may change its sign following tuning of the dot parameters, as well as gate voltages. In this case, one speaks about triplet singlet crossover. When the triplet state is involved in Kondo tunneling at finite energy but the zero bias anomaly disappears at low temperatures  $kT \ll |\lambda|$  [19,49], one even speaks of a quantum phase transition.

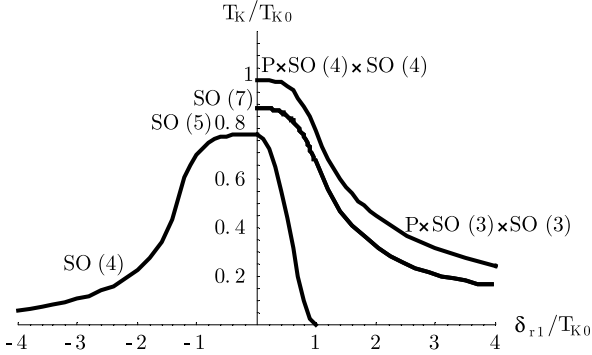
An even more multifarious situation arises in linear TQD [33]. For example, in a charge sector  $N = 4$  the electron distribution in a TQD with strong Coulomb blockade in the central dot and weak blockade in the side dots (Fig. 2a, b) has a shell-like structure: the first electron occupies the deepest level in the central dot, the two additional electrons which are shared between the two side dots form a sort of closed shell and the last electron dwells in a double ( $l$ - $r$ ) well. The low-lying part of the spin manifold consists of two triplets and two singlets. The relative positions of these levels may be changed by varying the gate voltages applied to the side dots as well as by tuning other experimentally controllable parameters of the device.

In the most symmetric situation where the  $l$ - $r$  reflection symmetry is preserved, the symmetry of the TQD is  $P_2 \otimes SO(4) \otimes SO(4)$ . This is the symmetry of two singlet/triplet pairs degenerate under left-right permutation. It can be reduced to the  $SO(7)$  dynamical symmetry, which is realized for the manifold consisting of two triplets and one singlet. The set of 21 operators, which describe transitions within this manifold form a closed algebra  $o(7)$ . These operators are grouped in six vectors and three scalars. The pattern of commutation relations has a more complicated form than (3). One may also construct five Casimir operators describing kinematical constraints on spin variables in the TQD. Another dynamical symmetry is the  $SO(5)$  symmetry of a manifold containing two singlets and one triplet. The corresponding algebra  $o(5)$  is formed by ten generators grouped in three vectors and one scalar (the latter describes transition between two singlets). Three Casimir operators constrain this dynamical symmetry. Other symmetries realizable in TQD are  $SO(4)$ ,  $P_2 \otimes SO(3) \otimes SO(3)$  and  $SO(3)$ .

Thus, because of interplay between discrete  $l$ - $r$  reflection symmetry and dynamical  $SO(n)$  symmetries of spin multiplets, the linear TQD possesses a quite complicated phase diagram, where the index  $n$  may be changed. In accordance with the general law (12), the Kondo energy depends on the value of the gaps  $\delta$  separating the lowest excitation of the ground state of the system. Figure 3 illustrates this variation for several parts of the phase diagram.

Thus, the parameters of the dynamical group straightforwardly influence the tunnel conductivity of the TQD.

The problem of Kondo tunneling through a TQD in cross and fork geometries (Fig. 2c, f) was analyzed in [35]. It was found that these TQDs with discrete mirror symmetry at odd occupation  $N = 1, 3$  possess properties, which were observed earlier only in QD with even occupation  $N = 2$ . In particular, the Kondo tunneling may be absent in the ground spin-doublet state due to the special symmetry of the electron wave function in the TQD, but, on the



**Tunneling Through Quantum Dots with Discrete Symmetries, Figure 3**

Variation of Kondo temperature with the difference between gate voltages  $\delta_{r1} = v_{gr} - v_{gl}$  applied to the left and right dot

other hand, Kondo-active states are involved in tunneling at higher excitation energies. This behavior is similar to that encountered in DQD with  $N = 2$  in a singlet ground state with low-lying triplet excitation [19].

A system consisting of TQD in the fork geometry looks promising also for studying the physics of spin entanglement [52]. In a charge sector  $N = 4$ , the spins of the electrons located in the side dots are entangled in a singlet state via the doubly occupied central dot.

Consider now the discrete point symmetry  $C_{3v}$  of an equilateral triangular TQD shown in Fig. 2d, e. It is superimposed on the spin symmetry of electrons localized in the three wells [9,34]. In the charge sector  $N = 1$  an electron may occupy three equivalent positions, so that the total symmetry of the TQD is  $SU(6)$  (spin + three colors) but this degeneracy is lifted by the inter-dot tunneling  $W$ , so that the orbital triplet is split into a singlet ( $a$ ) and a doublet ( $b$ ). The splitting energy is  $\Delta_{tr} = E_a - E_b = 3W$ . Since  $W < 0$ , the ground state is the orbital doublet, and the discrete degrees of freedom are quenched, provided  $E_K \ll \Delta_{tr}$ , which is usually the case. However, the orbital

degeneracy may be induced by an external magnetic field (see next section).

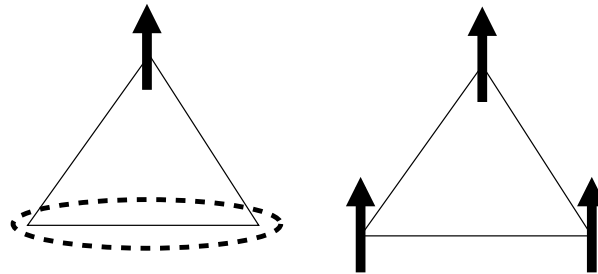
Two-electron states in the charge sector  $N = 2$  are formed by electrons occupying adjacent wells in triplet and singlet states, and this classification is supplemented by the same orbital classification as in the case of  $N = 1$ . The manifold consists of four levels ordered as  $E_{sa} < E_{tb} < E_{ta} < E_{sb}$ . As a result, the possibility opens for dynamical mixing of singlet and triplet states with different quantum numbers.

In the three-electron sector  $N = 3$ , the discrete symmetry of the triangle opens a new possibility of ordering spins. Since the indirect exchange interaction induced by virtual inter-dot tunneling is always antiferromagnetic, there is no possibility of energy minimization by means of simultaneous anti-parallel orientation of all three spins. Because of these frustrations the low-energy spin manifold is formed by three configurations of spin 1/2 localized in one of the three dots, while two other spins are coupled into a spin singlet (see Fig. 4). Only two of these states are linearly independent, so that the ground state is doubly degenerate. This degeneracy (stemming from the discrete point symmetry of a triangle) is lifted by the contact with source and drain (in the geometries of Fig. 2d, f).

The effective spin Hamiltonian describing these configurations has non-Heisenberg form [53]

$$H_d = \sum_{ij} J_{ij} \mathbf{S}_i \cdot \mathbf{S}_j + \sum_{ijk} D_{jkl} \mathbf{S}_i \cdot [\mathbf{S}_j \times \mathbf{S}_k]. \quad (13)$$

Here  $J_{ij}$  is the indirect inter-dot exchange integral,  $D_{jkl}$  is triple co-tunneling constant describing cyclic electron exchange between three sites. The quantity  $\chi_{ijk} = \mathbf{S}_i \cdot [\mathbf{S}_j \times \mathbf{S}_k]$  is characterized by a quantum number referred to as scalar spin chirality. Lead-dot co-tunneling processes involve also the excited high spin state with  $S = 3/2$ . As a result of interplay between discrete point symmetry and con-



**Tunneling Through Quantum Dots with Discrete Symmetries, Figure 4**

Possible configurations of three spins in TQD: two spin-polarized states with  $S = 1/2$  and high spin state with  $S = 3/2$

tinuous spin symmetry of the TQD, dynamical symmetry includes spin variables, spin chirality number and pseudo-spin describing the position of an un-compensated spin in the low-spin state of the TQD.

### Tunneling in External Fields

External electric field  $E$  and magnetic field  $B$  influence tunneling through CQD in two different ways. First, these fields introduce additional energy scales into the dynamics of co-tunneling processes. These are the electrostatic potential associated with the electric field and the Zeeman splitting energy  $g\mu_B B$  ( $g$  is the gyromagnetic ratio for the electron and  $\mu_B$  is the Bohr magneton). Second, these fields lower the symmetry of the system. Application of an electric field for achieving confinement in two-dimensional systems introduces an additional vector  $\mathbf{n}$  normal to the confinement plane (Rashba vector [7]). The magnetic field breaks spin rotational symmetry of CQD and introduces chirality of orbital states in case of closed configurations (by breaking time reversal invariance).

At finite bias  $eV$  applied to the source and drain, electron tunneling becomes a non-equilibrium process, but in weak enough fields, one may neglect non-equilibrium repopulation of dot states and consider the problem within a quasi equilibrium approximation, where the two chemical potentials  $\mu_s$  and  $\mu_d$  are introduced for the source and drain reservoirs which otherwise remain in an equilibrium state, so that  $eV = \mu_s - \mu_d$ . The energy acquired by an electron in the source electrode and then accelerated by the potential  $eV$  may compensate the exchange energy gap  $\lambda < 0$  and thus activate a Kondo tunneling processes which is otherwise quenched in the singlet ground state of the CQD at  $kT \ll |\lambda|$ . As a result a finite bias anomaly at  $eV = \lambda$  shows up in the tunneling conductance instead of zero bias anomaly [30]. Such behavior was observed experimentally in tunneling through single-wall carbon nanotubes [44].

The exchange energy gap may also be compensated by the Zeeman splitting energy under an external magnetic field. Indeed, in a “resonance” field defined as  $g\mu_B B_r - |\lambda| \sim E_K$  transitions between the singlet and “up” projection of the triplet are involved in the dynamical symmetry, so that one may introduce operators  $P^+ = \sqrt{2}|T_\uparrow\rangle\langle S|$ ,  $P_z = |T_\uparrow\rangle\langle T_\uparrow| - |S\rangle\langle S|$  in analogy with (4). In these resonance conditions, the effective co-tunneling Hamiltonian has the form  $H_{\text{cot}} = J\mathbf{P} \cdot \mathbf{s}$ . It describes Kondo tunneling due to triplet-singlet transitions induced by an external magnetic field [48]. This type of Kondo effect was also observed in single-wall carbon nano-tubes [43].

In cases where the tunneling channels and/or the CQD configuration form closed loops [Figs. 1b, 2b, d], the “which pass” situation arises in the trajectories leading from the source to the drain. In the absence of a magnetic field, the currents through the tunneling channels simply add, but in the presence of a perpendicular magnetic field  $B^\perp$  a more complicated superposition pattern arises due to the Aharonov–Bohm effect. Besides, the field  $B^\perp$  introduces charge chirality into the scheme of classification of electron states of the CQD [9,34]. This type of chirality arises because an electron acquires a  $U(1)$  gauge phase  $\phi = \Phi/3$  at each tunneling hopping event. In an anticlockwise direction between the vertices of the equilateral triangle the hopping amplitude  $W$  is then modified as,

$$W \rightarrow W \exp(i\phi), \quad (14)$$

where  $\Phi$  is the magnetic flux through the TQD. As a result, the electron spectrum displays a rich pattern of degenerate levels. Instead of the levels  $E_{a,b}$  there is now a magnetic field dependent spectrum  $E_K(\phi)$ . For example, in a charge sector  $N = 1$ ,

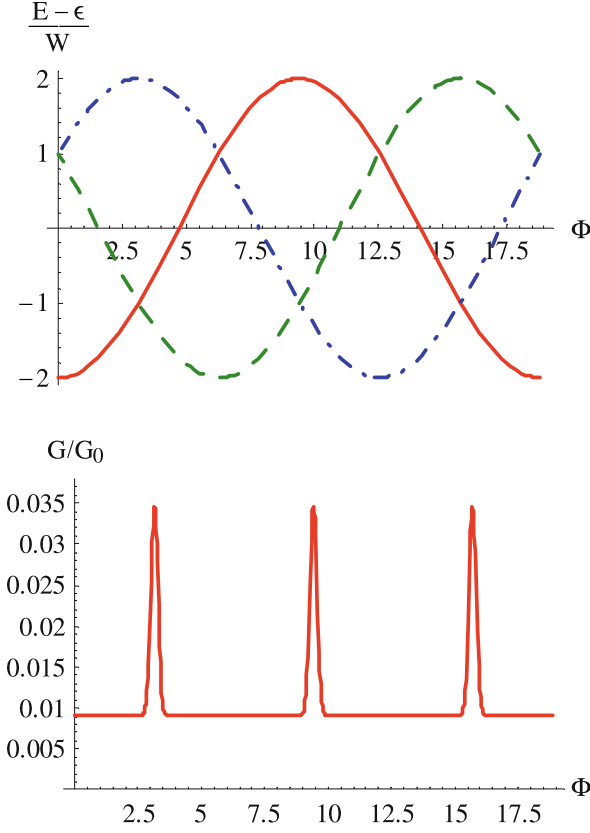
$$E_K(\phi) = \varepsilon - 2W \cos(K - \phi), \quad K = 0, \pm 1. \quad (15)$$

As a result, accidental orbital degeneracy arises at all  $\Phi_r = (2n + 1)/2$  (half-integer magnetic flux quanta) as can be seen in Fig. 5 (upper panel). This degeneracy transforms the spin symmetry  $SU(2)$  of the TQD into a spin+orbital symmetry  $SU(4)$ . In accordance with the general law (11), the Kondo temperature, and hence the tunnel conductance  $G \sim G_0 \ln^{-2}(T/T_K)$  at  $T \gg T_K$  have peaks around  $\Phi_r$  (see Fig. 5, lower panel).

A similar level crossing scenario takes place at  $N = 2$ , but in this case one of the orbital degeneracies is accompanied by the singlet/triplet degeneracy [9]. At  $N = 3$ , the amplitude of energy oscillations in the magnetic field,  $E_{N=3}(\Phi)$  are weak because all the orbital states are occupied and oscillations of different states nearly cancel each other. However, the level crossing between the states with  $S = 3/2$  and  $S = 1/2$  split due to the Zeeman effect [9,17] results in magnetic-field-induced Kondo tunneling similar to that discussed in [43].

Aharonov–Bohm interference arises, when an electron injected from the source electrode passes through two channels in the presence of a magnetic flux  $\Phi$  and the two trajectories interfere in the drain electrode. The interplay between the Kondo and the Aharonov–Bohm effects may be directly seen in the tunneling conductance. Such a regime may be realized both in DQD [21] (Fig. 1b without  $l$ - $r$  tunneling channel) and TQD [34] (Fig. 2d). In the former case the states in the DQD possessing a discrete  $l$ - $r$





**Tunneling Through Quantum Dots with Discrete Symmetries, Figure 5**

**Upper panel:** Evolution of the energy levels  $E_0$  (solid line) and  $E_{\pm}$  (dashed and dot-dashed lines, respectively). **Lower panel:** The corresponding evolution of the tunnel conductance (here  $G_0 = \pi e^2/\hbar$  is a quantum of conductance)

reflection symmetry are classified as even (e) and odd (o) relative to this reflection plane, so that the zero-temperature value (unitarity limit of the Kondo effect) of the tunnel conductance is determined by the difference between transparencies of the two channels,

$$G(T=0)/G_0 = \sin^2 \left\{ \frac{\pi}{2} [\langle n_e \rangle - \langle n_o \rangle] \right\}, \quad (16)$$

where  $\langle n_{e,o} \rangle$  are the ground state occupation numbers in the even and odd channels, respectively. Since different  $U(1)$  gauge phases (14) are acquired by the electron hopping parameters  $W$  in the two channels, interference of these channels results in a magnetic field dependence conductance  $G_{\phi}(T=0)$ . Besides, the Kondo energy  $E_K$  itself becomes a function of the magnetic flux, because the effective exchange parameter depends on the magnetic field ( $J \sim W^2$ ) in  $H_{\text{cot}}$  (7). The resulting Aharonov–Bohm dependence is  $J \sim \cos \Phi$ , and this destructive in-

terference suppresses Kondo tunneling at magnetic fields corresponding to a half-integer magnetic flux quantum. In the TQD shown in Fig. 2d, the ground state is an orbital singlet, and the most striking manifestation of an Aharonov–Bohm destructive interference pattern is due to the dependence  $J(\Phi)$ , which has a slightly more complicated form than that in DQD. In any case, electron tunneling through the “Kondo–Aharonov–Bohm interferometer” is completely suppressed at some values of the magnetic flux  $\Phi$ .

Another mechanism exhibiting the influence of external fields on tunneling properties of CQD is the so-called Thomas–Rashba precession of the magnetic quantization axis initiated by an external electric field that confines the electrons in two dimensions [7,50]. In a system with relativistic spin-orbit interaction, the generic form of the Thomas–Rashba interaction is

$$H_{\text{TR}} = \alpha \mathbf{n} \cdot [\mathbf{S} \times \mathbf{p}] \quad (17)$$

where  $\alpha$  is a spin-orbit coupling constant,  $\mathbf{n}$  is a unit vector characterizing the direction of confining electric field and  $\mathbf{p}$  is the electron momentum operator. If this relativistic interaction exists in the leads (reservoirs), it results in the appearance of a  $\mathbf{p}$ -dependent spin quantization axis and splitting of the electron dispersion law  $\varepsilon_s(\mathbf{p})$  into two branches characterized by “spirality” quantum number  $s$ . This may cause an indirect Ruderman–Kittel–Katsuya–Yoshida (RKKY) interaction exchange between two electrons in DQD with  $N=2$  due to lead-dot tunneling which results in a complicated spin-coupling Hamiltonian of the form [20],

$$H_{\text{RKKY}} = F_{12} \mathbf{S}_1 \cdot \mathbf{S}_2(\theta_{12}), \quad (18)$$

where  $F_{12}$  is the RKKY coupling constant and  $\theta_{12}$  is the angle between two twisted quantization axes. The origin of this twisting is the Thomas–Rashba precession in the leads. The mismatch between two quantization axes may be represented as an effective Dzyaloshinskii–Moriya interaction [cf. (13)] plus an additional Ising-type term

$$\begin{aligned} \mathbf{S}_1 \cdot \mathbf{S}_2(\theta_{12}) &= \mathbf{S}_1 \cdot \mathbf{S}_2 + \sin \theta_{12} [\mathbf{S}_1 \times \mathbf{S}_2] \\ &+ (1 - \cos \theta_{12}) S_1^y S_2^y. \end{aligned} \quad (19)$$

If a perpendicular magnetic field  $B^{\perp}$  is then switched on, two types of quantum interference arise in the Kondo tunneling through DQD due to Aharonov–Bohm charge gauge phase  $\phi$  induced by the magnetic field and Aharonov–Casher effect due to spin gauge phase  $\chi$  induced by the electric field. Both phases control the conductance  $G(\phi, \chi)$  through the DQD in this regime [1].

### Connections with Other Nano-objects and Future Prospects

The physics of tunnel spectroscopy of CQD was developed in parallel with that of tunnel spectroscopy of single molecular devices. Generic similarities between strong correlation effects in bulk materials, molecular complexes and quantum dots were noticed by many authors (see, e. g., [13,26,40,47]). Experimentally, tunnel devices may be prepared in many ways: tunnel spectra of molecules adsorbed on metallic layers may be measured by means of scanning tunneling microscopy. In this case, the nano-tip of the microscope and metallic substrate play the role of source and drain electrodes. Otherwise, the molecule may be suspended between two metallic electrodes by electrochemical methods. All characteristic features of tunneling through quantum dots (Coulomb blockade, Kondo effect, Fano effect etc) are observed in molecular tunneling spectroscopy as well.

Single-wall carbon nanotubes (SWCN) may serve as part of a “bridge” between artificial and natural molecular complexes. These long cylindrical macromolecules usually behave as quasi one-dimensional metallic or semiconductor wires, but if one confines the electrons in a segment of SWCN by imposing a pair of electrodes on it, the electron wave functions within this segment are quantized in all directions and the system as a whole behaves like a quantum dot with characteristic single-electron tunneling behavior due to the Coulomb blockade [54] and a Kondo-like zero bias anomaly due to spin screening in the metallic electrodes [43]. Quantum dots formed in SWCN possess the discrete symmetry inherited from two sub-lattice crystal structures of prototype graphene sheets. The spatially quantized states in these QDs are doubly degenerate [2] and this degeneracy manifests itself as a  $SU(2) \rightarrow SU(4)$  crossover in Kondo tunneling similar to that in DQD [38]. A pair of nanotubes may form a two-channel device in the form of a ring, where both the Fano effect and Aharonov–Bohm effect are involved in electron tunneling [3].

Other representatives of the carbon family, like various modifications of the fullerene molecules also demonstrate those features characteristic for nano-size quantum dots. Coulomb blockade features [45] and Kondo anomalies [59] were observed in conventional  $C_{60}$  molecules deposited onto pair of connected gold electrodes. There are some proposals to form double quantum dots from  $C_{140}$  molecule having the form of a dumb-bell [46] or pair of endofullerene molecules, like  $GdC_{82}$  [5], where a behavior similar to that of DQD is expected. It should be noticed, however that in real molecules (unlike artificial

ones), phonon emission/absorption processes assist single electron tunneling.

Molecular trimers, whose behavior should remind that of TQD may be absorbed on metallic surfaces. The first example of such system is chromium trimers on gold surfaces [22]. These trimers may be both linear and triangular, so from the point of view of Kondo tunneling their behavior should be similar to that of magnetic TQD with  $N = 3$ .

Another family of real molecules possessing discrete point symmetries was studied both experimentally and theoretically within the same context. These are molecular complexes containing transition and rare-earth metal ions secluded in cages formed by CH, CN and other organic radicals. These molecular cages are in direct contact with the electrodes, whereas the magnetic ions form spin multiplets. Endofullerenes and lanthanocene families are the simplest examples of such complexes. It was noticed [26,27] that their behavior in quantum tunneling should be similar to that of DQD in T-shape geometry (Fig. 1c) There are also molecular complexes containing several cages, which form  $2 \times 2$ ,  $3 \times 3$  and more complicated grids, which also may be incorporated into tunnel spectroscopy [57]. In these complexes the complicated structure of the spin multiplet is influenced by magnetic anisotropy due to the discrete symmetry of a grid.

Finally, one should mention the family of single-molecule magnets, namely, large molecular complexes containing about ten transition-metal atoms ( $Fe_8$ ,  $Mn_{12}$ , etc). In these molecules, the magnetic moments of individual spins are added, forming a nano-object with high spin and strong magnetic anisotropy. The discrete molecular symmetry predetermines the complicated quantum dynamics of this high spin object. Complicated spin selection rules for the Kondo effect in some cases suppress two-electron co-tunneling, and reversal  $M_S \rightarrow -M_S$  of spin projection appears only as a fourth order process [58]. Kondo tunneling through magnetic molecules in an external magnetic field displays very complicated structure. Moreover, spin reversal is also characterized by a Berry phase due to multiple excited level crossings [14]. Tunnel transport measurements can directly probe the Berry phase of individual single-molecule magnets.

Future prospects of quantum tunneling through nano-objects with discrete symmetries are closely connected with modern advance in quantum electronics and the quest for manipulating spin systems (spintronics). The most important challenge for theory, experiment and material science engineering concerns the fabrication and study of macroscopic networks formed from quantum dots and molecular complexes. In these composite struc-

tures, the dynamical and tunneling properties of each individual nano-object will serve as building blocks for complex systems, in which the information may be stored and read out on the quantum level in a controllable way.

## Bibliography

### Primary Literature

- Aono T (2007) Two-impurity Kondo problem under Aharonov–Bohm and Aharonov–Casher effects. *Phys Rev B* 76:073304
- Babic B, Kontos T, Schonenberger C (2004) Kondo effect in carbon nanotubes at half filling. *Phys Rev B* 70:235419
- Babic B, Schonenberger C (2004) Observation of Fano resonances in single-wall carbon nanotubes. *Phys Rev B* 70:195408
- Bawendi MG et al (1990) Electronic structure and photoexcited-carrier dynamics in nanometer-size CdSe clusters. *Phys Rev Lett* 65:1623–1626
- Bergeret FS et al (2006) Interplay between Josephson effect and magnetic interactions in double quantum dots. *Phys Rev B* 74:132505
- Bimberg D, Grundman M, Ledentsov NN (1999) *Quantum Dot Heterostructures*. Wiley, New York
- Bychkov YA, Rashba EI (1984) Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. *J Phys C: Solid State Phys* 17:6039–6045
- Craig NJ et al (2004) Tunable nonlocal spin control in coupled-quantum-dot system. *Science* 304:565–567
- Delgado F et al (2007) Theory of spin, electronic and transport properties of the lateral triplet quantum dot molecule in a magnetic field. *Phys Rev B* 76:115332
- Ekimov AI, Onushchenko AA (1984) Size quantization of the electron energy spectrum in a microscopic semiconductor crystal. *JETP Lett* 40:1136–1139
- Eto M (2001) Mean-field theory of the Kondo effect in quantum dots with an even number of electrons. *Phys Rev B* 64:085322
- Eto M (2005) Enhancement of Kondo effect in Multilevel quantum dots. *J Phys Soc Jpn* 74:95–102
- Fulde P (1995) *Electron Correlations in Molecules and Solids*, Chap 12, 3rd edn. Springer, Heidelberg
- Garg A (1993) Topologically quenched tunnel splitting in spin systems without Kramers degeneracy. *Europhys Lett* 22: 205–210
- Glazman LI, Raikh ME (1988) Resonant Kondo transparency of a barrier with quasilocal impurity states. *JETP Lett* 47: 452–455
- Guadreau L et al (2006) Stability diagram of a few-electron triple dot. *Phys Rev Lett* 97:036807
- Nojiri H, Ishikawa E, Yamase T (2005) Effect of spin chirality on magnetization in triangular rings. *Progr Theor Phys Suppl* No 159:292–296
- Hofmann F et al (1995) Single electron switching in a parallel quantum dot. *Phys Rev B* 51:13872–13875
- Hofstatter W, Schoeller H (2002) Quantum phase transitions in a multilevel dot. *Phys Rev Lett* 88:016803
- Imamura H, Bruno P, Utsumi Y (2004) Twisted exchange interaction between localized spins embedded in a one- and two-dimensional electron gas with Rashba spin-orbit coupling. *Phys Rev B* 69:121303
- Izumida W, Sakai O, Shimizu Y (1997) Many body effects on electron tunneling through quantum dots in an Aharonov–Bohm circuit. *J Phys Soc Jpn* 66:717–726
- Jamneala T, Madhavan V, Crommie MF (2001) Kondo response of a single antiferromagnetic chromium trimer. *Phys Rev Lett* 87:256804
- Joault B, Santoro G, Tagliacozzo A (2000) Sequential magnetotunneling in a vertical quantum dot tuned at the crossing to higher spin states. *Phys Rev B* 61:10242–10246
- Kang K et al (2001) Anti-Kondo resonance in transport through a quantum wire with a side coupled quantum dot. *Phys Rev B* 63:113304
- Kastner MA (1993) Artificial atoms. *Phys Today* 46:24–31
- Kikoin K, Avishai Y (2001) Kondo tunneling through real and artificial molecules. *Phys Rev Lett* 86:2090–2093
- Kikoin K, Avishai Y (2002) Kondo singlet versus Zhang-Rice and Heitler–London singlets. *Physica B* 312–313:165–166
- Kikoin K, Avishai Y, Kiselev MN (2006) Dynamical symmetries in nanophysics. In: *Nanophysics, Nanoclusters and Nanodevices*. Nova Publishers, New York, pp 39–86
- Kim TS, Hershfield S (2001) Suppression of current in transport through parallel double quantum dots. *Phys Rev B* 63:245326
- Kiselev MN, Kikoin K, Molenkamp LW (2003) Resonance Kondo tunneling through a double quantum dot at finite bias. *Phys Rev B* 68:155323
- Kondo J (1964) Resistance minimum in dilute magnetic alloys. *Progr Theor Phys* 32:37–49
- Kowenhoven LP, Austing DG, Tarucha S (2001) Few-electron quantum dots. *Rep Progr Phys* 64:701–736
- Kuzmenko T, Kikoin K, Avishai Y (2004) Kondo effect in systems with dynamical symmetries. *Phys Rev B* 69:195109
- Kuzmenko T, Kikoin K, Avishai Y (2006) Magnetically tunable Kondo–Aharonov–Bohm effect in triangular quantum dot. *Phys Rev Lett* 96:046601
- Kuzmenko T, Kikoin K, Avishai Y (2006) Tunneling through triple quantum dots with mirror symmetry. *Phys Rev B* 73:235310
- Leonhard D et al (1993) Direct formation of quantum-size dots from uniform coherent islands of InGaAs on GaAs surfaces. *Appl Phys Lett* 63:3203–3205
- Loss D, DiVincenzo DP (1998) Quantum computation with quantum dots. *Phys Rev A* 57:120–126
- Makarovski A et al (2007)  $SU(4)$  and  $SU(2)$  Kondo effects in carbon nanotube quantum dots. *Phys Rev B* 75:241407
- Matveev KA, Larkin AI (1992) Interaction-induced threshold singularities in tunneling via localized levels. *Phys Rev B* 46:15337–15347
- Mitra A, Aleiner I, Millis AJ (2004) Phonon effects in molecular transistors: quantum and classical treatment. *Phys Rev B* 69:245302
- Molenkamp LW et al (1995) Scaling of the Coulomb energy due to quantum fluctuations in the charge on a quantum dot. *Phys Rev Lett* 75:4282–4285
- Ng TK, Lee PA (1988) On-site Coulomb repulsion and resonant tunneling. *Phys Rev Lett* 61:1768–1771
- Nugard J et al (2000) Kondo physics in carbon nanotubes. *Nature* 408:342–345
- Paaske J et al (2006) Nonequilibrium singlet-triplet Kondo effect in carbon nanotubes. *Nat Phys* 2:460–464
- Park H et al (2000) Nanomechanical oscillations in a single- $C_{60}$  transistor. *Nature* 407:57–60

46. Pasupathy AN et al (2005) Vibration assisted electron tunneling in  $C_{140}$  single electron transistors. *Nano Lett* 5:203–207
47. Plihal M, Gadzuk JW (2001) Nonequilibrium theory of scanning tunneling spectroscopy via adsorbate resonances. *Phys Rev B* 63:085404
48. Pustilnik M, Avishai Y, Kikoin (2000) Quantum dots with even number of electrons: Kondo effect in a finite magnetic field. *Phys Rev Lett* 84:1756–1759
49. Pustilnik M, Glazman LI (2001) Kondo effect induced by magnetic field. *Phys Rev B* 64:45328
50. Rashba EI (2006) Modern trends in semiconductor Spintronics. In: Ivanov AL, Tikhodeev SG (eds) *Problems of Condensed Matter Physics*. Clarendon Press, Oxford, pp 188–198
51. Read MA (1993) Quantum dots. *Sci Amer* 268:118–123
52. Saraga DS, Loss D (2003) Spin-entangled currents created by a triple quantum dot. *Phys Rev Lett* 90:166803
53. Scarola VW, Das Sarma S (2005) Exchange gate in solid-state spin quantum computation: the applicability of the Heisenberg model. *Phys Rev A* 71:032340
54. Tans S et al (1998) Electron-electron correlations in carbon nanotubes. *Nature* 394:761–764
55. Teichert C (2002) Self-organization of nanostructures in semiconductor heteroepitaxy. *Phys Repts* 365:335–432
56. Vidan A et al (2004) Triple quantum dot charging rectifier. *Appl Phys Lett* 85:3602–3604
57. Waldmann O, Zhao L, Thompson LK (2002) Field-dependent anisotropy change in a supramolecular  $Mn(II)$ -[3 × 3] grid. *Phys Rev Lett* 88:066401
58. Wegewijs M et al (2007) Magnetotransport through single-molecule magnets: Kondo peaks, zero-bias dips, molecular symmetry and Berry's phase. *New J Phys* 9:344
59. Yu LH, Natelson D (2004) The Kondo effect in  $C_{60}$  single-molecule transistors. *Nano Lett* 4:79–83

## Books and Reviews

- Cuniberti G, Fagas G, Richter K (eds) (2005) *Molecular electronics. Lecture Notes in Physics*, vol 680. Springer, New York
- Englefield MJ (1972) *Group Theory and the Coulomb Problem*. Wiley, New York
- Kouwenhoven LP, Glazman LI (2001) Revival of the Kondo effect. *Phys World* 14:33–38
- Natelson D (2006) *Handbook of Organic Electronics and Photonics*. American Scientific Publishers, Valencia
- Sachrajda PH, Ciorga M (2003) *Nano-spintronics with lateral quantum dots*. Kluwer, Boston

## Two-Sided Matching Models

MARILDA SOTOMAYOR<sup>1,2</sup>, ÖMER ÖZAK<sup>2</sup>

<sup>1</sup> Department of Economics, University of São Paulo/SP, São Paulo, Brazil

<sup>2</sup> Department of Economics, Brown University, Providence, USA

## Article Outline

### Glossary

## Definition of the Subject

### Introduction

### Discrete Two-Sided Matching Models

### Continuous Two-Sided Matching Model

### with Additively Separable Utility Functions

### Hybrid One-to-One Matching Model

### Incentives

### Future Directions

### Bibliography

## Glossary

**Two-sided matching model** is a game theoretical model whose elements are (i) two disjoint and finite sets of agents:  $F$  with  $m$  elements, and  $W$  with  $n$  elements, referred to as the sides of the matching model; (ii) the structure of agents' preferences and (iii) the agents' quotas.

The rules of the game determine the feasible outcomes. The main activity of the agents from one set is to form partnerships with the agents on the other set. Players derive their payoffs from the set of partnerships they form. The agents belonging to  $F$  and  $W$  are called  $F$ -agents and  $W$ -agents, respectively.

**Quota of an agent** in a two-sided matching model is the maximum number of partnerships an agent is allowed to form. When every participant can form one partnership at most the matching model is called one-to-one. If only the players of one of the sides can form more than one partnership the matching model is said to be many-to-one. Otherwise the matching model is many-to-many.

**Allowable set of partners** for  $f \in F$  with quota  $r(f)$  is a family of elements of  $F \cup W$  with  $k$  distinct  $W$ -agents,  $0 \leq k \leq r(f)$ , and  $r(f) - k$  repetitions of  $f$ .

**Discrete two-sided matching model** In the discrete two-sided matching models agents have preferences over allowable sets of partners. The allowable sets of partners for  $f$  of the type  $\{w, f, \dots, f\}$  are identified with the individual agent  $w \in W$  and the allowable set of partners  $\{f \dots f\}$  is identified with  $f$ . Under this identification, agent  $w$  is *acceptable* to agent  $f$  if and only if  $f$  likes  $w$  as well as him/her/it self. Similar definitions and identifications apply to an agent  $w \in W$ . These preferences are transitive and complete, so they can be represented by ordered lists of preferences. The model can then be described by  $(F, W, P, r, s)$ , where  $P$  is the profile of preferences and  $r$  and  $s$  are the arrays of quotas for the  $F$ -agents and  $W$ -agents, respectively.

**Continuous two-sided matching model** In this model the structure of preferences is given by utility functions

which are continuous in some money variable which varies continuously in the set of real numbers. A particular case is obtained when agents place a monetary value on each possible partner or on each possible set of partners.

**Hybrid two-sided matching model** is a unification of the discrete and the continuous models. It is obtained by allowing the agents of both markets to trade with each other in the same market.

**Matching  $\mu$**  in a two-sided matching model with sides  $F$  and  $W$  is a function that maps every agent into an allowable set of partners for him/her/it, such that  $f$  is in  $\mu(w)$  if and only if  $w$  is in  $\mu(f)$ , for every  $(f, w) \in F \times W$ . If we relax this condition the function is called a *pre-matching*.

A matching describes the set of partnerships of the type  $(f, w)$ ,  $(f, f)$  or  $(w, w)$ , with  $f \in F$  and  $w \in W$ , formed by the agents. We say that a player that does not enter any partnership is unmatched. Agents compare two matchings by comparing the two allowable sets of partners they obtain.

**Feasible assignment** for a two-sided matching model with sides  $F$  and  $W$  is an  $m \times n$  matrix  $x = (x_{fw})$  whose entries are zeros or ones such that  $\sum_f x_{fw} \leq s(w)$  for all  $w \in W$  and  $\sum_w x_{fw} \leq r(f)$  for all  $f \in F$ . We say that  $x_{fw} = 1$  if  $f$  and  $w$  form a partnership and  $x_{fw} = 0$  otherwise. A feasible assignment  $x$  corresponds to a matching  $\mu$  which matches  $f$  to  $w$  if and only if  $x_{fw} = 1$ . Thus, if  $\sum_f x_{fw} = 0$  then  $w$  is unassigned at  $x$  or, equivalently, unmatched at  $\mu$ , and if  $\sum_w x_{fw} = 0$ , then  $f$  is likewise unassigned at  $x$  or, equivalently, unmatched at  $\mu$ .

**Responsive preference** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  has a responsive preference relation over allowable sets of partners if whenever (i)  $A$  and  $B$  are two allowable sets of partners for player  $f$ ; (ii)  $j$  and  $k$  are two elements of  $W \cup \{f\}$  and (iii)  $A = B \cup \{w\} \setminus \{w'\}$  with  $w \notin B$  and  $w' \in B$ , then  $f$  prefers  $A$  to  $B$  if and only if  $f$  prefers  $w$  to  $w'$ . Similarly we define the responsive preference for  $w \in W$ .

**$r(f)$ -Separable preference** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  with a quota of  $r(f)$  has a  $r(f)$ -separable preference relation over allowable sets of partners if whenever  $A = B \cup \{w\} \setminus \{f\}$  with  $w \notin B$  and  $f \in B$ , then  $f$  prefers  $A$  to  $B$  if and only if  $f$  prefers  $w$  to  $f$ . Similarly we define  $s(w)$ -separable preference for  $w \in W$ .

**Maximin preferences** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  with a quota of  $r(f)$  has a maximin preference relation over allowable

sets of partners if whenever two allowable sets  $C$  and  $C'$  contained in  $W$ , such that  $f$  prefers  $C'$  to  $C$  and no  $w$  in  $C$  is unacceptable to  $f$ , then a) all of  $C'$  are acceptable to  $f$  and b) if  $|C| = r(f)$ , then the least preferred worker in  $C' - C$  is preferred by  $f$  to the least preferred worker in  $C - C'$ . Similarly we define maximin preference for  $w \in W$ .

**Choice set of  $f \in F$  from  $A \subseteq W(Ch_f(A))$**  in a discrete two-sided matching model with sides  $F$  and  $W$ . Let  $B = \{A' | A' \text{ is an allowable set of partners for } f \text{ and } A' \cap W \text{ is contained in } A\}$ . Then,  $A' \in Ch_f(A)$  if and only if  $A' \in B$  and  $f$  likes  $A'$  at least as well as  $A''$ , for all  $A'' \in B$ . Similarly we define  $Ch_w(A)$  for  $w \in W$  and  $A \subseteq F$ .

**Substitutable preferences** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  has a substitutable preference relation over allowable sets of partners if whenever  $A \subseteq W$  and  $B \subseteq W$  are such that  $A \cap B = \emptyset$  then (i) for all  $S' \in Ch_f(A \cup B)$  there is some  $S \in Ch_f(A)$  such that  $S' \cap A \subseteq S$  and (ii) for all  $S \in Ch_f(A)$  there is some  $S' \in Ch_f(A \cup B)$  such that  $S' \cap A \subseteq S$ . If an agent's preference is responsive then it is substitutable.

When preferences are strict, conditions (i) and (ii) are equivalent to requiring that if  $Ch_f(A \cup B) = S'$  then  $S' \cap A \subseteq Ch_f(A)$ . This concept is similarly defined for  $w \in W$ .

**Strongly substitutable preferences** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  has a strongly substitutable preference relation over allowable sets of partners if for every pair of allowable sets of partners  $A$  and  $B$  such that  $A \succ_f B$ , if  $w \in Ch_f(W \cap A \cup \{w\})$ , then  $w \in Ch_f(W \cap B \cup \{w\})$ . This is a stronger condition than substitutability and responsiveness.

**Additively separable preferences** in a discrete two-sided matching model with sides  $F$  and  $W$ . Agent  $f \in F$  has additively separable preferences if he/she/it assigns a nonnegative number  $a_{fw}$  to each  $w \in W$  and assigns the value  $v(A) = \sum_{w \in A} a_{fw}$  to each allowable set  $A$  of partners for  $f$ . Agent  $f$  compares two allowable sets by comparing the values of these sets. This concept is similarly defined for  $w \in W$ . If the agents have additively separable preferences we can think that if a partnership  $(f, w) \in F \times W$  is formed then the partners participate in some joint activity that generates a payoff  $a_{fw}$  for player  $f$  and  $b_{fw}$  for player  $w$ . These numbers are fixed, i. e., they are not negotiable. If the preferences of the agents are additively separable then they are responsive. The converse is not true (see Kraft, Pratt and Seidenberg [47]).



**T-map** in a discrete two-sided matching model with sides  $F$  and  $W$  is defined as follows. For every pre-matching  $\mu$ , let  $T(\mu(f)) = Ch_f(U(f, \mu))$  for all  $f \in F$ , where  $U(f, \mu) = \{w \in W \mid f \in Ch_w(\mu(w) \cup \{f\})\}$ . Similarly,  $T(\mu(w)) = Ch_w(U(w, \mu))$  for all  $w \in W$ , where  $U(w, \mu) = \{f \in F \mid w \in Ch_f(\mu(f) \cup \{w\})\}$ .

**Lattice property** A set  $L$  endowed with a partial order relation  $\geq$  has the lattice property if  $\sup\{x, y\} \equiv x \vee y$  and  $\inf\{x, y\} \equiv x \wedge y$  are in  $L$ , for all  $x, y \in L$ . The lattice is complete if all its subsets have a supremum and an infimum. (See Birkhoff [11]).

**Pareto-optimal matching** A feasible matching  $\mu$  is Pareto-optimal if there is no feasible matching which is weakly preferred to  $\mu$  by all players and it is strictly preferred by at least one of them.

**Outcome** For the discrete two-sided matching models the outcome is a matching or at least corresponds to a matching; for the continuous two-sided matching models the outcome specifies a payoff for each agent and a matching.

**Stable outcome** It is the natural solution concept for a two-sided matching model. It is also referred as set-wise-stable outcome. See the definition below.

**F-optimal stable matching (respectively, payoff)** for a discrete (respectively, continuous) two-sided matching model is the stable matching (respectively, payoff) which is weakly preferred by every agent in  $F$ . Similarly we define the  $W$ -optimal stable matching (respectively, payoff).

**Achievable mate** for agent  $y$  in a discrete two-sided matching model is any  $y$ 's of partner under some stable matching.

**Matching mechanism** For the discrete two-sided matching models, a matching mechanism is a function  $h$  whose range is the set of all possible inputs  $X = (F, W, P, r, s)$ , and whose output  $h(X)$  is a matching for  $X$ .

**Stable matching mechanism** It is a matching mechanism  $h$  such that  $h(X)$  is always stable for the market  $X$ . If  $h(X)$  always produces the  $F$ -optimal stable matching for  $X$  then it is called an  $F$ -optimal stable matching mechanism, and so on.

**Revelation mechanism** Given the discrete two-sided matching model  $(F, W, P, r, s)$ , a revelation mechanism is the restriction of a matching mechanism  $h$  to the set of discrete two-sided matching markets  $(F, W, Q, r, s)$  where the sets of agents and quotas are fixed.

**Revelation game** It is the strategic game induced by a revelation mechanism for the discrete two-sided matching  $(F, W, P, r, s)$ : the set of players is given by the

union of  $F$  and  $W$ ; a strategy of player  $j$  is any possible list of preferences  $Q(j)$  that player  $j$  can state; the outcome function is given by the mechanism  $h$  and the preferences of the players over the set of outcomes are determined by  $P$ .

**Sincere strategy** for a player  $j$  in a revelation game is the true list of preferences  $P(j)$ .

**Manipulable mechanism** A mechanism  $h$  is *manipulable* or it is *not strategy-proof* if in some revelation game induced by  $h$ , stating the true preferences is not a dominant strategy for at least one player. A mechanism  $h$  is *collectively manipulable* if in some revelation game induced by  $h$ , there is a coalition whose members can be better off by misrepresenting its preferences.

**Rematching proof equilibrium** is a Nash equilibrium profile from which no pair of players  $(f, w) \in F \times W$  can profitably deviate given that the other players do not change their strategies.

**Truncation** Let  $P(a)$  be the  $a$ 's preference list over individuals for a discrete two-sided matching model. For  $a \in F \cup W$ , the list of preferences  $Q(a)$  over individuals is a truncation of  $P(a)$  at some agent  $b$  if  $Q(a)$  keeps the same ordering as  $P(a)$  but ranks as unacceptable all agents which are ranked below  $b$ .

## Definition of the Subject

This article describes the basic elements of the cooperative and non-cooperative approaches for two-sided matching models and analyzes the fundamental differences and similarities between some of these models.

## Basic Definitions

**Feasible outcome** is an outcome that is specified by the rules of the game. In the discrete case, a feasible outcome is a feasible matching  $\mu$  or at least corresponds to a feasible matching. The usual definition is the following. *The matching  $\mu$  is feasible if it matches every agent to an allowable set of partners and  $\mu(f) \in Ch_f(\mu(f) \cap W)$  and  $\mu(w) \in Ch_w(\mu(w) \cap F)$  for every  $(f, w) \in F \times W$ .* Then, if preferences are responsive, every matched pair is mutually acceptable. An implication of this definition is that a feasible outcome is always individually rational.

In the continuous case, the rules of the game may specify, for example, whether the agents negotiate their payoffs individually within each partnership or if they negotiate in blocks. In the former case a feasible outcome specifies an array of individual payoffs for each player, indexed according to the partnerships formed under the corresponding matching. In the latter case

the feasible outcome only specifies a total payoff for each agent.

**Corewise-stability** is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *dominated* by any feasible outcome via a coalition.

The feasible outcome  $x$  *dominates* the feasible outcome  $y$  via the coalition  $S \neq \emptyset$ , if (i) every player in  $S$  prefers  $x$  to  $y$  and (ii) if  $j \in S$  then all of  $j$ 's partners under  $x$  belong to  $S$ . Coalition  $S$  is said to *block* the outcome  $y$ .

It is a natural solution concept for the one-to-one matching models and for the continuous many-to-one matching models.

**Strong-corewise-stability** is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *weakly dominated* by any feasible outcome via a coalition.

The feasible outcome  $x$  *weakly dominates* the feasible outcome  $y$  via the coalition  $S \neq \emptyset$ , if (i) all players in  $S$  weakly prefer  $x$  to  $y$  and at least one of them strictly prefers  $x$  to  $y$ ; (ii) if  $j \in S$  then all of  $j$ 's partners under  $x$  belong to  $S$ . Coalition  $S$  is said to *weakly block* the outcome  $y$ .

Strong-corewise-stability is a natural solution concept for the discrete many-to-one matching models.

**Pairwise-stability** is a solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *quasi-dominated* by any feasible outcome via a pair of agents  $(f, w) \in F \times W$ .

For the discrete case an outcome can be identified with a matching. We say that the feasible matching  $x$  *quasi-dominates* the feasible matching  $y$  via the coalition  $S \neq \emptyset$ , if (i) every player in  $S$  prefers the matching  $x$  to the matching  $y$  and (ii) if  $j \in S$  and  $k \in x(j)$  then  $k \in S \cup y(j)$ .

For the continuous case in which agents negotiate their payoffs individually with each partner, the feasible outcome  $x$  *quasi-dominates* the feasible outcome  $y$  via the coalition  $S \neq \emptyset$ , if (i) every player in  $S$  gets a higher total payoff under  $x$  than under  $y$  and (ii) if  $j \in S$  and  $k \in x(j)$  then  $k \in S$  or  $k \in y(j)$ , in which case,  $k$  keeps the same individual payment obtained with  $j$  under  $y$ . For the continuous case in which agents cannot negotiate their individual payments, the definition of *quasi-dominance* is equivalent to that of dominance. In any case coalition  $S$  is said to *destabilize* the outcome  $y$ . Thus, for the discrete models, for example, a matching  $\mu$  is pairwise-stable if it is feasible and it is not destabilized by any pair  $(f, w) \in F \times W$ . The pair of agents  $(f, w)$  destabilizes the matching  $\mu$  if these

agents prefer each other to some of their current mates. Pairwise-stability is a natural solution concept for the marriage and the college admission models as well as for continuous matching models in which the agents negotiate their individual payoffs.

**Setwise-stability** is the solution concept that assigns to each two-sided matching model the set of feasible outcomes which are not *quasi-dominated* by any feasible outcome via a coalition.

It is a generalization of the group stability concept defined by Roth [65] for the college admissions model. An attempt to extend the concept of group stability to a discrete many-to-many matching market with substitutable and strict preferences was not successful in Roth [62]. The concept introduced in that paper is equivalent to pairwise stability. It turns out that pairwise-stable matchings may be blocked by coalitions of size greater than two in this model. In fact, an example of a pairwise-stable matching that is corewise-unstable is presented in Blair [12] and another one in Sotomayor [80]. Thus, the pairwise-stability concept cannot be regarded as the natural cooperative solution concept for this model.

Setwise-stability regarded as a new cooperative equilibrium concept, different from the core concept, was obtained for the first time in Sotomayor [75] in a many-to-many matching model with additively separable utilities. In this model the set of setwise-stable outcomes equals the set of pairwise-stable outcomes and may be smaller than the core.

Setwise-stability concept was introduced in Sotomayor [80] as a refinement of the core concept and as the natural cooperative equilibrium concept for a two-sided matching model. In the discrete many-to-many matching models it is stronger than pairwise-stability plus corewise-stability (see Example 3 below). Since the essential coalitions in the marriage and in the college admission models are pairs of players made up of one agent of each side of the market, a setwise-stable matching is a pairwise-stable matching in these models. If the preferences are strict, the pairwise-stable matchings are the corewise-stable matchings in the marriage model and are the strong corewise-stable matchings in the college admissions model.

## A Brief Historical Account

The theory of the two-sided matching markets was born in 1962, year of the publication of the seminal paper by Gale and Shapley, who formulated the stable matching problem for the marriage and the college admissions markets. The

problem of “college admissions” as described in that paper involves a set of colleges and a set of students. Each student lists in order of strict preference those colleges he/she wishes to attend and each college lists in order of strict preference those students it is willing to admit. Furthermore, each college has a quota, representing the maximum number of students it can admit. The problem is then to devise some allocation procedure of students to colleges in a way that takes account of their respective preferences. More specifically, given any set of colleges and students, together with their preferences and quotas, can one find a stable matching?

The answer to this question is affirmative.

For the existence proof Gale and Shapley constructed a simple deferred-acceptance algorithm, which, starting from the given data, leads to a stable matching in a finite number of steps. The matching obtained in this way is the unique stable matching (there may be many) which is preferred by all students to any other such matching. For this reason, it is called optimal stable matching for the students.

In the 1962s paper they remark that: “...even though we no longer have the symmetry of the marriage problem, we can still invert our admissions procedure to obtain the unique “college optimal” assignment. The inverted method bears some resemblance to a fraternity “crush week” it starts with each college making bids to those applicants it considers most desirable, up to its quota limit, and then the bid-for students reject all but the most attractive offer, and so on.”

In the paper mentioned above the authors express some reservation on the possibilities of application of their algorithm. It turns out that since 1951, a mathematically equivalent algorithm was being used by the National Resident Matching Program (NRMP), located in Evanston, Illinois. The NRMP has the task each year of assigning graduates of all medical schools in the United States to hospitals where they are required to serve a year’s residency. The algorithm used by the NRMP was mathematically the equivalent to the one described in Gale and Shapley [30] to produce the optimal stable matching for the colleges. Thus, in this algorithm, each hospital applies to its quota of students. The confirmation of this fact was obtained by David Gale in 1975. In a letter from December 8, 1975, in response to a letter from Gale to the NRMP, Elliott Peranson, consultant to NRMP, responsible for the technical operation of the matching program, says the following:

*“However, I might point out that the NRMP algorithm in fact uses the inverse procedure and produces the unique “college optimal” assignment rather than this “student optimal” assignment. This procedure more closely parallels the*

*actual admissions process where a matching algorithm is not used. In this case students apply to all hospitals they would consider (not just their first choice), each hospital then selects the most desirable students, up to its quota limit, from amongst all applicants, then the “bid-for” students reject all but the most desirable offer, and so on”.*

Hence, the proof that the NRMP was yielding a stable matching, which was the optimal stable matching for the colleges, is that such an outcome is always obtained by the Gale and Shapley’s algorithm with the colleges proposing.

The discovery that the two algorithms were mathematically equivalent was first spread orally and later reported in Gale and Sotomayor [32] with an equivalent description of the NRMP algorithm. (Roth [63] also presents the NRMP algorithm). This was the first application of matching theory of which we have knowledge. The algorithms proposed by Gale and Shapley are described below. Their description is quoted from Gale and Sotomayor [32].

#### **Gale-Shapley-Algorithm with the Colleges Proposing to the Applicants**

Each hospital  $H$  tentatively admits its quota  $q_H$  consisting of the top  $q_H$  applicants on its list. Applicants who are tentatively admitted to more than one hospital *tentatively accept* the one they prefer. Their names are then removed from the lists of all other hospitals which have tentatively admitted them. This gives the *first tentative matching*. Hospitals which now fall short of their quota again admit tentatively until either their quota is again filled or they have exhausted their list. Admitted applicants again reject all but their favorite hospital, giving the *second tentative matching*, etc. The algorithm terminates when, after some tentative matching, no hospitals can admit any more applicants either because their quota is full or they have exhausted their list. The tentative matching then becomes permanent.

#### **Gale-Shapley-Algorithm with the Applicants Proposing to the Colleges**

Each applicant *petitions* for admission to his/her favorite hospital. In general, some hospitals will have more petitioners than allowed by their quota. Such oversubscribed hospitals now reject the lowest petitioners on their preference list so as to come within their quota. This is the *first tentative matching*. Next, rejected applicants petition for admission to their second favorite hospital and again oversubscribed hospitals reject the overflow, etc. The algorithm terminates when every applicant is tentatively admitted or has been rejected by every hospital on his list.

The fact that the matching produced by the NRMP algorithm is stable stands for one of the most important applications of game theory to Economics. During about fifty years the allocation procedures used to assign interns to hospitals in the United States produced unstable matchings. Unsuccessful procedures were often proposed by the Association of American Medical Colleges. This sort of events culminated with a centralized mechanism that employed the NRMP algorithm. Such a centralized mechanism lasted for almost fifty years, suggesting that interns and hospitals had reached an equilibrium. And the paper written by Gale and Shapely corroborated that the game-theoretical predictions were, once more, correct.

Before the publication of Gale and Sotomayor [32], a few, but important, contributions were made to the theory of two-sided matching markets. The famous paper of 1972 by Shapley and Shubik establishes the assignment game via the introduction of money, as a continuous variable, into the marriage model. The book *Marriages Estables* by Knuth was published in 1976. In this volume, the proof, attributed to Conway, that the set of stable matchings for the marriage model is a lattice is presented. The assignment game was generalized by Kelso and Crawford [42] to a model where the utilities satisfy some gross substitute condition. Another generalization, which considers continuous utility functions, non-necessarily linear, was presented in Demange and Gale [18].

Nevertheless, among the contributions of this period one of them caused considerable impact. This was the non-manipulability theorem by Dubins and Freedman [21]. In a stable revelation mechanism, for every profile of preferences that can be selected by the agents, some algorithm that yields a stable matching is used. These authors prove that the revelation mechanism that produces the optimal stable matching for a given side of the marriage market is not collectively manipulable by the agents of that side. Also, this non-manipulability result holds for the college admission market when the mechanism yields the student-optimal stable matching. An analogue to this result was proved in Demange and Gale [18] for the continuous model through a key lemma that became known in the literature as the Blocking Lemma. The main challenge that motivated Gale and Sotomayor [32] was to prove the discrete version of the blocking lemma, which allowed to prove the non-manipulability theorem in just three lines. Two simple and short proofs (one with the use of the algorithm and the other one without the use of the algorithm) of Dubins and Freedman's theorem were presented as an alternative to the original proof by the authors which was about twenty pages long. An example in Dubins and Freedman [21], where some woman can be better off by

falsifying her preference list when the man-optimal stable matching is to be employed, motivated Gale and Sotomayor [31]. This paper proves that such a mechanism can almost always be manipulated by the women and then treats the strategic possibilities for these agents in the corresponding strategic game. Another paper of this period was Roth [60], which proves, via an example, that any rule for selecting a stable matching is manipulable (either by some man or some woman).

The existence theorem of manipulability by the women, the Impossibility Theorem and the non-manipulability theorem attracted the authors' interest toward a fruitful line of investigation concerning the incentives facing the agents when an allocation mechanism is employed. Algorithms have been developed for this purpose for several matching models. The games induced by such mechanisms are played non-cooperatively by the agents, and in general, their self-enforcing agreements lead to a stable outcome. In these cases a non-cooperative implementation of the set of stable outcomes is provided. Analyzing the strategic behavior of the agents in such games has been an important subject of research of several authors in an attempt to get precise answers to the strategic questions raised. In this direction we can cite Roth [60,61], Gale and Sotomayor [31], Perez-Castrillo and Sotomayor [56], Sotomayor [87,88,94], Kamecke [38], Kara and Sönmez [40,41], Alcalde [5], Alcalde, Perez-Castrillo and Romero-Medina [6], among others.

Over all these years the popularity of the matching theory has spread among mathematicians and economists, mainly due to the publication in 1990 of the first edition of the book *Two-sided matchings. A study in game theoretical modeling and analysis*, by Roth and Sotomayor, which attempts a comprehensive survey of the main results on the two-sided matching theory until that date. (An extensive bibliography can also be found in <http://kuznets.fas.harvard.edu/~aroth/bib.html#matchbib>).

The stable matching problem has been generalized to several two-sided matching models, which have been widely modeled and analyzed under both cooperative and non-cooperative game theoretic approaches. Through these models a variety of markets has become better understood, which has considerably contributed to their organization. The deferred acceptance algorithm of Gale and Shapley and adaptations of it have been applied in the reorganization of admission processes of many two-sided matching markets. And the design of these mechanisms has also raised new theoretical questions. (In this connection see e. g. Balinski and Sönmez [8], Ergin and Sönmez [28], and Pathak and Sönmez [55], Abdulkadiroglu and Sönmez [1] and Bardella and Sotomayor [9]).

## Introduction

The two-sided matching theory is the study of game theoretical models in which the set of players is partitioned into the disjoint union of two finite sets, and the main activity of the agents from one set is to form partnerships with the agents on the other set. In addition, a structure of preferences is available for the players, as well as an array of quotas, one quota for each participant, representing the maximum number of partnerships that he/she/it is allowed to form.

The rules of the game determine the feasible outcomes. These models are called two-sided matching models. They are suitable for modeling a great variety of labor markets of firms and workers, markets of buyers and sellers, markets of students and schools, etc. They are grouped into three categories: the discrete, the continuous and the hybrid matching models. The distinctions between them are based on the kind of preference structure they are endowed with. Within each category the models vary as to the possibilities of the agents to form multiple partnerships, the rules of the game and some variation of the structure of the preferences.

Given the sets of agents, the structure of preferences and quotas and the rules of the game, the question that emerges is:

- A. Which partnerships should be formed?
- B. If a partnership is formed, what payoff should be awarded to each agent?

The prediction should be outcomes that cannot be upset by any coalition. The idea is that, in games where players form partnerships, it should be allowed for coalitions to be formed in which its members keep some of their current partners if they wish, replace some others with new partners belonging to the coalition. Furthermore, by doing this they get a preferable outcome. This intuitive idea is captured by a refinement of the core concept introduced in Sotomayor [80], called setwise-stability (stability, for short). For the marriage model and the college admission model with responsive preferences this concept coincides with the one introduced by Gale and Shapley.

This review is an attempt to understand some of the differences and similarities between some matching models. We do that by analyzing both the cooperative and the non-cooperative structure of these models. Some future directions will then be presented.

The rest of the article is organized as follows. Section “Discrete Two-Sided Matching Models” is devoted to the cooperative approach of the discrete matching models with responsive preferences. It introduces the basic

cooperative one-to-one, many-to-one and many-to-many matching models and discusses the main properties that characterize the set of stable outcomes for these models. This kind of analysis is also provided in Sect. “Continuous Two-Sided Matching Model with Additively Separable Utility Functions” for the continuous matching models where the utilities are additively separable. Section “Hybrid One-to-One Matching Model” discusses some fundamental similarities and differences between a one-to-one hybrid model and its corresponding non-hybrid matching models. Strategic questions are treated in Sect. “Incentives”. Section “Future Directions” presents some future directions and open problems.

## Discrete Two-Sided Matching Models

There are two finite and disjoint sets of agents,  $F$  and  $W$ , which we may think of as being sets of men and women, colleges and students, firms and workers, etc. To fix ideas let us describe the model in terms of firms and workers. Then,  $F$  is a set of  $m$  firms and  $W$  is a set of  $n$  workers. Salaries cannot be negotiated and are part of the job description. Each worker  $w$  has a quota  $s(w)$  representing the maximum number of jobs in different firms she can take. Each firm  $f$  has a quota  $r(f)$  representing the maximum number of workers it can hire. Set  $r$  and  $s$  the array of quotas of the firms and the workers, respectively.

Since the agents form partnerships, they always have preferences over potential individual partners; that is, over allowable sets of partners with only one agent belonging to the opposite side of the market. These preferences are assumed to be strict. They are transitive and complete and so, they can be represented by ordered lists of preferences. Thus, the individual preference relation of firm  $f$  can be represented by an ordered list of preferences  $P(f)$  on the set  $W \cup \{f\}$ ; the individual preference relation of worker  $w$  can be represented by an ordered list of preferences  $P(w)$ , on the set  $F \cup \{w\}$ . The array of these preferences will be denoted by  $P$ . Then, an agent  $w$  is acceptable to an agent  $f$  if  $w \geq_f f$ . Similarly, an agent  $f$  is acceptable to an agent  $w$  if  $f \geq_w w$ .

If an agent may form more than one partnership then his/her/its preferences are not restricted to the individual potential partners. That is, agents have preferences over any allowable set of partners. Given two allowable sets of partners,  $A$  and  $B$ , for agent  $y \in F \cup W$ , we write  $A >_y B$  to mean  $y$  prefers  $A$  to  $B$ , and  $A \geq_y B$  to mean  $y$  likes  $A$  at least as well as  $B$ . In order to fix ideas we will assume that these preferences are *responsive* to the agents' individual preferences and are not necessarily strict.



The rules of the market are that any firm and worker pair may sign an employment contract with each other if they both agree; any firm may choose to keep one or more of its positions unfilled, and any worker may choose not to fill his or her quota of jobs if he or she wishes to do so.

When the quota of every agent is one, we have the *marriage model*. In this case an allowable set of partners for any agent is a singleton. Therefore, every agent only has preferences over individuals.

If only the agents of one of the sides are allowed to have a quota greater than one, then we have the so called *college admission model* with responsive preferences.

In both models it is a matter of verification that the sets of setwise-stable matchings, pairwise-stable matchings and strong corewise-stable matchings coincide. For the many-to-many case this is not always true. The strong corewise-stability concept is not a natural solution concept for this model as Example 1 shows.

*Example 1 (Sotomayor [80]) (The corewise-stability concept is not a natural solution concept for the many-to-many case)*

Consider two firms  $f_1$  and  $f_2$  and two workers  $w_1$  and  $w_2$ . Each firm may employ and wants to employ both workers; worker  $w_1$  may take, at most, one job and prefers  $f_1$  to  $f_2$ ; worker  $w_2$  may work and wants to work for both firms. If the agents can communicate with each other, the outcome that we expect to observe in this market is obvious:  $f_1$  hires both workers and  $f_2$  hires only worker  $w_2$ . Of course this outcome is in the strong core. Since  $f_1$  has a quota of two and  $w_1$  prefers  $f_1$  to  $f_2$  we cannot expect to observe the strong corewise-stable outcome where  $f_1$  hires only  $w_2$  and  $f_2$  hires both workers. That is, both outcomes are in the strong core, but only the first one is expected to occur. Our explanation for this is that only the first outcome is setwise-stable.

On the other hand, the pairwise-stability concept is not a refinement of the core for the discrete many-to-many matching models. See Example 2.

*Example 2 (Sotomayor [80]) (A pairwise-stable matching which is not in the core)*

Consider the following labor market with a set of firms  $F = \{f_1, f_2, f_3, f_4\}$  and a set of workers  $W = \{w_1, w_2, w_3, w_4\}$ , where each firm can hire two workers, and each worker can work for two firms. If firm  $f_i$  hires worker  $w_j$  then  $f_i$  gets the profit  $a_{ij}$  and  $w_j$  gets the salary  $b_{ij}$ . The pairs of numbers  $(a_{ij}, b_{ij})$  are given in Table 1.

Consider the matching  $\mu$  where  $f_1$  and  $f_2$  are matched to  $\{w_3, w_4\}$  and  $f_3$  and  $f_4$  are matched to  $\{w_1, w_2\}$ . (The payoffs of each matched pair are presented in boldface). This matching is pairwise stable. In fact,  $f_3$  and  $f_4$  do not

**Two-Sided Matching Models, Table 1**

Payoff Matrix of Example 2. Each row represents a firm and each column a worker. The values in the cell represent the payoffs to the corresponding firm (*first value*) and worker (*second value*)

10,1	1,10	<b>4,10</b>	<b>2,10</b>
1,10	10,1	<b>4,4</b>	<b>2,4</b>
<b>10,4</b>	<b>4,4</b>	2,2	1,2
<b>10,2</b>	<b>4,2</b>	2,1	1,1

**Two-Sided Matching Models, Table 2**

Payoff Matrix of Examples 3 and 4. Each row represents a firm and each column a worker. The values in the cell represent the payoffs to the corresponding firm (*first value*) and worker (*second value*)

13,1	<b>14,10</b>	<b>4,10</b>	1,10	0,0	0,0	3,10
1,10	0,0	0,0	10,1	<b>4,10</b>	<b>2,10</b>	0,0
<b>10,4</b>	0,0	0,0	0,0	0,0	0,0	0,0
<b>10,2</b>	0,0	0,0	0,0	0,0	0,0	0,0
0,0	<b>9,9</b>	0,0	<b>10,4</b>	0,0	0,0	0,0
0,0	0,0	0,0	<b>10,2</b>	0,0	0,0	0,0

belong to any pair which causes instability, because they are matched to their two best choices:  $w_1$  and  $w_2$ ; ( $f_1, w_1$ ) and ( $f_1, w_2$ ) do not cause instabilities since  $f_1$  is the worst choice for  $w_1$  and  $w_2$  is the worst choice for  $f_1$ ; ( $f_2, w_1$ ) and ( $f_2, w_2$ ) do not cause instabilities since  $w_1$  is the worst choice for  $f_2$  and  $f_2$  is the worst choice for  $w_2$ . Nevertheless,  $f_1$  and  $f_2$  prefer  $\{w_1, w_2\}$  to  $\{w_3, w_4\}$  and  $w_1$  and  $w_2$  prefer  $\{f_1, f_2\}$  to  $\{f_3, f_4\}$ . Hence this matching is not in the core, since it is blocked by the coalition  $\{f_1, f_2, w_1, w_2\}$ .

Example 3 shows that setwise stability is a strictly stronger requirement than pairwise-stability plus strong corewise-stability. It presents a situation in which the set of setwise-stable matchings is a proper subset of the intersection of the strong core with the set of pairwise-stable matchings.

*Example 3 (Sotomayor [80]) (A strong corewise-stable matching which is pairwise-stable and is not setwise-stable)*

Consider the following labor market with a set of firms  $F = \{f_1, f_2, f_3, f_4, f_5, f_6\}$  and a set of workers  $W = \{w_1, w_2, w_3, w_4, w_5, w_6, w_7\}$ , where  $r_1 = 3, r_2 = r_5 = 2, r_3 = r_4 = r_6 = 1, s_1 = s_2 = s_4 = 2$  and  $s_3 = s_5 = s_6 = s_7 = 1$ . If firm  $f_i$  hires worker  $w_j$  then  $f_i$  gets profit  $a_{ij}$  and the worker gets a salary  $b_{ij}$ . The pairs of numbers  $(a_{ij}, b_{ij})$  are given in Table 2.

Let  $\mu$  be the matching given by

$$\begin{aligned} \mu(f_1) &= \{w_2, w_3, w_7\}, \quad \mu(f_2) = \{w_5, w_6\}, \\ \mu(f_3) &= \mu(f_4) = \{w_1\}, \quad \mu(f_5) = \{w_2, w_4\}, \quad \text{and} \\ \mu(f_6) &= \{w_4\}. \end{aligned}$$

The associated payoffs are shown in bold in Table 2. This matching is in the strong core. In fact, if there is a matching  $\mu'$  which weakly dominates  $\mu$  via some coalition  $A$ , then, under  $\mu'$ , no player in  $A$  is worse off and at least one player in  $A$  is better off. Furthermore, matching  $\mu'$  must match all members of  $A$  among themselves. By inspection we can see that the only players that can be better off are  $f_1, f_2, w_1$  and  $w_4$ , for all remaining players are matched to their best choices. However, if  $A$  contains one player of the set  $\{f_1, f_2, w_1, w_4\}$ , then  $A$  must contain all four players. In fact, If  $f_1 \in A$ , then  $f_1$  must form a new partnership with  $w_1$ , so  $w_1 \in A$ . If  $w_1 \in A$ , then  $w_1$  must form a new partnership with  $f_2$ , so  $f_2 \in A$ . If  $f_2 \in A$ , then  $f_2$  must form a new partnership with  $w_4$ , so  $w_4 \in A$ . Finally, if  $w_4 \in A$  then  $w_4$  must form a new partnership with  $f_1$ , so  $f_1 \in A$ . Thus, if  $\mu'$  weakly dominates  $\mu$  via  $A$ , then  $f_1, f_2, w_1$  and  $w_4$  are in  $A$  and  $f_1$  and  $f_2$  form new partnerships with  $w_1$  and  $w_4$ . Nevertheless,  $f_1$  must keep his partnership with  $w_2$ , his best choice. Then  $w_2$  must be in  $A$ , so she cannot be worse off and so  $f_5$  must also be in  $A$ . But  $f_5$  requires the partnership with  $w_4$ , who has quota of 2 and has already filled her quota with  $f_1$  and  $f_2$ . Hence  $f_5$  is worse off at  $\mu'$  than at  $\mu$  and then  $\mu'$  cannot weakly dominate  $\mu$  via  $A$ .

The matching  $\mu$  is clearly pairwise stable. Nevertheless, the coalition  $\{f_1, f_2, w_1, w_4\}$  causes an instability in  $\mu$ . In fact, if  $f_1$  is matched to  $\{w_1, w_2, w_4\}$  and  $f_2$  is matched to  $\{w_1, w_4\}$  then  $f_1$  gets 28 and the rest of the players in the coalition get 11 instead of 6. Hence, the matching  $\mu$  is not setwise-stable.

The question is then to know if, given any two sets of agents with their respective preferences and quotas, one can always find a setwise-stable matching. The answer is affirmative for the one-to-one matching model and for the many-to-one matching models with substitutable preferences. The existence of setwise-stable matchings for the marriage model was first proved by Gale and Shapley [30]. Sotomayor [76] also provides a simple proof of the existence of stable matchings for the marriage model that connects stability with a broader notion of stability with respect to unmatched agents. Gale and Shapley construct a deferred acceptance algorithm described below and prove that it yields a stable matching in a finite number of steps.

**The deferred acceptance algorithm with the men making the proposals.** Each man begins by proposing to his favorite woman (the first woman on his preference list). Each woman rejects the proposal of any man unacceptable to her, and in case she gets several proposals, she keeps only her most preferred one. If a man is not rejected at

this step he is kept engaged. At any step, any man who was rejected at the previous step proposes to his next choice (his most preferred woman among those who have not rejected him), as long as there remains an acceptable woman to whom he has not yet proposed (if at one step a man has been rejected by all of his acceptable women, he issues no further proposals). The algorithm terminates after any step in which no man is rejected, because then every man is either engaged to some woman or has been rejected by every woman on his list of acceptable women. Women who did not receive any acceptable proposals, and men rejected by all women acceptable to them remain single.

To see that the matching yielded by this algorithm is stable first observe that no agent has an unacceptable partner. In addition, if there is some man  $f$  and woman  $w$  not matched to each other and such that  $f$  prefers  $w$  to his current partner, then woman  $w$  is acceptable to man  $f$  and so he must have proposed to her at some step of the algorithm. But then he must have been rejected by  $w$  in favor of someone she prefers to  $f$ . Therefore,  $w$  is matched to a man whom she prefers to  $f$ , by the transitivity of the preferences and so  $f$  and  $w$  do not destabilize the matching.

For the college admission model with responsive preferences Gale and Shapley defined a related marriage model in which each college is replicated a number of times equal to its quota, so that in the related model, every agent has a quota of one. If  $f_1, \dots, f_{r(f)}$  are the  $r(f)$  copies of college  $f$  then each of these  $f_i$ 's has preferences over individuals that are identical with those of  $f$ . Each student's preference list is changed by replacing  $f$ , wherever it appears on his/her list, by the string  $f_1, \dots, f_{r(f)}$  in that order of preference. Therefore, the stable matchings of the related marriage market are in natural one-to-one correspondence with the stable matchings of the college admission market. By using the existence theorem for the marriage model we obtain the corresponding result for the college admission model.

The existence proof for the many-to-one case with strict and substitutable preferences was first given by Kelso and Crawford [42] through a variant of the deferred-acceptance algorithm.

If the market does not have two-sides or the many-to-one matching model does not have substitutable preferences then the set of setwise-stable matchings may be empty. Gale and Shapley [30] present an example of a one-sided matching model that does not have any stable matchings. This model was called by these authors the "roommate problem". In this example there are four agents  $\{a, b, c, d\}$ , such that  $a$ 's first choice is  $b$ ,  $b$ 's first choice is  $c$ , and  $c$ 's first choice is  $a$ , and  $d$  is the last choice of all the other agents. Of course if some agent is un-

matched then there will be two unmatched agents and they will destabilize the matching. If every one is matched, the agent who is matched to  $d$  will form a blocking pair with the agent who lists him at the head of his list. Therefore, there is no stable matching in this example.

A small amount of literature has grown around the issues of finding conditions in which the set of stable matchings is non-empty for the roommate problem, and the performance of algorithms that can produce them when they exist. (See Abeledo and Isaak [3], Irving [37], Tan [95], Chung [13] and Sotomayor [89].)

If preferences are not substitutable Example 2.7 of Roth and Sotomayor [69] shows that setwise-stable matchings may not exist in the many-to-one case.

Even in the simplest case of preferences representable by additively separable utility functions, setwise-stable matchings may not exist for the many-to-many case. See the example below.

*Example 4 (Sotomayor [80]) (Nonexistence of stable matchings)* Consider again the matching model of Example 2. We are going to show that the set of setwise-stable matchings is empty. First, observe that  $f_3$  prefers  $\{w_1, w_2\}$  to any other set of players and  $f_3$  is the second choice for  $w_1$  and  $w_2$ ;  $w_3$  prefers  $\{f_1, f_2\}$  to any other set of players and  $w_3$  is the second choice for  $f_1$  and  $f_2$ . Then, in any stable matching  $\mu$ ,  $f_3$  must be matched to  $w_1$  and  $w_2$ , while  $w_3$  must be matched to  $f_1$  and  $f_2$ . Separate the cases by considering the possibilities for the second partner of  $w_1$ , under a supposed stable matching  $\mu$ :

- Case 1: ( $w_1$  is matched to  $\{f_2, f_3\}$ .) Then  $f_2$  is not matched to  $w_4$  and we have that  $\{f_2, w_4\}$  causes an instability in the matching, since  $f_2$  prefers  $w_4$  to  $w_1$  and  $f_2$  is the second choice for  $q_4$ .
- Case 2: ( $w_1$  is matched to  $\{f_3, f_4\}$ .) Then the following possibilities occur:
  - (i) ( $w_2$  is matched to  $\{f_3, f_4\}$ .) Then  $\{f_1, f_2, w_1, w_2\}$  causes an instability in the matching. This matching is pairwise-stable, but it is not in the core.
  - (ii) ( $w_2$  is matched to  $\{f_3, f_1\}$ .) Then  $\{f_1, w_4\}$  causes an instability in the matching, since  $f_1$  is the first choice for  $w_4$  and  $f_1$  prefers  $w_4$  to  $w_2$ .
  - (iii) ( $w_2$  is matched to  $\{f_3, f_2\}$  or  $\{f_3\}$ .) Then  $\{f_4, w_2\}$  causes an instability in both cases, since  $w_2$  is the second choice for  $f_4$  and  $w_2$  prefers  $f_4$  to  $f_2$  and prefers  $f_4$  to have an unfilled position.
- Case 3: ( $w_1$  is matched to  $\{f_1, f_3\}$  or  $\{f_3\}$ .) Then  $\{f_4, w_1\}$  causes an instability in both cases, since  $w_1$  is the

first choice for  $f_4$  and  $w_1$  prefers  $f_4$  to  $f_1$  and prefers  $f_4$  to have an unfilled position.

Hence, there are no stable matchings in this example.

Pairwise-stable matchings always exist when the preferences are substitutable. When preferences are strict, Roth [62] presents an algorithm that finds a pairwise-stable matching for a many-to-many matching model with substitutable preferences. Sotomayor [80] provides a simple and non-constructive proof of the existence of pairwise-stable matchings for the general discrete many-to-many matching model with substitutable and not-necessarily strict preferences. Martínez et al. [51] construct an algorithm, which allows finding the whole set of pairwise-stable matchings, when they exist, for the many-to-one matching model.

Authors have looked for sufficient conditions on the preferences of the agents for the existence of setwise-stable matchings in the many-to-many cases. Sotomayor [88] proves that if the preferences of the firms satisfy the *maximin* property, then the set of pairwise-stable matchings coincides with the set of setwise-stable matchings. An example in that paper shows that the set of setwise stable matchings may be empty if this condition is not satisfied. It is assumed there that the preferences are responsive and it is conjectured that the result above extends to the case of substitutable preferences.

Echenique and Oviedo [24] also address this problem with a different condition. They show that if agents on one side of the market have strongly substitutable preferences, while the other side has substitutable preferences, then the set of setwise stable matchings coincides with the set of pairwise-stable matchings.

Konishi and Ünver [46] give conditions on the preferences of the agents in a many-to-many matching market under which a pairwise-stable matching cannot be quasi-dominated by a pairwise-unstable matching via a collusion.

Eeckhout [26], under the assumption of strict preferences and that every man (woman) is acceptable to every woman (man), presents a sufficient condition for uniqueness of the stable matchings in the marriage market. The condition on preferences is simple: for every  $f_i \in F = \{f_1, f_2, \dots, f_m\}$ ,  $w_i \succ_{f_i} w_k$  for all  $k > i$ , and for every  $w_j \in W = \{w_1, w_2, \dots, w_n\}$ ,  $f_j \succ_{w_j} f_k$  for all  $k > i$ .

One line of investigation that has been developed in the theory of two-sided matchings concerns the mathematical structure of the set of stable matchings, because it captures fundamental differences and similarities between the several kinds of models. For the marriage model and the college admission model with responsive preferences,

assuming that the preferences over individuals are strict, the set of setwise-stable (stable for short) matchings have the following characteristic properties:

- A1. *Let  $\mu$  and  $\mu'$  be stable matchings. Then  $\mu \geq_F \mu'$  if and only if  $\mu' \geq_W \mu$ .*  
That is, there exists an opposition of interests between the two sides of the market along the whole set of stable matchings.
- A2. *Every agent is matched to the same number of mates under every stable matching.*  
Consequently, *if an agent is unmatched under some stable matching then he/she/it is unmatched under any other stable matching.*  
When preferences are strict, there are two natural partial orders on the set of all stable matchings. The partial order  $\geq_F$  is defined as follows:  $\mu \geq_F \mu'$  if  $\mu(f) \geq_f \mu'(f)$  for all  $f \in F$ . The partial order  $\geq_W$  is analogously defined. The fact that these partial orders are well defined follows from A1. Then,
- A3. *The set of stable matchings has the algebraic structure of a complete lattice under the partial orders  $\geq_F$  and  $\geq_W$ .*

The lattice property means the following: if  $\mu$  and  $\mu'$  are two stable matchings, then some workers (respectively firms) will get a preferable set of mates under  $\mu$  than under  $\mu'$  and others will be better off under  $\mu'$  than under  $\mu$ . The lattice property implies that there is then a stable matching which gives each agent the most preferable of the two sets of partners and also one which gives each of them the least preferred set of partners. That is, if  $\mu$  and  $\mu'$  are stable matchings the lattice property implies that the functions  $\lambda, \nu, \eta$  and  $\tau$  are stable matchings, where  $\lambda = \mu \vee_F \mu'$  is defined by  $\lambda(f) = \max\{\mu(f), \mu'(f)\}$  and  $\lambda(w) = \min\{\mu(w), \mu'(w)\}$ ; the function  $\eta = \mu \vee_W \mu'$  is analogously defined; the function  $\nu = \mu \wedge_F \mu'$  is defined by  $\nu(f) = \min\{\mu(f), \mu'(f)\}$  and  $\nu(w) = \max\{\mu(w), \mu'(w)\}$ . Analogously we define  $\tau = \mu \wedge_W \mu'$  (notice that  $\mu \vee_F \mu'$  is the same as  $\mu \wedge_W \mu'$  and  $\mu \vee_W \mu'$  is the same as  $\mu \wedge_F \mu'$ ).

The fact that the lattice is complete implies the existence and uniqueness of a maximal element and a minimal element in the set of stable payoffs, with respect to the partial order that is being considered. Thus, there exists one and only one stable matching  $\mu_F$  and one and only one stable matching  $\mu_W$  such that  $\mu_F \geq_F \mu$  and  $\mu_W \geq_W \mu$  for all stable matchings  $\mu$ . Property A1 then implies that  $\mu \geq_W \mu_F$  and  $\mu \geq_F \mu_W$ . That is,

- A4. *There is an F-optimal stable matching  $\mu_F$  with the property that for any stable matching  $\mu$ ,  $\mu_F \geq_F \mu$  and*

*$\mu \geq_W \mu_F$ ; there is a W-optimal stable matching  $\mu_W$  with symmetrical properties.*

Property A1 was first proved by Knuth [45] for the marriage model. The result for the college admission model with responsive preferences was proved in Roth and Sotomayor [69] by making use of the following proposition of Roth and Sotomayor [68]: *Suppose colleges and students have strict individual preferences, and let  $\mu_1$  and  $\mu_2$  be stable matchings for the college admission model such that  $\mu_1(f) \neq \mu_2(f)$ . Let  $\mu_1^*$  and  $\mu_2^*$  be the stable matchings corresponding to  $\mu_1$  and  $\mu_2$  in the related marriage model. If  $\mu_1^*(f_i) >_f \mu_2^*(f_i)$  for some position  $f_i$  of  $f$  then  $\mu_1^*(f_j) \geq_f \mu_2^*(f_j)$  for all positions  $f_j$  of  $f$ .*

Property A2 was proved by Gale and Sotomayor [32] for both models. For the college admission model with responsive preferences Roth [66] added that *if a college does not fill its quota at some stable matching then it has the same set of mates at every stable matching*. The restriction of property A2 to the marriage model where every pair of partners is mutually acceptable was proved by McVitie and Wilson [52].

For the many-to-one case with substitutable preferences, Martínez et al. [50] presents an example in which there are agents who are unmatched under some stable matching and are matched under another one. By introducing quotas in the model with substitutable preferences of Roth and Sotomayor [69], these authors prove that if the preferences of the colleges are strict, substitutable and  $r(f)$ -separable for every college  $f$ , then property A2 holds. Furthermore Roth's result mentioned above also applies.

The lattice property of the set of stable matchings for the marriage model is attributed by Knuth [45] to Conway. The existence of the optimal stable matchings for each side of the marriage market and the college admission market with responsive preferences was first proved in Gale and Shapley [30] by using the deferred acceptance procedure. The idea of their elegant proof is to show that a proposer is never rejected by an achievable mate, so he/she/it ends up with his/hers/its best achievable mate.

The lattice property for the college admission model with responsive preferences was obtained in Roth and Sotomayor [69]. To show that the functions  $\lambda, \nu, \eta$  and  $\tau$  above are well defined these authors used the following theorem from Roth and Sotomayor [68]: *If colleges and students have strict preferences over individuals, then colleges have strict preferences over those groups of students that they may be assigned at stable matchings. That is, if  $\mu_1$  and  $\mu_2$  are stable matchings, then a college  $f$  is indifferent between  $\mu_1(f)$  and  $\mu_2(f)$  only if  $\mu_1(f) = \mu_2(f)$ .*



This result is an immediate consequence of the proposition mentioned above, due to the responsiveness of the preferences. Therefore, if  $\mu_1$  and  $\mu_2$  are two stable matchings then  $f$  prefers  $\mu_1(f)$  to  $\mu_2(f)$  if and only if the  $r(f)$  most preferred students by  $f$  in the set formed by the union of  $\mu_1(f)$  and  $\mu_2(f)$  are those ones in  $\mu_1(f)$ .

For the many-to-many matching market with strict and substitutable preferences, Blair [12] proved that the set of pairwise-stable matchings (not necessarily setwise-stable) has the lattice structure under some partial order relation that is not defined by the preferences of the agents. The definition of the partial order  $\geq_F$  uses that if  $\mu_1$  and  $\mu_2$  are pairwise-stable matchings then  $\mu_1 \geq_F \mu_2$  if and only if  $Ch_f(\mu_1(f) \cup \mu_2(f)) = \mu_1(f)$ , for every agent  $f \in F$ . Similarly the partial order  $\geq_W$  is defined. As remarked above the partial order defined by Blair coincides with the one defined by the preferences of the players in the college admission model with responsive preferences.

Adachi [4] introduces a map, which is called a  $T$ -map, defined over the set of pre-matchings, in order to show that the set of stable matchings is a non-empty lattice in the marriage market under strict preferences. Adachi defines the  $T$ -map as follows: given a pre-matching  $\mu$ ,  $T(\mu(f))$  is  $f$ 's most preferred worker in  $\{w \in W | f \geq_w \mu(w)\} \cup \{f\}$  for all  $f \in F$ , and similarly,  $T(\mu(w))$  is  $w$ 's most preferred firm in  $\{f \in F | w \geq_f \mu(f)\} \cup \{w\}$  for all  $w \in W$ . Clearly, any fixed point of the  $T$ -map is a matching, and Adachi showed it has to be stable. Using the partial order  $\geq_F$  defined by the agents' preferences, he showed that the set of pre-matchings endowed with this partial order is a complete lattice and the  $T$ -map is an isotone function (order preserving). Thus, Tarski's fixed-point theorem implies that the set of fixed points of the  $T$ -map, which is the set of stable matchings, is a non-empty complete lattice.

**Theorem 1 (Tarski's Theorem [96])** *Let  $E$  be a complete lattice with respect to some partial order  $\geq$ , and let  $f$  be an isotone function from  $E$  to  $E$ . Then the set of fixed points of  $f$  is non-empty and is itself a complete lattice with respect to the partial order  $\geq$ .*

In this same vein, Echenique and Oviedo [23,24] extend Adachi's [4] approach and the  $T$ -map in order to analyze the many-to-one and many-to-many models, respectively. Again, any fixed point of the  $T$ -map is a matching. Echenique and Oviedo [23] show that for the many-to-one model the set of fixed points of the  $T$ -map is equal to the set of stable matchings. By making successive iterations of the  $T$ -map, starting from some specific pre-matching, until a fixed point is reached, this map can be used to find all the stable matchings, as long as they exist. This procedure is called the  $T$  algorithm. These authors show that *as long*

*as the strong core is non-empty, the  $T$  algorithm always converges, and if the strong core is empty, it cycles.* They present an example of a situation in which the preferences are not substitutable and the  $T$  algorithm finds strong core allocations, but the algorithm with firms proposing according to their preference lists over allowable sets of workers does not do so. Finally, they give a bound on the computational complexity of the  $T$  algorithm and show how it can be used to calculate both the supremum and infimum under Blair's partial order, which under non-substitutable preferences might not be easily computed. Furthermore, under substitutability, the set of pre-matchings endowed with the partial order defined by Blair is again a complete lattice and the  $T$ -map is isotone, so Tarski's theorem implies that *the strong core is a non-empty lattice under the partial order introduced in Blair [12].*

For the many-to-many model, the set of fixed-points of the  $T$ -map studied by Echenique and Oviedo [24] is shown to be, under substitutability, equal to the set of pairwise-stable matchings, and a superset of the set of setwise-stable matchings. Furthermore, the set of pre-matchings endowed with Blair's partial order is again a complete lattice. Then Tarski's theorem applies and the set of pairwise-stable matchings is a non-empty complete lattice. *If both sides of the market satisfy the strong substitutability property then the set of setwise stable matchings is a complete lattice, both for Blair's partial order and for the partial order defined by the agents' preferences.*

Martínez et al. [51] propose an algorithm which allows them to calculate the whole set of pairwise-stable matchings under substitutability.

Echenique and Yenmez [25] study the college admission problem when students have preferences over colleagues. Using the  $T$ -map they construct an algorithm, which finds all the core allocations, as long as the core is non-empty. In a similar setup, Pycia [57] finds necessary and sufficient conditions for the existence of stable matchings. Dutta and Massó [22] studied a many-to-one version of the model of Echenique and Yenmez [25]. They showed that under certain conditions on preferences the core is non-empty.

Hatfield and Milgrom [36] present a general many-to-one model in which, the central concept is that of a contract, which allows a different formalization of a matching. In their model there are a finite set of firms, a finite set of workers and a finite set of wages offered by firms. Each contract  $c$  specifies the pair  $(f, w)$  involved and the wage the worker  $w$  gets from firm  $f$ , so that the set of contracts is  $C = F \times W \times WP$  (where  $WP$  is the set of wages offered by the firms). Clearly, if each firm offers a unique wage level to all workers, their model is a college admis-



sion model. Agents have preferences over the contracts in which they could be involved. In this model, a feasible allocation is a set of contracts  $C' \subseteq C$  in which each worker  $w$  appears at most in one contract  $c \in C'$  and each firm  $f$  appears in at most  $r(f)$  contracts  $c_1, \dots, c_{r(f)} \in C'$  and such that for each agent  $a \in F \cup W$  we have that  $Ch_a(C_a) = C_a$ , where  $C_a$  is the set of contracts in  $C'$  in which agent  $a$  appears. According to Hatfield and Milgrom a feasible allocation  $C'$  is stable if there does not exist an alternative feasible allocation which is strictly preferred by some firm  $f$  and weakly preferred by all of the workers it hires. Making use of Tarski's fixed point theorem, they prove that if preferences are strict and satisfy substitutability over the set of contracts, then the set of stable allocations is a non-empty lattice. They introduce the condition of the law of aggregate demand on preferences, which requires that for all allowable sets  $X, Y$  if  $X \subseteq Y$  then  $|Ch_f(Y)| \geq |Ch_f(X)|$ . By assuming that firms' preferences satisfy substitutability and the law of aggregate demand, they prove some characteristic results on the structure of the set of stable allocations and analyze the incentives facing the agents when a mechanism which produces the  $W$ -optimal stable allocation is adopted. They show that under this mechanism, it is a dominant strategy for the workers to state their true preferences.

Ostrovsky [54] generalizes the model presented in Hatfield et al. [36] to a  $K$ -sided discrete many-to-many matching model. A set of contracts, which allows the production and consumption of some goods, is called a network. This author considers supply networks where goods are sold and transformed through many stages, starting from the suppliers of initial inputs, going through different intermediaries until they reach the final consumer. He generalizes the concept of pairwise stability to this setting and calls it *chain stability*, which requires that there does not exist a chain of contracts such that all members of this chain are better off. A chain of contracts specifies a sequence of agents, each of whom is the seller in one contract and the buyer in the next one. Under certain conditions on the preferences of the agents he proves the existence of chain stable networks and, by using fixed point methods, he shows that the set of chain stable networks is a non-empty lattice. Furthermore, he proves that there exists a consumer optimal network and an initial supplier optimal network, similar to the  $F$ -optimal and  $W$ -optimal matchings in other models. Finally, for the case in which each agent can be the seller (buyer) in at most one contract he shows that the set of chain stable networks is equal to the core.

Crawford [15] proposes to allow offers in the NRMP mechanism to include salaries and demonstrates how the

resulting market can generate stable outcomes, which might Pareto dominate the ones in the current form of the NRMP. This model can be seen as an application of the Hatfield and Milgrom [36] paper.

Another line of investigation that has grown in the last decade concerns the special case of the college admission model in which colleges have fixed preferences, known nowadays as the school choice model. The seminal paper is Sotomayor [77] which was motivated by the admission market of economists to graduate centers of economics in Brazil. The students take some tests and each institution places weights on each of these tests in order to rank the students according to the weighted average of the tests. See Ergin and Sönmez [28], Pathak and Sönmez [55] and Balinski and Sönmez [8].

### Continuous Two-Sided Matching Model with Additively Separable Utility Functions

The two-sided matching model with *additively separable utility functions* involves two finite and disjoint sets of players which will be denoted by  $B$ , with  $m$  elements, and  $Q$ , with  $n$  elements. Each  $b \in B$  has a quota  $r(b)$  and each  $q \in Q$  has a quota  $s(q)$  representing the maximum number of partnerships they can form. The main characteristic of this model is that agents are able to negotiate their individual payoffs: If a partnership  $(b, q)$  is formed, the partners undertake an activity together that produces a payoff  $v_{bq}$  which is divided between them into the payoffs  $u_{bq}$  for  $b$  and  $w_{bq}$  for  $q$  respectively, as a result of a negotiation process. Therefore, an outcome for this game is a matching, along with individual payoffs  $u_{bq}$ 's and  $w_{bq}$ 's. Dummy players, denoted by 0, are included for technical convenience in both sides of the market. We have that  $v_{b0} = v_{0q} = 0$  for all  $b \in B$  and  $q \in Q$ . As for the quotas, a dummy player may form as many partnerships as needed to fill the quotas of the non-dummy players. Then, an allowable set of partners for agent  $b$ , with only  $r(b) - k$  elements of  $Q$ , has  $k$  copies of the dummy  $Q$ -agent introduced in the model.

A matching  $\mu$  is feasible if each player is matched to an allowable set of partners. A feasible outcome, denoted by  $(u, w; \mu)$ , is a feasible matching  $\mu$  and a pair of payoffs  $(u, w)$ , where the individual payoffs of each  $b \in B$  and  $q \in Q$  are given by the arrays of numbers  $u_{bq}$ , with  $q \in \mu(b)$ , and  $w_{bq}$ , with  $b \in \mu(q)$ , respectively, such that  $u_{bq} + w_{bq} = v_{bq}$ ,  $u_{bq} \geq 0$  and  $w_{bq} \geq 0$ . Consequently,  $u_{b0} = u_{0q} = w_{b0} = w_{0q} = 0$  in case these payoffs are defined. We say that the matching  $\mu$  is compatible with the payoff  $(u, w)$ .

The value of  $\mu$  is  $\sum_{q \in Q, b \in \mu(q)} v_{bq}$ . The matching  $\mu$  is *optimal* if it attains the maximum value among all feasible matchings.

This model generates the following game in coalitional function form with side payments. The set of players is  $N = B \cup Q$ , and the characteristic function  $v$  satisfies the following:

- (a)  $v(\emptyset) = 0$ ,
- (b)  $v(S) = 0$  if  $S \subseteq B$  or  $S \subseteq Q$ ,
- (c)  $v(S) \leq v(T)$  if  $S \subseteq T$ ,
- (d)  $v(b, q) = v_{bq}$  for all  $(b, q) \in B \times Q$ .

For every  $b \in B$  and for all sets  $S \subseteq Q$  with  $|S| \geq r(b)$ ,

- (e)  $v(b \cup S) = \max\{v(b \cup S'); S' \subseteq S \text{ and } |S'| = r(b)\}$ .
- (c) and (d) imply that for all sets  $S \subseteq W$  with  $|S| \geq r(b)$ ,  
 $v(b \cup S) = v(b \cup S')$ , for some  $S' \subseteq S$  with  $|S'| = r(b)$

Analogously we define  $v(q \cup S)$  for every  $q \in Q$  and all sets  $S \subseteq B$  with  $|S| \geq s(q)$ .

The condition that the game has *additively separable utilities* means that for every coalition  $S = R \cup T$ ,  $R \subseteq B$  and  $T \subseteq Q$ ,

- (f)  $v(R \cup T) = \max\{\sum_{(b,q) \in R \times T} x_{bq} v_{bq}, \text{ for every feasible assignment } x\}$ . Consequently, for all  $T \subseteq Q$  with  $|T| \leq r(b)$  and for all  $R \subseteq B$  with  $|R| \leq s(q)$ ,  $v(b \cup T) = \sum_{q \in T} v_{bq}$  and  $v(q \cup R) = \sum_{b \in R} v_{bq}$ .

When the quota of any agent is one, the model is the well known assignment game introduced in Shapley and Shubik [72]. In this case the set of individual payoffs of an agent is a singleton, so these payoffs need not be indexed according to the partnerships formed under the matching. Furthermore, the concepts of setwise-stability, pairwise-stability and corewise-stability are equivalent. *The outcome  $(u, w; \mu)$  is stable if it is feasible and  $u_b + w_q \geq v_{bq}$  for all pairs  $(b, q)$ .*

The existence of stable payoffs for the assignment game was proved in Shapley and Shubik [72] with the use of linear programming. A different, but also simple proof, is obtained in Sotomayor [81] using only combinatorial arguments. Crawford and Knoer [16] consider a discrete version (as well as the continuous version) of the assignment game and develop a version of the deferred acceptance algorithm of Gale and Shapley to prove the non-emptiness of the set of pairwise-stable payoffs.

The main properties that characterize the set of stable outcomes of the assignment game of Shapley and Shubik are the following:

- B1. *Let  $(u, w)$  be some stable payoff. Then  $\mu$  is an optimal matching if and only if it is compatible with  $(u, w)$ .*

This result means that the set of stable payoffs is the same under every optimal matching. Then we can concentrate on the payoffs of the agents rather than on the underlying matching.

- B2. *Let  $(u, w)$  and  $(u', w')$  be stable payoffs. Then  $u \geq u'$  if and only if  $w' \geq w$ .*

That is, there exists an opposition of interests between the two sides of the market along the whole set of stable payoffs.

- B3. *If an agent is unmatched under some stable outcome then he/she gets a zero payoff under any other stable outcome.*

This means, for example, that if a worker is unemployed under some stable outcome then this worker will get a zero salary under any other stable outcome.

- B4. *The set of stable payoffs forms a convex and compact lattice under the partial orders  $\geq_B$  and  $\geq_Q$ .*

The partial order  $\geq_B$  on the set of stable payoffs is defined as follows:  $(u, w) \geq_B (u', w')$  if  $u_b \geq u'_b$  for all  $b \in B$ . Property B2 implies that  $w_q \leq w'_q$  for all  $q \in Q$ , so this partial order is well defined. The partial order  $\geq_Q$  is symmetrically defined. Then,  $(u, w) \vee_B (u', w') = (\max\{u, u'\}, \min\{w, w'\})$  and  $(u, w) \wedge_B (u', w') = (\min\{u, u'\}, \max\{w, w'\})$ .

The lattice property implies that there exist a maximal element and a minimal element in the set of stable payoffs. The fact that the lattice is complete implies the uniqueness of these extreme points. Thus, there exists one and only one stable payoff  $(u_*, w_*)$  and one and only one stable payoff  $(u_*, w_*)$  such that  $(u_*, w_*) \geq_B (u, w)$  and  $(u_*, w_*) \geq_Q (u, w)$  for all stable payoffs  $(u, w)$ . That is,

- B5. *There is a B-optimal stable payoff  $(u_*, w_*)$  with the property that for any stable payoff  $(u, w)$ ,  $u_* \geq u$  and  $w_* \leq w$ ; there is a Q-optimal stable payoff  $(u_*, w_*)$  with symmetrical properties.*

- B6. *The set of stable payoffs equals the core and the set of competitive equilibrium payoffs.*

Excluding property B3 which follows from property 1 of Demange and Gale [18], all the other properties were first proved in Shapley and Shubik [72].

The general quota case is a version of the model studied in Crawford and Knoer [16] and was first presented in Sotomayor [75] in the context of a labor market of firms and workers. Under this approach, the number  $r(b)$  is the maximum number of workers firm  $b$  can hire; the number  $s(q)$  is the maximum number of jobs worker  $q$  can

take and the number  $v_{bq}$  is the productivity of worker  $q$  in firm  $b$ . The natural cooperative solution concept is that of setwise-stability which is shown to be equivalent to the concept of pairwise-stability. Then, *the feasible outcome*  $(u, w; \mu)$  is *setwise-stable* if  $u_b(\min) + w_q(\min) \geq v_{bq}$  for all pairs  $(b, q)$  with  $q \notin \mu(b)$ , where  $u_b(\min)$  is the smallest individual payoff of firm  $b$ ;  $w_q(\min)$  is the smallest individual payoff of worker  $q$ .

The existence of setwise-stable payoffs for this model was proved in Sotomayor [75,79] through the use of linear programming.

Another interpretation of this model considers a buyer-seller market:  $B$  is a set of buyers and  $Q$  is a set of sellers. Buyers are interested in sets of objects owned by different sellers and each seller owns a set of identical objects. The number  $r(b)$  is the number of objects buyer  $b$  is allowed to acquire; the number  $s(q)$  is the number of identical objects seller  $q$  owns and the number  $v_{bq}$  is the amount of money buyer  $b$  considers to pay for an object of seller  $q$ . We say that  $v_{bq}$  is the value of object  $q$  (object owned by seller  $q$ ) to buyer  $b$ . An artificial *null-object*, 0, owned by the dummy seller, whose value is zero to all buyers and whose price is always zero is introduced for technical convenience. Under this approach a buyer will be assigned to an allowable set of objects at a feasible allocation, meaning that she is matched to the set of sellers who own the objects in the given set.

Given a price vector  $p \in R_+^\sigma$ , with  $\sigma \equiv \sum_{q \in Q} s(q)$ , the preferences of buyers over objects are completely described by the numbers  $v_{bq}$ 's: For any two allowable sets of objects  $S$  and  $S'$ , buyer  $b$  prefers  $S$  to  $S'$  at prices  $p$  if her total payoff when she buys  $S$  is greater than her total payoff when she buys  $S'$ . She is indifferent between these two sets if she gets the same total payoff with both sets. Usually, given the prices of the objects, buyers demand their favorite allowable sets of objects at those prices. The set of such allowable sets is called the demand set of buyer  $b$  at prices  $p$ . *An equilibrium is reached if every buyer is assigned to an allowable set of objects of her demand set, every seller with a positive price sells all of his items and the number of objects in the market is enough to meet the demand of all buyers.* The solution concept that captures this intuitive idea of equilibrium is that of *competitive equilibrium payoff* defined in Sotomayor [91] as an extension of the concept of competitive equilibrium price for the Assignment game given in Demange, Gale and Sotomayor [19]. Formally,  $(u, p; \mu^*)$  is a **competitive equilibrium outcome** if (i) it is feasible, (ii)  $\mu^*$  is a feasible allocation such that, if  $\mu^*(b) = S$  then  $S$  is in the demand set of  $b$  at prices  $p$  for all  $b \in B$  and (iii)  $p_q = 0$  if object  $q$  is left unsold.

If  $(u, p; \mu^*)$  is a competitive equilibrium outcome we say that  $(u, p)$  is a competitive equilibrium payoff,  $(p, \mu^*)$  is a competitive equilibrium and  $p$  is a *competitive equilibrium price* or an *equilibrium price* for short.

One characteristic of the additively separable utility function is that if a buyer demands a set  $A$  of objects at prices  $p$  and some of these objects have their prices raised, then the buyer will continue to want to buy the objects in  $A$  whose prices were not changed. That is, the function  $v(\{b\} \cup A)$  over all allowable sets  $A$  of partners for  $b$  satisfies the *gross substitute condition*. Kelso and Crawford [42] formulated a discrete and a continuous many-to-one matching model where the functions  $v(\{b\} \cup A)$  satisfy the gross substitute condition and are not necessarily additively separable. In this model, the core, the set of setwise-stable payoffs and the set of competitive equilibrium payoffs coincide and are non-empty. These authors prove, through an example, that without this condition the core may be empty.

A consequence of the competitive equilibrium concept for the many-to-many case with additively separable utility functions is that sellers do not discriminate buyers under a competitive equilibrium payoff, as they might do under a stable outcome. The competitive equilibrium payoffs for this model are characterized as *the setwise-stable payoffs where every seller has identical individual payoffs*. It is interesting to point out that if the identical objects are owned by different sellers, they need not be sold at the same price unless the two sellers have the same number of objects and the selling price is the minimum competitive equilibrium price. (Sotomayor [91]).

A stable (respectively, competitive equilibrium) payoff is called a  $B$ -optimal stable (respectively, competitive equilibrium) payoff if every agent in  $B$  weakly prefers it to any other stable (respectively, competitive equilibrium) payoff. That is, the  $B$ -optimal stable (respectively, competitive equilibrium) payoff gives to each agent in  $B$  the maximum total payoff among all stable (respectively, competitive equilibrium) payoffs. Similarly we define a  $Q$ -optimal stable (respectively, competitive equilibrium) payoff.

The existence and uniqueness of the  $B$ -optimal and of the  $Q$ -optimal stable payoffs are proved in Sotomayor [79] by showing that the set of stable payoffs is a lattice under two conveniently defined partial orders. This result runs into the difficulties of defining a partial order relation in the set of stable payoffs, due to the fact that, on the one hand the arrays of individual payoffs are unordered sets of numbers indexed according to the current matching and on the other hand, the agents' preferences do not define a partial order relation, since they violate the anti-symmetric property.

To solve this problem Sotomayor [79] defines a partnership  $(b, q)$  to be nonessential if it occurs in some but not all optimal matchings and essential if it occurs in all optimal matchings. Then, two matchings differ only by their nonessential partnerships. According to Theorem 1 of that paper, (i) *in every stable outcome a player gets the same payoff in any nonessential partnership; furthermore this payoff is less than or equal to any other payoff the player gets under the same outcome*; (ii) *given a stable outcome  $(u, w; \mu)$  and a different optimal matching  $\mu'$ , we can reindex the  $u_{bq}$ 's and  $w_{bq}$ 's according to  $\mu'$  and still get a stable outcome*.

Therefore, the array of individual payoffs of a player can be represented by a vector in a Euclidean space whose dimension is the quota of the given player. The first coordinates are the payoffs that the player gets from his essential partners (if any), following some given ordering. The remaining coordinates (if any) are equal to a number which represents the payoff the player gets from all his nonessential partners. This representation is clearly independent of the matching, so any optimal matching is compatible with a stable payoff. Hence, by ordering the players in  $B$  (respectively,  $Q$ ), we can immerse the stable payoffs of these players in a Euclidean space, whose dimension is the sum of the quotas of all players in  $B$  (respectively,  $Q$ ). Then, the natural partial order relation of this Euclidean space induces the partial order relation  $\geq_B$  (respectively  $\geq_Q$ ) in the set of stable payoffs. We say that  $(u, w) \geq_B (u', w')$  if the vector of individual payoffs of any buyer, under  $(u, w)$ , is greater than or equal to her vector of individual payoffs under  $(u', w')$ . Similarly we define  $(u, w) \geq_Q (u', w')$ .

The main results of Sotomayor [79] are that, under the vectorial representation of the stable payoffs, properties B1, B2, B3, B4 and B5 hold for the general many-to-many case.

An implication of property B2 is the conflict of interests that exists between the two sides of the market with respect to two comparable stable payoffs. That is, if payoffs  $(u, w)$  and  $(u', w')$  are stable and comparable, then for all  $(b, q) \in B \times Q$  we have that  $b$ 's total payoff under the first outcome is greater than  $b$ 's total payoff under the second outcome if and only if  $q$ 's total payoff under the second outcome is greater than  $q$ 's total payoff under the first outcome. From property B3, if a seller has some unsold object under a stable outcome, then one of his individual payoffs will be zero under any other such outcome.

Even though the preferences of the players do not define the partial orders  $\geq_B$  and  $\geq_Q$ , the property stated in B4 is of interest because the two extreme points of the lattice have an important meaning for the model. The

extreme points of the lattice are precisely the  $B$ -optimal and the  $Q$ -optimal stable payoffs. Also every buyer weakly prefers any stable payoff to the  $Q$ -optimal stable payoff and any seller weakly prefers any stable payoff to the  $B$ -optimal stable payoff.

Indeed, the set of competitive equilibrium payoffs is a sub lattice of the set of stable payoffs. This connection is given by the following theorem of Sotomayor [91] that states that *the set of competitive equilibrium payoffs is contained in the set of stable payoffs and is a non-empty and complete lattice under the partial order  $\geq_B$  (respectively  $\geq_Q$ ) whose supremum (respectively infimum) is  $B$ -optimal and whose infimum (respectively supremum) is  $Q$ -optimal*.

The idea of the proof is that the set of competitive equilibrium payoffs can be obtained by “shrinking” the set of stable payoffs through the application of a convenient isotone (order preserving) map  $f$  whose fixed points are exactly the competitive equilibrium payoffs. The desired result is concluded via the *algebraic fixed point theorem* due to Alfred Tarski [96].

It is also proved in Sotomayor [91] that *the  $B$ -optimal stable payoff is a fixed point of  $f$ , so the  $B$ -optimal stable payoff is the  $B$ -optimal competitive equilibrium payoff*.

As for property B6, Sotomayor [84] shows that the core coincides with the set of pairwise-stable payoffs in the many-to-one case where sellers have a quota of one. Since a seller only owns one object then he cannot discriminate the buyers, so the core coincides with the set of competitive equilibrium payoffs in this model. Thus, the set of competitive equilibrium payoffs is a lattice in this model. The same result is reached in Gül and Stacchetti [33] for the many-to-one case in which the utilities satisfy the gross substitute condition.

However, in the general quota case under additively separable utilities, the core may be bigger than the set of stable payoffs, which in its turn contains and may contain properly the set of competitive equilibrium payoffs, as it is illustrated in the example below from Sotomayor [91]. This example also shows that the core may not be a lattice and, the polarization of interests, observed in the sets of stable payoffs and of competitive equilibrium payoffs, does not always carry over to the core payoffs: The best core payoff for the buyers is not necessarily the worst core payoff for the sellers.

*Example 5 (Sotomayor, [91])* Consider the following situation. The  $B$ -players will be called firms and the  $Q$ -players will be called workers. There are two firms,  $b$  and  $b'$ , and two workers  $q$  and  $q'$ . Each firm may employ and wants to employ both workers; worker  $q$  may take, at most, one job and worker  $q'$  may work and wants to work for both



firms. The first row of matrix  $v$  is  $(3, 2)$  and the second one is  $(3, 3)$ .

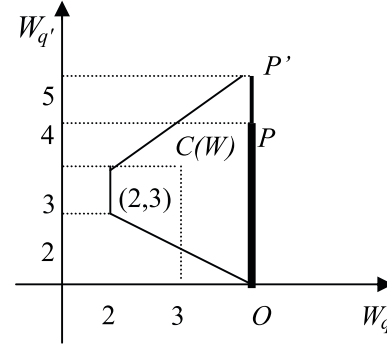
There are two optimal matchings:  $\mu$  and  $\mu'$ , where  $\mu(b) = \{q, q'\}$ ,  $\mu(b') = \{q', 0\}$  and  $\mu'(b') = \{q, q'\}$ ,  $\mu'(b) = \{q', 0\}$ . The core is described by the set of individual payoffs  $(u, w)$  whose total payoffs  $(U, W)$ , satisfy the following system of inequalities:  $0 \leq U_b \leq 2$ ,  $0 \leq U_{b'} \leq 3$ ;  $W_q + W_{q'} \geq 3$ ,  $W_{q'} - W_q \leq 2$ ,  $1 \leq W_q \leq 3$ . It is not hard to see that the outcome  $(u, w)$  is stable if and only if seller  $q$  always gets payoff  $w_q = 3$ , seller  $q'$  gets individual payoffs  $w_{bq'} \in [0, 2]$  and  $w_{b'q'} \in [0, 3]$ ; the individual payoffs of buyers  $b$  and  $b'$  are given by  $(u_{bq} = 0, u_{bq'} = 2 - w_{bq'})$  and  $(u_{b'q'} = 3 - w_{b'q'}, u_{b0} = 0)$ , respectively.

To see that corewise-stability is not adequate to define the cooperative equilibrium for this market, let  $(u, w; \mu)$  be such that  $u_{bq} = 1$ ,  $u_{bq'} = 1$ ,  $u_{b'q'} = 1$ ,  $u_{b'0} = 0$ ;  $w_{bq} = 2$ ,  $w_{bq'} = 1$ ,  $w_{b'q'} = 2$ . That is, firm  $b$  hires workers  $q$  and  $q'$ , obtains from each one of them a profit of one and pays two to  $q$  and one to  $q'$ ; firm  $b'$  hires worker  $q'$  at a salary of two and obtains a profit of one. Observe that  $b'$  has quota of two, so it has one unfilled position. It happens that  $b'$  can pay more than two to  $q$ . Thus, if agents can communicate with each other and behave cooperatively, this outcome will not occur, because worker  $q$  will not accept to receive only two from firm  $b$ , since she knows that she can get more than two by working with firm  $b'$ . Hence, this outcome cannot be a cooperative equilibrium. Observe that  $2 = u_{b'0} + w_{bq} < v_{b'q} = 3$ , so this outcome is not stable. On the other hand, it is in the core. In fact, if there is a blocking coalition then it must contain  $\{b', q\}$ . These agents cannot increase their total payoffs by themselves;  $b'$  needs to hire both workers. However  $\{b', q, q'\}$  does not block the outcome, because  $q'$  is worse off by taking only one job. Nevertheless, the coalition of all agents does not block the outcome, since  $b$  loses worker  $q$ , so it will be worse off.

Now, consider the outcome  $(u', w'; \mu)$ , where  $u'_{bq} = 0$ ,  $u'_{bq'} = 1$ ,  $u'_{b'q'} = 1$ ,  $u'_{b'0} = 0$ ;  $w'_{bq} = 3$ ,  $w'_{bq'} = 1$ ,  $w'_{b'q'} = 2$ . Firm  $b'$  cannot offer more than three to worker  $q$ , so the structure of the outcome cannot be ruptured. Then, although both outcomes  $(u, w; \mu)$  and  $(u', w'; \mu)$  are corewise-stable, only the second one can be expected to occur, so only this outcome is a cooperative equilibrium. Our explanation for this is that only  $(u', w'; \mu)$  is stable.

The connection between the core, the set of stable payoffs and the set of competitive payoffs, exhibited in this example, can be better understood via Fig. 1. Now, if the reader prefers, sets  $B$  and  $Q$  are better interpreted as being the set of buyers and the set of sellers, respectively.

In Fig. 1,  $C(W)$  is the set of seller's total payoffs, which



Two-Sided Matching Models, Figure 1

Core, stable and competitive equilibrium payoffs in Example 5

can be derived from any core payoff. The segment  $OP'$  is the set of sellers total payoffs which can be derived from any stable payoff. That is,  $(W_q, W_{q'}) \in OP'$  if and only if there is a stable outcome  $(u, w; \mu)$  such that  $W_q = w_{bq}$  and  $W_{q'} = w_{bq'} + w_{b'q'}$ . The segment  $OP$  is the set of seller's total payoffs, which can be derived from any competitive equilibrium price. That is,  $(W_q, W_{q'}) \in OP$  if and only if there is a competitive equilibrium price  $p$  such that  $W_q = p_q$  and  $W_{q'} = p_{q'} + p_{q'}$ .

We can see that  $C(W)$  is bigger than  $OP'$  which, in its turn, is bigger than  $OP$ . The point  $(2, 3) \in C(W) - OP'$ . It corresponds to the outcome  $(u, w, \mu)$  described above that is in the core but is not stable.

It is clear in Fig. 1 that  $C(W)$  is not a lattice, so the core is not a lattice under neither  $\geq_B$  nor  $\geq_Q$ . In fact, the outcome which corresponds to the point  $(3, 0)$  gives the individual payoff of 3 to  $q$  and two individual payoffs of zero to  $q'$ . On the other hand, the core outcome that corresponds to  $(2, 3)$  gives the individual payoff of two to  $q$ . Then, the infimum (respectively supremum) of these two core payoffs under  $\geq_Q$  (respectively  $\geq_B$ ) gives payoff two to seller  $q$  and two individual payoffs of zero to seller  $q'$ . This payoff corresponds to the vector of total payoffs  $(2, 0)$ , which is not in  $C(W)$ .

It is evident that the stable payoff corresponding to point  $P' = (3, 5)$  is the  $Q$ -optimal stable payoff. Seller  $q$  receives three and seller  $q'$  receives two from  $b$  and three from  $b'$ . Buyers get zero from the sellers. By applying the function  $f$  we obtain the  $Q$ -optimal competitive equilibrium payoff, corresponding to point  $P = (3, 4)$ , where  $q$  receives three and  $q'$  receives two from both buyers.

Point  $O = (3, 0)$  corresponds to the outcome  $(u'', w''; \mu)$ , where  $u''_{bq} = 0$ ,  $u''_{bq'} = 2$ ,  $u''_{b'q'} = 3$ ,  $u''_{b'0} = 0$ ;  $w''_{bq} = 3$ ,  $w''_{bq'} = 0$ ,  $w''_{b'q'} = 0$ . Payoff  $(u'', w'')$  is the  $B$ -optimal stable payoff, the  $B$ -optimal competitive equilibrium



payoff and the  $B$ -optimal core payoff. It is the worst stable payoff and the worst competitive equilibrium payoff for the sellers. However, it is not the worst core payoff for the sellers since it is the best core payoff for  $q$ . Indeed, as it can be observed, there is no minimum core payoff for the sellers.

A fruitful line of investigation has been the design of mechanisms to produce a competitive equilibrium price. Demange, Gale and Sotomayor [19] propose a generalization of the English auction for the assignment game, which yields the minimum competitive equilibrium price in a finite number of steps. In the same spirit, Sotomayor [83] presents a descending bid auction mechanism which leads the maximum competitive equilibrium price as a generalization of the Dutch auction. Gül and Stacchetti [34] obtain the minimum competitive equilibrium price through a generalization of the auction of Demange et al. [19] to the many-to-one case in which the utility functions satisfy the gross substitute condition. Sotomayor [90] obtains the same result by considering a dynamic mechanism for the many-to-many case with additively separable utility functions. This mechanism leads to the  $B$ -optimal stable payoff. Using the symmetry of the model, the  $Q$ -optimal stable payoff is obtained by reverting the roles between buyers and sellers in the mechanism.

A number of works related to the assignment game can be found in the literature. Demange and Gale [18] generalize the assignment game by allowing agents' preferences to be represented by any continuous utility function in the money variable. For that model, Roth and Sotomayor [67] generalize a previous result of Rochford [58], using Tarski's fixed point theorem, and show that the set of fixed points of a "rebargaining" function is a subset of the core, which maintains the lattice structure of the core. Another approach of the many-to-many assignment game with additively separable utilities is treated in Sotomayor [75]. There, agents are not allowed to negotiate their individual payments and act in blocks. An outcome only specifies their total payoffs. In this model, the core coincides with the set of stable payoffs and is not a lattice. Also, pairwise-stability is not equivalent to corewise-stability. Sotomayor [84] considers an extension of the assignment game to a many-to-many matching model in the context of firms and workers in which the quotas of the agents are not the number of partnerships they are allowed to form. Instead, they are given by the units of labor time they can supply or employ. Bikhchandani and Mamer [10] analyze the existence of market clearing prices in an exchange economy in which agents have interdependent values over several indivisible objects. Although an agent can

be both a buyer and a seller, such an exchange economy can be transformed into a many-to-one matching market where each seller owns only one object and buyers want to buy a bundle of objects, and can be viewed as an extension of the assignment game. See also Demange [17], Leonard [48], Perez-Castrillo and Sotomayor [56], Sotomayor [86], Thompson [97] and Kaneko [39].

### Hybrid One-to-One Matching Model

The hybrid one-to-one matching model is the name given in the literature to a unified model due to Eriksson and Karlander [29] and inspired in the unification proposed in Roth and Sotomayor [70]. Agents from the marriage market and the assignment market are put together so that they can trade with each other in the same market. We can interpret the hybrid model as being a labor market of firms and workers:  $P$  is the set of firms and  $Q$  is the set of workers. There are two classes of agents in each set: rigid agents and flexible agents. For each pair  $(p, q) \in P \times Q$  there is a number  $c_{pq}$  representing the productivity of the pair. If a firm  $p$  hires worker  $q$  and both agents are flexible then the number  $c_{pq}$  is allocated into salary  $v_q$  for the worker and profit  $u_p$  for the firm as a result of a negotiation process. If one of the agents is rigid, then the payoffs of the agents are pre-set and fixed and are part of the job description. In this case the profit of  $p$  and the salary of  $q$  will be  $a_{pq}$  and  $b_{pq}$ , respectively. The definitions of feasible outcome, corewise-stable outcomes and setwise-stable outcomes are straightforward extensions from the respective concepts for the non-hybrid models. Therefore, the concepts of setwise-stability and corewise-stability are equivalent.

This model is motivated by the fact that, in practice, a wide range of real-world matching markets are neither completely discrete nor completely continuous. In the United States, for example, new law school graduates may enter the market for associate positions in private law firms, which negotiate salaries, or they may seek employment as law clerks to federal circuit court judges, which are civil service positions with predetermined fixed salaries. In the market for academic positions and professors, for example, the American universities compete with each other in terms of salaries, while the French public universities offer a preset and fixed salary. In Brazil, new professors may enter the market for permanent positions (with preset and fixed salaries) in federal universities, or they may seek employment in private universities, which do not offer such positions, but compensate the entrants with better and negotiable salaries.

Eriksson and Karlander [29] present an algorithm to find a stable outcome under the assumption that the num-

bers  $a_{pq}$ ,  $b_{pq}$  and  $c_{pq}$  are integer numbers of some unit. A non-constructive proof of the existence result without imposing any restriction is provided in Sotomayor [81]. In this paper it is proved that, under the assumption that the core,  $C$ , is equal to the strong core,  $C^*$ , the main properties that characterize the stable payoffs of the marriage and of the assignment models carry over to the hybrid model. That is, for the hybrid model,

- C1. Let  $(u, w)$  and  $(u', w')$  be stable payoffs for the hybrid model. If  $C = C^*$  then  $u \geq u'$  if and only if  $w' \geq w$ .
- C2. If  $C = C^*$  and an agent is unmatched under some stable outcome then he/she gets a zero payoff under any other stable outcome.
- C3. If  $C = C^*$  then the set of stable payoffs forms a complete lattice under the partial orders  $\geq_P$  and  $\geq_Q$ .
- C4. If  $C = C^*$  then there is a  $P$ -optimal stable payoff  $(u_*, w_*)$  with the property that for any stable payoff  $(u, w)$ ,  $u_* \geq u$  and  $w_* \leq w$ ; there is a  $Q$ -optimal stable payoff  $(u_*, w_*)$  with symmetrical properties.

Sotomayor [92] studies the hybrid model without imposing the assumption that the core is equal to the strong core. Instead she assumes that the preferences of the rigid agents, as well as the preferences of the flexible agents over rigid agents, are strict. It is shown there that the core of the hybrid model has a non-standard algebraic structure given by the disjoint union of complete lattices endowed with the properties above. The extreme points of the lattices of the core partition are called *quasi-optimal stable payoffs for firms* and *quasi-optimal stable payoffs for workers*. When the workers are always flexible, then the marriage market is obtained when the flexible firms leave the hybrid market and the assignment game is obtained when the rigid firms leave the hybrid market.

Each subset of the core partition of the hybrid model is obtained as follows. For any matching  $\mu$  which is compatible with some stable payoff, decompose the market participants into two disjoint subsets. One subset contains all rigid firms and their mates at  $\mu$  and the other one contains all flexible firms, their mates at  $\mu$  and the unmatched workers. Now, fix such a partition of the agents. The desired subset  $C(\mu)$  of the core partition is formed with the core payoffs  $(u, w; \mu')$  such that all agents in the first set are matched among themselves under  $\mu'$  and all agents in the second set are matched among themselves under  $\mu'$ .

Clearly, as the rigid firms exit the hybrid market, the core partition for the corresponding assignment market is reduced to only one set, since any stable matching is compatible with any core payoff by property B1. An analogous result holds as the flexible firms leave the hybrid market, due to the fact that the matched agents in the mar-

riage market are the same at every stable matching, which is implied by property A2. Therefore, as all flexible firms leave the hybrid market, or as all rigid firms leave the hybrid market, the restriction of the algebraic structure to the core of the resulting non-hybrid market is that of a complete lattice. Then, the extreme points of the resulting lattice are exactly the firm-optimal and the worker-optimal stable payoffs.

This algebraic structure was used in Sotomayor [92] to investigate the comparative effects on the quasi-optimal stable payoffs for firms and on the quasi-optimal stable payoffs for workers caused by the entrance of rigid firms into the assignment market or by the entrance of flexible firms into the marriage model. The results of that paper can be summarized as follows: *Whether agents are allocated according to a quasi-optimal stable payoff for firms or according to a quasi-optimal stable payoff for workers, it will always be the case that if flexible firms enter the rigid market, no rigid firm will be made better off and no worker will be made worse off; if rigid firms enter the flexible market, no flexible firm will be made better off and no worker will be made worse off.*

Comparative static results of adding agents from the same side to the marriage market or to the assignment market have been obtained in the literature under the assumption that the agents are allocated according to the optimal-stable outcome for firms or according to the optimal-stable outcome for workers. However, in the approach considered in Sotomayor [92],

1. *The firms that are added are different from the firms, which are already in the market.* For example, in the marriage market, where utility is non-transferable, the comparative static adds firms with flexible wages who can transfer utility.
2. *The points which are compared belong to cores with quite distinct algebraic structures.*
3. *There may exist several quasi-optimal stable outcomes for firms and several quasi-optimal stable outcomes for workers in the hybrid model. Despite the multiplicity of these outcomes, all of them reveal the same kind of comparative static effects.*

Therefore, the result above has no parallel in the non-hybrid models.

It is argued in Sotomayor [92] that if the resulting core partition is not reduced to only one set when, say, the flexible firms leave the hybrid model, the comparative statics may be meaningless. This happens, for example, if we define a set of the core partition as the set of all stable payoffs compatible with some given matching. Then each lattice of the core partition for the marriage market has only one

stable matching, which is both the supremum and the infimum of the lattice. Of course the distinctions between, say, the best stable payoff for workers of some lattice of the core partition of the hybrid market and an arbitrary core point of the marriage model, cannot be attributed to the entrance of the flexible firms into the marriage market.

Results of comparative statics were originally obtained by Gale and Sotomayor [32] for the marriage model and the college admission model: *If agents from the same side of the market enter the market then no agent from this side is better off and no agent of the opposite side is worse off, if any of the two optimal stable matchings prevails.* A similar result was proved by Demange and Gale [18] for a continuous one-to-one matching model that includes the assignment game.

For the assignment game, Shapley [71] showed that the optimal stable payoff for an agent weakly decreases when another agent is added to the same side and weakly increases when another agent is added to the other side. Still with regard to the assignment game, Mo [53] showed that if the incoming worker is allocated to some firm in some stable outcome for the new market, there is a set of agents such that every firm is better off and every worker is worse off in the new market than in the previous one. A symmetric result holds when the incoming agent is a firm. An analogous result is demonstrated by Roth and Sotomayor [69] for the marriage market.

For the many-to-one matching markets with substitutable preferences, Kelso and Crawford [42] showed that, within the context of (flexible) firms and workers, *the addition of one or more firms to the market weakly improves the workers' payoffs, and the addition of one or more workers weakly improves the firms' payoffs, under the firm-optimal stable allocation.* Similar conclusions were obtained by Crawford [14] for a many-to-many matching model with strict and substitutable preferences, by comparing pairwise-stable outcomes instead of setwise-stable outcomes.

## Incentives

The strategic questions that emerge when a stable revelation mechanism is adopted concerns its non-manipulability and, for the games induced by the mechanism, the existence of strategic equilibria and the implementability of the set of stable matchings via such equilibria. For the marriage model with strict preferences the equilibrium analysis of a game induced by a stable matching mechanism leads to the following results:

1. (Impossibility Theorem) (Roth and Sotomayor, [69]) *When any stable mechanism is applied to a marriage market in which preferences are strict and there is more*

*than one stable matching, then at least one agent can profitably misrepresent his or her preferences, assuming the others tell the truth. (This agent can misrepresent in such a way as to be matched to his or her most preferred achievable mate under the true preferences at every stable matching under the mis-stated preferences).*

2. (Limits on successful manipulation.) (Demange, Gale, and Sotomayor [20]). *Let  $P$  be the true preferences (not necessarily strict) of the agents, and let  $P'$  differ from  $P$  in that some coalition  $C$  of men and women misstate their preferences. Then there is no matching  $\mu$ , stable for  $P'$ , which is preferred to every stable matching under the true preferences  $P$  by all members of  $C$ .*

A corollary of this result is due to Dubins and Freedman [21] which states that *the man-optimal stable matching mechanism is non-manipulable, individually and collectively, by the men.*

3. (Gale and Sotomayor [31]) *When all preferences are strict, let  $\mu$  be any stable matching for  $(F, W, P)$ . Suppose each woman  $w$  in  $\mu(F)$  chooses the strategy of listing only  $\mu(w)$  on her stated preference list of acceptable men (and each man states his true preferences). This is a Nash equilibrium in the game induced by the man-optimal stable matching mechanism (and  $\mu$  is the matching that results).*
4. (Roth, [61]) *Suppose each man chooses his dominant strategy and states his true preferences, and the women choose any set of strategies (preference lists)  $P'(w)$  that form a Nash equilibrium for the revelation game induced by the man-optimal stable mechanism. Then the corresponding man-optimal stable matching for  $(F, W, P')$  is one of the stable matchings for  $(F, W, P)$ .*
5. (Gale and Sotomayor, [31]) *Suppose each man chooses his dominant strategy and states his true preferences, and the women truncate their true preferences at the mate they get under the woman-optimal stable mechanism. This profile of preferences is a strong equilibrium for the women in the game induced by the man-optimal stable mechanism (and the woman-optimal stable matching under the true preferences is the matching that results).*

Results (3) and (4) imply that the man-optimal stable mechanism implements the core correspondence via Nash equilibria.

For the college admission model with responsive and strict preferences, the theorem of Dubins and Freedman implies that *the student-optimal stable mechanism is non-manipulable individually and collectively by the students.*

Roth [65] shows through an example that *the college-optimal stable mechanism is manipulable by the colleges* due to the fact that the colleges may have a quota greater than one.

Sotomayor [78,82,94] analyzes the strategic behavior of the students in a school choice model where participants have strict preferences over individuals. This paper proves that *the college-optimal stable mechanism implements the set of stable matchings via the Nash equilibrium concept*. When some other stable mechanism is used, an example shows that the strategic behavior of the students may lead to unstable matchings under the true preferences. A sufficient condition for the stability of the Nash equilibrium outcome is then proved to be that the set of stable matchings for the Nash equilibrium profile is a singleton. A random stable matching mechanism is proposed and the Nash equilibrium concept ex-ante is shown to be equivalent to the Nash equilibrium concept ex-post of the game induced by such a mechanism. This refinement of the Nash equilibrium concept is called *Nash equilibrium in the strong sense*. *Under this equilibrium concept any stable matching mechanism (and in particular the random stable matching mechanism) implements the set of stable matchings*. Also, *if the students only play truncations of the true preferences, any stable matching mechanism implements the student-optimal stable matching via strong equilibrium in the strong sense and Nash equilibrium in the strong sense*.

Ergin and Sönmez [28] and Pathak and Sönmez [55] analyze the Boston mechanism, which is used to assign students to schools in many cities in the US and show that students' parents do not have incentives to report preferences truthfully.

Ma [49] analyzes the strategic behavior of both students and colleges in the college admission model with responsive preferences. This author proves that *the set of stable matchings is implemented by any stable mechanism via rematching proof equilibrium and strong equilibrium in truncation strategies at the match point*.

The implementability of the set of stable matchings through stable and non-necessarily stable mechanisms has also been investigated by several authors. Alcalde [5] presents a mechanism for the marriage market closely related to the algorithm of Gale and Shapley, which implements the core correspondence in undominated equilibria. Kara and Sönmez [40] analyze the problem of implementation in the college admission market. They show that the set of stable matchings is implementable in Nash equilibrium. Nevertheless, no subset of the core is Nash implementable. Romero-Medina [59] studies the mechanism employed by the Spanish universities to distribute the students to colleges, which can produce unsta-

ble matchings for the stated preferences. However, when students play in equilibrium, only stable allocations are reached. Sotomayor [85] investigates a mechanism for the marriage model which is not designed for producing stable matchings. Here also the equilibrium outcomes are stable matchings under the true preferences.

For the discrete many-to-one matching model with responsive preferences, Alcalde and Romero-Medina [7] analyze the following mechanism: firms announce a set of workers they want to hire. Then each worker selects the firm she wants to work for. This paper proves that such a mechanism implements the set of stable matchings in subgame perfect Nash equilibrium. For the many-to-many case, Sotomayor [88] shows that this result does not carry over. This paper proves that subgame perfect Nash equilibria always exist, while strong equilibria may not exist. The subgame perfect Nash equilibrium outcomes are precisely the pairwise-stable matchings, which may be out of the core when the preferences of the agents in one of the sides are not maximin. Under this condition, the equilibrium outcomes are the setwise-stable matchings and every subgame perfect Nash equilibrium is a strong equilibrium.

By assuming non-strict preferences, Abdulkadiroğlu, Pathak, Roth and Sönmez [2] show that no mechanism (stable or not, and Pareto optimal or not), which is better for the students than the student proposing deferred acceptance algorithm with tie breaking, can be strategy proof.

Sönmez [73] analyzes a model which he calls the *generalized indivisible allocation problem* that includes the roommate and the marriage markets. He looks for conditions which explain the differences on strategy-proofness results that have been generated in the literature. He shows how some of the results in the literature can be seen as corollaries of his results.

Ehlers and Massó [27] study Bayesian Nash equilibria of stable mechanisms (such as the NRMP) in matching markets under incomplete information. They show that truth-telling is an equilibrium of the Bayesian revelation game induced by a common belief and a stable mechanism if and only if all the profiles in the support of the common belief have singleton cores.

For the continuous matching models the idea is to use competitive equilibrium as an allocation mechanism to produce outcomes with the desirable properties of fairness and efficiency. It involves having agents specify their supply and demand functions. The competitive equilibria are then calculated and allocations are made accordingly. Demange [17] and Leonard [48] considered the assignment game of Shapley and Shubik and, independently, proved that the allocation mechanism that yields the minimum



competitive equilibrium prices is individually non-manipulable by the buyers. Demange and Gale [18] consider a one-to-one matching model in which the utilities are continuous functions in the money variable and not necessarily additively separable. These authors prove a sort of non-manipulability theorem which states that *if the mechanism which produces the buyer-optimal stable payoff is adopted then no coalition of buyers by falsifying demands can achieve, only through the mechanism, higher payoffs to all of its members.* We added *only through the mechanism* because this model allows monetary transfers within any coalition. As in the marriage model, this result is an immediate consequence of a more general theorem due to Sotomayor [74] which states the following: *Let  $(u', w'; \mu)$  be any stable outcome for the market  $M'$  where  $B' \cup Q'$  is the set of agents who misrepresent their utility functions. Let  $(u^*, w^*)$  be the true payoff under  $(u', w'; \mu)$ . Then, there exists a stable payoff  $(u, w)$  for the original market such that  $u_b \geq u^*_b$  for at least one  $b$  in  $B'$  or  $w_q \geq w^*_q$  for at least one  $q$  in  $Q'$ .*

Demange and Gale [18] also addresses the strategic behavior by the sellers when the mechanism produces the buyer-optimal stable payoff. These authors show that by specifying their supply functions appropriately the sellers can force, by strong Nash equilibrium strategies, the payoff to be given by the maximum rather than the minimum equilibrium price. Under the assumption that the sellers only manipulate their reservation prices then, if a profile of strategies does not give the maximum equilibrium price allocation, then either some seller is using a dominated strategy or the strategy profile is not a Nash equilibrium. For this model Sotomayor [74] proves that the outcome produced by a Nash equilibrium strategy is stable for the original market.

Sotomayor [87] considers, for the assignment game of Shapley and Shubik, the strategic games induced by a class of market clearing price mechanisms. In these procedures, buyers and sellers, in different stages reveal their demand and supply functions and a competitive equilibrium is produced by the mechanism. For each vector of reservation prices selected by the sellers, the buyers play the subgame that starts and can force the buyer-optimal stable payoff through Nash equilibrium strategies. However sellers can reverse this outcome by forcing the subgame perfect equilibrium allocation to be the seller-optimal stable payoff for the original market.

Kamecke [38] and Perez-Castrillo and Sotomayor [56] consider the assignment game of Shapley and Shubik. The former paper presents two mechanisms for this market. In the first one, agents act simultaneously. In the second game, the strategies are chosen sequentially. These mech-

anisms implement the social choice correspondences that yield the core and the optimal stable payoff for the sellers, respectively. The second paper analyzes a sequential mechanism, which implements the social choice correspondence that yields the optimal stable payoff for the sellers.

## Future Directions

In this section we present some directions for future investigations and some open problems which have intrigued matching theorists.

1. The discrete two-sided matching models with non-necessarily strict preferences have been explored very little in the literature. In the discrete models under strict preferences and in the continuous models, due to the fact that there is no weak blocking pairs, the set of Pareto-stable outcomes coincides with the set of setwise-stable outcomes. However, under weak preferences, setwise-stable matchings may not be Pareto-optimal. Sotomayor [93] proposes that in this case the Pareto-stability concept, which requires that the matching is stable and Pareto optimal, should be considered the natural solution concept. The justification for the Pareto-stability concept relies in the argument that in a decentralized setting, where agents freely get together in groups, recontracts between pairs of agents already allocated according to a stable matching leading to a (weak) Pareto improvement of the original matching should be allowed. Thus, weak blockings can upset a matching once they come from the grand coalition. We think that the study of the discrete two-sided matching models with non-necessarily strict preferences and the search for algorithms to produce the Pareto-stable matchings is a new and interesting line of investigation.
2. Consider the hybrid model where no worker is rigid. If rigid firms enter the flexible market or flexible firms enter the rigid market then no firm gains and no worker loses if a quasi-optimal stable outcome for one of the sides always prevails. Suppose now that some rigid firm becomes flexible or some flexible firm becomes rigid. What kind of comparative static effect is caused by this change in the market?
3. One line of investigation not yet explored in the literature concerns the incentives faced by the agents in the hybrid model when some stable allocation mechanism is used.
4. Consider the discrete many-to-many matching market with substitutable preferences where a matching  $\mu$  is



feasible if  $Ch_y(\mu(y)) = \mu(y)$  for every agent  $y$ . Is the core always non-empty for this model?

5. Consider the assignment game of Shapley and Shubik in the context of buyers and sellers. Consider a sealed bid auction in which the buyers select a monetary value for each of the items. The auctioneer then chooses a competitive equilibrium price vector for the profile of selected values, according to some pre-set probability distribution. It is of theoretical interest the investigation of the buyers' strategic behavior.
6. We know that the core of the many-to-many assignment model of Sotomayor [75] in which the agents negotiate in blocks is not a lattice. However, a problem that is still open is to know if the optimal stable payoffs for each side of the market always exist.

## Bibliography

1. Abdulkadiroglu A, Sönmez T (2003) School Choice: A Mechanism Design Approach. *Am Econ Rev* 93(3):729–747
2. Abdulkadiroglu A, Pathak P, Roth A, Sönmez T (2006) Changing the Boston School Choice Mechanism: Strategy-proofness as Equal Access, working paper. Boston College and Harvard University, Boston
3. Abeledo H, Isaak G (1991) A characterization of graphs which assure the existence of stable matchings. *Math Soc Sci* 22(1):93–96
4. Adachi H (2000) On a characterization of stable matchings. *Econ Lett* 68(1):43–49
5. Alcalde J (1996) Implementation of Stable Solutions to Marriage Problems. *J Econ Theory* 69(1):240–254
6. Alcalde J, Pérez-Castrillo D, Romero-Medina A (1998) Hiring Procedures to Implement Stable Allocations. *J Econ Theory* 82(2):469–480
7. Alcalde J, Romero-Medina A (2000) Simple Mechanisms to Implement the Core of College Admissions Problems. *Games Econ Behav* 31(2):294–302
8. Balinski M, Sönmez T (1999) A Tale of Two Mechanisms: Student Placement. *J Econ Theory* 84(1):73–94
9. Bardella F, Sotomayor M (2006) Redesign and analysis of an admission market to the graduate centers of economics in Brazil: a natural experiment in market organization, working paper. Universidade de São Paulo, São Paulo
10. Bikhchandani S, Mamer J (1997) Competitive Equilibrium in an Exchange Economy with Indivisibilities. *J Econ Theory* 74(2):385–413
11. Birkhoff G (1973) Lattice theory, vol v. 25. Colloquium publications. American Mathematical Society, Providence
12. Blair C (1988) The Lattice Structure of the Set of Stable Matchings with Multiple Partners. *Math Oper Res* 13(4):619–628
13. Chung K (2000) On the Existence of Stable Roommate Matchings. *Games Econ Behav* 33(2):206–230
14. Crawford V (1991) Comparative statics in matching markets. *J Econ Theory* 54(2):389–400
15. Crawford V (2008) The Flexible-Salary Match: A Proposal to Increase the Salary Flexibility of the National Resident Matching Program. *J Econ Behav Organ* 66(2):149–160
16. Crawford V, Knoer E (1981) Job Matching with Heterogeneous Firms and Workers. *Econometrica* 49(2):437–450
17. Demange G (1982) Strategyproofness in the assignment market game, working paper. Ecole Polytechnique, Laboratoire D'Econometrie, Paris
18. Demange G, Gale D (1985) The Strategy Structure of Two-Sided Matching Markets. *Econometrica* 53(4):873–888
19. Demange G, Gale D, Sotomayor M (1986) Multi-Item Auctions. *J Political Econ* 94(4):863–872
20. Demange G, Gale D, Sotomayor M (1987) A further note on the stable matching problem. *Discret Appl Math* 16(3):217–222
21. Dubins L, Freedman D (1981) Machiavelli and the Gale-Shapley Algorithm. *Am Math Mon* 88(7):485–494
22. Dutta B, Massó J (1997) Stability of Matchings When Individuals Have Preferences over Colleagues. *J Econ Theory* 75(2):464–475
23. Echenique F, Oviedo J (2004) Core many-to-one matchings by fixed-point methods. *J Econ Theory* 115(2):358–376
24. Echenique F, Oviedo J (2006) A theory of stability in many-to-many matching markets. *Theor Econ* 1(2):233–273
25. Echenique F, Yenmez M (2007) A solution to matching with preferences over colleagues. *Games Econ Behav* 59(1):46–71
26. Eeckhout J (2000) On the uniqueness of stable marriage matchings. *Econ Lett* 69(1):1–8
27. Ehlers L, Massó J (2004) Incomplete Information and Small Cores in Matching Markets, working paper. CREA, Barcelona
28. Ergin H, Sönmez T (2006) Games of school choice under the Boston mechanism. *J Public Econ* 90(1–2):215–237
29. Eriksson K, Karlander J (2000) Stable matching in a common generalization of the marriage and assignment models. *Discret Math* 217(1):135–156
30. Gale D, Shapley L (1962) College Admissions and the Stability of Marriage. *Am Math Mon* 69(1):9–15
31. Gale D, Sotomayor M (1985) Ms. Machiavelli and the Stable Matching Problem. *Am Math Mon* 92(4):261–268
32. Gale D, Sotomayor M (1983, 1985) Some remarks on the stable matching problem. *Discret Appl Math* 11:223–232
33. Gül F, Stacchetti E (1999) Walrasian Equilibrium with Gross Substitutes. *J Econ Theory* 87(1):95–124
34. Gül F, Stacchetti E (2000) The English Auction with Differentiated Commodities. *J Econ Theory* 92(1):66–95
35. Gusfield D (1988) The Structure of the Stable Roommate Problem: Efficient Representation and Enumeration of All Stable Assignments. *SIAM J Comput* 17:742–769
36. Hatfield J, Milgrom P (2005) Matching with Contracts. *Am Econ Rev* 95(4):913–935
37. Irving R (1985) An efficient algorithm for the stable roommates problem. *J Algorithms* 6:577–595
38. Kamecke U (1989) Non-cooperative matching games. *Int J Game Theory* 18(4):423–431
39. Kaneko M (1982) The central assignment game and the assignment markets. *J Math Econ* 10(2–3):205–232
40. Kara T, Sönmez T (1996) Nash Implementation of Matching Rules. *J Econ Theory* 68(2):425–439
41. Kara T, Sönmez T (1997) Implementation of college admission rules. *J Econ Theory* 9(2):197–218
42. Kelso Jr A, Crawford V (1982) Job Matching, Coalition Formation, and Gross Substitutes. *Econometrica* 50(6):1483–1504
43. Kesten O (2004) Student placement to public schools in the US: Two new solutions, working paper. University of Rochester, Rochester

44. Kesten O (2006) On two competing mechanisms for priority-based allocation problems. *J Econ Theory* 127(1):155–171
45. Knuth D (1976) *Marriage Stables*. Les Presses de l'Université de Montréal, Montréal
46. Konishi H, Ünver MU (2006) Credible group stability in many-to-many matching problems. *J Econ Theory* 127(1):57–80
47. Kraft C, Pratt J, Seidenberg A (1959) Intuitive Probability on Finite Sets. *Ann Math Stat* 30(2):408–419
48. Leonard H (1983) Elicitation of Honest Preferences for the Assignment of Individuals to Positions. *J Political Econ* 91(3):461–479
49. Ma J (2002) Stable matchings and the small core in Nash equilibrium in the college admissions problem. *Rev Econ Des* 7(2):117–134
50. Martínez R, Massó J, Neme A, Oviedo J (2001) On the lattice structure of the set of stable matchings for a many-to-one model. *Optimization* 50(5):439–457
51. Martínez R, Massó J, Neme A, Oviedo J (2004) An algorithm to compute the full set of many-to-many stable matchings. *Math Soc Sci* 47(2):187–210
52. McVitie D, Wilson L (1970) Stable marriage assignment for unequal sets. *BIT Numer Math* 10(3):295–309
53. Mo J (1988) Entry and structures of interest groups in assignment games. *J Econ Theory* 46(1):66–96
54. Ostrovsky M (2008) Stability in Supply Chain Networks. *Am Econ Rev* 98(3):897–923
55. Pathak P, Sönmez T (2006) Leveling the Playing Field: Sincere and Strategic Players in the Boston Mechanism, working paper. Boston College and Harvard University, Boston
56. Pérez-Castrillo D, Sotomayor M (2002) A Simple Selling and Buying Procedure. *J Econ Theory* 103(2):461–474
57. Pycia M (2007) Many-to-One Matching with Complementarities and Peer Effects, working paper. Penn State Working Paper, Pennsylvania
58. Rochford S (1984) Symmetrically Pairwise-Bargained Allocations in an Assignment Market. *J Econ Theory* 34(2):262–281
59. Romero-Medina A (1998) Implementation of stable solutions in a restricted matching market. *Rev Econ Des* 3(2):137–147
60. Roth A (1982) The Economics of Matching: Stability and Incentives. *Math Oper Res* 7(4):617–628
61. Roth A (1984) Misrepresentation and Stability in the Marriage Problem. *J Econ Theory* 34(2):383–387
62. Roth A (1984) Stability and Polarization of Interests in Job Matching. *Econometrica* 52(1):47–58
63. Roth A (1984) The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory. *J Political Econ* 92(6):991–1016
64. Roth A (1985) Conflict and Coincidence of Interest in Job Matching: Some New Results and Open Questions. *Math Oper Res* 10(3):379–389
65. Roth A (1985) The College Admissions Problem is not Equivalent to the Marriage Problem. *J Econ Theory* 36(2):277–288
66. Roth A (1986) On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets. *Econometrica* 54(2):425–427
67. Roth A, Sotomayor M (1988) Interior points in the core of two-sided matching markets. *J Econ Theory* 45(1):85–101
68. Roth A, Sotomayor M (1989) The College Admissions Problem Revisited. *Econometrica* 57(3):559–570
69. Roth A, Sotomayor M (1990) Two-Sided Matching: A Study in Game-Theoretic Modeling and Analysis, vol 18. In: *Econometric Society Monographs*. Cambridge University Press, New York
70. Roth A, Sotomayor M (1996) Stable outcomes in discrete and continuous models of two-sided matching: A unified treatment. *Braz Rev Econom* 16:1–4
71. Shapley L (1962) Complements and substitutes in the Optimal Assignment Problem. *Navals Res Logist Q* 9:45–48
72. Shapley L, Shubik M (1972) The Assignment Game I: The core. *Int J Game Theory* 1(1):111–130
73. Sönmez T (1999) Strategy-Proofness and Essentially Single-Valued Cores. *Econometrica* 67(3):677–689
74. Sotomayor M (1986) On incentives in a two-sided matching market, working paper. Depart of Mathematics, PUC/RJ, Rio de Janeiro
75. Sotomayor M (1992) The multiple partners game in Equilibrium and dynamics. In: Majumdar M (ed) *Essays in honour of David Gale*. MacMillan Press Ltd, New York, pp 322–336
76. Sotomayor M (1996) A Non-constructive Elementary Proof of the Existence of Stable Marriages. *Games Econ Behav* 13(1):135–137
77. Sotomayor M (1996) Admission mechanisms of students to colleges. A game-theoretic modeling and analysis. *Braz Rev Econom* 16(1):25–63
78. Sotomayor M (1998) The strategy structure of the college admissions stable mechanisms. In: *Annals of Jornadas Latino Americanas de Teoria Económica*, San Luis, Argentina, 2000; *First World Congress of the Game Theory Society*, Bilbao, 2000; *4th Spanish Meeting*, Valencia, 2000; *International Symposium of Mathematical Programming*, Atlanta, 2000; *World Congress of the Econometric Society*, Seattle, 2000, [http://www.econ.fea.usp.br/marilda/artigos/roommates\\_1.doc](http://www.econ.fea.usp.br/marilda/artigos/roommates_1.doc)
79. Sotomayor M (1999) The lattice structure of the set of stable outcomes of the multiple partners assignment game. *Int J Game Theory* 28(4):567–583
80. Sotomayor M (1999) Three remarks on the many-to-many stable matching problem. *Math Soc Sci* 38(1):55–70
81. Sotomayor M (2000) Existence of stable outcomes and the lattice property for a unified matching market. *Math Soc Sci* 39(2):119–132
82. Sotomayor M (2000) Reaching the core through college admissions stable mechanisms. In: *Annals of abstracts of the following congresses: International Conference on Game Theory*, 2001, Stony Brook; *Annals of Brazilian Meeting of Econometrics*, Salvador, Brazil, 2001; *Latin American Meeting of the Econometric Society*, Buenos Aires, Argentina, 2001, [http://www.econ.fea.usp.br/marilda/artigos/reaching\\_%20core\\_random\\_stable\\_allocation\\_mechanisms.pdf](http://www.econ.fea.usp.br/marilda/artigos/reaching_%20core_random_stable_allocation_mechanisms.pdf)
83. Sotomayor M (2002) A simultaneous descending bid auction for multiple items and unitary demand. *Revista Brasileira Economia* 56:497–510
84. Sotomayor M (2003) A labor market with heterogeneous firms and workers. *Int J Game Theory* 31(2):269–283
85. Sotomayor M (2003) Reaching the core of the marriage market through a non-revelation matching mechanism. *Int J Game Theory* 32(2):241–251
86. Sotomayor M (2003) Some further remark on the core structure of the assignment game. *Math Soc Sci* 46:261–265
87. Sotomayor M (2004) Buying and selling strategies in the assignment game, working paper. Universidade São Paulo, São Paulo

88. Sotomayor M (2004) Implementation in the many-to-many matching market. *Games Econ Behav* 46(1):199–212
89. Sotomayor M (2005) The roommate problem revisited, working paper. Universidade São Paulo, São Paulo. [http://www.econ.fea.usp.br/marilda/artigos/THE\\_ROOMMATE\\_PROBLEM\\_REVISITED\\_2007.pdf](http://www.econ.fea.usp.br/marilda/artigos/THE_ROOMMATE_PROBLEM_REVISITED_2007.pdf)
90. Sotomayor M (2006) Adjusting prices in the many-to-many assignment game to yield the smallest competitive equilibrium price vector, working paper. Universidade de São Paulo, São Paulo. [http://www.econ.fea.usp.br/marilda/artigos/A\\_dynam\\_stable-mechan\\_proof\\_lemma\\_1.pdf](http://www.econ.fea.usp.br/marilda/artigos/A_dynam_stable-mechan_proof_lemma_1.pdf)
91. Sotomayor M (2007) Connecting the cooperative and competitive structures of the multiple partners assignment game. *J Econ Theory* 134(1):155–174
92. Sotomayor M (2007) Core structure and comparative statics in a hybrid matching market. *Games Econ Behav* 60(2):357–380
93. Sotomayor M (2008) The Pareto-Stability concept is a natural solution concept for the Discrete Matching Markets with indifference, working paper. Universidade São Paulo, São Paulo. [http://www.econ.fea.usp.br/marilda/artigos/ROLE\\_PLAYED\\_SIMPLE\\_OUTCOMES\\_STABLE\\_COALITI\\_3.pdf](http://www.econ.fea.usp.br/marilda/artigos/ROLE_PLAYED_SIMPLE_OUTCOMES_STABLE_COALITI_3.pdf)
94. Sotomayor M (2007) The stability of the equilibrium outcomes in the admission games induced by stable matching rules. *Inter Jour Game Theory* 36(3–4):621–640
95. Tan J (1991) A Necessary and Sufficient Condition for the Existence of a Complete Stable Matching. *J Algorithms* 12(1): 154–178
96. Tarski A (1955) A lattice-theoretical fixpoint theorem and its applications. *Pac J Math* 5(2):285–309
97. Thompson G (1980) Computing the core of a market game. In: Fiacco AA, Kortane K (eds) *Extremal Methods and Systems Analysis*, vol 174. Springer, New York, pp 312–324