Network Traffic as a Self-Organized Critical Phenomenon

Kai Nagel
Steen Rasmussen
Christopher L. Barrett

SFI WORKING PAPER: 1996-08-056

SFI Working Papers contain accounts of scientific work of the authors and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author’s copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu
Network Traffic as a Self-Organized Critical Phenomena

Kai NAGEL, Steen RASMUSSEN, and Christopher L. BARRETT,

TSA-DO/SA MS-M997 and CNLS MS-B258
Los Alamos National Laboratory
Los Alamos, New Mexico, 87545
USA
Santa Fe Institute
1399 Hyde Park Road
Santa Fe, New Mexico, 87501
USA

February 22, 1996

Submitted to the Proceedings of the International Twin Conference on

Abstract
This paper presents an overview of an approach to address complexity issues and real-life engineering problems in large, urban transportation systems. In this context we discuss the fundamental problem of designing a metropolitan transportation system which is both efficient and controllable.
1 Urban transportation systems

More and more metropolitan areas worldwide suffer from a transportation demand which exceeds capacity. In many cases, it is not possible or even not desirable to extend capacity to meet the demand [1]. In consequence, a consistent management of these large, distributed transportation systems has become more and more important. Examples of such activities include the construction of fast mass transit systems, the introduction of local bus lines, design of traveler informational systems and car pooling to improve the use of current capacity, introduction of congestion pricing, and in the long term also guidance of the urban planning process towards an evolution of urban areas with lower transportation needs.

At the level of a metropolitan region, the transportation dynamics is the aggregated result of thousands or, in some cases, millions of individual trip-making decisions for the movement of people and goods between origins and destinations. Every decision is based on incomplete information of the state of the transportation system as a whole. Since complete global knowledge of every relevancy of the current (and future) state(s) of a transportation system is impossible to obtain, future information based control strategies might, to a large extent, be based strategies exploiting self-organizing properties of the systems. That would still not remove the inherent tension between global and local transportation optima. This essential tension is one of many reasons why predictability is very difficult in such systems.

One method of approaching these and other inherent complexities of the large transportation systems is to represent the systems and generate their dynamics through simulation in order to assess them. The most straightforward way seems to be a bottom-up microsimulation of the dynamics of all travelers and loads at the level of where the transport decisions are made. Starting with a generation of travel demands derived from synthesized traveler populations and consequent trip planning decisions, over production of associated traffic and eventually the consequences for congestion, travel time, air quality, and other dynamical system properties, can be all be generated, and thus analyzed. This is the approach used by the TRANSIMS project [2], which this work also is a part of.

2 TRANSIMS

The Transportation Analysis and Simulation System (TRANSIMS) is part of the multi-track Travel Model Improvement Program sponsored by the U.S. Department of Transportation and the Environmental Protection Agency. Los Alamos National Laboratory is leading its development. TRANSIMS will address issues resulting from the Intermodal Surface Transportation and Efficiency Act of 1991, such as considerations of land use policies, intermodal connectiv-
ity, and enhanced transit service. It will support analysis of potential responses to the stringent air-quality requirements of the Clean Air Act Amendments of 1990.

The TRANSIMS project objective is to develop a set of mutually supporting realistic simulations, models, and data bases that employ advanced computational and analytical techniques to create an integrated regional transportation systems analysis environment. By applying forefront computational technologies and advanced methods relevant to complex systems analysis, it will simulate the dynamic details that contribute to the complexity inherent in today's and tomorrow's transportation issues. The integrated results from the various detailed simulations will support transportation planners, engineers, and others who must address environmental pollution, energy consumption, traffic congestion, land use planning, traffic safety, intelligent vehicle efficacies, and the transportation infrastructure effect on the quality of life, productivity, and economy.

Fig. 1 illustrates the TRANSIMS architecture [2]. The TRANSIMS methods deal with individual behavioral units and proceed through several steps to estimate travel.

TRANSIMS predicts trips for individual households, residents, freight loads, and vehicles rather than for zonal aggregations of households. The Travel Demand Module (module 1 in fig. 1) generates the households and commercial activities through the creation of regional synthetic populations from census and other data. Using activity-based methods and other techniques, it then produces a travel representation of each household and traveler.

The Intermodal Route Planner (module 2 in fig. 1) involves using a demographically defined travel cost decision model particular to each traveler. Vehicle and mode availability are represented and mode choice decisions are made during route plan generation. The method estimates desired trips not made (latent demand), induced travel, and peak load spreading. This allows evaluation of different transportation control measures and travel demand measures on trip planning behaviors.

The Traffic Microsimulation (module 3 in fig. 1) executes the generated trips on the transportation network to predict the performance of individual vehicles and the transportation system. It attempts to execute every individual's travel itinerary in the region. For example, every passenger vehicle has a driver whose driving logic attempts to execute the plan, accelerates or decelerates the car, or passes as appropriate in traffic on the roadway network.

The Traffic Microsimulation produces traffic information for the Air Quality Module (module 4 in fig. 1) to estimate such things as motor vehicle fuel use, source emissions, dispersion, transport, air chemistry, meteorology, visibility, and resultant air quality. The emissions model accounts for both moving and stationary vehicles. The regional meteorological model for atmospheric circulation is supplemented by a model for local effects. The dispersion model is used for directly emitted contaminants and handles both local and urban scale
Figure 1: The TRANSIMS architecture. Viewing a metropolitan transportation system as a large dynamical system enables us to isolate dynamics of different time scale. Urban evolution which operates over years - the left part of the figure - is currently not a part of the TRANSIMS project. TRANSIMS is currently primarily concerned with the shorter time scale dynamics. It assumes a certain land use and transportation infrastructure and starts out by estimating the travel demand on a day-to-day or week-to-week basis. This estimated travel demand is then being routed including mode choices and trans shipment processes. For instance note that the transportation of fuel oil and school children poses very different constraints on the the mode of transportation. Once the routing is completed a microsimulation of the actual trips occurs which deals with the second-to-second dynamics. As an important side effect of urban traffic the mobile source pollution generation can be computed and the resulting air quality impact estimated.

problems. The air chemistry model includes dispersion, but is designed to deal with secondary pollutant production on larger scales.

An important aspect to note at this point is that all these modules describe different time scales, but always reference individual travelers. Activity planning operates on a daily or even weekly (e.g. shopping) basis; trip planning on
a link traversal time basis constraints changing approximately on a daily basis; the microsimulation on a second-by-second basis; a typical time-scale for meteorological model is of the order of 5 minutes. Yet, these modules are not only connected "downwards" as described above. Various feedbacks couple these modules. Unplannable trips will change the weekly activities of individuals; trips which in the microsimulation take much longer than planned will need replanning; etc. In the most extreme case, all submodules feed back into urban evolution and settlement patterns: Bad air quality, traffic jams or unfulfillable transportation demands all make people or businesses to relocate.

3 Travel time variance and unpredictability

The advantage of a microsimulation approach is that the system dynamics is being generated through the simulation with all its emergent properties without any explicit assumptions or aggregated models for these properties. The major disadvantages of a complete microsimulation are extremely high computational demands on one side and perhaps explanatory problems on the other. The inclusion of many details of reality may be excellent for generating a dynamics which is close to the system under investigation, but it does not necessarily lead to a better understanding of the basic (minimal) mechanisms that cause the dynamics. Therefore the TRANSIMS project also includes the investigation of much simpler and computationally less demanding models and simulations, for example the one we are going to discuss here.

One of the important issues both for analysis and for realistic simulation of transportation systems is their high variability and the effect this has on predictability. Here, we want to concentrate on one particular source of unpredictability which may very well become important in a foreseeable future: Assume that traffic management measures and modern information technology (see, e.g., [4, 5]) succeed in moving the transportation system closer towards higher efficiency. Then we face an interesting problem because in transportation systems (and presumably also in many other large, distributed man-made systems), there is a "critical" regime around maximum capacity, where the system is very sensitive to small perturbations. Small perturbations will generate large fluctuations in congestion formation and thus travel times.

To investigate this phenomenon we can initially concentrate on an extremely simplified transportation system. We only include vehicular traffic, and we assume that all vehicles are of the same type. Our system includes only single lane traffic on a circular road, and the driving dynamics is generated by only a few very basic rules. Using a cellular automata of the form (2) and (3) and a parallel update functional $U$ we can obtain a very simple dynamic traffic system – a simulation – of the form

$$\{S_t(t + 1)\} = U\{S_t(t)\}$$  \hspace{1cm} (1)
where
\[ S_i = S_i(x_i, y_i, z_i, f_{ij}, t) \]  
(2)
is the \( i \)th car (object), \( x_i \) its position on a 1-D lattice, \( y_i \) its current state (velocity), \( z_i \) its neighborhood in front (gap to next car, which is object \( j \)), and \( f_{ij} \) their interaction rules
\[ f_{ij}(x_i(t), y_i(t), z_i(t)) \rightarrow (x_i(t + 1), y_i(t + 1)), \]  
(3)
which changes the location and the internal state of current car (object). For a detailed discussion of the dynamics we refer to [3, 9, 10]. The algorithm for the dynamics is for completeness also listed in the appendix. For a general discussion of some of the mathematical properties of dynamical systems of the form (1) we refer to [14].

The critical regime effect can be seen in Figure 2. The top plot shows flow as a function of density. The middle plot shows the average time, \( t_i \), that a vehicle in the simulation needs in order to travel \( l = 750 \) meters. And the bottom plot shows the relative variance of this travel time, i.e.
\[ \sigma(t_i) := \frac{\sqrt{\langle (t_i - \langle t_i \rangle)^2 \rangle}}{t_i}. \]  
(4)
where \( \langle \ldots \rangle \) denotes the average over all cars during the simulation; \( \langle t_i \rangle \) therefore is the average travel time for all cars during the simulation. — Note the explosion of the variance near maximum flow.

On an intuitive level, this is fairly straightforward to understand: If, in light traffic, some short temporary disturbance happens (e.g. a minor accident), the queue caused by this disturbance will be dissolved very quickly after the accident has been cleared away. If the same happens in very dense traffic, it will not have any grave effect because there is congestion all over the system anyway and it just shifts the pattern. However, in between these two regimes there is a traffic density, where there are only few jams in the system, but the new jam caused by the accident has difficulties to dissolve. This is the traffic density when small disturbances, such as a minor accident, have maximum influence. — Technically, one can use the language of a directed percolation phase transition to precisely describe what happens [5]. A first order (critical) phase transition exists in the system.

4 Simple adaptive agents

Obviously, any traveler would like to avoid congestion if possible. Given a transportation network, travelers will try to route the trips around congested areas if alternative routes are not too long or too costly. To see what this routing behavior does with the overall dynamics in a transportation network we can formulate a minimal traffic network. Here the travelers have individual
Figure 2: Throughput, travel time, and variations of travel time as a function of density. Note the explosion of the variance near maximum flow.
Figure 3: Schematic sketch of the network used for the simulations. Vehicles drive from A to B and can choose between the direct route and the much longer alternative route. On the direct route they encounter a bottleneck. Other vehicles drive from C to D.

Routing plans and can make decisions about which route they want to take depending on knowledge of congestion. They can also re-plan depending on their earlier experiences of travel time.

Imagine (see Fig. 3) a road from A to B with capacity \( q_{\text{max}} \), with a bottleneck with capacity \( q_{\text{bn}} \) shortly before B. Further imagine that there exists an alternative, but longer route between A and B. On the direct route from A to B additional travelers from C have to go to destination D. First assume that there are no travelers with origin in C.

If many drivers are heading from A to B, they will, without knowing anything about the overall traffic situation, all enter the direct road. In consequence, a queue builds up from the bottleneck.

A Nash-Equilibrium is defined as a situation where no agent (= driver) can lower his or her cost (= decrease travel time) by unilaterally changing behavior. Assuming that the drivers have complete information, this implies that the waiting time in the queue exactly compensates for the additional driving time on the alternative route.

Now assume that there are additional travel demands from C to D (see Fig. 3), the exit for the latter lying shortly before the bottleneck. Obviously, this traffic is suffering from the bottleneck queue upstream (= left) of the bottleneck, and from these travelers’ point of view it would be much better if the queue were located to the left of the ramp that the travelers from C use to enter the link. Note that moving the queue further upstream does not make any difference for the drivers originating in A. — This example illustrates that one easily finds situations where there are better overall solutions than the NE.
A way to push a traffic system from a NE towards a better overall solution is to keep the density on each road at or below $\rho^*$, the density of maximum throughput. Then there would not exist queues anywhere in the system, thus ensuring that additional traffic could proceed undisturbed. Note that this could for instance imply (in the limit of a perfect implementation) that drivers have to wait to enter the road network until sufficient capacity is available for them.

One possible way to achieve this is to introduce a congestion-dependent toll ("congestion pricing" as opposed to "road pricing"), and this toll is simply increased until the density on the respective link has dropped to the desired level.

This is exactly the system that we simulated.

In our simple network, there are only two different types of travelers: travelers from A to B, and travelers from C to D. Travelers from A to B can choose between the direct and the longer alternate route. In order to make decisions, each AB-driver remembers his or her last travel-time on each of the two routes.

A traveler calculates expected costs [7] according to

$$cost_{direct} = toll + \alpha \cdot t_{direct}$$  \hspace{1cm} (5)

and

$$cost_{alt} = \alpha \cdot t_{alt}$$  \hspace{1cm} (6)

where $cost_{direct}$ and $cost_{alt}$ are the expected costs for the two route choices, $toll$ is the toll for the current day (see below), $t_{direct}$ and $t_{alt}$ are the remembered travel times for each route, and $\alpha$ is a conversion factor which reflects trade-off between time and money. $\alpha$ could be different for each driver, but is uniformly equal to one in this work. ($\alpha$ reflects "standard values of time"; VOT, which can be looked up for traffic systems.)

Then, each driver chooses the cheaper route, except that there is a 5% probability of error (which gives each driver a chance from time to time to update her information about the other possibility).

As long as the traffic dynamics is deterministic and completely uniform, this scheme leads to a Nash equilibrium [7]. However, in our case of stochastic traffic dynamics, this is no longer true: There might well be a decision rule different from the one above where at least one traveler is better off, for example by triggering some kind of day-to-day oscillation between the two routes and taking advantage of it. In other words, by dealing with stochastic traffic dynamics, the notions of economic equilibrium theory have to be used with care.

We describe 200 consecutive days of a simulation where the toll was kept at zero during the first 100 days, and in addition all A-B-travelers were forced to use the direct route during the first 50 days.

Fig. 4a shows results for the trip times and the adaptive toll, Fig. 4b the vehicle-to-vehicle variations of the trip time (as defined earlier), and Fig. 4c the day-averaged density, on selected road sections. These sections are: (i) the section where the density for the toll adaptation is measured, (ii) the section of
Figure 4: Simulation output for 200 iteration of the simple corridor network model. Time-steps 1-50: No adaptation; 51-100: drivers can choose alternative route; 101-200: drivers can choose alternative route, and the toll adapts in order to keep the density at an efficient level. Top: Average trip times for the direct and for the alternative route from A to B as well as for the route from C to D, and toll for the direct route from A to B. Middle: Vehicle-to-vehicle fluctuations of trip time for the direct and for the alternative route from A to B. Bottom: Densities on the segment shared by A-B-direct travelers and C-D-travelers, on the segment shortly before the bottleneck used for determination of the toll, and on the alternative route from A to B.

the main road between the on-ramp from C and the off-ramp towards D, and (iii) the alternative route.

Even when allowed (i.e. after day 50), not many of the A-B drivers use the new option of the alternative route. This is to be expected, since it is more
than six times longer than the direct one. In consequence, travel times and
fluctuations do not change much.

After day 100, the adaptive tolling starts and fairly quickly reaches a sta-

tionary value around 260. As the “toll” line in Fig. 4c indicates, this keeps
indeed $\rho_{toll}$ near the “efficient” range between $\rho = 0.05$ and 0.10. In addition,
the density on the main segment (used by both A-B and C-D travelers) drops
to around 0.11, above, but close to the density of maximum throughput.

Travel times for C-D and for A-B-direct travelers go down (Fig. 4a); and the
toll just offsets the time gain for use of the direct route: $time_{direct} + \alpha \cdot toll \approx time_{alternat}$. (recall that $\alpha = 1$).

Vehicle-to-vehicle fluctuations (Fig. 4b) for the use of the alternative road go
up from ca. 2% to around 12%, and for the use of the direct road from ca. 11%
to around 42%. Moreover, the day-to-day fluctuations also seem to go up in all
measurements.

One should distinguish between two different kinds of fluctuations: Fluctua-
tions due to the dynamics, and fluctuations due to the learning. The fluctuations
in the latter might be due to the specifics of the chosen learning scheme, es-
pecially the lack of historic information beyond the last day. More realistic
assumptions about the learning and en-route information are claimed to avoid
that [8]. However, the results for the vehicle-to-vehicle fluctuations (i.e. the $\sigma$
as defined in the text) only depend on the fact that the traffic density is driven
towards the critical value. A less fluctuating learning scheme should therefore
even increase our values for $\sigma$. For more details, see [9, 10].

The above work has to a large extent been motivated by Arthur's “bar prob-
lem” [11]: Assume that people want to be in a bar when it is neither too empty
nor too crowded, similar to the wish of not wanting to spend too much time
traveling which results in a choice of either the direct or the alternative route
(going to the bar or not). Also, the individual decision dynamics is fairly equiva-
 lent: Individuals make their choice, then execute their decision, the outcome
of this is added to each individual's personal experience, and the cycle starts
again. A fairly important difference between Arthur's work and ours, is that
Arthur needs many different, albeit simple, decision rules for each individual to
stabilize the outcome. In the traffic case this has not been necessary. Arguably,
in the traffic case, the dynamics itself already provides enough fluctuations that
individuals, even when faced with the seemingly same problem, make different
decisions.

5 Self-organized criticality in traffic networks

Now we are in a position to justify our initial claim that traffic management
measures will lead to higher fluctuations and thus lower predictability in traffic
networks. The last section clearly shows that traffic management measures
will tend to "equilibrate" traffic patterns, that is, to make overused parts of the
system less overused, and to make underused parts of the system less underused. Quite in general thus, the wholes system tends to operate closer to the point of maximum efficiency. But as stated in Section 3 and also recovered in Section 4, this regime is the regime where variations are highest, or, in other words, the system naturally evolves into a state where fluctuations are highest, which can be seen as yet another example of self-organized criticality [12].

That means that, for an individual driver, it is really impossible to predict how long a certain trip along a certain route will actually take. Which means in return that neither a driver nor an omniscient traffic management system can decide which of several possible paths might be the fastest or best. In this way, it is the increasing unpredictability caused by the traffic management which eventually impedes further improvement.

It is clear that this argument would benefit from further simulations in realistic traffic networks. Although this has not yet been done, the following section shows an agent based simulation which has all the ingredients for such an investigation.

6 A realistic network

In this section, we want to explain how the above methods can be extended to simulate traffic in realistic networks. As a practical example, we use the freeway network of the German land Northrhine-Westfalia (NRW). We only show results of a single-lane implementation. Multi-lane implementations are straightforward [13] but only make sense when one expects that the possibility of passing introduces additional effects.

The important elements of the approach are (i) individual trip plans, i.e. each "driver" knows before the start of the simulation which route he/she wants to take, and (ii) the use of individual decision rules based on past "simulated" experience. Roughly, for a plot like Fig. 5, the following was done in the simulation:

- At the beginning of each "period" (≈ rush "hour"), there were 20 ordered queues of vehicles with drivers waiting to enter into the network. Each queue consisted of 2000 vehicles.
- Each simulated driver had an individual destination, and the set of the 10 shortest paths to that destination to select from.
- Each driver randomly selected a yet untried path; or in case all paths had been tried, he/she selected the path which had performed best in the past (with a small random chance to try something else again).
- The simulation is executed, with each driver following his/her path. If too many vehicles attempt to use the same road section, this creates congestion such as in Fig. 5.
Figure 5: Simulated traffic jams in a single-lane implementation of the freeway network of the German land Northrhine-Westfalia (NRW). Situation at "day 16" after 6000 iterations (100 minutes). Free traffic is denoted by dots, critical traffic by light gray 'x', and jammed traffic by triangles.

This scheme was executed for several consecutive periods, until the congestion pattern "relaxed" to a pattern which did not change much from one period to the next. This was usually reached after simulating 15 periods; note that 10 periods were necessary until each driver had tried each of his/her options.

Obviously, it will be necessary to replace the arbitrary origin/destination pattern of these simulations by more realistic data. Yet, some of the network bottlenecks seem generic with respect to transit traffic through NRW: The jams between Wuppertal and Kreuz Kamen are well known, and, as one sees, a consequence of the missing extension of the freeway A4 beyond Olpe. This extension has since long been planned; but it leads through environmentally
sensitive areas, and it is thus under discussion if it will ever be built. Note that
the simulation methodology presented here can be used to evaluate the utility
of such an extension, or what is needed to replace it by improvements along
existing paths. Or, which traffic streams have to be reduced in order to manage
with the currently existing infrastructure, and how this can be achieved.

The problems near Krefeld are due to the same bottleneck in North-East/South-
West direction. It is also known that the K"olner Ring presents a bottleneck.

For further details, see [10].

7 Conclusion

All the above is in agreement with our intuition that traffic management can
indeed make traffic more efficient, but may in addition lead to higher fluctua-
tions and, as a consequence, lower predictability, since the system is driven
closer to capacity and thus to criticality. In summary we seem to have reached
a paradox: In wanting to obtain a better control of the transportation system,
by the introduction of a traffic management system, we actually produce a more
unpredictable traffic dynamics. This happens because the traffic management
system in essence moves traffic from more congested roads to less congested
roads, and thus as a whole, forces the transportation system into the critical
regime where small perturbations have a large influence on the microscopic dy-
namics. Network traffic produced with adaptive drivers and traffic management
systems is therefore an example of self-organized critical dynamics.

Since air pollution as well as serious accidents also are maximal where ac-
celeration and de-acceleration is maximal, the critical regime, in addition to its
non-controlability, produces these highly non-desirable side effects.

Are Traffic Management systems then not desirable? This is probably the
wrong way to look at it. For example in stock markets, modern information tech-
nology has brought the market fluctuations to much higher levels than before,
and traders just have learned to live with that (and have introduced additional
financial instruments which insure against the risks of fluctuations). It is also
unclear if society will accept a completely efficient way of traffic management —
for example, congestion pricing (unfortunately often confused with road pricing)
seems to evoke strong opposition by many people. And then there is always the
possibility that, when we are aware of the risks, we are able to design traffic
management systems which circumvent the problems — maybe by having less
efficient flows — flows below maximal capacity and thus criticality.
Appendix

The simple, single lane 1-D cellular automata used in the simulations for this paper are given by the following four rules [3]:

FOR all vehicles $i \in \{1, \ldots, N\}$ DO
(1) IF $\text{velocity}[i] < V_{\text{max}}$ AND $\text{gap}[i] > \text{velocity}[i]$
 THEN $\text{velocity}[i] = \text{velocity}[i] + 1$.
(2) IF $\text{gap}[i] < \text{velocity}[i]$ THEN $\text{velocity}[i] = \text{gap}[i]$.
(3) IF $\text{velocity}[i] > 0$ THEN with probability $0.5$ $\text{velocity}[i] = \text{velocity}[i] - 1$.
(4) $\text{position}[i] = \text{position}[i] + \text{velocity}[i]$.
END

where $\text{position}[i]$ is current position of vehicle $i$, $\text{velocity}[i]$ current velocity of vehicle $i$, \text{gap}[i] distance to nearest vehicle ahead of vehicle $i$, and $V_{\text{max}}$ the maximum velocity of each vehicle. Note that, because all values are integer, relations like $\text{gap} < \text{velocity}$ and $\text{gap} + 1 \leq \text{velocity}$ are equivalent. For more details we refer to [3].

References


TRANSIMS Travelogue. January 1995. Travel model improvement program newsletter Issue No. 2 (Texas Transportation Institute, 1600 E Lamar Blvd #112, Arlington TX 76011), p. 9
TRANSIMS Travelogue. July 1995. Travel model improvement program newsletter Issue No. 3, p. 10


