Hypercongestion

UCI-ITS-WP-97-2

Kenneth A. Small [1]
Xuehao Chu [2]

[1] Department of Economics and Institute of Transportation Studies
University of California, Irvine; Irvine, CA 92697-5100, U.S.A., ksmall@uci.edu

[2] Center for Urban Transportation Research
University of South Florida; Tampa, FL 33620, U.S.A., xchu@cutr.eng.usf.edu

March 1997

Institute of Transportation Studies
University of California, Irvine
Irvine, CA 92697-3600, U.S.A.
http://www.its.uci.edu

HYPERCONGESTION

Kenneth A. Small and Xuehao Chu

ABSTRACT

The standard economic model for analyzing traffic congestion, due to A.A. Walters, incorporates a relationship between speed and traffic flow. Empirical measurements indicate a region, known as hypercongestion, in which speed increases with flow. We argue that this relationship is unsuitable as a supply curve for equilibrium analysis because hypercongestion occurs as a response to transient demand fluctuations. We then present tractable models for handling such fluctuations, both for a uniform expressway and for a dense street network such as in a central business district (CBD). For the CBD model, we consider both exogenous and endogenous time patterns for demand, and we make use of an empirical speed-density relationship for Dallas, Texas to characterize both congested and hypercongested conditions.

Paper prepared for the annual meeting of the American Real Estate and Urban Economics Association, New Orleans, January 1997. Financial support from the U.S. Department of Transportation and the California Department of Transportation, through the University of California Transportation Center, is gratefully acknowledged. We thank Richard Arnott, John Bates, and Fred Hall for very insightful discussions.

HYPERCONGESTION

Kenneth A. Small and Xuehao Chu

1. Introduction

It has been a third of a century since A.A. Walters (1961) established what is now the standard way economists think about congestion.[1] Walters used the functional relationship between travel time on a given length of highway and the traffic flow rate, which looks like the curve AC in Figure 1. This relationship is well known in traffic theory as a variant of the *fundamental diagram of traffic flow*, equivalently expressed as a speed-density, speed-flow, or flow-density relationship.[2] It was the genius of Walters to recognize that by identifying flow as quantity, the engineering relationship could be viewed as an average cost curve (with a suitable transformation from travel time to cost), and then combined with a demand curve to analyze equilibria and optima, the latter being decentralized by administering a Pigovian charge known as a congestion toll. This model has proved extraordinarily fruitful, not only in transportation but more widely in the literature on local public finance and clubs.[3]

---

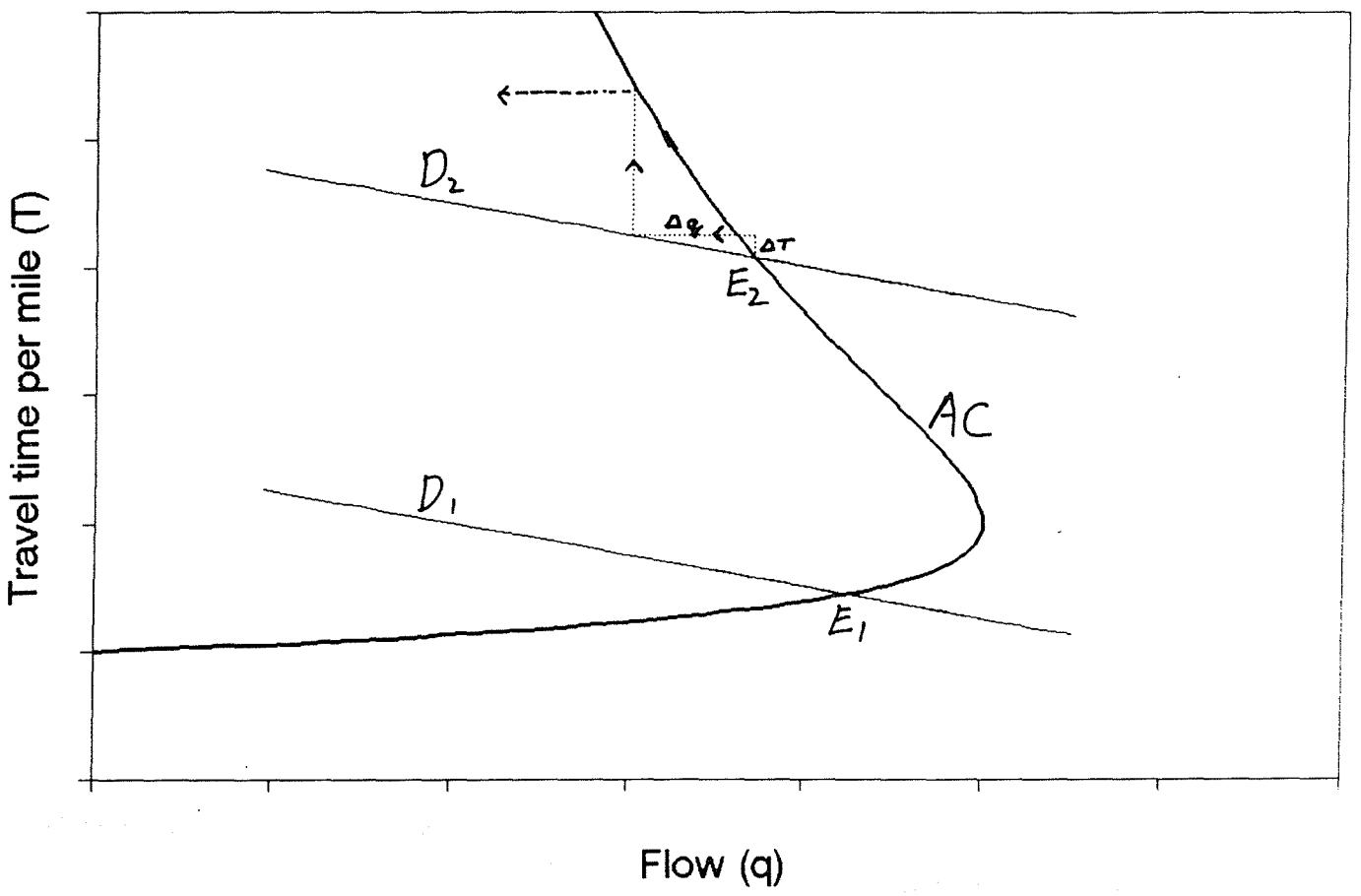
[1]See Walters (1987), Newbery (1990), Small (1992b), or Button (1993) for current practice. Important predecessors of Walters' formulation include Pigou (1920), Knight (1924), and Beckmann et al. (1955).

[2]See Haight (1963). The equivalence among the three is due to the definitional identity equating flow to speed times density. See also Mun (1994), Figure 2.

[3]To mention just some key developments appearing in the general interest economics journals: William Vickrey (1963, 1969) has tirelessly developed theoretical refinements and implementation techniques that enhance the practicality of congestion tolls. Mohring (1970) integrated the pricing analysis rigorously with investment analysis, placing them both squarely in the realm of peakload pricing as developed by Boiteux (1949), Vickrey (1955), Williamson (1966), and others. Lévy-Lambert (1968) and Marchand (1968) worked out second-best pricing rules if substitute facilities cannot be priced. DeVany and Saving (1980) added uncertain demand. Edelson (1971) treated a monopoly road supplier, and David Mills (1981) combined monopoly with heterogeneous values of time. Applications to local public finance and clubs include Oakland (1972), Arnott (1979), and

(continued...)

Figure 1.
A Model of Travel Time Versus Flow



One feature, however, gave Walters some difficulty and has caused endless trouble ever since. This is the non-unique relationship between travel time and flow depicted in Figure 1, and in particular the possibility for a second equilibrium, such as E_2 in Figure 1, where the average cost curve is downward-sloping. This branch of the curve is known in the economics literature as the region of "hypercongestion," in contrast to the lower branch which depicts ordinary congestion.[4] Walters described equilibrium E_2 as "The Bottleneck Case" (p. 679), a terminology which is not fully explained but which, as we shall see, is highly appropriate in the case of a straight road.

What are we to make of point E_2? On the face of it, E_2 is just an especially inefficient equilibrium. This is precisely the conventional interpretation, encouraged if not specifically stated by Walters, from which it follows that first-order welfare gains are possible by somehow shifting the equilibrium down to the lower branch of the curve. Indeed serious arguments about Pareto-improving tolls have been based on just this argument, both for highways and by analogy for renewable resources (De Meza and Gould (1987)).

But something is wrong. For one thing, E_2 is not obviously a stable equilibrium. Consider a simple quantity adjustment mechanism on the demand side in response to a small upward fluctuation ΔT in travel time, due for example to an influx of inexperienced drivers. Quantity demanded would be reduced by $-\Delta q$ in Figure 1; according to the hypercongested "supply" curve, this in turn would cause a further *increase* in travel time; and so forth. This is shown in the figure. We could perhaps eliminate the instability by making curve D_2 steeper, so as to cut AC from above; or by defining other dynamic adjustment mechanisms. But the story is wrong altogether: how could curtailing the

---

[3](...continued)
Berglas and Pines (1981).

[4]Unfortunately engineering terminology differs, with the lower branch called "uncongested" or "free flow", and the upper called "congested flow." We use the economics terminology here.

quantity demanded make traffic worse? Or to put it differently, how can drivers confer a positive externality on each other, as implied by the downward-sloping portion of the average cost curve? This violates common sense.

Walters recognized the these problems better than many writers who followed, and was careful not to call the backward-bending portion of curve AC a "supply curve." Rather he called it "the equilibrium relation between flow and unit cost when density has been taken into account" (p. 680). This equilibrium was analyzed through a verbal dynamic argument involving underlying demand and supply relationships in different spaces. Specifically: demand relates trip time to the *inflow* of entrants to the bottleneck section, whereas the supply relationship is between speed and *density* of vehicles.[5] Speed times density is the *outflow* from the bottleneck. Whenever demand changes, a set of adjustments takes place until inflow equals outflow, their equalized value defining one point of Walters' AC curve.[6]

But how likely is it that a steady state will prevail under such conditions? There is, after all, a reason why heavy congestion is called "peak congestion": it doesn't last very long. For example, consider the afternoon rush hour near a large factory. In that case quantity demanded is expressed as a flow of vehicles entering the roadway for the purpose of making a trip. When this inflow is large, it cannot be equal to the throughput of vehicles at some intermediate points on the roadway, nor to the outflow on its exits; throughput

---

[5]This supply relationship is single-valued and monotonic. For example, curve AC in Figure 1 is based on a linear relationship between speed and density; see section 4, equations (9) and (15).

[6]Hills (1993) also makes the point that "demand should be measured in terms of the number of vehicles *wishing to embark* on trips during a given period of time" (p. 96). McDonald and d'Ouville (1988) make a similar distinction between inflow and outflow by defining them as inputs and outputs, respectively, of a production function. Else (1981) and Alan Evans (1992) try to rescue the static analysis by redefining quantity demanded as "the number of vehicles on the road" (Else, p. 221; Evans, p. 212), that is, as density. Evans correctly notes that "consumers do not choose the traffic flow given the price" (p. 212), so he rejects a curve like D_1 in Figure 1 as a valid demand curve; but he seems oblivious to the fact that consumers do not choose density either. Rather, density is a stock variable that depends on past inflows and on capacity and other parameters of the flow-density relationship.

and outflow are governed by roadway capacity and other characteristics as well as demand conditions.

When inflow and outflow differ, density will not remain constant along the road, and the engineering relationship underlying curve AC in Figure 1 no longer applies. Instead, dynamic approaches come into play, notably the "kinematic" theory of Lighthill and Whitham (1955) and various car-following theories based on micro driver behavior.[7] These theories explain, for example, the stop-and-go conditions so familiar to expressway drivers: they can be viewed as density waves emanating from a dynamic disturbance, analogous to the waves of air pressure that we perceive as sound.

Or consider what happens if the demand curve is steeper than those shown in Figure 1 so that it crosses the AC curve twice. Each crossing represents a possible steady state in Walters' analysis; but what determines which one occurs? The answer can come only from a dynamic analysis. As stated succinctly by Arnott (1990): "hypercongestion occurs as a transient response of a non-linear system to a demand spike" (p. 200).

This paper is about how to deal with such demand spikes. Congestion is a peak-load problem, with effects that are inherently dynamic. For ordinary congestion, ignoring the spikiness of demand causes only inaccuracies, not fundamental inconsistencies.[8] But for hypercongestion, spikiness is the whole game.

In the sections that follow, we demonstrate more fully the truth of Arnott's characterization of hypercongestion (Section 2). We then develop the consequences for two quite different situations: a uniform stretch of roadway (Section 3), and a dense street network (Section 4). Our goal is to provide tractable models for economic analysis, so we

---

[7]For a recent example using micro-simulation, see Nagel and Rasmussen (1994).

[8]See Agnew (1977), whose formulation is similar to ours of Section 4A except that his demands occur as step functions with infinitely long duration (and with elasticities applying independently at each instant), rather than as spikes. Also, he solves for optimal time paths but not for the unpriced time paths. The optimal paths all tend toward a steady state in which hypercongestion is eliminated, but transient hypercongestion may occur as part of an optimal solution.

make some simplifications compared to a fully developed engineering model of traffic flow. In the case of a uniform expressway, our simplification leads to a piecewise-linear average cost function based on simple queueing theory, in which hypercongestion occurs but is largely irrelevant because it affects flow only inside the queue. In the case of a dense street network, our two alternate simplifications allow hypercongestion to occur and to create large costs, even though it is not necessarily optimal to eliminate it.

Our model of dense street networks (Section 4) is applicable to many situations where congestion not only entails costs but interferes with throughput. Telephone networks may break down when switching equipment is overtaxed. Storm drains clog when high water flow carries debris that under normal flow remained in the gutters. Bureaucrats unable to process an unusually high flow of paperwork may be besieged by irate calls, thus lowering their ability to process throughput just when they need it most. Or think of it as the messy desk syndrome: you normally can handle work as it comes in, but the backlog from a temporary overload so clutters your desk that you can't find the papers you need, reducing your rate of productivity and causing the backlog to increase even faster. All these situations can, depending on parameters, lead either to temporary problems (busy signals, moderate flooding) or to a complete breakdown known as gridlock.

2. Queueing Makes the Supply Curve Upward-Sloping

In this section, we provide hard evidence for what every urban driver already knows: when more vehicles try to use the road, their travel times go up. We do this because so much economic analysis has been based on the contrary assumption, due to viewing the hypercongested region of Figure 1 as a supply curve suitable for supply-demand analysis. The consequences for modeling congestion are postponed to Sections 3 and 4.

The basic argument is simple. Hypercongestion occurs when too many cars try to occupy the same places at the same time. More precisely, a capacity limit is exceeded

somewhere in the system. As a result, local queueing begins, which becomes more severe the more cars are added to the input flows. Queueing adds time to trips beyond what is portrayed by the instantaneous speed-flow relationship.

Of course, this condition cannot persist indefinitely — for then travel time would rise without limit, eventually choking off demand. So demand must at some point fall back below capacity — i.e., below the level that caused queueing in the first place. Once that happens, queues dissipate and the system reverts to one exhibiting ordinary congestion. Hence hypercongestion is, as Arnott said, a transient response to a demand spike.

We now explain in more detail what happens when traffic exceeds the capacity of some bottleneck in the system. In the case of a simple bottleneck, say at the egress of an otherwise uniform stretch of expressway, the queue is easily identified and can be analyzed precisely. We show in subsection A that hypercongestion then occurs, but only in the queue itself, where it has little bearing on the average cost of an entire trip. On complex street networks or highways with lots of exits and entrances, queues may pop up in many places and the analysis becomes messier. For example queues at one intersection may block another, causing reduced flow across a wide region. Such a condition, called *oversaturation*, is analyzed in subsection B. In this situation hypercongestion does add considerably to cost, but a suitably defined supply curve is still normally upward sloping.[9] An extreme form of oversaturation is gridlock, in which flows cease entirely.

[9]An exception could occur on an unpriced network serving more than one origin-destination (o-d) pair. Increasing the demand flow for one o-d pair could decrease travel time for another o-d pair, possibly even decreasing total travel time. An example of such a "paradox" is given by Fisk (1979). The crux of the paradox is that the extra congestion on the route serving the first o-d pair causes other travelers, who originally shared a part of this route, to switch to alternative routes which happen to have lower external costs.

A. Straight Uniform Highways

The fundamental diagram of traffic flow describes an instantaneous relationship between variables measured over a very short section of roadway. Of course, actual measurements must be made over finite distances and times. Even for a straight uniform freeway segment, there is considerable uncertainty over the exact shape of the fundamental diagram. Consider for example the *Highway Capacity Manual*, the standard reference for highway design in the United States, developed over decades by blue-ribbon committees of the Transportation Research Board (TRB (1992)). As recently as 1992 the speed-flow curves for such "straight pipes" were quite drastically revised. The new curves portraying ordinary congestion exhibit a higher capacity than the old ones, and they are flatter up to flows quite close to that capacity.[10] In other words, serious congestion sets in only at higher volumes than previously thought, and then it sets in quite quickly. The same is true for rural and suburban multilane highways.

One reason for the difficulty in measuring the relationship is the paucity of data from the region near capacity. It is actually rare to observe flow near capacity on a short uniform highway segment, and the resulting speed-flow plots tend to show an enormous scatter. An example is shown in Figure 2, where the hypercongested region is the lower part of the figure. Even worse, there are both history-dependence and, sometimes, a discontinuity in the speed-flow curve. As a result, the hypercongested branch of the speed-flow curve is rather ill-defined and not necessarily connected to the normal branch.[11] Small (1992b, p. 66) presents a case from the economics literature for which attempting to fit a single function through the broad scatter of points drastically overstates

---

[10]TRB (1992), p. 3-i.

[11]See Banks (1989) for a very clear discussion. Two other examples illustrating these problems are shown by Small (1992b, pp. 64-65). History dependence is seen clearly in Hall and Hall (1990), and is formulated in terms of catastrophe theory by Dillon and Hall (1990).

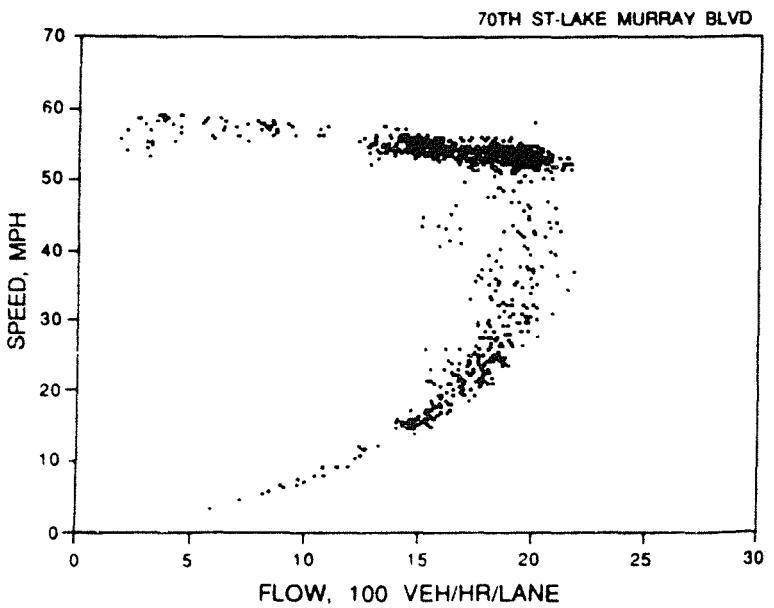


Figure 2. Speed-Flow Observations on Interstate 8 at 70th Street, San Diego
Reproduced from Banks (1989), Figure 3

the slope of the congested part of the speed-flow curve throughout most of its range — precisely what seems to have happened in the data used for previous editions of the *Highway Capacity Manual*.

A series of papers by Fred Hall and several associates using data for Toronto, and another by James Banks using data for San Diego, have now established a primary reason for these problems in the case of urban expressways.[12] Traffic that is in or near a condition of hypercongestion is almost always influenced by a nearby bottleneck. Because of entrance ramps and variations in the roadway, the ratio of flow to capacity is never constant across distance. Instead, local bottlenecks occur where capacity is exceeded, and these affect adjacent sections: upstream of a bottleneck traffic tends to form a queue, while downstream it is metered to a level well below the capacity of that section. Within the queue, the speed-flow relationship is hypercongested; its backward-bending shape is apparent, although the precise relationship is sensitive to the timing and circumstances governing the bottleneck that created the queue.[13]

The upshot, then, is that hypercongestion on an urban expressway usually occurs as part of a queue. Where it occurs, the flow rate is governed not by quantity demanded but rather by downstream bottleneck capacity — a point made explicitly by Branston (1976, p. 224) and in more detail by Mun (1994). The density accompanying that flow rate is more or less irrelevant to the trip time, which is governed mainly by the number of vehicles in the queue and the rate at which they can flow through the bottleneck.

In the simplest and most commonly applied picture, the downstream bottleneck capacity is independent of flow conditions in the queue. For example, all the "link capacity functions" reviewed by Branston (1976) have this property. This would suggest that point bottleneck theory is a better basis than speed-flow curves for analyzing severe freeway

---

[12]For an excellent synthesis and interpretation, see Hall et al. (1992).

[13]This sensitivity is demonstrated quite clearly by Branston (1976) and by Hall and Hall (1990).

congestion; and furthermore that hypercongestion is irrelevant. May and Keller (1966) long ago concluded that the bottleneck provided a good analytical simplification, and many simulation models analyze a freeway as a set of point bottlenecks connected by segments exhibiting ordinary congestion (May(1987)). In economics, the fixed-capacity bottleneck model has recently become a standard tool for what amounts to an alternative economic analysis of congestion,[14] which we describe in Section 3B below.

In a more complex model the capacity of a bottleneck might depend on upstream conditions or on the history of past flow conditions — as suggested by our earlier observations about discontinuities and history-dependence in speed-flow curves. In that case hypercongestion is relevant to trip time, in much the same way as in the model for city streets to be discussed shortly. It is theoretically possible for a steady state to exist in such a situation, which is proposed by Newbery (1990, p. 28) as the best interpretation of the backward-bending portion of the speed-flow curve. Like Newbery, we have doubts about the stability of such a steady state. But more importantly, we believe that such a steady state is empirically rare and thus not very useful as a basic model for economic analysis.

Returning to the fixed-capacity case, we should explain why queue density is only "more or less irrelevant" rather than entirely irrelevant to total trip time. If the queue has low density, it will extend further back in space from the bottleneck, so the traveler waiting his turn in the queue will be covering some of the distance he would have had to cover anyway. Hence the less dense the traffic in the queue, the less the total trip time.[15] The full relationship is established by Mun (1994) using the kinetic traffic flow theory of Lighthill and Whitham (1955). The edge of the queue forms a shock wave moving backward at a

---

[14]See Vickrey (1969), Arnott et al. (1990, 1993), or Small (1992a,b).

[15]This presents an interesting paradox: expanding the capacity *behind* a bottleneck, say by adding a lane, actually lengthens trip time because it allows the queued vehicles to be stored in a shorter space, requiring them to take slightly longer to reach the storage area. This assumes the original queue was not blocking an exit ramp.

velocity determined by the hypercongested speed-density relationship. (Density waves within the queue itself are assumed to be absent.) The number of vehicles $Q(t)$ stored in the queue grows at rate $\lambda(t)$-q_b, where $\lambda(t)$ is the entry flow and q_b is bottleneck capacity. The resulting travel time for a vehicle entering a highway segment of length L at time t is:[16]

$$T(t) = \frac{L - J(t)}{v_1[\lambda(t)]} + \frac{J(t)}{v_2(q_b)} \qquad (1)$$

where $J(t)$ is the spatial length of the queue once it is encountered by this vehicle, and $v_1(\bullet)$ and $v_2(\bullet)$ describe the congested and hypercongested branches of the speed-flow relationship governing the part of roadway behind the bottleneck.[17]

By differentiating equation (1) with respect to input flow, Mun is able to show formally that the average travel time is a non-decreasing function of current and past flows. He explicitly interprets this as indicating a rising supply curve: "the backward-bending section never appears in the cost curve for travel along a long road" (p. 369). This vindicates the claim made by Hills (1993, p. 97) that once demand is properly defined the average cost curve does not bend backward.

---

[16]Mun (1994), p. 369, equation (9). There is actually some ambiguity in Mun's article about the meaning of time t because he never explicitly takes account of the gap between when flow enters the roadway and when it reaches the queue, both of which are denoted by t. Hence our interpretation that $J(t)$ refers to the physical queue length at the time it is encountered by a vehicle that entered the roadway at time t. This interpretation is also required to justify Mun's equation (10), which makes use of the derivative of queue length $J(t)$ with respect to the instantaneous inflow rate $\lambda(t)$. The same ambiguity occurs in Agnew (1977).

[17]Mun implicitly assumes that the speed of a vehicle prior to reaching the queue is governed by the flow rate at the time it entered the highway, and is unaffected by inhomogeneities in traffic density along the part of the highway subjected only to ordinary congestion. This is a simplification because with changing $q(t)$, cohorts of vehicles will encroach on or disperse from each other, causing the density they encounter to change in the course of their trip. This same simplification is the basis of the models by Henderson (1981) and Mahmassani and Herman (1984), which together are adapted to create the model we present in Section 4B below.

If the physical length of the queue is negligible compared to L, and queueing speed v_2 is much less than approach speed v_1, then equation (1) is well approximated by the free-flow time L/v_1 plus the queueing delay J/v_2, or:[18]

$$T(t) = \frac{L}{v_1[\lambda(t)]} + \frac{Q(t)}{q_b} . \qquad (2)$$

Hence the proper application of traffic flow theory to a uniform straight roadway with a bottleneck leads to a rising supply curve. Hypercongestion exists, but is irrelevant in the simplified version of equation (2) and barely relevant in the more complete version of equation (1). In no case is there an equilibrium remotely described by point E_2 in Figure 1.

B. Dense Street Networks

The fundamental diagram of traffic flow is not expected to apply to an entire network of streets. Instead, analysis typically proceeds by simulation using queueing theory at each intersection.[19] Considerable work has been done trying to characterize the "oversaturation delay" in such a situation, both for single intersections and for groups of intersections. One purpose of such work has been to develop relationships between travel time and input-flow characteristics for use in intersection design, as for example in the *Highway Capacity Manual* (TRB (1992), ch. 9). These relationships have the property one expects from ordinary experience: greater inflows cause greater delays. As a particularly simple example, consider the deterministic queueing delay caused by a flat demand spike

[18]Equation (2) uses the two definitional identities $Q \equiv Jk_2$ (where k_2 is the density in the queue) and $q_b = k_2 v_2$ (since the queue's flow rate is q_b.)

[19]See for example Dewees (1979), Williams et al. (1987), or Arnott (1990).

of height λ and duration $(t_2\text{-}t_1)$ at a single intersection of capacity q_b. The number of queued vehicles starts at zero and rises to a maximum of $(\lambda\text{-}q_b)(t_2\text{-}t_1)$, for an average delay:

$$T = \tfrac{1}{2}\left(\frac{\lambda}{q_b}-1\right)(t_2-t_1) \ . \tag{3}$$

Other patterns give more complex formulas.[20]

Dewees (1978, 1979) uses a standard simulation package to estimate delays on two real street networks in the Toronto area, one suburban and one downtown. The simulations take into account the interactions among traffic flows on different streets due to their network interconnections. Starting with a base set of flows representing actual rush hour conditions, Dewees made marginal increments, one at a time, to the flow entering each major street in order to determine the effects on average travel time. In all cases average travel time rises with entering traffic, despite the fact that many intersections were oversaturated and many links may well have been operating in conditions of hypercongestion.

Small (1992b, p. 70) shows that Dewees' data for travel times on one suburban street, subjected to a variety of alternative input flows λ in his simulations, are approximated quite well by a power law of a form used by Vickrey (1963) and many others:

[20]See, for example, Rouphail and Akçelik (1992), whose equation (34) is the case just mentioned. As they note on p. 32, stochastic delay is quickly swamped by deterministic delay in typical oversaturated conditions, leading to common use of deterministic queueing as a good approximation.

$$T = T_0 + T_1(\lambda/q_b)^{\varepsilon} \tag{4}$$

with ε=4.08. One reason this power law has been popular for applied work on congestion pricing is precisely because it is single-valued, monotonically increasing, and defined for all input flows, thereby conveniently bypassing the conceptual problems we have described. The point here is that this is an entirely reasonable finesse because as long as we insist on a static model, a rising supply curve such as equation (4) represents time-averaged travel times better than an instantaneous relationship like that in Figure 1.

3. Modeling Hypercongestion on a Straight Uniform Highway

We now turn to the search for dynamic models that deal with hypercongestion, yet are tractable enough to be part of a toolkit for broader economic analysis. We provide examples for the two cases just discussed: the straight uniform highway in this section, and the dense street network in Section 4.

The application of kinetic traffic theory by Mun, described in the previous section, provides a rigorous model for examining hypercongestion with and without pricing on a long uniform roadway. Mun himself accomplishes this, deriving among other things the optimal time-varying congestion toll. It contains a term representing the additional delays imposed on all subsequent travelers because the queue's length is increased by the vehicle in question.

Mun also provides numerical examples using a demand function of the form:

$$q(t) = q_0(t)\, e^{-aP(t)} \tag{5}$$

where $P(t)$ is the "full price," including time cost and toll, at the time when the queue is encountered. The results show that the congestion toll is high at the beginning of the queueing period, then falls as the queue builds up. This property is well known in other economic analyses of queueing, and reflects the fact that early in the rush hour there are more subsequent travelers who are delayed by exogenously adding a car now to the queue.

Note that with this demand function, demand at any time is independent of the full price at other times. (This is a sharply different from the assumption behind models of endogenous trip timing discussed in Sections 3B and 4B below.) As a result, it is entirely possible for queueing, hence hypercongestion, to exist in Mun's optimum. This is because marginal cost, properly defined, rises throughout the possible range of input flows but never becomes infinite;[21] so if demand is high enough the optimum can involve input flows that cause temporary hypercongestion. Mun indeed simulates two such cases.

All this is easier to see in the approximation given by equation (2), in which we ignore the physical queue length. Adding one vehicle to the queue at time t causes every subsequently queued vehicle to incur an additional delay $1/q_b$, which is the inverse of the queue discharge rate. The external cost associated with the queue is simply the value of this delay multiplied by the number of vehicles entering from time t until the queue is discharged.

These ideas become clearer by assuming specific forms for demand. In the next two subsections we make two very different assumptions about demand: (A) that the timing is exogenous, and (B) that the timing is determined by scheduling costs.

---

[21] As Agnew (1977) notes, "traffic jams are not infinitely bad," so it is unrealistic to allow an infinite marginal cost at an achievable demand quantity, as Walters' model does when flow equals capacity.

A. Exogenous Demand Spike

Suppose demand is a pulse of height λ over the interval $[t_1, t_2]$. If t_1 and t_2 are fixed, then the relevant average cost curve is simply a function of λ. It can be found by taking the time average of equation (2), with $Q(t)=(\lambda-q_b)(t-t_1)$:

$$AC(\lambda) = \begin{cases} \alpha\dfrac{L}{v_1(\lambda)} & \text{if } \lambda \le q_b \\ \alpha\dfrac{L}{v_1(\lambda)} + \tfrac{1}{2}\alpha\left(\dfrac{\lambda}{q_b}-1\right)(t_2-t_1) & \text{if } \lambda > q_b \ . \end{cases} \tag{6}$$

where α is the unit value of travel delay. (This holds so long as λ is less than the capacity of the road behind the bottleneck.) This cost function is shown in Figure 3, along with the associated marginal cost. In the region of ordinary congestion ($\lambda<q_b$), average cost is rising modestly; marginal cost exceeds it as in the conventional analysis. In the hypercongested region, marginal cost exceeds average cost by an additional amount $\frac{1}{2}\alpha(\lambda/q_b)(t_2-t_1)$, which is also the optimal time-invariant toll associated with the queue.[22] At $\lambda=q_b$ marginal cost is discontinuous. It is clear from Figure 3 that depending on the location of demand curve, optimal inflow λ could be less than, equal to, or greater than bottleneck capacity.[23]

We can further simplify by making $v_1(t)$ constant, justified by the empirical findings described earlier which indicate that the speed-flow relationship is quite flat in the region or ordinary congestion. Then the curves in Figure 3 become piecewise linear and perfectly flat in the region of ordinary congestion. Such a cost function is shown by Small (1992b)

---

[22]Small (1992b), p. 122.

[23]The discontinuity forces a reinterpretation in the case where demand crosses the marginal cost curve on its vertical portion. The external cost can then be interpreted as the value of trips foregone by the capacity constraint which is effectively imposed in this optimum.

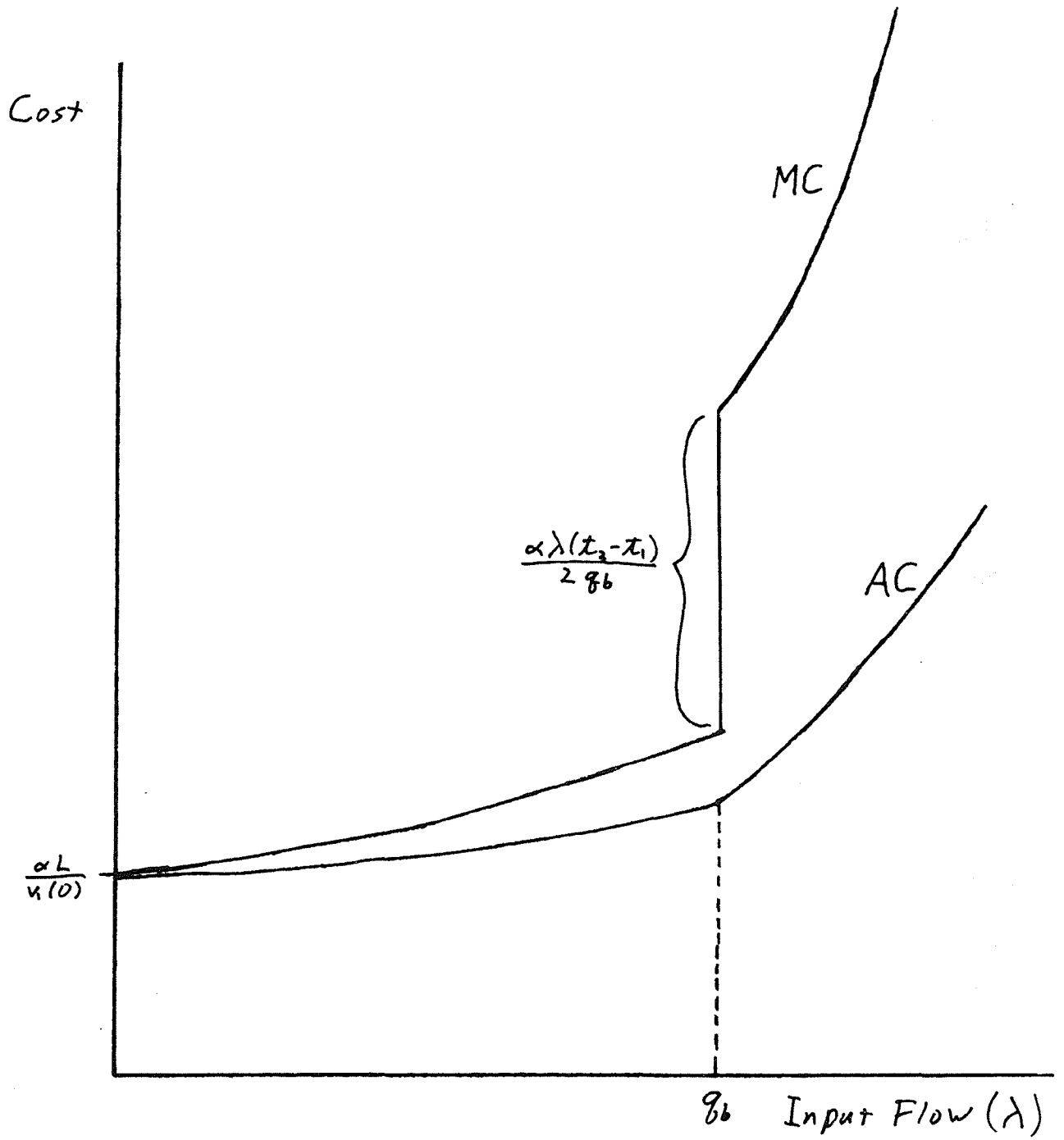


Figure 3. Cost Versus Inflow: Long Uniform Road Ending in Bottleneck, with Flat Demand Pulse

to give a reasonably good fit to Dewees' Toronto arterial data discussed earlier, as well as to some additional data from Boston expressways. This model was used in an application to a San Francisco Bay Area freeway by Small (1983). It turned out that optimal flow was frequently equal to capacity and never exceeded it, indicating that the vertical section of the marginal cost curve (i.e. the discontinuity in marginal cost) was quite high.

B. Endogenous Demand Pattern

Given that the road is well approximated by a point bottleneck, the endogenous trip scheduling analysis of Vickrey (1969), Fargier (1983), Newell (1987), and Arnott et al. (1990, 1993) provides an attractive alternative specification of how demand becomes expressed as entering flow rates. This work is summarized concisely by Small (1992a). The simplest version postulates N identical travelers, each with preferred trip-completion time t^* and per-minute costs β or γ for being earlier or later than that.[24] That is, average cost is:

$$c(t) = \begin{cases} \alpha T(t) + \beta(t^* - t) & \text{if } t \le t^* \\ \alpha T(t) + \gamma(t - t^*) & \text{if } t \ge t^* \end{cases} \tag{7}$$

[24]For technical reasons it is normal to assume $\beta<\alpha<\gamma$, which is supported empirically by Small (1981).

where t is the time the trip through the bottleneck is completed. Equilibrium requires that this cost be equalized at all times in the interval $[t_i, t_u]$ during which travel occurs. This condition implies:

$$\frac{dT(t)}{dt} = \begin{cases} \beta/\alpha & \text{for } t_i < t < t^* \\ -\gamma/\alpha & \text{for } t^* < t < t_u \end{cases} . \tag{8}$$

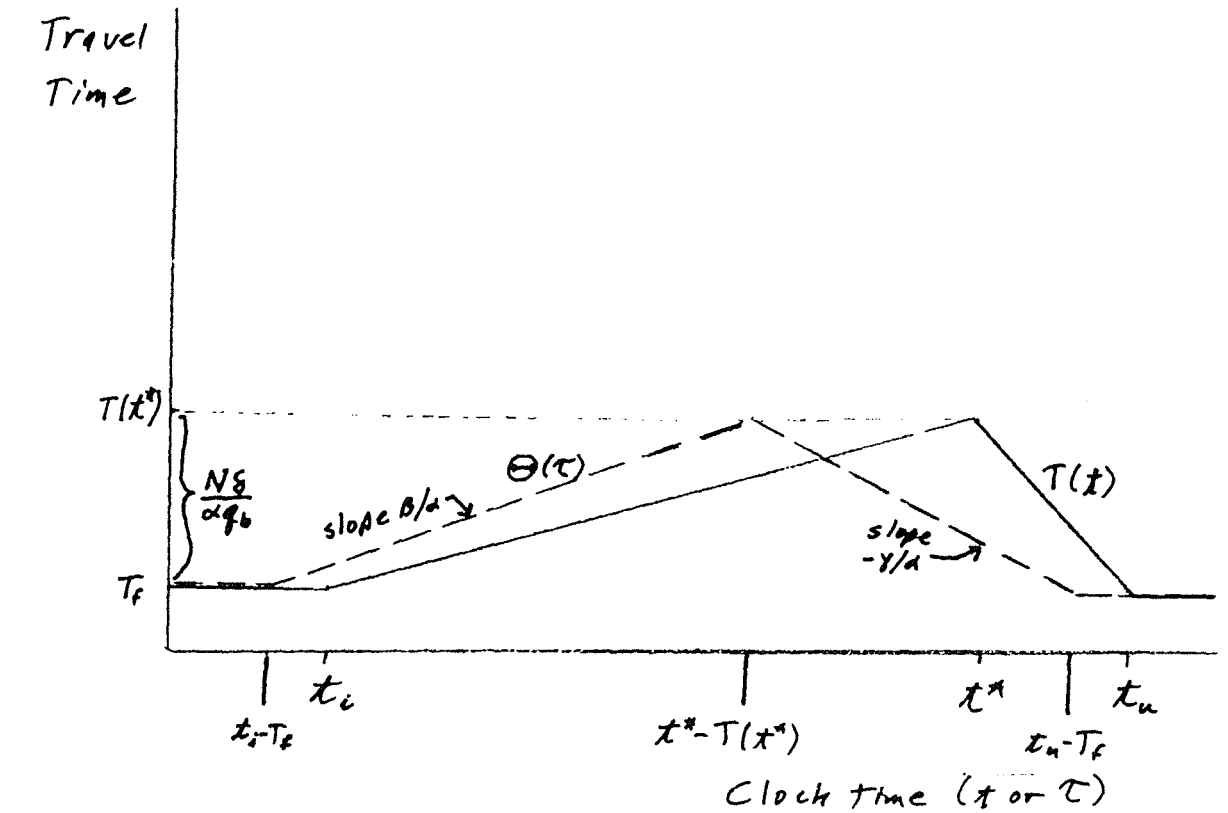
The equilibrium travel-time pattern therefore is that shown by the solid line in Figure 4a, with queueing delay beginning and ending at times t_i and t_u that are determined endogenously.

Travel time can be described equally well as a function of trip completion time, t, or as a function of trip start time, $\tau \equiv t - T(t)$. The latter function is shown as the dashed line in Figure 4a.[25] It is convenient because from it we can work backward to find the inflow pattern by applying deterministic queueing theory. Assuming a constant free-flow travel time, T_f, the inflow pattern is shown as the dashed line in Figure 4b. Inflow has two levels, the first greater than bottleneck capacity and the second less than capacity. Outflow, also shown, is equal to capacity throughout the period of travel. For ease of interpretation inflow is shown as a function of trip start time, while outflow is shown as a function of trip completion time.

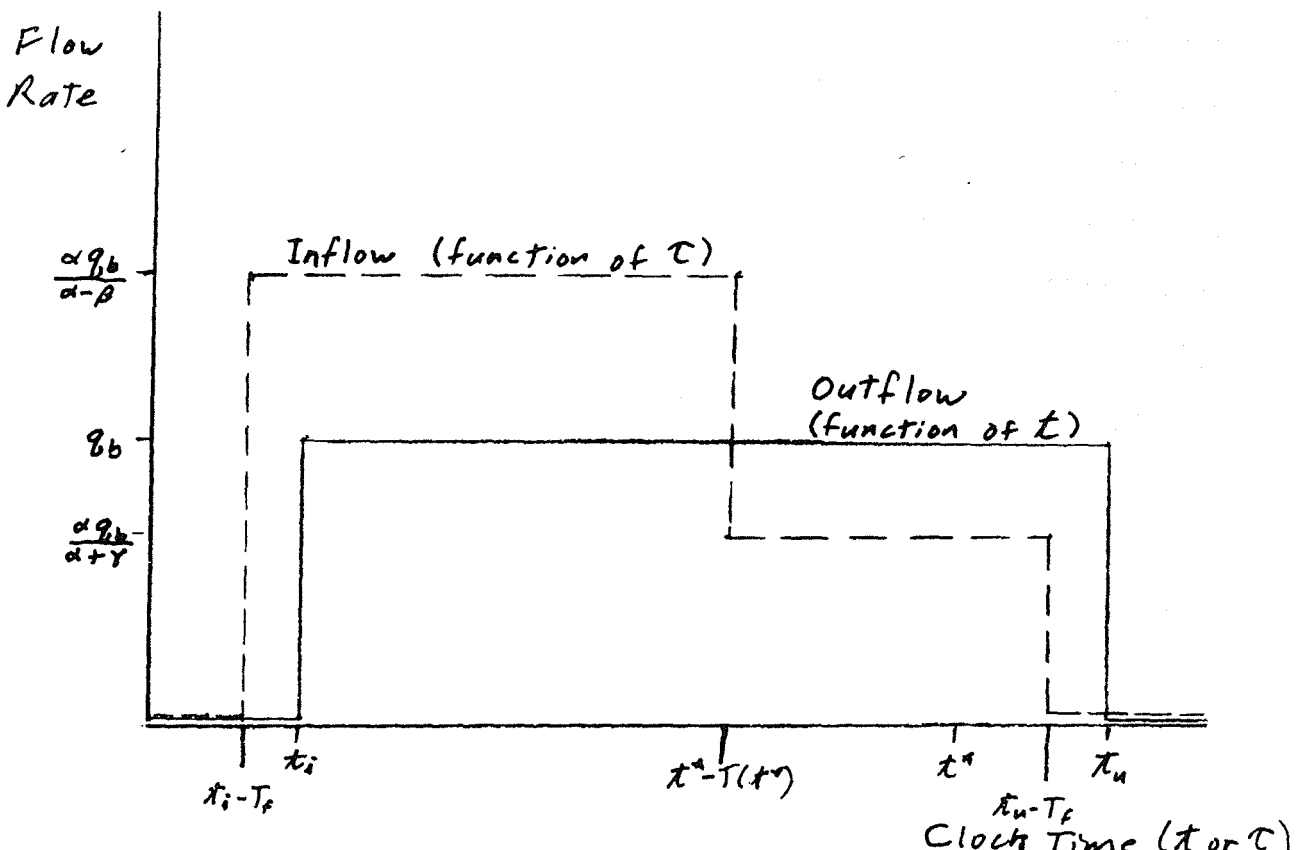
It turns out that the duration $t_u - t_i$ of the period of queueing is proportional to N. So is the equalized "net" average cost, $c(t) - \alpha T_f$, caused by the bottleneck. Remarkably, this equalized net average cost does not depend on α, the unit value of travel time. This result is particular to this case, but it shows how drastically our notions of the supply curve must be altered when endogenous scheduling is accounted for.

---

[25]Formally, this function, $\Theta(\tau)$, is defined as the solution to $T(\tau+\Theta)=\Theta$. Its slope is $\beta/(\alpha-\beta)$ for $t<\tilde{t}$ and $-\gamma/(\alpha+\gamma)$ for $t>\tilde{t}$, where $\tilde{t} = t^* - T(t^*)$. It is the function derived by Arnott et al. (1990).



(a) Travel Time as Function of Trip Start Time (τ) or Completion Time (t)



(b) Inflow and Outflow Rates

Figure 4. Equilibrium Travel Times, Inflow Rate, and Outflow Rate: Endogenous Trip Timing with Bottleneck

For fixed N, the optimal time-varying congestion toll equals the equilibrium travel delay multiplied by α, plus an arbitrary constant. The result of imposing it is that everyone's trip-completion time is unaltered, while queueing is completely eliminated.[26] Perceived price, equal to cost plus toll, is again equalized across travelers. Using this equalized perceived price as the basis for defining a demand curve, Arnott et al. (1993) show that the conventional static analysis is completely recovered by defining an appropriate average cost as a function of N, thereby determining the arbitrary constant in the time-varying toll. This average cost is increasing in N, once again showing that the relevant supply curve for static analysis is upward-sloping.

4. Modeling Hypercongestion in a Dense Street Network

Networks of city streets, unlike freeways, are prone to slowdowns in which various flows interfere with each other so much as to drastically slow traffic. As already noted this phenomenon implies the existence of numerous local queues; but in this case it seems implausible that they would obey the laws of deterministic queueing at an isolated bottleneck. Rather, individual queue discharge rates are likely to depend on traffic density in neighboring parts of the network due to cross traffic. We should therefore imagine a system in which density builds up when total input flow exceeds total exit flow, with the latter depending on the average density within the network.

To formalize this notion, we adapt a framework used by Agnew (1977) and Mahmassani and Herman (1984). We may think of our model as applying to a central business district (CBD). Trips inside the CBD begin and end either at its borders or at

[26]The more realistic case where preferred trip-completion times are uniformly distributed over a fixed time interval $[t_1,t_2]$ provides broadly similar results: see Small (1992b), pp. 89-93. Costs now depend on both the height and the duration of this desired demand spike. Because no queueing occurs until N exceeds $q_b \cdot (t_2-t_1)$, the average cost curve now is piecewise linear with a flat region followed by a much steeper region, like Figure 3 if v_1 were to be held constant.

parking spaces within. The CBD contains M lane-miles of streets, and average trip length is L. Vehicles enter the CBD streets at some rate $\lambda(t)$ vehicles per hour. At any time t traffic is characterized by two spatially aggregated variables: per-lane density $k(t)$ (vehicles per lane-mile), and average speed $v(t)$ (miles per hour). We make the following assumptions:

A1: Vehicles exit the streets at rate $(M/L)q(t)$, where

$$q(t) = k(t)v(t) \ . \tag{9}$$

Equation (9) defines $q(t)$ (measured in vehicles per hour per lane) as an average per-lane flow rate. One way this assumption could be realized is if all flows in the system contain the same fraction M/L of vehicles that are reaching their destinations.

A2: Average speed is related instantaneously to density by a functional relationship $V(\bullet)$:

$$v(t) = V[k(t)] \ . \tag{10}$$

That is, the fundamental diagram applies instantaneously in the aggregate. The function V is assumed to satisfy $V'<0$ and $kV''+2V'<0$ for all k. Hence flow kv is a single-humped function of k, rising from zero (at $k=0$) to a maximum q_m at some value k_m, then falling to zero at a density k_j known as the "jam density;" the region where it is falling is known as hypercongestion.

A3: Vehicles do not appear or disappear except through the entry and exit flows already identified. That is, the number of vehicles in the system changes according to:

$$M\frac{dk(t)}{dt} = \lambda(t) - (M/L)q(t) \ . \tag{11}$$

For the speed-density relationship, we adopt the empirical relationship measured by Ardekani and Herman (1987) from combined ground and air observations of the central business districts of Austin and Dallas, Texas. The functional form comes from the "two-fluid" theory of Herman and Prigogine (1979), in which moving vehicles and stopped vehicles follow distinct laws of motion. Letting K denote the "normalized density" k/k_j, the Ardekani-Herman (AH) formula is:

$$v(t) = v_f\left[1 - K(t)^{\pi}\right]^{1+\rho} \tag{12}$$

where v_f is the free-flow speed and π and ρ are additional parameters. This formula implies a maximum flow of:

$$q_m = v_f k_j \frac{(1+\rho)^{1+\rho}}{(2+\rho)^{2+\rho}} \tag{13}$$

occurring at density

$$k_m = \frac{k_j}{2+\rho} \ . \tag{14}$$

A special case of the AH formula, for $\pi=1$ and $\rho=0$, is the Greenshields (1935) linear speed-density relationship:

$$v = v_f(1-K) \ . \tag{15}$$

This simple relationship implies that flow is a quadratic function of speed; the congested and hypercongested branches are the two roots of the quadratic. The Greenshields relationship is used frequently in both the engineering and economics literature on congestion, and was used to draw Figure 1. Maximum flow $q_m = \frac{1}{4} v_f k_j$ occurs at density $k = \frac{1}{2} k_j$.

We now proceed to apply this supply model to the same two demand models considered in the previous section: first an exogenous demand spike, then an endogenous demand pattern generated by linear scheduling costs.

A. Exogenous Demand Spike

Assume again that commuters enter the CBD network at a uniform rate λ over a fixed peak period $[t_1, t_2]$. We use the Greenshields special case of the speed-density relationship, equation (15). By substituting (9) and (15) into (11), we obtain a differential equation in normalized density that applies for $t_1 < t < t_2$:

$$T_f \frac{dK}{dt} = \frac{\lambda}{4\mu_m} - K(1-K) \tag{16}$$

where $\mu_m \equiv (M/L) q_m \equiv \frac{1}{4}(M/L) v_f k_j$ is the maximum possible exit flow (completed trips per hour) and $T_f \equiv L/v_f$ is the free-flow average trip time. The boundary condition is $K(t_1) = 0$. After t_2, the same equation applies but with λ replaced by zero and with boundary condition that density be continuous at t_2.

Once $K(t)$ is determined, $v(t)$ follows from (10), and the distance traveled through any small time interval dt can be calculated as $dt/v(t)$. Travel time $\Theta(\tau)$ for a trip of length L beginning at time τ is then the solution of the following equation:

$$\int_{\tau}^{\tau+\Theta(\tau)} \frac{dt'}{v(t')} = L \ . \tag{17}$$

Equivalently, the travel time $T(t)$ for a trip *exiting* the system at time t is the solution to:

$$\int_{t-T(t)}^{t} \frac{dt'}{v(t')} = L \ . \tag{17a}$$

The solution for $K(t)$ is provided in the Appendix and is shown in Figure 5. Its broad properties can be inferred just by inspecting equation (16). Recall that the term $-K(1-K)$ is zero when $K=0$ or $K=1$, and it reaches its most negative value when $K=½$. The curve portraying density $K(t)$ therefore starts upward at time t_1 with initial slope $\lambda/(4\mu_m T_f)$. As time progresses the curve becomes flatter due to the term $-K(1-K)$, then becomes steeper again when and if K increases beyond ½. At time t_2, the slope undergoes a discontinuity and becomes negative, the curve being steepest near $K=½$ but then flattening and approaching zero asymptotically. Thus we have a period of density buildup during time interval $[t_1, t_2]$ followed by a gradual relaxation back toward free-flow conditions.

The solution has different regimes depending on the value of λ. If $\lambda \leq \mu_m$, normalized density builds asymptotically to a value less than or equal to ½. Thus hypercongestion does not occur, and the inflows can be maintained indefinitely. (The dashed curves in the figure show the paths that would be taken if the inflow were not ended at t_2.) But if $\lambda > \mu_m$, the system reaches maximum outflow q_m with inflow still exceeding q_m. At this

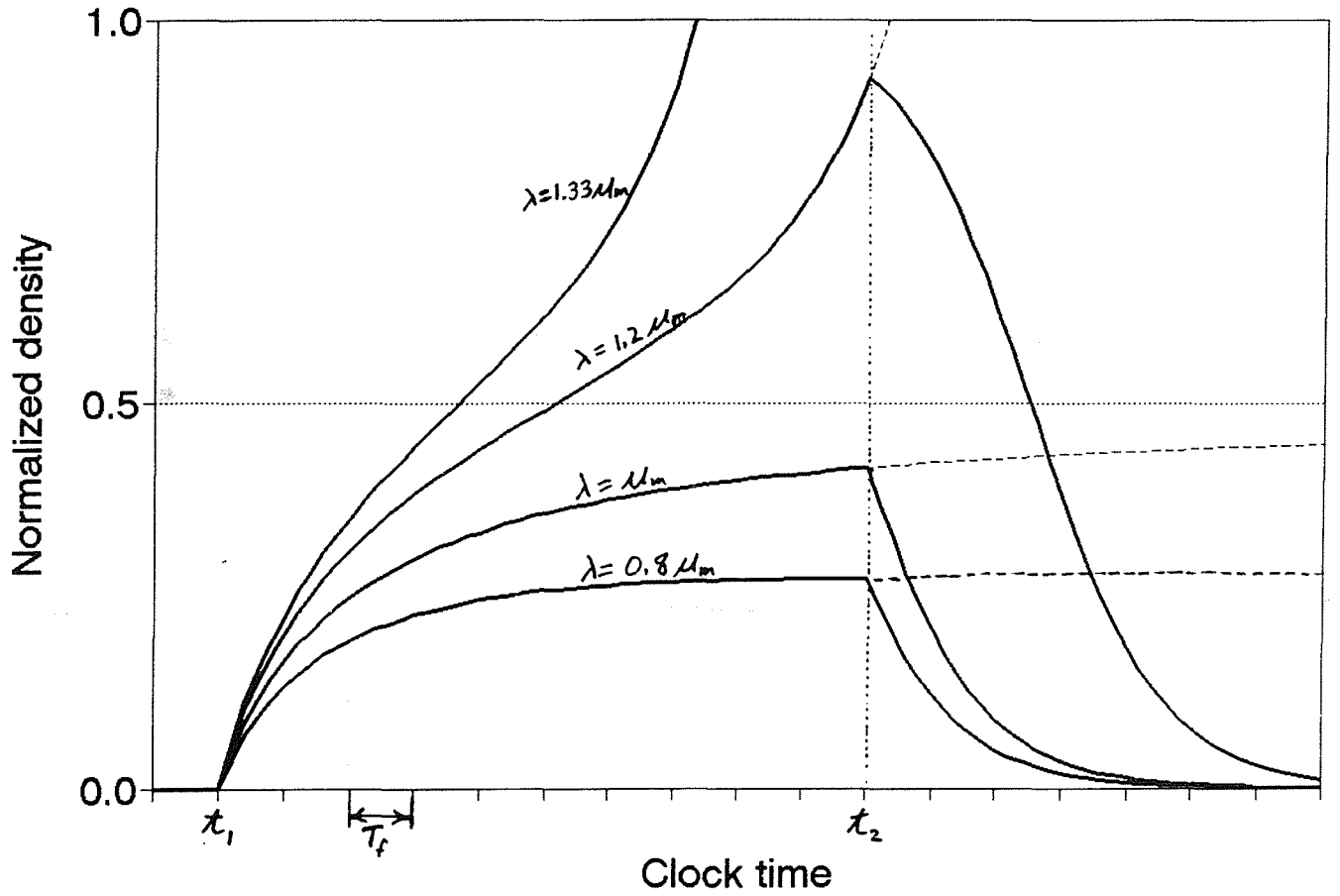


Figure 5. Dense Street Network, Exogenous Demand Spike

point density builds up precipitously. This is the region of hypercongestion, and outflow declines steadily. If t_2 comes soon enough, a rather long discharge period begins: it is especially long if K has nearly reached one so that $K(1\text{-}K)$ in (16) is small, indicating a condition where outflow is nearly blocked. But if t_2 exceeds a "jam time" t_j whose value is given in the Appendix, density reaches jam density ($K=1$) and the system breaks down: no more vehicles can enter. This applies to the leftmost curve in Figure 5.

Some numbers are helpful. If $\lambda=1.05\bullet\mu_m$, hypercongestion is reached at time $t_h=t_1+12.08\bullet T_f$, and jam density is reached at time $t_j=t_1+25.63\bullet T_f$. Since T_f is the free-flow time for the average trip, probably just a few minutes in a typical CBD, these times are not unreasonable for the duration of a typical urban rush hour. But if $\lambda=1.33\bullet\mu_m$, conditions deteriorate fairly quickly: hypercongestion is reached after a time interval of just $3.63\bullet T_f$ and breakdown occurs after $8.57\bullet T_f$.

Thus, this model does not seem to be able to handle rush hours of realistic duration unless inflow λ is limited to just a little above capacity μ_m. In real cities, where "rush hours" last for hours, some form of demand elasticity must be maintaining a rather delicate balance between λ and μ_m. One possibility is that people rearrange their trip schedules in response to the sharp changes in travel times over the rush hour.[27] This possibility is formally modeled in the next subsection, using the same demand structure as before.

B. Endogenous Demand Pattern

We have seen in Section 3B that the model of endogenous scheduling, subject to linear scheduling costs, has proven highly productive when applied to a single deterministic bottleneck. Applying it to a hypercongested street network adds serious

[27]Nagel and Rasmussen (1994) show through stochastic simulation of a car-following model that such a balance can produce a very high travel-time variance when driver behavior contains random elements.

complications. But we can overcome them by means of an approximation used by Henderson (1981) and Mahmassani and Herman (1984), in which we assume the speed for an entire trip is determined solely by conditions encountered at one point in the trip.

This assumption is clearly a drastic one, and its use by Mahmassani and Herman (MH) provoked vehement objection by Newell (1988). However, it dramatically simplifies the problem. Furthermore, its accuracy as an approximation can be determined by recomputing travel time, after the solution is obtained, through equation (17a) above. MH in fact determined trip times this way, apparently without realizing that a different pattern of trip times was already implicit in their solution method.

One other important modification must be made to the formulations of Henderson and of MH. For technical reasons, it is the speed at the *end* of the trip, not the beginning as assumed in both of these papers, that must determine trip time. Otherwise we get a discontinuity in travel time at the end of the rush hour that causes two inconsistencies: it is incompatible with equilibrium for the last traveler, and it implies that a vehicle can overtake another which started earlier.[28] Note that the approximation involved is no more severe using trip completion times than using trip start times; formally, there is a one-to-one translation between them and all dynamic equilibrium conditions can be stated equally well as functions of the times vehicles enter the system or of the times they exit.

We therefore assume in this section that the travel time of a trip *completed* at time t is:

[28]These technical problems, explained fully by Chu (1995, 1994), are related to the fact that the scheduling costs are assumed to depend on trip completion time. Presumably the situation would be reversed if scheduling costs were determined by deviation from a desired trip start time, as one might wish to postulate for an afternoon rush hour.

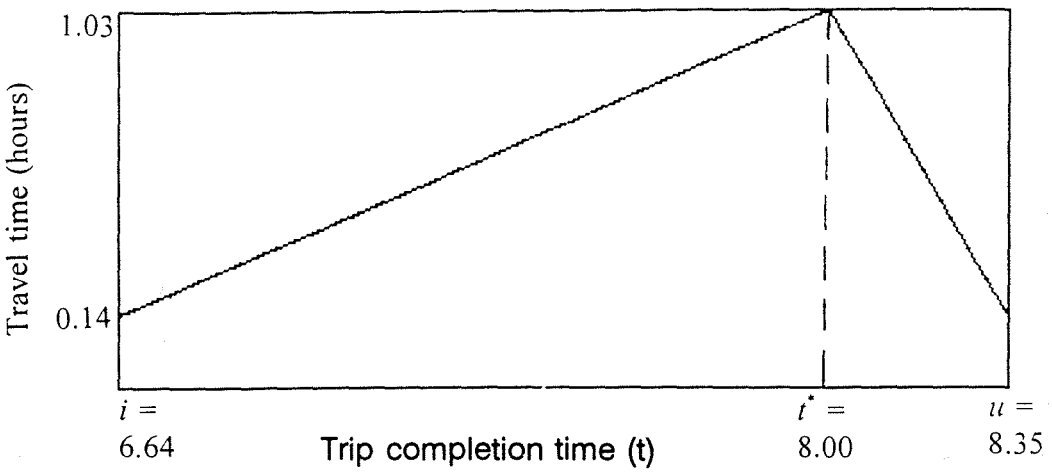
$$T(t) = \frac{L}{v(t)} \,. \tag{18}$$

The reason this assumption simplifies the problem so much is that the shape of the equilibrium travel-time pattern is completely determined by the requirement that total trip cost be the same for everyone. This is just as in the bottleneck model of section 3b, and the result is again a triangular shaped pattern — just like that in Figure 4a, with $T(t)$ obeying equation (8). The pattern of flows, however, is quite different than in the bottleneck model. Hence so are the duration of the rush hour and the maximum travel delay. This is because the exit rate of vehicles during the period of congestion is no longer constant at bottleneck capacity, but instead varies with conditions.

We solve the system in the Appendix, using the Ardekani-Herman speed-density relationship, equation (12), for the special case $\pi=1$. Figures 6 and 7 show the results for a particular set of parameters pertaining to the Dallas CBD,[29] along with scheduling and travel-time cost parameters from Arnott et al. (1990) and certain arbitrarily chosen values: $t^*=8.00$ hours, $L=4$ miles, and $N=10{,}000$ vehicles. Figure 6 shows the triangular equilibrium travel-time pattern and the endogenous pattern of entry flows required to generate it. Figure 7 shows how density and outflow vary as part of the solution.

We see that density follows a concave time path, rising at a decreasing rate from the start of the rush hour (at time $t_i=6.64$ hours) up to t^*, then falling at an increasing rate until it reaches zero at time $t_u=8.35$. Hypercongestion occurs between times 7.07 and 8.24. Vehicle outflow rises to the maximum possible value q_m, as given by equation (13), which is reached at the time when hypercongestion begins. Outflow is below the

[29]These are: $M=117$ lane-miles; $v_f=27.54$ miles/hour; $k_j=100$ vehicles/lane-mile; and $\rho=1.67$. The implied maximum internal flow is $q_m=233$ vehicles/lane-hour, which occurs at density 27 vehicles per lane-mile and speed 8.5 miles/hour.



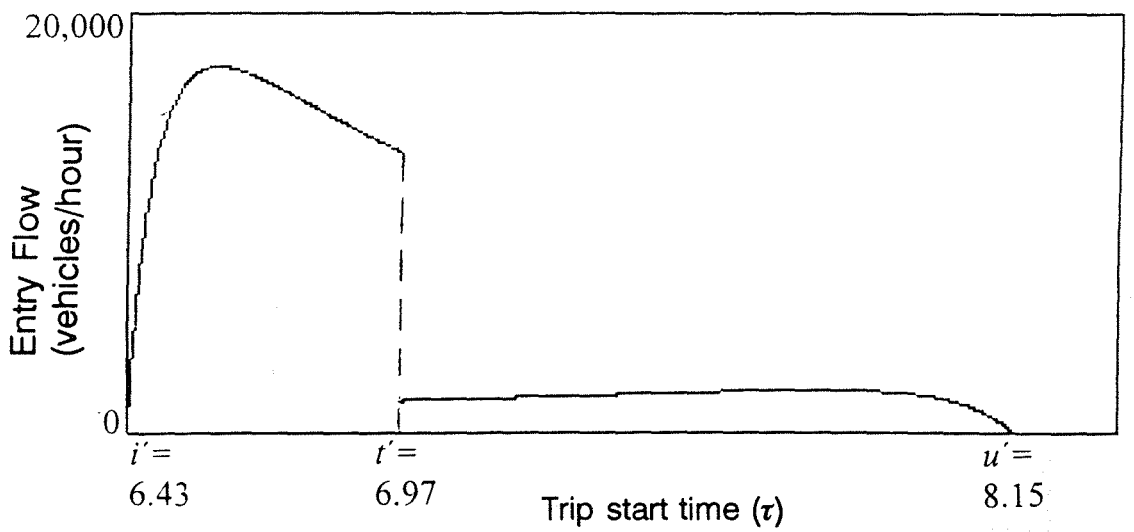
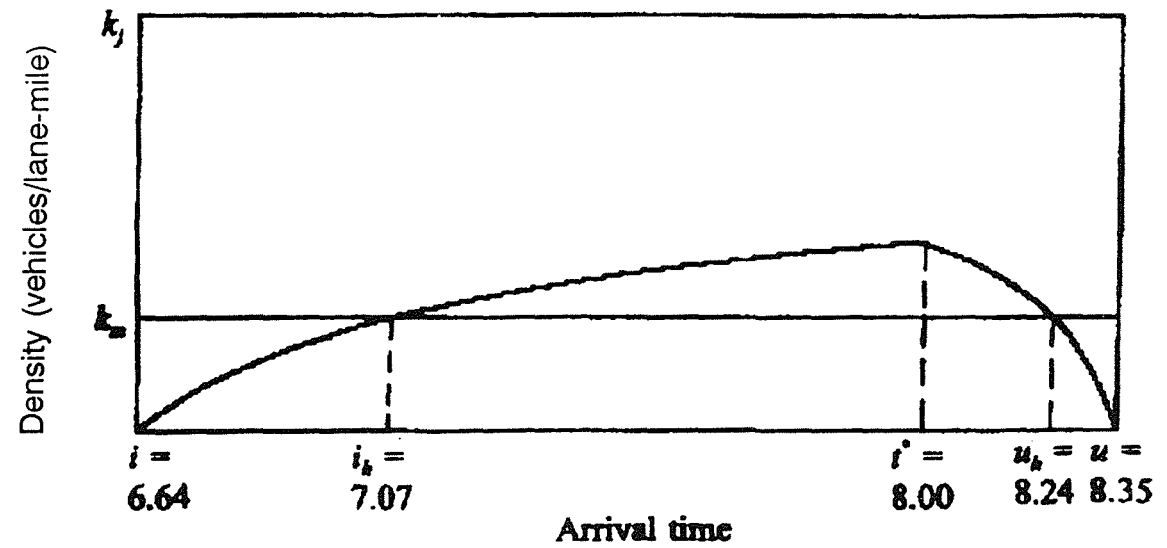


Figure 6. Travel Time and Entry Flow:
Endogenous Trip Timing on a Hypercongested Street Network



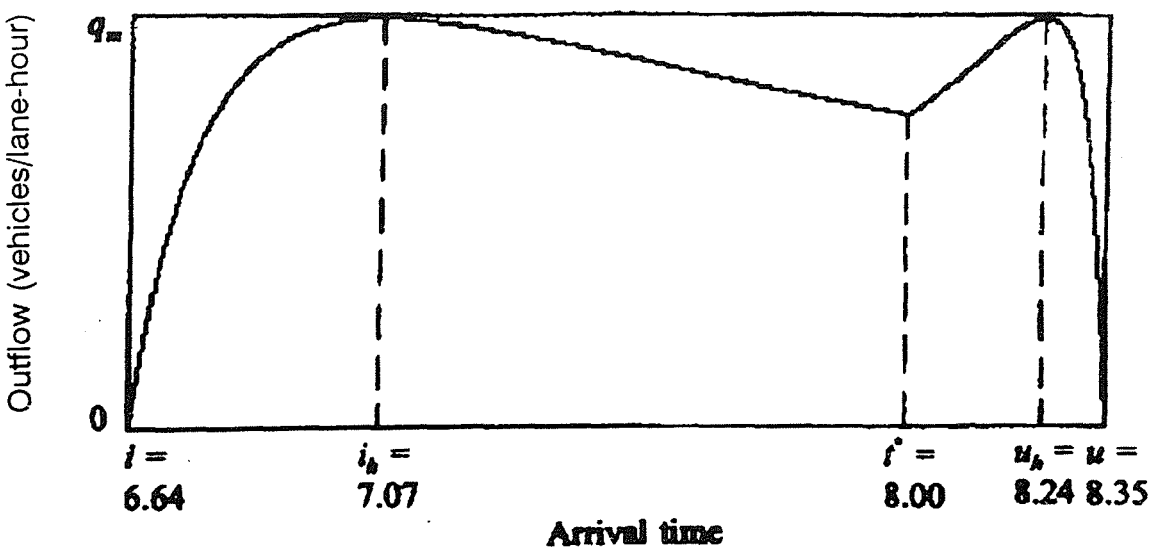


Figure 7. Density and Outflow Rate:
Endogenous Trip Timing on a Hypercongested Street Network

maximum during the hypercongested period, falling to its lowest value just at t^*. It reaches the maximum again when hypercongestion ends, then falls rapidly to zero.

The existence of hypercongestion imposes additional scheduling costs because even fewer people can complete their trips near the desired time t^* than would be true for a point bottleneck with capacity q_m. Table 1 shows the numerical results for the simulation just described and also for one with a smaller N, for which no hypercongestion occurs. Among the results reported is the ratio of schedule delay costs to total costs. This is the proportion of the costs in equation (7), added up over all travelers, that is accounted for by the terms involving β or γ. In the bottleneck model with $T_f=0$, this proportion is exactly one-half, and in Henderson's model it is always less than this, approaching it as the exponent in (4) becomes larger (Chu (1995)). In our simulations, the proportion is 50 percent in the low-demand case (without hypercongestion) and 60 percent in the high-demand case (with hypercongestion), even though we include free-flow time in total cost. This suggests that in hypercongested situations it is especially important to account for endogenous scheduling and its associated costs.

5. Conclusion

Hypercongestion is a real phenomenon, potentially creating inefficiencies and imposing considerable costs. However, it cannot be understood within a steady-state analysis because it does not persist as a steady state. Rather, hypercongestion occurs as a result of transient demand surges and can be fully analyzed only within a dynamic model. Even if the dynamic model is converted to a static one through the use of time averaging, the appropriate specification of average cost depends on the underlying dynamics. In virtually all circumstances that specification will portray average cost as a rising function even when hypercongestion occurs.

Table 1. Simulated Equilibrium for Dense Street Network

Number of Commuters [N]	2,500	10,000
Duration of peak period, hours (t_u-t_i)	0.49	1.71
Duration of hypercongestion, hours (t_{uh}-t_{ih})	0	1.16
Fraction of trips which encounter hypercongestion	0	0.71
Peak congestion delay, hours [$T(t^*)$]	0.30	0.89
Peak normalized density [$K(t^*)$]	0.25	0.46
Peak density as fraction of critical density for hypercongestion [$K(t^*)\bullet(2+\rho)$]	0.92	1.69
Flow at peak density as fraction of max flow [$q(t^*)/q_m$]	0.99	0.76
Average travel-time cost, \$/trip ($\alpha\bar{T}$)	\$1.41	\$3.24
Average scheduling cost, \$/trip ($\bar{c}$-$\alpha\bar{T}$)	\$1.39	\$4.79
Average total cost, \$/trip ($\bar{c}$)	\$2.80	\$8.03
Ratio of scheduling to total cost [($\bar{c}$-$\alpha\bar{T}$)/$\bar{c}$]	0.50	0.60

In one important case, that of a uniform length of highway ending in a bottleneck, hypercongestion turns out not to be very important. This is because hypercongestion just describes the density of vehicles within the queue, whose discharge rate is governed by bottleneck capacity. We do not really need to know the density of vehicles in the queue unless we are worried about the queue backing up and blocking another entrance or exit, or unless we want to account for the rather small difference the queue's physical length makes to the free-flow travel time required to reach it from further upstream.

In another important case, that of a dense street network, it is plausible to model flow within a well-defined area as subject to hypercongestion. We have shown that a dynamic model incorporating this feature can be constructed and solved at least for special cases of the demand pattern. Doing so explains features that we observe in real cities: the gradual buildup of vehicle density during a rush hour, with dramatic and quite sudden slowdowns possible if density reaches the hypercongested region. A state of total breakdown, where speed falls to zero, is theoretically possible: this is gridlock in its literal meaning, with the various local queues on the network totally blocking each other. There is no way out of gridlock within the model. However, severe congestion short of gridlock ultimately dissipates once the demand surge abates.

One promising way to model these demand surges is by means of the endogenous scheduling models that have worked their way prominently into both the economics and engineering literatures. We show how ideas from two of these models, developed previously for situations lacking hypercongestion, can be applied to a dense street network subject to hypercongestion.

References

Agnew, Carson E. (1977): "The Theory of Congestion Tolls," *Journal of Regional Science*, 17, 381-393.

Ardekani, Siamak, and Robert Herman (1987): "Urban Network-Wide Traffic Variables and Their Relations," *Transportation Science*, 21, 1-16.

Arnott, Richard (1979): "Optimal City Size in a Spatial Economy," *Journal of Urban Economics*, 6, 65-89.

Arnott, Richard (1990): "Signalized Intersection Queuing Theory and Central Business District Auto Congestion," *Economics Letter*, 33, 197-201.

Arnott, Richard, Andre de Palma, and C. Robin Lindsey (1990): "Economics of A Bottleneck," *Journal of Urban Economics* 27, 111-130.

Arnott, Richard, Andre de Palma, and C. Robin Lindsey (1993): "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83, 161-179.

Banks, James H. (1989): "Freeway Speed-Flow-Concentration Relationships: More Evidence and Interpretations," *Transportation Research Record*, 1225, 53-60.

Beckmann, Martin, C.B. McGuire, and Christopher B. Winsten (1955): *Studies in the Economics of Transportation*. New Haven: Yale University Press.

Berglas, Eitan, and David Pines (1981): "Clubs, Local Public Goods and Transportation Models," *Journal of Public Economics*, 15, 141-162.

Boiteux, M. (1949): "La Tarification des Demandes en Pointe: Application de la Théorie de la Vente au Coût Marginal," *Revue Générale de l'Électricité*, August. Reprinted in English translation as "Peak-Load Pricing," *Journal of Business*, 33 [1960], 157-179.

Branston, David (1976): "Link Capacity Functions: A Review, *Transportation Research*, 10, 223-236.

Button, Kenneth J. (1993): *Transport Economics*, Aldershot, UK: Edward Elgar.

Chu, Xuehao (1995): "Endogenous Trip Scheduling: The Henderson Approach Reformulated and Compared with the Vickrey Approach," *Journal of Urban Economics*, 37, 324-343.

Chu, Xuehao (1994): "A Structural Model of Hypercongestion for Street-Network Traffic," Center for Urban Transportation Research.

De Meza, David, and J.R. Gould (1987): "Free Access versus Private Property in a Resource: Income Distributions Compared," *Journal of Political Economy*, 95, 1317-1325.

DeVany, Arthur S., and Thomas R. Saving (1980): "Competition and Highway Pricing for Stochastic Traffic," *Journal of Business*, 53, 45-60.

Dewees, Donald N. (1978): "Simulations of Traffic Congestion in Toronto," *Transportation Research*, 12, 153-165.

Dewees, Donald N. (1979): "Estimating the Time Costs of Highway Congestion," *Econometrica*, 47, 1499-1512.

Dillon, Dan S., and Fred L. Hall (1987): "Freeway Operations and the Cusp Catastrophe: An Empirical Analysis," *Transportation Research Record*, 1132, 66-76.

Edelson, Noel M. (1971): "Congestion Tolls Under Monopoly," *American Economic Review*, 61, 873-882.

Else, P.K. (1981): "A Reformulation of the Theory of Optimal Congestion Taxes," *Journal of Transport Economics and Policy*, 15, 217-232.

Evans, Alan W. (1992): "Road Congestion: The Diagrammatic Analysis," *Journal of Political Economy*, 100, 211-217.

Evans, Andrew W. (1993): "Road Congestion Pricing: When Is It a Good Policy? A Rejoinder," *Journal of Transport Economics and Policy*, 27, 99-105.

Fargier, Paul-Henri (1983): "Effects of the Choice of Departure Time on Road Traffic Congestion," in *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory*, ed. by V.F. Hurdle, E. Hauer, and G.N. Steuart. Toronto: University of Toronto Press, 223-263.

Fisk, Caroline (1979): "More Paradoxes in the Equilibrium Assignment Problem," *Transportation Research*, 13B, 305-309.

Greenshields, B.D. (1935): "A Study of Traffic Capacity," *Highway Research Board Proceedings*, 14, Part I, 448-477.

Haight, Frank (1963): *Mathematical Theories of Traffic Flow*. New York: Academic Press.

Hall, Fred L., and Lisa M. Hall (1990): "Capacity and Speed Flow Analysis of the QEW in Ontario," *Transportation Research Record*, 1287, 108-118.

Hall, Fred L., V.F. Hurdle, and James H. Banks (1992): "Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways," *Transportation Research Record*, 1365, 12-18.

Henderson, J. Vernon (1981): "The Economics of Staggered Work Hours," *Journal of Urban Economics*, 9, 349-364.

Herman, Robert, and Ilya Prigogine (1979): "A Two-Fluid Approach to Town Traffic," *Science*, 204, 148-151.

Hills, Peter (1993): "Road Congestion Pricing: When Is It a Good Policy? A Comment," *Journal of Transport Economics and Policy*, 27, 91-99.

Lévy-Lambert, H. (1968): "Tarification des Services à Qualité Variable — Application aux Péages de Circulation," *Econometrica*, 36, 564-574.

Lighthill, M.H., and G.B. Whitham (1955): "On Kinematic Waves, II: A Theory of Traffic Flow on Long Crowded Roads," *Proceedings of the Royal Society* (London), A229, 317-345.

Mahmassani, Hani, and Robert Herman (1984): "Dynamic User Equilibrium Departure Times and Route Choice on Idealized Traffic Arterials," *Transportation Science* 18, 362-384.

Marchand, Maurice (1968): "A Note on Optimal Tolls in an Imperfect Environment," *Econometrica*, 36: 575-581.

May, Adolf D. (1987): "Freeway Simulation Models Revisited," *Transportation Research Record*, 1132, 94-98.

May, Adolf D., and Hartmut E.M. Keller (1966): "A Deterministic Queueing Model," paper presented at the Operations Research Society of America, Twenty-Ninth National Meeting, Santa Monica, California. Berkeley: University of California, Institute of Transportation Studies.

McDonald, John F., and Edmond L. d'Ouville (1988): "Highway Traffic Flow and the 'Uneconomic' Region of Production," *Regional Science and Urban Economics*, 18, 503-509.

Mills, David E. (1981): "Ownership Arrangements and Congestion-Prone Facilities," *American Economic Review*, 71, 493-502.

Mohring, Herbert (1970): "The Peak Load Problem with Increasing Returns and Pricing Constraints," *American Economic Review*, 60, 693-705.

Mun, Se-Il (1994): "Traffic Jams and the Congestion Toll," *Transportation Research*, 28B, 365-375.

Nagel, Kai, and Steen Rasmussen (1994): "Traffic at the Edge of Chaos," in *Artificial Life IV*, ed. by Rodney A. Brooks and Pattie Maes. Cambridge, Massachusetts: MIT Press, pp. 222-235.

Newbery, David M. (1990): "Pricing and Congestion: Economic Principles Relevant to Pricing Roads," *Oxford Review of Economic Policy*, 6, 22-38.

Newell, Gordon F. (1987): "The Morning Commute for Nonidentical Travelers," *Transportation Science*, 21, 74-88.

Newell, Gordon F. (1988): "Traffic Flow for the Morning Commute," *Transportation Science* 22, 47-58.

Oakland, William H. (1972): "Congestion, Public Goods and Welfare," *Journal of Public Economics*, 1, 339-357.

Rouphail, Nagui M., and Rahmi Akçelik (1992): "Oversaturation Delay Estimates with Consideration of Peaking," *Transportation Research Record*, 1365, 71-81.

Small, Kenneth A. (1982): "The Scheduling of Consumer Activities: Work Trips," *American Economic Review*, 72, 467-79.

Small, Kenneth A. (1983): "Bus Priority and Congestion Pricing On Urban Expressways," in *Research in Transportation Economics*, Vol. 1, ed. by Theodore E. Keeler. Greenwich, Connecticut: JAI Press, 27-74.

Small, Kenneth A. (1992a): "Trip Scheduling in Urban Transportation Analysis," *American Economic Review, Papers and Proceedings*, 82(2), 482-486.

Small, Kenneth A. (1992b): *Urban Transportation Economics*, Vol. 51 of *Fundamentals of Pure and Applied Economics* series. Chur, Switzerland: Harwood Academic Press.

Transportation Research Board (1992): *Highway Capacity Manual*, 3rd edition. TRB *Special Report* 209. Washington: National Academy Press.

Vickrey, William S. (1955): "Some Implications of Marginal Cost Pricing for Public Utilities," *American Economic Review, Papers and Proceedings*, 45(2), 605-620.

Vickrey, William S. (1963): "Pricing in Urban and Suburban Transport," *American Economic Review, Papers and Proceedings*, 53, 452-465.

Vickrey, William (1969): "Congestion Theory and Transport Investment," *American Economic Review* 59, 251-261.

Vickrey, William (1991): "Congestion in Midtown Manhattan in Relation to Marginal Cost Pricing," mimeo, Columbia University, May.

Walters, A.A. (1961): "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, 29, 676-699.

Walters, A.A. (1987): "Congestion," in *The New Palgrave: A Dictionary of Economics*, Macmillan, New York.

Williams, James C., Hani S. Mahmassani, and Robert Herman (1987): "Urban Traffic Network Flow Models," *Transportation Research Record*, 1112, 78-88.

Appendix: Solutions for Section 4

A. Exogenous Demand Spike

Equation (16) can be solved for $K(t)$ by writing it as:

$$\frac{dK}{(K-\frac{1}{2})^2 - \frac{1}{4}\left(1 - \frac{\lambda}{\mu_m}\right)} = \frac{dt}{T_f} . \tag{A1}$$

We describe here the solution for $t_1<t<t_2$, which allows us to characterize the response to hypercongestion. The boundary condition is $K(t_1)=0$.

The integral of the left-hand side of (A1) depends on the sign of the second term in the denominator. If $\lambda<\mu_m$, this term is negative and integrating (A1) yields:

$$\frac{1}{2B} \ln \left| \frac{B-(K-\frac{1}{2})}{B+(K-\frac{1}{2})} \right| = \frac{t}{T_f} + C \tag{A2}$$

where C is a constant of integration. Applying the boundary condition to determine C and solving for K yields:

$$K(t) = \frac{1}{2} - B\frac{g(t) + 1}{g(t) - 1} \tag{A3}$$

where

$$B = \frac{1}{2}\sqrt{1 - \frac{\lambda}{\mu_m}} \tag{A4}$$

and

$$g(t) = \frac{½ + B}{½ - B} e^{2B(t-t_1)/T_f} . \quad \text{(A5)}$$

As $t \to \infty$, $K \to ½$-B, a positive constant less than ½. Thus hypercongestion does not occur.

If $\lambda = \mu_m$, the integral of the land-hand side of (A1) is $-(K-½)^{-1}$, and the solution is:

$$K(t) = ½ - \frac{1}{2 + (t-t_1)/T_f} . \quad \text{(A6)}$$

This can also be derived as the limit of (A3) as $B \to 0$. It approaches ½ as $t \to \infty$.

If $\lambda > \mu_m$, the integral of the left-hand side of (A1) is $(1/\bar{B})\arctan[(K-½)/\bar{B}]$, where:

$$\bar{B} = ½ \sqrt{\frac{\lambda}{\mu_m} - 1} . \quad \text{(A7)}$$

The solution is:

$$K(t) = ½ + \bar{B} \tan \left[\frac{\bar{B}(t-t_1)}{T_f} - \arctan \left(\frac{1}{2\bar{B}} \right) \right] . \quad \text{(A8)}$$

Normalized density K reaches ½, the onset of hypercongestion, at time t_h given by:

$$t_h = t_1 + \frac{T_f}{\bar{B}} \arctan \left(\frac{1}{2\bar{B}} \right) \quad \text{(A9)}$$

and jam density is reached at time

$$t_j = t_h + \frac{T_f}{\bar{B}} \arctan\left(\frac{2}{\bar{B}}\right) . \qquad \text{(A10)}$$

For $t>t_j$ the model breaks down and no solution can be given; in reality outside intervention is needed to halt inflow until the built-up density can be discharged.

Assuming $K(t_2)$ is defined, the solution for $t>t_2$ is obtained by using (A2) with $B=1$ and with the constant of integration chosen to make $K(t)$ continuous at t_2. The result is $K(t)=[1+a\bullet e^{\theta}]^{-1}$ where $\theta=(t-t_2)/T_f$ and $a=(1-K_2)/K_2$, K_2 being the value $K(t_2)$ from the solution during the time interval $[t_1, t_2]$. This solution is downward-sloping, approaches zero asymptotically, and has an inflection point when and if $K=½$.

B. Endogenous Demand Pattern

First, we formulate the differential equation that will determine normalized traffic density $K(t)$. Substituting (12) into (18), we differentiate the result to obtain the time derivative, dT/dt, that is consistent with the flow dynamics:

$$\frac{dT}{dt} = T_f(1+\rho)\pi(1-K^{\pi})^{-(2+\rho)}\frac{dK}{dt} . \qquad \text{(A11)}$$

But (8) gives the time derivative that is consistent with equilibrium in scheduling. Equating the two yields:

$$\frac{dK}{dt} = \frac{\sigma}{(1+\rho)\pi} K^{1-\pi}(1-K^{\pi})^{2+\rho} \qquad \text{(A12)}$$

where $\sigma=\beta/(\alpha T_f)$ for $t<t^*$ and $\sigma=\gamma/(\alpha T_f)$ for $t>t^*$. We solve this differential equation for the case $\pi=1$, which is very close to the value of 0.95 measured for Dallas. The boundary conditions for the two regions $t<t^*$ and $t>t^*$ are, respectively, $K(t_i)=0$ and $K(t_u)=0$, which can be written equivalently as $T(t_i)=T(t_u)=T_f$. Here t_i and t_u are the times the first and last trips are completed and their values are still to be determined. We can eliminate one of them by noting that $K(t)$ and hence $T(t)$ must be continuous at t^*; using (8) this means that

$$\beta(t^*-t_i) = \gamma(t_u-t^*) \equiv s\alpha T_f \tag{A13}$$

where s is thereby defined as the equalized ratio of schedule delay cost to travel-time cost for the first and last travelers. (We could have deduced this equality directly from the fact that the first and last travelers suffer no travel delay and so must have equal scheduling costs.)

Equation (A12) with $\pi=1$ is solved by writing it as:

$$\frac{dK}{(1-K)^{2+\rho}} = \frac{\sigma dt}{1+\rho} . \tag{A14}$$

The solution is:

$$K(t) = \begin{cases} 1 - \left(1 + \dfrac{\beta(t-t_i)}{\alpha T_f} \right)^{-1/(1+\rho)} & \text{for } t_i < t < t^* \\ 1 - \left(1 + \dfrac{\gamma(t_u-t)}{\alpha T_f} \right)^{-1/(1+\rho)} & \text{for } t^* < t < t_u \end{cases} . \tag{A15}$$

The final unknown is eliminated by integrating the exit flow $(M/L)q(t)$ over the duration of the rush hour and setting the result equal to N, the exogenous number of travelers. This yields:

$$\left(\frac{\delta}{Mk_j\alpha} \right) N = \ln(1+s) + (1+\rho)\left[(1+s)^{-1/(1+\rho)} - 1\right] \tag{A16}$$

where $\delta \equiv \beta\gamma/(\beta+\gamma)$ is a kind of average measure of scheduling costs which plays a key role in the analysis of Arnott et al. (1990, 1993). Equation (A16) can be solved numerically for s, hence for t_i and t_u.

Hypercongestion occurs if K reaches its maximum value of $1/(2+\rho)$ as given by (14), which occurs if

$$s > \left(\frac{2+\rho}{1+\rho} \right) - 1 . \tag{A17}$$

If hypercongestion occurs, it begins and ends at:

$$t_{ih} = t_i + \frac{\alpha T_f}{\beta}\left[\left(\frac{2+\rho}{1+\rho}\right)^{1+\rho} - 1\right]$$

$$t_{uh} = t_u - \frac{\alpha T_f}{\gamma}\left[\left(\frac{2+\rho}{1+\rho}\right)^{1+\rho} - 1\right]. \qquad \text{(A18)}$$

For the special case $\rho=0$, which is the Greenshields linear speed-density relationship, the occurrence of hypercongestion is coincident with the condition $s>1$, that is, that scheduling costs exceed travel-time costs for the first and last travelers.