Equilibrium Traffic Dynamics in a Bathtub Model: A Special Case

Richard Arnott*, Anatolii Kokoza† and Mehdi Naji‡

November 7, 2016

Abstract

For a quarter century, a top priority in transportation economics has been to develop models of rush-hour traffic dynamics that incorporate hypercongestion – situations of heavy congestion where throughput decreases as traffic density increases. Unfortunately, even the simplest models along these lines appear in general to be analytically intractable, and none of the models that have made approximations in order to achieve tractability has gained widespread acceptance. This paper takes a different tack, developing an analytical solution for a special case – a no-toll equilibrium in an isotropic downtown area with identical commuters, the Greenshields' congestion technology, and the $\alpha - \beta$ cost function (no late arrivals permitted). This is followed by a discussion of directions for future research.

JEL Code: L91, R41

Keywords: equilibrium, rush hour, traffic congestion

Acknowledgements:

We would like to thank Joshua Buli for helpful comments related to the existence and uniqueness of equilibrium in the model, Matthew Fitzgerald for research assistance, seminar participants at the University of Lille for helpful comments and criticisms, and Mogens Fosgerau, Nikolas Geroliminis, Raphael Lamotte, Lewis Lehe, Robin Lindsey, and Kenneth Small for very useful comments on an earlier version of the paper.

*Department of Economics, University of California, Riverside, 4106 Sproul Hall, Riverside, CA 92521-0427; richard.arnott@ucr.edu

†Department of Economics, University of Arizona, 1130 E Helen St, 401 McClelland Hall, Tucson, AZ 85721-0108; kokoza@email.arizona.edu

‡School of International Relations, Niavaran, Tehran, Iran 19367-83114; mehdinaji@gmail.com

1 Introduction

Until a decade ago, there were no aggregate data on traffic congestion (flow, density, and velocity) at the level of a downtown neighborhood or of an entire downtown area. Then, in a landmark paper, Geroliminis and Daganzo (2008), using a combination of stationary and mobile (taxis) sensors, measured traffic flow and density over a neighborhood of Yokohama, Japan essentially continuously over a period of weeks. At this spatial scale, they found an inverse U-shaped relationship between traffic flow and density that was stable over the course of the day, and across days, which they termed the neighborhood's macroscopic fundamental diagram (MFD). Figure 1 plots the data that were collected. Subsequent research has documented the same qualitative result for downtown areas in other cities, though the MFD's vary across neighborhoods and across the downtown areas of different cites.

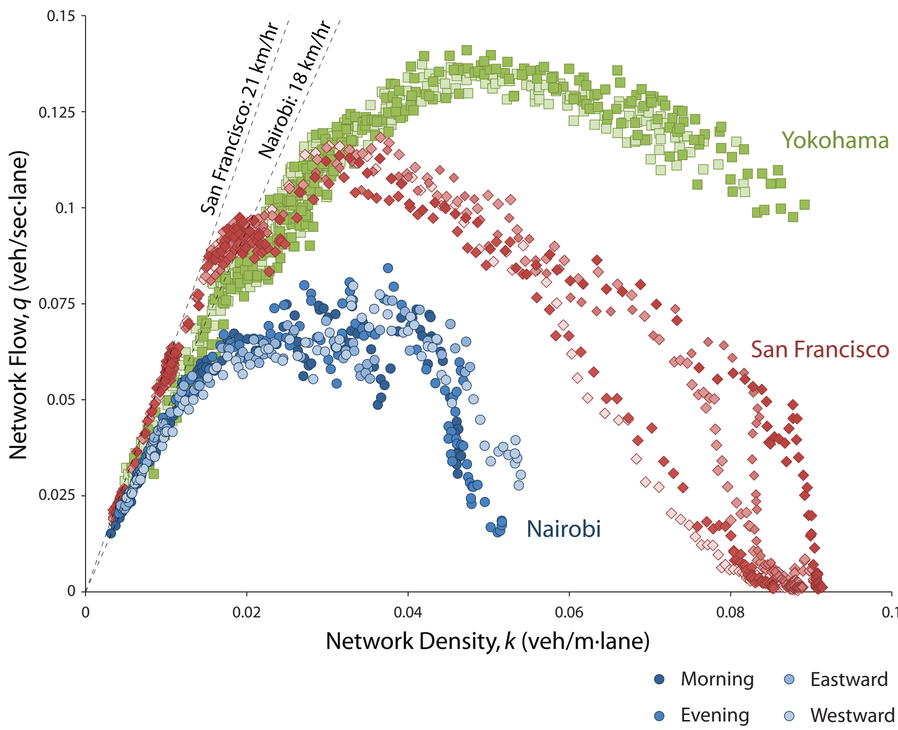


Figure 1: Macroscopic fundamental diagram for three cities
Note: The diagram is reproduced from Gonzales et al. (2011) The MFD for Yokohama is drawn using the data of Geroliminis and Daganzo (2008); the MFD's for Nairobi and San Francisco are obtained from traffic microsimulations.

This research confirmed what many urban transportation economists and transportation scientists had long suspected, that hypercongestion – traffic jam situations where traffic flow is *negatively* related to traffic density – is a pervasive and quantitatively important feature of equilibrium rush-hour traffic dynamics at the scale of a downtown area.

William Vickrey's bottleneck model (1969) has been the workhorse model of metropolitan rush-hour traffic dynamics for a quarter century.[1] While it has proved very adaptable and has generated a host of useful insights, as a model of downtown traffic congestion it is flawed since it rules out hypercongestion, assuming instead that under congested conditions aggregate traffic flow is constant. Urban transportation economists have been searching for a model of rush-hour traffic dynamics that admits hypercongestion without sacrificing the elegant simplicity of the bottleneck model.

There is a natural and intuitive alternative model of equilibrium morning rush-hour traffic dynamics that accommodates hypercongestion. This alternative model has three essential elements. First, the traffic congestion technology is described by an MFD relating traffic flow to traffic density. Second, Vickrey's equilibrium trip-timing condition, that no commuter can reduce her trip price by altering her departure time, is imposed. Third, the downtown area is isotropic, so that commuting trip origins (residences) and destinations (workplaces) are uniformly distributed over space. The third essential element eliminates the complexity introduced by spatial heterogeneity (as does Vickrey's bottleneck model). Models with these elements are now being referred to as *bathtub models*.[2]

Unfortunately, bathtub models are generally analytically intractable.[3] Some modelers have addressed this intractability by making approximating assumptions, but none of these models has been widely accepted, partly because, without solution of the "proper" model, the accuracy of the approximations is unknown.

This paper takes a different tack. Rather than attempt to solve for equilibrium in a general bathtub model, it develops a special case that is analytically tractable. The underlying modeling philosophy is that understanding simple cases will hopefully generate some general insights and lay the groundwork for understanding more general cases. In this special case, all commuters are identical, including having the same commuting distance and the same desired arrival time at work, and the trip cost function is the $\alpha-\beta-\gamma$ trip cost function, in which a commuter's trip cost is linear in travel time and schedule delay (time early or time late, relative to a desired arrival time), with the additional simplification that late arrival is not permitted. Thus, the special case is identical to the basic bottleneck model with no late arrivals permitted except for the specification of the congestion technology.

Consider the situation where the street system can accommodate all commuters at one time with little congestion. There is an equilibrium in which all commuters depart at the same time in a single departure

[1] Small (2015) reviews the literature on Vickrey's bottleneck model.

[2] Water in a bathtub flows in such a way that the height of water in the bathtub is constant. Similarly here, traffic flows in such a way that the density of traffic is constant over space. The term bathtub model is drawn from hydrology, where a bathtub model is a model of a water table that assumes that water flows in such a way that the height of water in the water table is constant.

[3] The source of the intractability is discussed in an earlier version of this paper, Arnott et al. (2015). In all cases, it arises from the model naturally generating delay differential equations, whose study is considerably less advanced than that of ordinary or partial differential equations.

mass, travel on the city streets together, and arrive exactly on time in an arrival mass. A commuter who departs after the mass arrives late, which is not permitted. A commuter who departs before the mass has a lower travel time than those in the mass since at the beginning of his commute there is no traffic on the road, but he also arrives early, incurring schedule delay cost. Since there is little congestion, the schedule delay cost he incurs more than offsets his travel cost savings. Since no commuter can reduce her trip cost by departing at a different time from the mass, all commuters departing in a single departure mass and arriving on time is an equilibrium

Now gradually increase the population of commuters, holding fixed the street system. In due course, the number of commuters reaches the point where the level of congestion is sufficiently high that the travel cost saving from departing before the mass equals the schedule delay cost incurred. When the number of commuters is slightly increased above this level, we show that there is an equilibrium with two departure masses, the earlier of which arrives just before the later mass, which arrives on time, departs. Thus, a characteristic of equilibrium is contiguous travel time intervals.

This paper provides a complete analytical solution for this type of equilibrium, which we term the *restricted equilibrium*, and investigates its properties. We believe the equilibrium to be unique since we have been unable to construct an equilibrium without contiguous travel time intervals, but have not proved uniqueness. The paper will also present an analytical solution for the restricted social optimum (which also has contiguous departure masses), and compare it to the restricted equilibrium, but only for the population interval with one or two departure masses

Section 2 provides a brief review of the relevant literature. Section 3 presents the model. Section 4 provides the complete analytical solution of the restricted equilibrium and a thorough treatment of its properties. Section 5 presents preliminary results for the restricted social optimum and compares them to those for the restricted equilibrium. Section 6 discusses extensions, and Section 7 concludes.

2 Literature Review

We shall review that branch of the literature on rush-hour traffic dynamics that accommodates hypercongestion, which includes bathtub models as defined earlier. We shall use the term "proper" model to refer to a model that is consistent internally, consistent with individual rationality, and consistent with the laws of physics. While somewhat imprecise, we use the term to refer to a model that makes no awkward compromises in order to achieve tractability.

While it does not explicitly deal with traffic congestion or consider equilibrium, the first paper in this sub-literature is arguably Agnew (1976), which examines the optimal control of congestion-prone systems in which throughput is a strictly concave function of load. In the context of rush-hour traffic dynamics, throughput is the exit rate or the arrival rate at the destination and load is traffic density, so that the assumption that throughput is a function of the load is that the exit rate depends on only traffic density and not on the time pattern of entries. Vickrey (1991) sketches a model that essentially adapts Agnew's model to traffic congestion in Manhattan.[4] Arnott (2013) provides a formalization of Vickrey's model that incorporates schedule delay costs, and in the equilibrium variant applies Vickrey's trip-timing equilibrium condition. It shows that Agnew's assumption that the exit rate depends only on traffic density holds if trip distances have the same negative exponential distribution throughout the rush hour, but also argues that this condition will generally not hold since those commuters with shorter trips would travel closer to the peak.

Several papers have modeled rush-hour traffic dynamics by extending the bottleneck model so that the capacity of the bottleneck depends on the length of the queue behind it. Most recently, Fosgerau and Small (2013) sidesteps the analytical intractability of a proper model by treating a discrete number of capacity levels. Yet other papers avoid the intractability of the proper bathtub model by making the approximation that a commuter's trip duration depends on the density of cars when the commuter begins (Geroliminis and Levinson, 2009) or ends (Small and Chu, 2003) her trip.

All of the above models make approximating assumptions. None has gained widespread acceptance. One may reasonably object to an approximating assumption on three grounds. First, a model may be inconsistent with rational economic behavior; this is a sound criticism since rational behavior is necessary for sound welfare analysis. Second, a model may be "inconsistent with the laws of physics"; this too is a sound criticism, though it needs to be made precise what laws of physics are being violated and how. And third, an approximating assumption may result in inaccuracy, both qualitative and quantitative; this too is a valid criticism, though to judge whether an approximating assumption results in inaccuracy, one needs to know the solution of the proper model, and, remarkably, only Fosgerau (2015) has solved a proper model numerically.

The model of this paper is fully consistent with rational economic behavior. It also entails no approximations. However, its congestion technology can be viewed as inconsistent with the laws of physics and with fluid dynamics. For one thing, when a car enters the street system, it is assumed to travel instantaneously at the velocity of the prevailing traffic, which entails infinite acceleration and hence infinite force; similarly, when it exits the street system, its speed is assumed to fall instantaneously to zero. The inconsistency is more severe in our model since a departure mass entails a mass of cars simultaneously entering the street

[4] Vickrey referred to the model as a "bathtub" model, using the term in a different sense than that in this paper. In Vickrey's sense, a bathtub model is one in which the exit rate from the bathtub (the exit rate from the street system) depends only on the height of water in the bathtub (the traffic density).

system and instantaneously traveling at the velocity consistent with the density of cars in the mass.

Fosgerau (2015) is the only paper in the literature that works with a proper bathtub model. In contrast to this paper, it specifies a smooth trip cost function, which results in a smooth pattern of departures. Also, in contrast to this paper, the two variants of the model it explores treat heterogeneous commuters. In the first, commuters are heterogeneous with respect to trip distance; in the second, they are heterogeneous with respect to trip distance and the analog in his model to desired arrival time. For both variants, an assumption is made that implies "regular sorting", in which commuters with longer trip distances both depart earlier and arrive later than those with shorter trips, which entails all commuters being on the road at the same time. In the first variant, the model comes up with the strong result that the equilibrium and the optimum coincide. In the second variant, among drivers with the same trip distance, a driver with an earlier desired arrival time departs and arrives earlier than a driver with a later desired arrival time, and tolling is effective. The paper presents numerical examples both with and without regular sorting. In none of the examples is there hypercongestion in the social optimum, but in some there is hypercongestion in the no-toll equilibrium. The paper's analysis suggests, paradoxically, that the mathematics with a continuum of commuters who differ in terms of a naturally ordered characteristic, such as trip distance or desired arrival time, may be easier than the mathematics with identical individuals.

To a reader who is unfamiliar with the method of microeconomic theory, the literature's preoccupation with developing models that check off all the boxes in terms of conceptual consistency may seem both odd and pedantic, especially since all the models entail gross simplifications in other respects. Microeconomics has been so successful largely because it has developed a rigorous body of theory that is enduring. Firm foundations take considerable effort to construct, but intellectual edifices built without them will crumble.

While the resistance of the transportation economics community to short cuts in model building is well justified, its resistance to numerical analysis is not. The community welcomes numerical examples that supplement and quantify theory but not those that substitute for theory. However, to break the current logjam in the study of rush-hour traffic dynamics with hypercongestion, for the moment at least numerical analysis of conceptually sound but analytically intractable models is at least a line well worth pursuing.

3 The Basic Model and the Properties of its Equilibrium

Throughout most of the paper we work with variables defined in terms of normalized units. However, in this section to introduce the basic model, and at the start of the next section to introduce the model's algebra, we employ variables defined in terms of unnormalized units. To avoid confusion, we shall denote variables defined in terms of unnormalized units with a $\widehat{}$ and the corresponding variables defined in terms of normalized units without a $\widehat{\ }$.

3.1 The Basic Model

The model is spatial and its space is an isotropic downtown area: home locations, job locations, and road capacity are uniformly distributed over space.[5] Per unit area, $\hat{N}$ commuters must travel from home to work over the morning rush hour. All commuters have the same commuting distance, L, and the same desired arrival time, $\hat{t}^*$.

The form of congestion is flow congestion in which commuters' velocity is decreasing in the density of commuters. To simplify the algebra, Greenshields' Relation is assumed, in which traffic velocity, $\hat{v}$, is negative linearly related to traffic density, $\hat{k}$:

$$\hat{v} = v_f \left(1 - \frac{\hat{k}}{\Omega}\right),$$

where v_f is free-flow velocity and Ω is jam density. Thus, travel time per unit distance is

$$\frac{1}{\hat{v}} = \left(\frac{1}{v_f}\right) \frac{\Omega}{\Omega - \hat{k}}.$$

Greenshields' Relation has the properties that maximum or capacity flow occurs when traffic density equals one-half jam density. Traffic is said to be congested at densities below this level and hypercongested at densities above this level.

The physical relationship that trip distance equals the integral of velocity over trip duration is

$$\int_{\hat{t}}^{\hat{t}+\hat{T}(\hat{t})} \hat{v}(\hat{k}(u))du = L, \tag{1}$$

where $\hat{t}$ is departure time, and $\hat{T}(\hat{t})$ is travel time with departure time $\hat{t}$.

The familiar $\alpha - \beta - \gamma$ trip cost function is employed:

$$\hat{c} = \alpha(\text{travel time}) + \beta(\text{time early}) + \gamma(\text{time late}),$$

[5]One may think of the downtown area as having a dense, symmetric grid network of streets extending infinitely far in both the east-west and north-south directions. The geometry of a finite, closed, symmetric grid network depends on what topological properties are imposed. We conjecture that the simplest geometry with the minimal topological properties required in this context entails streets wrapping round a square in both the north-south and east-west directions. The simplest geometry of a finite, closed symmetric grid network that is topologically equivalent (in the sense that a commuter could not distinguish whether he is driving on an infinite or closed and finite network would appear to be a Clifford torus – we thank Parker Williams for pointing this out.

where $\hat{c}(\hat{t})$ is trip cost as a function of departure time:

$$\hat{c}(\hat{t}) = \alpha\hat{T}(\hat{t}) + \beta\max(0, \hat{t}^* - \hat{t} - \hat{T}(\hat{t})) + \gamma\max(0, \hat{t} + \hat{T}(\hat{t}) - \hat{t}^*). \tag{2}$$

We impose the standard condition that the unit cost of travel time exceeds the unit cost of time early, $\alpha > \beta$. In the morning rush hour, each commuter chooses when to depart from home so as to minimize her trip price. A commuter's trip price equals her trip cost plus the toll she pays.

Four other conditions complete the model. The first is the accumulation equation, that density at time $\hat{t}$ equals cumulative entries at time $\hat{t}$ minus cumulative exits at time $\hat{t}$. Two other conditions are the boundary conditions that density equal zero right before the start of the rush hour and right after the end of the rush hour. Together these equations imply the conservation of vehicles. The fourth condition is that cumulative exits at time $\hat{t} + \hat{T}(\hat{t})$ equal cumulative entries at time $\hat{t}$. Along with (1), this condition ensures that all commuters have the same trip length.

In the basic bottleneck model, the length of the rush hour in both the no-toll equilibrium and the optimum equals $\hat{N}/s$ – the population of commuters divided by the flow capacity of the bottleneck. $\hat{N}/s$ measures relative demand, demand relative to capacity. An analog in the basic bathtub model is $\hat{N}/\Omega$, the ratio of the population of commuters per unit area to the jam density per unit area. The reciprocal of relative demand measures the adequacy of the downtown street network to accommodate rush-hour demand.

3.2 Equilibrium

A *morning rush hour equilibrium* is a time path of departures over the morning rush hour, and the induced time paths of traffic density and arrivals, such that no commuter can reduce her trip price by altering her departure time. We refer to the corresponding equilibrium when no toll is applied as a no-toll (morning rush-hour) equilibrium. "Trip price" in the definition of equilibrium may be replaced by trip cost.

In order to simplify the analysis, throughout most of the paper we assume that late arrival is not permitted. In section 6.2, which admits late arrival, we shall prove that, with a finite population, late arrival does not occur if the unit cost of time late (see (2)) is sufficiently high. Thus, we may replace the assumption that late arrival is not permitted with the conceptually more appealing assumption that the unit cost of time late is sufficiently high that late arrival does not occur in the equilibrium or social optimum, as the case may be.

In the no-toll equilibrium with identical commuters, the trip-timing condition reduces to the condition that the trip cost is the same at all times at which there are departures, and is at least as high at all times at which there are no departures. Where D denotes the set of times at which there are departures, this condition is

$$\hat{c}(\hat{t}) = \underline{\hat{c}} \quad \text{for all } \hat{t} \in D \qquad \text{and} \qquad \hat{c}(\hat{t}) \geq \underline{\hat{c}} \quad \text{for all } \hat{t} \notin D,$$

where $\hat{c}(\hat{t})$ is given by (2).

We noted earlier that the equilibria that we examine in this paper have the property that the departure pattern takes the form of contiguous travel time intervals, with a departure mass at the beginning of each interval and no departures other than those in the departure masses. We believe that all equilibria take this form. Since we have not proved this, however, we define a *restricted* morning rush-hour equilibrium to be an equilibrium with this property, and shall consider only such equilibria. Thus:

A *restricted (morning rush-hour) equilibrium* is a morning rush-hour equilibrium in which the departure pattern takes the form of contiguous travel time intervals, with a departure mass at the beginning of each interval and no departures other than those in the departure masses.

In what follows, we shall show by construction that a restricted equilibrium always exists and is unique.

4 (No-toll) Equilibrium with Identical Individuals

Section 4.1 examines the restricted no-toll equilibrium (which we shall refer to simply as the equilibrium for the remainder of the section) when the street system has generous capacity relative to the population of commuters so that equilibrium entails either one or two departure masses, and presents a numerical example. Section 4.2 provides a general solution of equilibrium. Section 4.3 examines the comparative static and dynamic properties of equilibrium. And section 4.4 extends the numerical example to more than two departure masses.

4.1 Equilibrium with One or Two Departure Masses

Consider a city with a small population density relative to its road capacity, in fact sufficiently small that in equilibrium all commuters depart at the same time in a single departure mass and arrive at work exactly on time. No commuter has an incentive to depart earlier since the decrease in travel time cost from doing so is more than offset by the increase in schedule delay cost. As population density increases, there is a critical value above which a commuter has an incentive to depart earlier than the mass. At this population density, equilibrium switches from having one departure mass to having two departure masses, and at a higher critical population density equilibrium switches from having two departure masses to three, etc.

Let m denote the number of departure masses, and i index a departure mass. Departure masses are indexed in reverse order of departure time; thus, the latest mass to depart, which arrives on time, has the index $i = 1$. This may seem counterintuitive, but the indexation is chosen so that the index of the departure mass that arrives on time does not change as the number of departure masses changes. Let $\hat{c}_i^m(\hat{N})$ be the trip cost of each commuter in mass i when there are m departure masses and the population density is $\hat{N}$, $\hat{n}_i^m(\hat{N})$ be the number of commuters in the ith departure mass with population density $\hat{N}$, $\hat{c}^e(\hat{N})$ be the equilibrium trip cost with population density $\hat{N}$, and $\hat{N}^e_{m,m+1}$ be the critical population density at which equilibrium switches from having m to $m+1$ departure masses.

4.1.1 One departure mass

Since there is only the one departure mass, $\hat{n}_1^1 = \hat{N}$. Also, since this departure mass arrives on time, commuters experience no schedule delay cost. Travel time is trip distance, L, divided by velocity,

$$\hat{v} = v_f \left(1 - \frac{\hat{N}}{\Omega}\right),$$

and trip cost equals travel time times the value of travel time, α. Thus,

$$\hat{c}_1^1(\hat{N}) = \frac{\alpha L}{\hat{v}} = \frac{\alpha L}{v_f \left(1 - \frac{\hat{N}}{\Omega}\right)} \tag{3}$$

and the departure time is

$$t^* - \frac{L}{v_f(1 - \frac{\hat{N}}{\Omega})}.$$

To avoid notational clutter, for the rest of the paper we employ several normalizations, but record results both with and without the normalizations. There are four units of measurement employed in the paper, those with respect to distance, time, money, and population per unit area. The normalizations are $L = 1$, $v_f = 1$, $\alpha = 1$, and $\Omega = 1$. Thus, the normalized distance is trip distance, the normalized time unit is the length of time it takes to travel the trip distance at free-flow velocity, the normalized money unit is the cost of travel per normalized time unit, and normalized population density is jam density. With these normalizations, (3) reduces to

$$c_1^1(N) = \frac{1}{1 - N}; \tag{4}$$

the velocity of the mass is $1 - N$, and its travel time is $\dfrac{1}{1-N}$. With this normalization, travel in a departure mass is congested if the size of the departure mass is less than 1/2 and hypercongested if the size of the departure mass is greater than 1/2. To convert from normalized units to unnormalized units, 1 normalized distance unit equals L unnormalized distance units, 1 normalized time unit equals $\dfrac{L}{v_f}$ unnormalized time units, 1 normalized money unit equals $\frac{\alpha L}{v_f}$ unnormalized monetary units, and 1 normalized population density unit equals Ω unnormalized population density units.[6]

Table 1: Normalizations

Unit	**1 Normalized unit = x Unnormalized units**	**Interpretation**
Distance	L	Trip distance
Time	L/v_f	Time required to travel trip distance at free-flow velocity
Money	$\alpha L/v_f$	Value of a unit of normalized travel time
Population density	Ω	Jam density
Velocity	v_f	Free-flow velocity

Notes: 1. Capacity flow is the maximum flow (per unit area) that can be supported by the street system. Under Greenshields' Relation: Capacity flow is $v_f\Omega/4$ in unnormalized units and 0.25 in normalized units; capacity density (the density associated with capacity flow) is $\Omega/2$ in unnormalized units and 0.5 in normalized units; regular congestion occurs when density is less than capacity density, and hypercongestion occurs when density exceeds capacity density.
2. Since the length of the rush hour is minimized at capacity flow, the minimum length of the rush hour is $4\hat{N}/(v_f\Omega)$ in unnormalized units or $4N$ in normalized units, where N is relative demand, as defined above.

To further simplify notation: t^* is set equal to zero, so that time is measured relative to the desired arrival time; $\theta \equiv \frac{\beta}{\alpha}$ equals the ratio of the value of time early to the value of travel time and is assumed to be less than one; and $\rho \equiv \frac{\gamma}{\alpha}$ equals the ratio of the value of time late to the value of travel time.

We now proceed with the analysis in normalized units. Consider an infinitesimal commuter who departs a period $\Delta t \leq 1$ earlier than the departure mass. Since normalized free-flow velocity equals 1, she travels a distance Δt before encountering the departure mass. She then travels the remaining distance $1 - \Delta t$ with the departure mass at the speed $1 - N$, arriving at work at

$$-\frac{1}{1-N} - \Delta t + \Delta t + \frac{1-\Delta t}{1-N} = \frac{-\Delta t}{1-N}.$$

Thus, her travel time is

$$\frac{-\Delta t}{1-N} + \Delta t + \frac{1}{1-N} = \frac{1}{1-N} - \frac{N\Delta t}{1-N}.$$

[6] It will be useful to provide some intuition for the magnitude of N. Let q denote flow, $q = kv$. Applying Greenshields' Relation, the relationship between flow and density is $q = kv(k) = k(1-k)$. Maximum or capacity flow is 1/4. Thus, with $N = 1$, the duration of the rush hour at capacity flow would be four normalized time units. In the extended example that we shall employ, we assume that $v_f = 15$ mph and $L = 5$ miles, so that the duration of a trip at free-flow speed, which is the normalized time unit, is 20 minutes. With these parameters and $N = 1$, the duration of the rush hour at capacity flow would be 80 minutes.

Her trip cost is therefore

$$c_1^1(N) - \frac{N\Delta t}{1-N} + \frac{\theta\Delta t}{1-N}.$$

Her trip cost is therefore lower when she departs earlier than the departure mass if $N > \theta$, and higher otherwise. Thus, the critical population density at which equilibrium switches from having one to two departure masses is $N_{1,2}^e = \theta$. Consistent with our definition of restricted equilibrium, when N is infinitesimally above θ equilibrium entails two departure masses, with the earlier departure mass arriving at the departure time of the departure mass that arrives on time, and trip cost being the same for both departure masses.[7]

Let $TC_{(m)}(N)$ denote total trip cost with population density N conditional on there being m departure masses, and $TC^e(N)$ denote trip cost with the equilibrium number of departure masses for population density N. From (4), when there is a single departure mass in equilibrium, and thus when $c^e(N) = c_1^1(N)$, total cost is

$$TC^e(N) = TC_{(1)}(N) = Nc^e(N) = \frac{N}{1-N}.$$

The corresponding marginal social cost and marginal congestion externality cost are therefore

$$MSC^e(N) = \frac{dTC_{(1)}^e}{dN} = \frac{1}{(1-N)^2}$$
$$MCE^e(N) = MSC^e(N) - c^e(N) = \frac{N}{(1-N)^2}.$$

Total trip cost may be decomposed into total travel time cost, TTC, and total schedule delay cost, SDC. With only one departure mass, since all commuters arrive exactly on time and therefore experience no schedule delay, all of the total trip cost is total travel time cost. In this case, the marginal congestion externality cost has a simple interpretation. It is the cost imposed on other commuters from increasing traffic density in the single departure mass by one unit. Define the severity of congestion, s, to be the ratio of the marginal congestion externality cost to the private trip cost. Then in equilibrium with one departure mass

$$s^e(N) = \frac{MCE^e(N)}{c^e(N)} = \frac{N}{1-N}.$$

We bring together the above results in

Proposition 1. *A restricted equilibrium with a single departure mass occurs when $N \leq \theta$. Over this interval of N, $c^e(N) = \frac{1}{1-N}$, $MSC^e(N) = \frac{1}{(1-N)^2}$, $MCE^e(N) = \frac{N}{(1-N)^2}$, and $s^e(N) = \frac{N}{1-N}$.*

[7] Stability is defined with respect to an adjustment process. Here the natural adjustment process is that, as N rises infinitesimally above θ, the first commuter to deviate will choose to depart at that time that minimizes her trip cost. Per the above reasoning, she will choose to depart such that she arrives "just before" the departure time of the departure mass that arrives on time. With this adjustment process, the restricted equilibrium is stable.

4.1.2 Two departure masses

Now we consider two departure masses, that is $m = 2$. To satisfy the trip-timing equilibrium condition, trip cost must be the same for each departure mass. Letting n_i^m denote the normalized number of commuters in departure mass i when there are m departure masses, equilibrium with two departure masses solves the following pair of equations:

$$n_1^2 + n_2^2 = N \tag{5}$$

$$\frac{1}{1-n_1^2} = c_1^2 = c_2^2 = \frac{1}{1-n_2^2} + \frac{\theta}{1-n_1^2}. \tag{6}$$

Departure mass 1 arrives on time, so that $c_1^2 = \dfrac{1}{1-n_1^2}$. Departure mass 2 arrives immediately before departure mass 1 departs, so that a commuter in departure mass 2 experiences travel time of $\dfrac{1}{1-n_2^2}$ and schedule delay of $\dfrac{1}{1-n_1^2}$. Solving (5) and (6) gives

$$\begin{aligned} {}^e n_1^2 &= \frac{N+\theta-N\theta}{2-\theta} \\ {}^e n_2^2 &= \frac{N-\theta}{2-\theta}. \end{aligned} \tag{7}$$

Two additional conditions are required for (7) to describe an equilibrium with two departure masses. The first is that each departure mass have a strictly positive density, which requires that $N > \theta$. The second is that a deviating commuter does not have an incentive to form a third departure mass. It is shown below that this condition is that $N \leq \theta(3-\theta)$. Thus, equilibrium entails two departure masses for $N \in (\theta, \theta(3-\theta))$. For N in this interval

$$c^e(N) = c_1^2 = \frac{2-\theta}{(2-N)(1-\theta)} \tag{8}$$

$$TC^e(N) = \frac{(2-\theta)N}{(2-N)(1-\theta)} \tag{9}$$

$$MSC^e(N) = \frac{2(2-\theta)}{(2-N)^2(1-\theta)} \tag{10}$$

$$MCE^e(N) = \frac{(2-\theta)N}{(2-N)^2(1-\theta)} \tag{11}$$

$$SDC^e(N) = \frac{{}^e n_2^2\theta}{1-{}^e n_1^2} = \frac{\theta(N-\theta)}{(2-N)(1-\theta)} \tag{12}$$

$$TTC^e(N) = TC^e(N) - SDC^e(N) = \frac{2N(1-\theta)+\theta^2}{(2-N)(1-\theta)}. \tag{13}$$

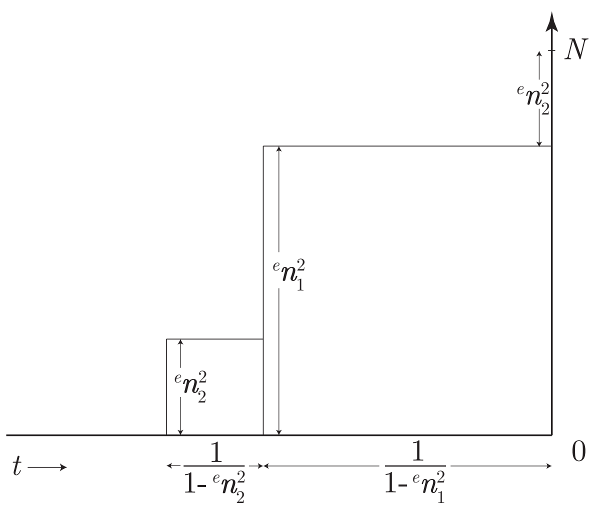
$N_{2,3}^e$ is that N for which a commuter is indifferent between departing in departure mass 2 and departing in departure mass 3 by herself. If she departs in departure mass 3 by herself, her travel time cost decreases by $\dfrac{1}{1-n_2^2}-1$ and her schedule delay cost increases by $\dfrac{\theta}{1-n_2^2}$. The decrease in travel time cost equals the increase in schedule cost when $\dfrac{1-\theta}{1-n_2^2}=1$, which is when $n_2^2=\theta$, implying $N^e_{2,3}=\theta(3-\theta)$.

Comparing (8) and (11) gives the severity of congestion

$$s^e(N)=\frac{N}{2-N}.$$

We bring together the results for two departure masses in

Proposition 2. *A restricted equilibrium with two departure masses occurs when $N\in(\theta,\theta(3-\theta))$. Over this interval of N:* ${}^e n_1^2=\dfrac{N+\theta-N\theta}{2-\theta}$, ${}^e n_2^2=\dfrac{N-\theta}{2-\theta}$, $c^e(N)=\dfrac{2-\theta}{(2-N)(1-\theta)}$, $MSC^e(N)=\dfrac{2(2-\theta)}{(2-N)^2(1-\theta)}$, $MCE^e(N)=\dfrac{(2-\theta)N}{(2-N)^2(1-\theta)}$, *and* $s^e(N)=\dfrac{N}{2-N}$.



(a)

$$N={}^e n_1^2+{}^e n_2^2$$
$${}^e c_1^2=\frac{1}{1-{}^e n_1^2}$$
$${}^e c_2^2=\frac{1}{1-{}^e n_2^2}+\frac{\theta}{1-{}^e n_1^2}$$
$${}^e c_1^2={}^e c_2^2\Rightarrow\frac{1-\theta}{1-{}^e n_1^2}=\frac{1}{1-{}^e n_2^2}$$

(b)

Figure 2: No-toll equilibrium with two departure masses
Note: The Figure is drawn using the functions and parameters of this section's numerical example, with $N=1$, ${}^e n_1^2=\frac{2}{3}$, and ${}^e n_2^2=\frac{1}{3}$.

Figure 2 displays the equilibrium with two departure masses graphically. The abscissa is the normalized time axis and the ordinate is normalized population density. Departure masses are numbered so that departure mass 1 arrives on time, and departure mass 2 arrives immediately before departure mass 1 departs. Since in equilibrium commuters in departure mass 1 have the same trip cost as commuters in departure mass 2, and since commuters in departure mass 1 arrive on time, experiencing no schedule delay cost, while those in departure mass 2 arrive early, experiencing schedule delay cost, travel time cost must be higher for commuters in departure mass 1 than those in departure mass 2. Thus, the size of the departure mass, and hence traffic density, must be higher in departure mass 1 than in departure mass 2. Travel speed is therefore lower for commuters in departure mass 1, resulting in a longer trip duration. The sum of the normalized population densities over the two departure masses gives the exogenous normalized population density, N. The duration of the rush hour equals the sum of the trip durations of the two departure masses.

To illustrate the results thus far, consider a numerical example in which $\theta = 0.5$, so that $N^e_{1,2} = 0.5$ and $N^e_{2,3} = 1.25$. The values of N considered are 0, 0.5, 1, and 1.25. To convert costs from normalized to unnormalized units, the following parameter values are assumed: α = \$20.00/hr, L = 5.0 miles, and v_f = 15.0 mph. L/v_f = 0.333 hrs is the assumed trip duration at free-flow speed. The numerical results are recorded in Table 2. Complementing Table 2 is Figure 3, which plots $TC^e(N)$, $c^e(N)$, and $MSC^e(N)$, for $N \in [0, 1.5)$.

Turn first to the three panels of Figure 3. The top one plots total cost against N, the middle one marginal social cost against N, and the bottom one trip cost against N. The three panels are aligned vertically.

Starting with the bottom panel, $c_{(1)}(N) = \dfrac{1}{1-N}$ equals trip cost as a function of N when the entire population travels in a single departure mass. Since the mass arrives on time, the entire trip cost is travel time cost. Trip cost is a convex function having the properties that $c_{(1)}(0) = 1.0$, $c_{(1)}(0.5) = 2.0$, and $c_{(1)}(1.0) = \infty$. Normalized trip cost is 1.0 when population is zero since trip cost equals travel time cost at free-flow speed, which is normalized to 1; is equal to 2.0 when normalized population is 0.5 since density equals one-half jam density, and is equal to ∞ when normalized population is $\Omega = 1.0$ since density equals jam density. The curve is drawn as a solid, bold line for $N \in [0, 0.5)$, the interval over which the equilibrium number of departure masses is 1, and as a dashed line outside this interval. $c_{(2)}(N) = \dfrac{2-\theta}{(2-N)(1-\theta)}$ is a convex function. $N = \theta$ is the lowest population density at which the equal trip cost condition for each departure mass is consistent with both departure masses having positive population density, while $N = 2$ corresponds to jam density. $c_{(2)}(N)$ has the properties that $c_{(2)}(\theta) = c_{(1)}(\theta) = \dfrac{1}{1-\theta}$, $c_{(2)}(1) = \dfrac{2-\theta}{1-\theta}$, and $c_{(2)}(2) = \infty$. The curve is drawn as a solid line for $N \in (0.5, 1.25)$, the interval over which the equilibrium number of departure masses is two, and as a dashed line outside this interval. $c_{(3)}(N) = \dfrac{3\theta - 3 - \theta^2}{(3-N)(1-\theta)^2}$. Since a switch occurs from m to $m+1$ departure masses when a commuter faces the same trip cost whether she departs in the existing departure masses, or deviates and departs in her own departure mass, the equilibrium trip cost function, $c^e(N)$, is the lower envelope of the trip cost functions for specific numbers of departure masses. As a result it has an escalloped shape, shown as a solid bold line.

The middle panel displays the marginal social cost functions with one, two, and three departure masses. If the bottom and the middle panels were combined, it would be seen that each marginal social cost function lies above the corresponding trip cost functions, with the vertical distance between the two functions measuring the congestion externality cost. The equilibrium marginal social cost function is not the lower envelope of

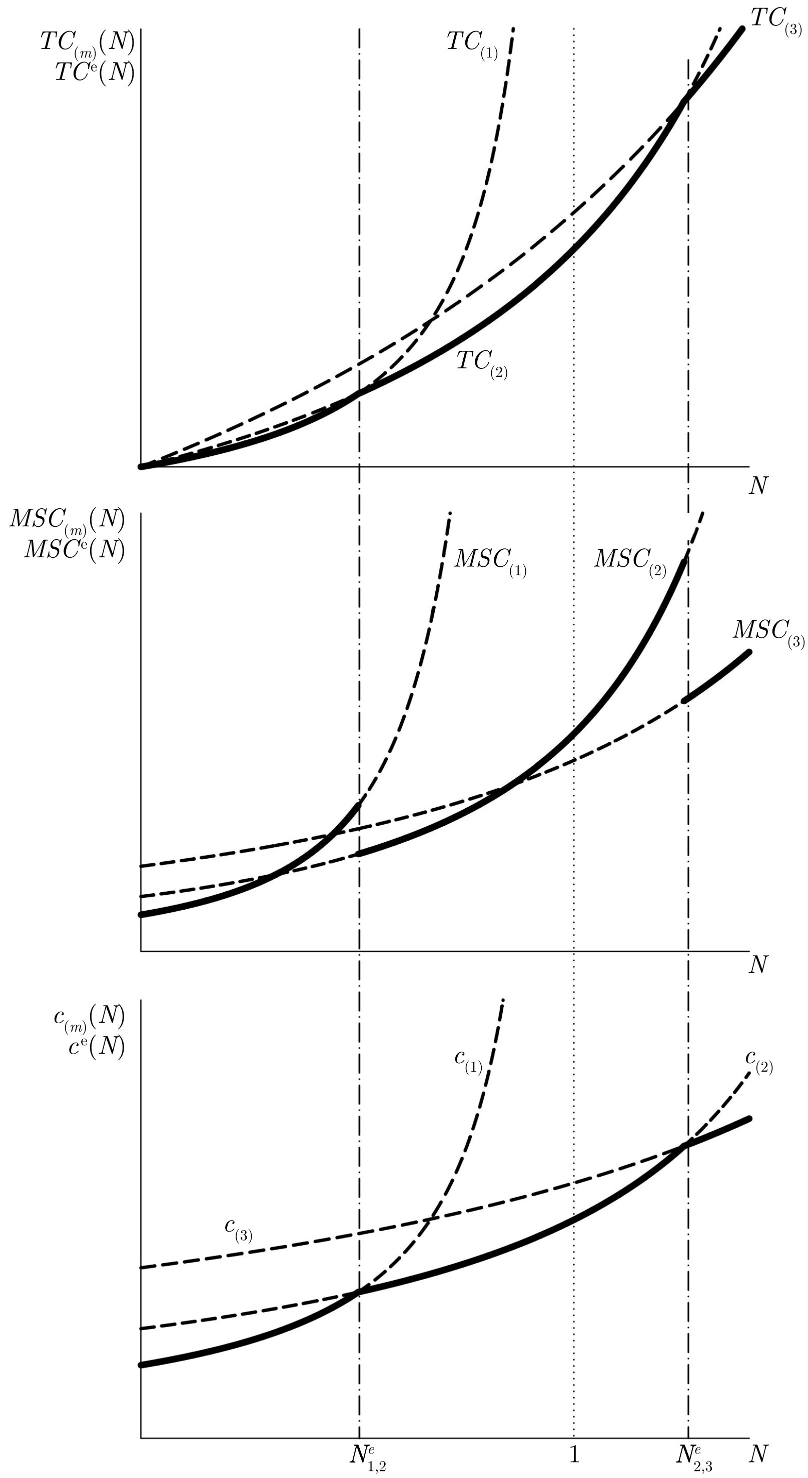


Figure 3

Note: The Figure is drawn using the functions and parameters of this section's numerical example.

the departure-mass specific marginal social cost functions. Instead, $MSC^e(N)$, which is drawn as the solid bold line, jumps downward at each critical population density at which there is a switch from m to $m+1$ departure masses in equilibrium. The reason is that at any of these critical population densities, a commuter imposes a lower congestion externality cost if she departs in her own departure "mass" than if she departs in the existing departure masses.

The top panel displays total cost as a function of population density with one, two, and three departure masses, as well as the equilibrium total cost function, which is the lower envelope of the total cost functions for specific numbers of departure masses.

Table 2: Numerical example for certain values of N and m

N Normalized population density	m Number of departure masses	c^e Normalized trip cost	MSC^e Normalized marginal social cost	s^e Severity of congestion	SDC^e/TTC^e Total schedule delay cost/Total travel time cost	$\hat{D}$ Rush hour length in hrs	$\hat{c}^e$ Trip cost in $
0	1	1.0	1.0	0.0	0.0	0.333	6.667
0.25	1	1.333	1.778	0.333	0.0	0.444	8.889
0.5	1	2.0	4.0	1.0	0.0	0.667	13.33
0.5	2	2.0	2.667	0.333	0.0	1.0	13.33
1.0	2	3.0	6.0	1.0	0.2	1.5	20.0
1.25	2	4.0	10.67	1.667	0.25	2.0	26.67
1.25	3	4.0	6.857	0.714	0.25	2.333	26.67
2.0	3	7.0	21.0	2.0	0.647	4.083	46.67
2.125	3	8.0	27.43	2.429	0.7	4.667	53.33
2.125	4	8.0	17.07	1.133	0.7	5.0	53.33
3	4	15.0	60.0	3.0	1.195	9.375	100.0
3.063	4	16.0	68.34	3.269	1.227	10.0	106.7
3.063	5	16.0	41.31	1.581	1.227	10.33	106.7
4.0	5	31.0	155.0	4.0	1.787	20.02	206.7
4.031	5	32.0	165.1	4.160	1.804	20.67	213.3
4.031	6	32.0	97.50	2.047	1.804	21.0	213.3

Notes: 1. The money normalization is that a trip at free-flow travel speed that arrives at the common work start time costs 1.0 unit. Since a trip has a length of $L = 5.0$ miles, since free-flow speed is 15.0 mph, since a trip that arrives at the common work start time entails no schedule delay cost, and since the value of travel time is \$20.00/hr, the (unnormalized) dollar cost of a trip at free-flow speed that arrives at the common work start time is \$6.67.
2. D is the length of the rush hour in normalized time, measured from the time of the first departure to the time of the last arrival, and $\hat{D}$ is the unnormalized length. The time normalization is that a trip at free-flow travel speed takes 1.0 time unit. Since a trip has a length of 5.0 miles and since the free-flow speed is 15.0 mph, the unnormalized time unit is 20.0 minutes.

Table 2 displays the quantitative properties of equilibrium for N = 0.0, 0.25, 0.5, 1.0, 1.25, as well as higher N, which are discussed in section 4.4. The new data presented in the table are the severity of congestion, the ratio of total schedule delay cost to total travel time cost, trip cost in dollars, and the length of the rush hour in hours. Observe that: i) the ratio of total schedule delay cost to total travel time cost appears to increase monotonically with population density; ii) the severity of congestion increases with population density over each population density interval for which the number of departure masses is constant, and decreases discontinuously at each critical population density where a departure mass is added; and iii) the length of the rush hour increases continuously with population density over population density intervals where the number of departure masses remains constant, and increases discontinuously at each critical population density where a departure mass is added.

Hypercongestion occurs when the normalized density of cars exceeds 0.5. For the population density interval over which there is one departure mass in equilibrium, hypercongestion occurs when $\theta > N > 0.5$, and does not occur when $\theta \leq 0.5$; with the assumed parameter value of $\theta = 0.5$, hypercongestion does not occur. For the population density interval over which there are two departure masses in equilibrium, hypercongestion occurs in departure mass 1 when $n_1^2 = \dfrac{N+\theta-N\theta}{2-\theta} > 0.5$ and in departure mass 2 when $n_2^2 = \dfrac{N-\theta}{2-\theta} > 0.5$; with the assumed parameter value of $\theta = 0.5$, hypercongestion occurs in departure mass 1 for $N \in (0.5, 1.25)$, but for no values of N in the second departure mass.

The stage is now set to work out equilibrium with three or more departure masses in equilibrium.

4.2 General Solution of Equilibrium

Fortunately, a recursive structure in the equilibrium size of adjacent departure masses permits neat, closed-form solution for the equilibrium in cities of all sizes and with any number of departure masses. The analysis below first solves for total trip cost, marginal social cost, and marginal congestion externality cost as functions of m and N, such that trip cost is the same in each departure mass (even though this can entail negative departure masses) and then determines the equilibrium m as a function of N.

With m departure masses,

$$c_j^m = \frac{1}{1-n_j^m} + \theta \sum_{i=1}^{j-1} \frac{1}{1-n_i^m}. \tag{14}$$

The trip-timing equilibrium condition implies that

$$1 - n_{j+1}^m = \frac{1-n_j^m}{1-\theta}. \tag{15}$$

Combining (15) with the condition that $\sum_{j=1}^m n_j^m = N$ yields a finite series expression for n_1^m. Rewriting the finite series expression as the difference between two infinite series, and then applying standard results on the sum of infinite series and solving for n_1^m gives

$$n_1^m = 1 - \frac{m-N}{\dfrac{1-\theta}{\theta}\dfrac{1-(1-\theta)^m}{(1-\theta)^m}}. \tag{16}$$

Combining (16) and (14) for $j = 1$, and noting that in equilibrium the trip cost is the same for all departure masses, yields the equilibrium trip cost

$$c^e_{(m)}(N) = \frac{1}{m-N}\frac{1-\theta}{\theta}\frac{1-(1-\theta)^m}{(1-\theta)^m}. \tag{17}$$

The total trip cost can then be calculated as $TC^e_{(m)}(N) = Nc^e_{(m)}(N)$:

$$TC^e_{(m)} = \frac{N}{m-N}\frac{1-\theta}{\theta}\left[\frac{1-(1-\theta)^m}{(1-\theta)^m}\right]. \tag{18}$$

Differentiation of $TC^e_{(m)}(N)$ with respect to N yields marginal social cost:

$$MSC^e_{(m)} = \frac{m}{N(m-N)}TC^e_{(m)}.$$

Marginal congestion externality cost can be calculated either as $MCE^e_{(m)}(N) = MSC^e_{(m)}(N) - c^e_{(m)}(N)$ or as $MCE^e_{(m)}(N) = N\left(\dfrac{\partial c^e_{(m)}(N)}{\partial N}\right)$:

$$MCE^e_{(m)} = \frac{1}{m-N}TC^e_{(m)}.$$

The equilibrium number of departure masses is now calculated as a function of N. By the equal trip cost condition, the switch from m to $m+1$ departure masses occurs for that N for which the trip cost with $m+1$ departure masses equals the trip cost with m departure masses: $TC^e_{(m+1)}(N^e_{m,m+1}) = TC^e_{(m)}(N^e_{m,m+1})$. Using (18), this reduces to

$$N^e_{m,m+1} = m - \frac{1-\theta}{\theta}\Big(1-(1-\theta)^m\Big). \tag{19}$$

This can be rewritten as a recursive relationship:

$$N^e_{m+1,m+2} = \theta(m+1) + (1-\theta)N^e_{m,m+1}. \tag{20}$$

Using (15) and (16), the duration of the rush hour with m departure masses and population density N is

$$D^e_{(m)} = \sum_{i=1}^{m}\frac{1}{1-{}^e n_i^m} = \left[\frac{1-(1-\theta)^m}{\theta}\right]^2\frac{(1-\theta)^{1-m}}{m-N}$$

We also have that

$$TTC^e_{(m)}(N;\theta) = \sum_{i=1}^{m} \frac{{}^{e}n_i^m}{1 - {}^{e}n_i^m} = \sum_{i=1}^{m} \left(\frac{1}{1 - {}^{e}n_i^m} - 1 \right)$$
$$= \left[\frac{1 - (1-\theta)^m}{\theta} \right]^2 \frac{(1-\theta)^{1-m}}{m - N} - m$$

and

$$SDC^e_{(m)}(N;\theta) = TC^e_{(m)}(N;\theta) - TTC^e_{(m)}(N;\theta)$$
$$= \frac{1 - (1-\theta)^m}{\theta^2} \frac{(1-\theta)^{1-m}}{m - N} \left[N\theta - 1 + (1-\theta)^{1-m} \right] + m$$

Table 3 brings together results in normalized form. Table 4 gives the corresponding results in unnormalized form.

Table 3: Algebraic results in normalized form: equilibrium with no late arrivals

Population of ith departure mass per unit area	$n_i^m(N;\theta) = 1 - \dfrac{m-N}{(1-\theta)^{i-1} A(m,\theta)}$
Trip cost per commuter	$c_{(m)}(N;\theta) = \dfrac{1}{m-N} A(m,\theta)$
Total trip cost per unit area	$TC_{(m)}(N;\theta) = \dfrac{N}{m-N} A(m,\theta)$
Marginal social cost per unit area	$MSC_{(m)}(N;\theta) = \dfrac{m}{(m-N)^2} A(m,\theta)$
Marginal congestion externality cost per unit area	$MCE_{(m)}(N;\theta) = \dfrac{N}{(m-N)^2} A(m,\theta)$
Ratio of marginal social cost to trip cost	$\dfrac{MSC_{(m)}(N;\theta)}{c_{(m)}(N;\theta)} = \dfrac{m}{m-N}$
Severity of congestion	$s^e_{(m)}(N;\theta) \equiv \dfrac{MCE_{(m)}(N;\theta)}{c^e(N;\theta)} = \dfrac{N}{m-N}$
Total travel time cost per unit area	$TTC_{(m)}(N;\theta) = \dfrac{1-(1-\theta)^m}{\theta} \dfrac{A(m,\theta)}{m-N} - m$
Total schedule delay cost per unit area	$SDC_{(m)}(N;\theta) = \dfrac{A(m,\theta)}{m-N} \left[N - \dfrac{1+(1-\theta)^m}{\theta} \right] + m$
Mass switching population densities	$N^e_{m,m+1} = m - (1-\theta)^m A(m,\theta)$
Duration of rush hour	$D_{(m)}(N;\theta) = \dfrac{1-(1-\theta)^m}{\theta} \dfrac{A(m,\theta)}{m-N}$

Note: Let $A(m,\theta) = \dfrac{1-\theta}{\theta} \left[\dfrac{1-(1-\theta)^m}{(1-\theta)^m} \right]$.

4.3 Comparative Static and Dynamic Properties of Equilibrium

The comparative static properties of the no-toll equilibrium are given in Table 5. The derivations[8] are given in Table 4 of Arnott et al. (2015). Comparing Tables 3 and 4, it can be seem that some of the

[8]Several of the derivations entail the use of polynomial inequalities for which we drew heavily on Hardy et al. (1952).

Table 4: Algebraic results in unnormalized form: equilibrium with no late arrivals

Population of ith departure mass per unit area	${}^{e}\hat{n}_i^m(\hat{N};\theta) = \Omega\Big[1 - \dfrac{m - \frac{\hat{N}}{\Omega}}{(1-\theta)^{i-1}A(m,\theta)}\Big]$
Trip cost per commuter	$\hat{c}^e_{(m)}(\hat{N};\theta) = \dfrac{\alpha L}{v_f\left(m - \frac{\hat{N}}{\Omega}\right)}A(m,\theta)$
Total trip cost per unit area	$\hat{TC}^e_{(m)}(\hat{N};\theta) = \dfrac{\alpha L\hat{N}}{v_f\left(m - \frac{\hat{N}}{\Omega}\right)}A(m,\theta)$
Marginal social cost	$\hat{MSC}^e_{(m)}(\hat{N};\theta) = \dfrac{\alpha Lm}{v_f(m - \frac{\hat{N}}{\Omega})^2}A(m,\theta)$
Marginal congestion externality cost	$\hat{MCE}^e_{(m)}(\hat{N};\theta) = \dfrac{\alpha L\frac{\hat{N}}{\Omega}}{v_f(m - \frac{\hat{N}}{\Omega})^2}A(m,\theta)$
Ratio of marginal social cost to trip cost	$\dfrac{\hat{MSC}^e_{(m)}(\hat{N};\theta)}{\hat{c}^e_{(m)}(\hat{N};\theta)} = \dfrac{m}{m - \frac{\hat{N}}{\Omega}}$
Severity of congestion	$\hat{s}^e_{(m)}(\hat{N};\theta) \equiv \dfrac{\hat{MCE}^e_{(m)}(\hat{N};\theta)}{\hat{c}^e_{(m)}(\hat{N};\theta)} = \dfrac{\hat{N}}{m\Omega - \hat{N}}$
Total travel time cost per unit area	$\hat{TTC}^e_{(m)}(\hat{N};\theta) = \frac{\alpha L\Omega}{v_f}\left(\frac{1}{m-\frac{\hat{N}}{\Omega}}\frac{1-(1-\theta)^m}{\theta}A(m,\theta) - m\right)$
Total schedule delay cost per unit area	$\hat{SDC}^e_{(m)}(\hat{N};\theta) = \frac{\alpha L\Omega}{v_f}\left[\frac{A(m,\theta)}{m-\frac{\hat{N}}{\Omega}}\left[\frac{\hat{N}}{\Omega} - \frac{1-(1-\theta)^m}{\theta}\right] + m\right]$
Mass switching population densities	$\hat{N}^e_{m,m+1} = \Omega\Big[m - (1-\theta)^m A(m,\theta)\Big]$
Duration of rush hour	$\hat{D}^e_{(m)}(\hat{N};\theta) = \dfrac{L}{v_f}\dfrac{1-(1-\theta)^m}{\theta}\dfrac{A(m,\theta)}{m - \frac{\hat{N}}{\Omega}}$

Notes: The normalized monetary unit is the cost of the time it takes to travel trip distance at free-flow velocity; thus, to convert to unnormalized units, multiply by $\dfrac{\alpha L}{v_f}$. The normalized time unit is the time it takes to travel the trip distance at free-flow velocity; thus, to convert to unnormalized units, multiply by $\dfrac{L}{v_f}$, trip distance divided by free-flow velocity. The normalized population density is relative to the jam density; thus, to convert to unnormalized units, multiply by jam density, Ω.

comparative static effects operate through the normalizations, and might therefore be called scale effects, while others operate via θ and N. The discreteness of departure masses raises difficulties for comparative static analysis, since an infinitesimal change in an exogenous variable can cause a change in the equilibrium number of departure masses, and when this occurs some endogenous variables change discontinuously. In all but the last row of the Table, the comparative static results presented are derived holding constant the equilibrium number of departure masses. When an increase in an exogenous parameter causes a change in the equilibrium number of departure masses, the last row of the table indicates whether the change is an increase or a decrease.

The signs of the comparative static derivatives are the same with three or more departure masses as they are with two. To more easily convey the intuition, we focus on the case when there are two departure masses in equilibrium, which was presented in Section 4.1.2.

Table 5: Some Comparative Static Properties of Equilibrium (Unnormalized)

	$\hat{N}$	Ω	prop $\uparrow$ in $\hat{N}$ and Ω	v_f	L	prop $\uparrow$ in L and v_f	θ with α fixed	α with β fixed	prop $\uparrow$ in α and β
${}^{e}\hat{n}_i^m$	+	+	+	0	0	0	+	-	0
$\hat{c}^e_{(m)}$	+	-	0	-	+	0	+	?	+
$M\hat{S}C^e_{(m)}$	+	-	0	-	+	0	+	?	+
$M\hat{C}E^e_{(m)}$	+	-	0	-	+	0	+	?	+
$\frac{M\hat{S}C^e_{(m)}}{\hat{c}^e_{(m)}}$	+	-	0	0	0	0	0	0	0
$s^e_{(m)}$	+	-	0	0	0	0	0	0	0
$T\hat{T}C^e_{(m)}$	+	-	+	-	+	0	+	?	+
$S\hat{D}C^e_{(m)}$	+	-	+	-	+	0	?	?	+
$\frac{T\hat{T}C^e_{(m)}}{S\hat{D}C^e_{(m)}}$	-	+	0	0	0	0	+	-	0
$\hat{N}^e_{m,m+1}$	0	+	+	0	0	0	+	-	0
$\hat{D}^e_{(m)}$	+	-	0	-	+	0	+	-	0
m^e	≥ 0	≤ 0	0	0	0	0	≤ 0	≥ 0	0

Note: "prop $\uparrow$" means "a proportional increase".

The only comparative static derivative with respect to population density worthy of remark is that the ratio of $\dfrac{T\hat{T}C}{S\hat{D}C}$ unambiguously decreases with $\hat{N}$. The intuitive reason is that schedule delay is experienced only by those in departure mass 2, and as $\hat{N}$ increases the proportion of the population in departure mass 2 increases (see (7)). The only comparative static derivative with respect to jam density worthy of remark is that the population in departure mass 1 increases with jam density. The reason is that, as jam density increases, the level of traffic congestion falls, so that trip costs are equalized for those traveling in the first and second departure mass when departure mass 1 receives a larger proportion of the population. When there is a proportional increase in $\hat{N}$ and Ω, the equilibrium is unchanged except for a scaling up; all per capita magnitudes remain unchanged.

The comparative static properties with respect to free-flow velocity derive from Greenshields' Relation. Speed increases proportionally for all densities. The size of each departure mass remains unchanged, but travel time and schedule delay shrink in the same proportion as free-flow travel time. A proportional increase in free-flow velocity and trip distance has no effect on the listed endogenous variables. In each departure mass, commuters travel double the distance at double the speed, resulting in no change in trip cost.

The comparative static properties with respect to α and β are a quantum level more complex. Consider the effects of an increase in θ, holding α constant, i.e. an increase in β holding α constant. This change causes commuters to attach more weight to reducing schedule delay, which tilts the distribution of commuters over departure masses towards masses that arrive less early. This increases the severity of congestion and hypercongestion in the masses that arrive less early, leading to some counterintuitive and anomalous comparative static results. One striking result is that an increase in β can lead to a decrease in total schedule delay costs, which implies that schedule delay falls more than in proportion to the rise in β. Recall that in the example presented in Section 4.1.1, a second departure mass starts to form when $N = \theta$. Now consider an initial situation when $N_0 = \theta_0 + \Delta$, where Δ is a small, positive number. From (7), the number of commuters in departure mass 2 is

$$\frac{N - \theta_0}{2 - \theta_0} = \frac{\Delta}{2 - \theta_0},$$

which is the number that experience schedule delay. An increase in β, holding α constant, causes an increase in θ from θ_0 to θ_1. This causes the number of commuters in departure mass 2 to shrink from $\dfrac{\Delta}{2 - \theta_0}$ to $\dfrac{\Delta - (\theta_1 - \theta_0)}{2 - \theta_1}$.

As θ_1 increases, a point is reached where the number of commuters in departure mass 2, and hence schedule delay costs, shrinks to zero, decreasing at an infinite rate, far exceeding the proportional increase in β. Another striking result is that an increase in β causes as *unambiguous increase* in the duration of the rush hour. It is paradoxical that a stronger desire to arrive closer to the desired arrival time results in a lengthening of the rush hour. Reconciling intuition with the result combines three observations. The first is that, holding the number of departure masses fixed, as is done in the table, an increase in β causes population to be redistributed from earlier to later departure masses, in this sense concentrating the arrival distribution, which is consistent with intuition. The second observation is that, because travel time is convex in congestion, again holding fixed the number of departure masses, an increasing concentration of commuters in later departure masses increases the total length of the rush hour. The third observation is that the result becomes ambiguous once the number of departure masses is allowed to vary.

Another result worthy of note is that an increase in α may cause equilibrium trip cost to *fall*. This is paradoxical since in other contexts an increase in the price of a factor of production (here the factor of production is travel time) increases costs. Here inputs are not combined in a cost-minimizing way because of externalities within the production process. More specifically, by deconcentrating departures across departure masses, the rise in α causes travel time in departure mass 1 to fall. If travel is severely hypercongested in departure mass 1, the proportional decrease in travel time in the mass may exceed the proportional increase in α. The result then follows from noting that, since travelers in departure mass 1 experience no schedule delay, their trip cost, which is the equilibrium trip cost, equals their travel time cost.

4.4 Numerical Example Extended to Higher N

N measures the population density of commuters relative to the capacity of the road network per unit area. In most US cities, N is modest. Even though hypercongestion may occur, the rush hour is relatively short. In the world's mega-cities, however, most of which are in developing countries, N can be much larger. For instance, a typical driver in Bangkok spent forty four hours per year sitting in gridlock (Gibbs, 1997). Earlier in the paper, in Table 2 (section 4.1.2), we presented a numerical example, but discussed only those cases where equilibrium entailed two departure masses. Now that we have generalized our analysis, we discuss the remaining cases, with larger N, which provides insight into the behavior of our model under heavily congested conditions. In the numerical example, we assumed that $\theta = 0.5$. From (16), we have that $N^e_{1,2} = 0.5$, $N^e_{2,3} = 1.25$, $N^e_{3,4} = 2.125$, $N^e_{4,5} = 3.063$, and $N^e_{5,6} = 4.031$.

Particularly striking is how rapidly trip cost increases with population density. At integer values of population density, $c^e(N) = 2^{N+1} - 1$. Since trip cost in departure mass 1 is entirely travel time cost, and since the cost of travel time is normalized at 1, travel time in departure mass 1 too is related to N according to $2^{N+1} - 1$, while travel speed as function of N is $\dfrac{1}{2^{N+1} - 1}$. In particular, speed in departure mass 1, and therefore at the peak of the rush hour, is 15 mph with $N = 0.0$, 5.0 mph with $N = 1.0$, 2.143 mph with $N = 2.0$, 1.0 mph with $N = 3.0$, and so on. The relationship between peak speed and N, which relates demand to capacity, is specific to Greenshields' Relation, but employing an empirically estimated relationship between velocity and density would give a qualitatively similar result.

With $\theta = 1/2$, as assumed in the example, for integer N, the number of departure masses in the restricted equilibrium is $m = N + 1$. The severity of congestion,

$$s^e(N) = \frac{MCE^e(N)}{c^e(N)} = \frac{N}{m - N},$$

then reduces to N. But a conceptually superior measure of the severity of congestion is the ratio of the congestion externality cost imposed by a commuter divided by the congestion cost experienced by the commuter, which we term the private congestion cost, $PCC(N)$. Private congestion cost is defined as trip cost minus trip cost with no congestion, which is normalized to one: $PCC(N) = c(N) - 1$. Defining this alternative measure of the severity of congestion as $S(N) \equiv \dfrac{MCE(N)}{PCC(N)}$,

$$S(N) = \frac{MCE(N)}{c(N) - 1} = \frac{N}{m - N}\frac{c(N)}{c(N) - 1},$$

which, with integer N and $\theta = 1/2$, reduces to

$$S(N) = \frac{Nc(N)}{c(N) - 1}.$$

Thus, $S(1) = 3/2$, $S(2) = 7/3$, $S(3) = 45/14$. As demand relative to capacity rises, the congestion cost experienced by a commuter is a decreasingly small fraction of the congestion cost she imposes on others. In this sense, as N rises not only does the absolute distortion due to unpriced congestion increase but so too does the relative distortion. A further point to note is how rapidly the length of the rush hour increases with N. For integer N and $\theta = \frac{1}{2}$,

$$\hat{D}(N) = \frac{(2^{N+1} - 1)(2 - (\frac{1}{2})^N)}{3}.$$

5 Social Optimum with Identical Individuals

Our focus in this paper is on the no-toll equilibrium. A later paper is planned that focuses on the social optimum. Here we restrict analysis of the social optimum to departure patterns in which all departures occur in departure masses with contiguous travel time intervals, referring to the resulting allocation of commuters to departure times as the restricted social optimum. This allows comparison with the restricted no-toll equilibrium treated in the previous section. While we have not proved that the global social optimum with no late arrivals and a cost function that is linear in travel time and schedule delay entails departure masses, the restricted social optimum that we consider is at least a local social optimum in the sense that the marginal social cost of no commuter can be reduced by altering her departure time. Furthermore, we restrict our analysis to two departure masses. We do this since, in contrast to the restricted no-toll equilibrium, we have been unsuccessful in deriving closed-form solution for more than two departure masses.[9] Thus, the results of this section are suggestive rather than exhaustive.

When population density is sufficiently low that both the restricted social optimum and the restricted equilibrium have a single departure mass, the restricted social optimum and the restricted equilibrium are identical. As population density increases, a critical population density is reached at which it becomes optimal for there to be two departure masses. Determination of the social optimum with two departure masses is analogous to that for the no-toll equilibrium except that the marginal social cost of trips in each of the two departure masses is equalized rather than the trip cost. The optimality conditions are

$$n_1^2 + n_2^2 = N,$$
$$MSC_1^2 = MSC_2^2.$$

Total social costs are

$$TC_{(2)} = \frac{n_1^2}{1-n_1^2} + \frac{n_2^2}{1-n_2^2} + \frac{\theta n_2^2}{1-n_1^2}.$$

The first term is the travel time cost of departure mass 1 commuters; the second is the travel time cost of departure mass 2 commuters; and the third is the schedule delay cost of mass 2 commuters. Thus

$$MSC_1^2 = \frac{1}{(1-n_1^2)^2} + \frac{\theta n_2^2}{(1-n_1^2)^2}, \tag{21}$$
$$MSC_2^2 = \frac{1}{(1-n_2^2)^2} + \frac{\theta}{1-n_1^2}. \tag{22}$$

The social cost of inserting an extra commuter in departure mass 1 equals the direct cost associated with the added commuter, $\frac{1}{1-n_1^2}$, plus the travel time externality cost imposed on other commuters in departure mass 1, $\frac{n_1^2}{(1-n_1^2)^2}$, plus the schedule delay externality cost imposed on commuters in departure mass 2, $\frac{\theta n_2^2}{(1-n_1^2)^2}$. The social cost of inserting an extra commuter in departure mass 2 equals the direct cost associated with the added commuter, $\frac{1}{1-n_2^2} + \frac{\theta}{1-n_1^2}$, plus the travel time externality cost imposed on other commuters in departure mass 2, $\frac{n_2^2}{(1-n_2^2)^2}$. Equating the marginal social costs for the two departure

[9] We were able to obtain closed-form solution for the restricted no-toll equilibrium by exploiting the recursive relationship (15), in particular by finding the infinite series solution and then calculating the finite series solution as the difference between two infinite series solution. Unfortunately, we have been unsuccessful in applying the same "trick" to obtain closed-form solution for the social optimum.

masses and substituting in the population condition yields

$$\frac{1-\theta(1-N)}{(1-n_1^2)^2} = \frac{1}{(1-N+n_1^2)^2},$$

which reduces to

$$^*n_1^2 = \frac{1-(1-N)B(N)}{1+B(N)}, \text{ where } B(N) = \sqrt{1-\theta(1-N)}.$$

Thus

$$^*c_1^2 = \frac{1}{1 \ - \ ^*n_1^2} = \frac{1+B(N)}{B(N)(2-N)} \tag{23}$$

$$^*n_2^2 = N \ - \ ^*n_1^2 = \frac{B(N)-(1-N)}{1+B(N)} \tag{24}$$

$$^*c_2^2 = \frac{1}{1 \ - \ ^*n_2^2} + \frac{\theta}{1 \ - \ ^*n_1^2} = \frac{(B(N)+\theta)(1+B(N))}{B(N)(2-N)} \tag{25}$$

$$\begin{aligned} ^*MSC_1^2 &= \frac{1+\theta n_2^2}{(1-n_1^2)^2} = \frac{(B(N)+1+\theta)(1+B(N))}{B(N)(2-N)^2} \\ ^*MSC_2^2 &= \frac{1}{(1-n_2^2)^2} + \frac{\theta}{1-n_1^2} = \frac{(B(N)+1+\theta)(1+B(N))}{B(N)(2-N)^2}. \end{aligned} \tag{26}$$

By setting $^*n_1^2 = N$, we obtain the critical population density at which there is a switch from one to two departure masses at the social optimum, $N_{1,2}^*$:[10]

$$N_{1,2}^* = \frac{(2+\theta)-(\theta^2+4)^{1/2}}{2}.$$

We next solve for $N_{2,3}^*$. To calculate this, we solve for the N for which $^*MSC^3(N_{2,3}^*) =^* MSC^2(N_{2,3}^*)$. Now, when $N = N_{2,3}^*$, $^*n_3^3 = 0$, and

$$^*MSC^3(N_{2,3}^*) = 1 + \frac{\theta}{1 \ - \ ^*n_2^2(N_{2,3}^*)} + \frac{\theta}{1 \ - \ ^*n_1^2(N_{2,3}^*)}. \tag{27}$$

The first term is the marginal social travel time of a commuter in departure mass 3 when $^*n_3^3 = 0$ and the last two terms are the marginal schedule delay cost of this commuter. Now, from (22),

$$^*MSC^3(N_{2,3}^*) = \ ^*MSC_2^2(N_{2,3}^*) = \frac{1}{(1- \ ^*n_2^2(N_{2,3}^*))^2} + \frac{\theta}{1- \ ^*n_1^2(N_{2,3}^*)}. \tag{28}$$

Comparing (27) and (28) permits a closed-form solution for $^*n_2^2(N_{2,3}^*)$. Substituting this into $MSC_2^2(N_{2,3}^*) = MSC_1^2(N_{2,3}^*)$ from (21) and (22) gives a closed-form solution, albeit an ugly one, for $^*n_1^2(N_{2,3}^*)$.

[10] There is a negative root and a positive root. The positive root has $^*n_1^2 > 1$, which does not make economic sense.

The results we have obtained for one or two departure masses are recorded below.

Result 1. ${}^{e}N_{1,2} = \theta > {}^{*}N_{1,2}$ *for* $\theta \in (0, 1)$

The critical population density at which there is a switch from one to two departure masses is lower in the restricted social optimum than in the restricted no-toll equilibrium, which is consistent with the downward shift in $MSC^{e}(N)$ at $N^{e}_{1,2}$ as population increases that was noted in the previous section. In the no-toll equilibrium with $N = N^{e}_{1,2}$, there are externalities associated with adding a commuter to either departure mass 1 or departure mass 2, but the externality cost of adding a commuter to departure mass 1 is higher than adding a commuter to departure mass 2. The travel time externality cost is higher by adding a commuter to departure mass 1 than to departure mass 2. Furthermore, adding a commuter to departure mass 1 generates a schedule delay externality while adding a commuter to departure mass 2 does not.

Result 2. $N^{*}_{1,2} < 1/2$ *for* $\theta \in (0, 1)$

For all sets of parameters,[11] in the restricted social optimum as population density increases departure mass 2 is created before hypercongestion arises in departure mass 1. In contrast, there is a single hypercongested departure mass in the no-toll equilibrium when N is greater than $\frac{1}{2}$ but less than θ.

Result 3. $N^{e}_{2,3} = 3\theta - \theta^{2} > N^{*}_{2,3}$ *for* $\theta \in (0, 1)$

On the basis of the intuition provided for Result 1, it is natural to conjecture that $N^{e}_{m,m+1} > N^{*}_{m,m+1}$.

Result 4. $\frac{1}{2} \;>\; {}^{*}n^{2}_{1} \;>\; {}^{*}n^{2}_{2}$ *for* $\theta \in (0, 1)$

It is natural to conjecture that ${}^{*}n^{m}_{i} \;>\; {}^{*}n^{m}_{i+1}$ for all m and for all i from 1 to $m-1$ as in the equilibrium, that traffic density increases monotonically during the morning rush hour. It is also natural to conjecture that hypercongestion does not occur in the restricted social optimum for $\theta \in (0, 1)$.

We have already identified the externalities associated with adding a commuter to departure mass 1 and then to departure mass 2. As is standard, the social optimum can be decentralized by imposing a congestion toll equal to the trip externality cost, evaluated at the social optimum. Thus,

$$ {}^{*}\tau^{2}_{i} = {}^{*}MSC^{2}_{i} - {}^{*}c^{2}_{i},\; i = 1, 2, $$

[11] We have imposed the restriction that $\theta \in (0, 1)$. As in the bottleneck model, this is a necessary condition for equilibrium to exist. This can be seen by comparing (5) and (6) in the case of two departure masses. Since the schedule delay of departure mass 2 equals the travel time of departure mass 1, with $\theta > 1$ the schedule delay cost alone of traveling in departure mass 2 would exceed the trip cost of traveling in departure mass 1, which comprises only travel time cost. Thus, all commuters would choose to travel in departure mass 1, but if $N > 1$ this would not be possible. In contrast, the restricted social optimum in the isotropic model is well defined with $\theta > 1$ (as it is in the bottleneck model). In this case, hypercongestion can arise in the social optimum. To see this, consider the limiting case in which θ approaches infinity. With $N < 1$, it would be optimal for all commuters to travel in a single mass since all would be on time.

which can be calculated from (23), (25), and (26):

$$^*\tau_1^2 = \frac{(1+B(N))(B(N)+N-1+\theta)}{B(N)(2-N)^2}$$
$$^*\tau_2^2 = \frac{(1+B(N))(1+(N-1)(B(N)+\theta))}{B(N)(2-N)^2}.$$

We now construct a numerical example with two departure masses in the social optimum, and compare the social optimum and the no-toll equilibrium. To obtain particular qualitative results, we assume that $\theta = 0.9$, in contrast to $\theta = 1/2$, which was assumed in the earlier example. Otherwise, the parameters are the same as in the earlier example. Per the procedure outlined above, we obtain $N^*_{2,3} = 0.776$, and assume $N = 0.75$ in order to obtain almost as congested as possible a social optimum with two departure masses. With the parameter values chosen, there are two departure masses in the social optimum but only one in the no-toll equilibrium. Table 6 compares numerically the social optimum and the no-toll equilibrium for these parameter values.

Table 6: Comparison of the no-toll equilibrium and social optimum with $N = 0.75$ and $\theta = 0.9$ (normalized units)

	m	n_1^2	n_2^2	c_1	c_2	MSC	TC	TTC	SDC	D
No-toll eq.	1	0.75	0	4	4	16	3	3	0	4
Social opt.	2	0.4148	0.3352	1.709	3.042	3.801	1.729	1.213	0.5155	3.213

Notes: Recall that a normalized time unit equals 20 minutes, that dollar trip cost equals \$6.66 times normalized trip cost, and that the number of commuters is measured relative to jam density, so that $N = 1$ corresponds to a rush hour lasting 80 minutes at capacity flow.

To put the example into perspective, recall that a normalized time unit is 20 minutes, and that, if traffic flow were at capacity throughout the rush hour, the duration of the rush hour would be $4N = 3$ normalized time units or one hour. Thus, we are considering a small city, not a mega-city. The unit schedule delay cost was chosen to be high so that rush hour in the no-toll equilibrium would be so concentrated that hypercongestion would develop, resulting in substantial efficiency gains from congestion tolling. Because commuters attach a high value to arriving at work close to on time, the no-toll equilibrium is highly congested. There is only a single departure mass, which travels at only 3.75 mph – severe hypercongestion – resulting in 80 minutes of travel time. Each commuter imposes 4 hours of delay on other commuters, resulting in a marginal social cost of a trip, in terms of travel time of 5 hours 20 mins. In the social optimum, in contrast, commuters distribute themselves between two departure masses. Travel speed in the more congested departure mass is 8.778 mph compared to a free-flow travel speed of 15 mph, while that in departure mass 2 is 9.972 mph. Commuters in departure mass 1 experience a travel time of 34.17 minutes, for a normalized cost of 1.709, while commuters in mass 2 have a 30.09 minute commute, which, along with the 34.17 minute schedule delay, results in a normalized cost of 3.042. The marginal social cost of a trip is 3.801 normalized time units, which is less than one-quarter of that in the no-toll equilibrium. Even though the optimum has two departure masses, its rush hour is 64.26 minutes, significantly shorter than that in the no-toll equilibrium. This example illustrates well a paradox of hypercongestion – even though commuters in the no-toll equilibrium ignore the high cost that their traveling at the peak of the rush hour imposes on others, which intuitively should result in concentration of the rush hour, the length of the rush hour is in fact higher than in the social optimum. The resolution of the paradox is that ignoring the external cost they impose on others causes commuters to concentrate their departure times (in fact, in the example they all depart at the same time), but the concentration of departure times creates such severe hypercongestion that the length of the rush hour increases. Congestion tolling, by causing commuters to face the external cost, results in them deconcentrating their departures, eliminating hypercongestion and shortening the rush hour.

Table 6 also illustrates another point: under circumstances where the no-toll equilibrium is highly congested, the efficiency gains from imposing the optimal congestion toll exceed the toll revenue raised! In the example, the optimal toll is 2.092 ($13.93) for commuters traveling in departure mass 1 and 0.7586 ($5.05) for those traveling in departure mass 2. In normalized units, the total revenue is 1.122, while the efficiency gain from congestion tolling is 1.271 ($11.30 per commuter). Thus, the example illustrates the very considerable efficiency gains achievable through congestion tolling even in a small city, albeit one highly prone to congestion.

A word of caution is in order. The parameters were chosen to keep the calculations simple (only two departure masses at the optimum) while at the same time illustrating the very substantial efficiency gains achievable under congestion tolling when the no-toll equilibrium is highly congested. With a more realistic choice of θ, commuter efficiency gains from congestion tolling of the magnitude in the example would occur only for considerably "larger" cities – cities with considerably longer rush hours.

6 Extensions

Sections 6.1 and 6.2 present straightforward extensions. Section 6.3 discusses directions for future research.

6.1 Price-sensitive Demand

As in the bottleneck model, the function relating trip cost to the number of commuters in the restricted no-toll equilibrium can be regarded as a reduced-form supply curve. The reduced-form supply function for the restricted no-toll equilibrium is given by (17) and (19). Eq. (19) identifies the population intervals over which there are various numbers of departure masses. For each of these population intervals, (17) relates trip cost to population. The reduced-form supply curve is upward sloping over the entire range of population. Adding to this a demand curve relating the number of commuters to trip cost permits solution of the restricted no-toll equilibrium with price-sensitive demand, as shown in Figure 4. An increase in demand results in movement up the upward-sloping, reduced-form supply curve, and hence to an increase in both the equilibrium trip cost and the equilibrium number of commuters. This stands in contrast to the stationary-state reduced-form supply curve derived in Arnott and Inci (2010), which is backward bending. There high *stationary-state* (flow) demand relative to capacity could be accommodated only through the trip price rising to a level entailing hypercongestion. Here there is the extra margin of the length of the rush hour that adjusts to achieve equilibrium when (stock) demand for trips over the rush hour is high relative to capacity. Thus, one may say of the no-toll equilibrium in this paper's isotropic model of rush-hour traffic dynamics that, while hypercongestion may exist at the peak of the rush hour (in the sense that traffic density exceeds capacity density) it does not occur at the aggregate level (in the sense that the aggregate supply curve for trips is upward sloping at all levels of population density).

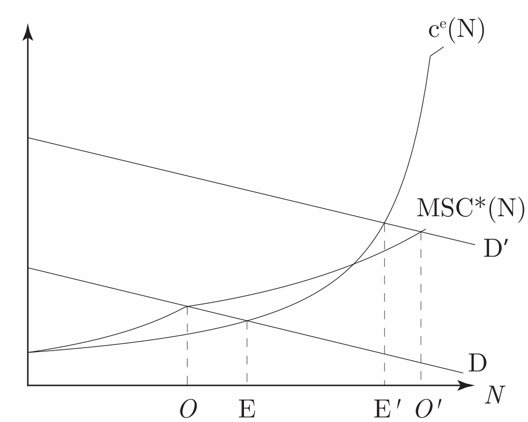


Figure 4: No-toll equilibrium trip cost and marginal social cost with price-sensitive demand
Note: $c^e(N)$ and $MSC^*(N)$ were drawn using the functions and parameters used in the construction of Table 6. The demand functions D and D' are qualitative, having the properties that D intersects $c^e(N)$ and $MSC^*(N)$ to the left of the intersection point of those two curves, and D' to the right.

The social optimum with price-sensitive demand and identical commuters occurs where the marginal social benefit of trips over the morning rush hour equals the marginal social cost. Diagrammatically, this corresponds to the point of intersection of the demand curve for trips and the marginal social cost of trips (O and O'), shown in Figure 4. It is important to note that the marginal social cost curve for the social optimum is different from that in the no-toll equilibrium. As in the bottleneck model, efficient pricing in the isotropic model has two effects. The first is to alter the timing of departures over the rush hour, holding the number of commuters fixed, and the second is to ensure the efficient number of trips.

A final point to note is that the reduced-form supply curve for trips in the no-toll equilibrium intersects the marginal social cost for trips in the social optimum, as shown in Figure 4.[12] For population levels below the population level where the two curves intersect, trip price is lower and the number of trips higher in the no-toll equilibrium than in the social optimum ($E > O$), and the revenue raised from the optimal time-varying toll is greater than the welfare gains it achieves. For population levels above the population level where the two curves intersect, trip price (trip cost plus the optimal time-varying toll) is lower and the number of trips higher in the social optimum than in the competitive equilibrium ($O' > E'$), and the revenue raised from the optimal time-varying toll is less than the welfare gain it achieves. Thus, as in Arnott and Inci (2010), in cities in which demand is high relative to capacity, optimal congestion tolling would be beneficial even if the toll revenue were completely squandered.

6.2 Late Arrivals

Earlier we simply asserted that the restricted no-toll equilibrium with no late arrivals permitted is the limiting case of the corresponding no-toll equilibrium with late arrivals permitted but an infinite value of time late. The proof proceeds in two steps. The first step is to prove that, with the $\alpha - \beta - \gamma$ cost function, there is an arrival mass at t^*. The proof proceeds by contradiction. Suppose not. Then a deviating commuter who arrives exactly on time will incur a lower trip cost than commuters in each of the departure masses that he travels with on his journey. Hence, an arrival mass at t^* is a necessary condition for equilibrium. The second step is to show that the no-toll equilibrium derived in section 4 is achieved as the limiting case of the equilibrium with the $\alpha - \beta - \gamma$ cost function as the unit time late cost approaches infinity. We now demonstrate this somewhat informally.

To keep the indexation consistent with that earlier in the paper, we shall denote by $i = -1$ the first late departure mass, $i = -2$ the second late departure mass, and so on, with the departure masses arriving early or on time are indexed as before.[13]

We shall again restrict our analysis to two departure masses. With a low population density, there is only a single departure mass which departs early and arrives on time. As population density grows, a critical population density is reached at which a deviating commuter will choose to depart either in a second early departure mass or in the first late departure mass, which departs at t^*. With departure in a second early departure mass, the deviating commuter's trip cost is ${}^{e}c_2^2 = 1 + \dfrac{\theta}{1-N}$. With late departure, her trip cost

[12]All the points made in this paragraph were made for the bathtub model in Arnott (2013).

[13]It would be more intuitive to index early departure masses with a $-$ and late departure masses with a $+$. We have not done so only to achieve notational consistency in the paper.

is ${}^{e}c^{2}_{-1} = 1 + \rho$ (recall that $\rho \equiv \frac{\gamma}{\alpha}$); she travels at free-flow speed, incurring one unit of travel time cost and one unit of time late cost. Earlier we proved that, when late arrivals are not permitted, a second departure mass starts to form when $N = \theta$. When late arrivals are permitted, the same result holds when the second departure mass to form is a second early departure mass. Thus, the second mass to form is a second early departure mass if $\frac{\theta}{1-\theta} < \rho$, and a first late departure mass if the inequality is reversed. Empirical work suggests that θ is around 1/2, while ρ is around 2.0, in which case the second departure mass to form would be a second early departure mass. On the assumption that this is the case, we can determine the third departure mass to form. With late departure, the trip cost of a deviating commuter remains ${}^{e}c^{2}_{-1} = 1 + \rho$. With departure in the third early departure mass and $\theta = 1/2$, the trip cost of a deviating commuter is 4 (see Table 2). Thus, with $\theta = 1/2$ and $\rho \in (1, 3)$, the third departure mass to form is the first late departure mass. In the limit as γ approaches ∞, as long as population density is finite, a late departure mass never forms, and the equilibrium is the same as when late arrivals are not admitted.

6.3 Directions for Future Research

The bathtub model appears to be the natural starting point for the study of downtown, rush-hour traffic dynamics with hypercongestion. Unfortunately, solving the model requires working with mathematics that is advanced and not well developed. The best way forward is unclear. There would appear to be four, non-exclusive options in how to proceed: i) pushing ahead with analytical exploration of the bathtub model, despite the mathematical difficulties; ii) developing computational methods to solve variants of the bathtub models numerically; iii) working with models that simplify the mathematics of the bathtub model through approximating assumptions; and iv) searching for an alternative to the bathtub model that addresses rush-hour traffic dynamics and admits hypercongestion but somehow sidesteps the mathematical difficulties of the bathtub model.

Perhaps the most promising way to proceed at the current time is to combine options i) to through iii). Explore the mathematical structure of each problem, then use it as a basis for constructing an algorithm for numerical solution, and finally compare the numerical solution to that obtained under various approximations. Unfortunately, because the mathematics of delay differential equations is poorly developed, to our knowledge there are no off-the-shelf results that can be employed to establish existence and uniqueness of equilibrium, or the uniqueness of optima. Furthermore, since there appears to be only a small literature that considers optimal control with delay differential equations, and no literature to our knowledge on economic equilibrium with delay differential equations, researchers will need to proceed with care. From the limited work done thus far by Fosgerau (2015), Arnott et al. (2015), and Arnott and Buli (2016), it appears that different solution methods are suited to different types of problems: equilibrium vs optimum, smooth vs non-smooth cost functions, and identical individuals vs a discrete number of types vs a continuum of types. Ordinarily, the simplest cases are the easiest to analyze, so an advisable course of action is to develop a thorough solution of the simplest case, and then to build on the knowledge acquired to investigate more complex cases. But for the bathtub model, even that general rule may not apply because working with a continuum of types may smooth discontinuities.[14]

Indeed, Lamotte and Geroliminis (2016) have very recently (during the period this paper was under review) distributed a working paper that extends this paper's model in several respects, including to the situation where commuters differ in terms only of trip length, which is continuously distributed in the population. They show analytically for "a large class of scheduling preferences that if users have continuously distributed characteristics, the network accumulation at equilibrium is a continuous function of time."[15] This result implies that working with a continuum of commuter types does indeed smooth out discontinuities at the aggregate level, though at the cost of added complexity.

This paper explored the bathtub model with identical individuals and the $\alpha - \beta - \gamma$ cost function. It provided a comprehensive analysis of the restricted equilibrium, in which all departures occur in masses with contiguous travel time intervals for each mass, constructively proving existence and uniqueness of this type of equilibrium. Whether there are equilibria other than the restricted equilibria remains an open issue, as is the stability of the restricted equilibria. The paper also presented a preliminary analysis of the restricted social optimum, in which all departures occur in masses with contiguous travel time intervals for each mass, which suggested but did not establish existence and uniqueness of this type of optimum. Whether there are optima other than the restricted optima is an open issue.

7 Conclusions

The bottleneck model has been the workhorse for the economic analysis of rush-hour traffic dynamics for a quarter century. It has served our community well, having proved amenable to a rich set of extensions and having provided a bounty of insights. However, the model has a serious deficiency. It does a bad job of modeling downtown traffic congestion when it is at its worst. In particular, it assumes that throughput is the same whether downtown traffic congestion is moderate or severe. Experts have long believed that throughput falls sharply when congestion becomes severe, but only recently was this confirmed empirically. Since downtown traffic congestion is a critical problem when it is severe, it is time to move beyond the

[14] In numerical solution of equilibrium in the bathtub model in the simplest case – with identical individuals and a smooth cost function – Arnott and Buli (2016) have found that the entry function is not in general continuous.

[15] The paper also examines analytically at length how commuter departure and arrival times are arranged by trip length, and undertakes extensive numerical simulation.

bottleneck model.

Urban transportation economists have long understood how the bottleneck can be extended or replaced to treat hypercongestion – severe congestion in which throughput falls as traffic density increases. The problem is that, even though easy to write down, all "proper" models – those that respect rational economic behavior and the physics of traffic congestion – have proven to be analytically intractable. All result in delay differential equations with an endogenous delay with both an initial and a terminal condition on the state variable, whose study is at the research frontier in applied mathematics.

Several papers have attempted to break this impasse by introducing approximations that restore tractability. None has gained wide acceptance since their approximations have been challenged on a priori grounds. This paper took a different tack, working out a special case that generates a closed-form solution for a no-toll equilibrium without making any approximations. The special case adapts the simplest bottleneck model, which has identical commuters, a trip cost function that is linear in travel time and schedule delay, and no late arrivals, to an isotropic downtown area where the congestion technology entails velocity being a negative linear function of traffic density. As in the bottleneck model, the central equilibrium condition is that no commuter can reduce her trip price by altering her departure time. This special case has a no-toll equilibrium in which all departures are in contiguous masses, which we referred to as a restricted no-toll equilibrium.

For this special case, the paper provided a comprehensive analysis of the no-toll equilibrium and presented some preliminary results on the optimum. Among the important insights generated by the paper are the following:

- In the basic bottleneck model, in both the equilibrium and the optimum, the length of the rush hour and average trip cost increase linearly with the ratio of population to capacity. In contrast, in the bathtub model in both the equilibrium and the optimum, the length of the rush hour (in excess of free-flow travel time) and average variable trip cost (trip cost in excess of trip cost with zero population) are convex functions of relative demand – the ratio of population density to the density of capacity (as measured by the jam density).

- Hypercongestion occurs when traffic flow decreases with traffic density. In the bottleneck model, the congestion technology is such that hypercongestion never occurs. In the bathtub model, in equilibrium hypercongestion occurs over part of the rush hour except when relative demand is low, and becomes increasingly severe as relative demand increases.

- At high levels of relative demand, hypercongestion can be so severe that the length of the rush hour is paradoxically higher in the equilibrium than in the social optimum.

- At high levels of relative demand, in equilibrium the efficiency cost due to unpriced congestion may be so high that the efficiency gains from optimal tolling exceed the revenue raised. Put alternatively, optimal tolling would benefit commuters even if the toll revenue were completely squandered.

The big advantages of the special case are that it entails no approximations, its properties can be explored analytically, and its economics and physics are intuitive. Thus, it provides a promising starting point. Whether or not this promise is realized will depend on the robustness of the special case, which remains to be explored. Robustness has several aspects: whether the restricted no-toll equilibrium is the unique equilibrium for the special case, whether its qualitative properties are robust, and how far the special case can be extended in the direction of realism without sacrificing closed-form solution or at least analytical tractability. The model of this paper has the property that departures occur in masses. Lamotte and Geroliminis (2016) have recently shown that this unrealistic property disappears when the assumption of homogeneous commuters is replaced by the assumption that there is a continuum of commuters who differ in some characteristic. Whether analytical tractability is preserved, and whether closed-form solutions can be obtained, with this alternative assumption remains to be seen.

Developing a theory of rush-hour traffic dynamics that is economically and physically sound, mathematically tractable, and admits hypercongestion has proved to be difficult. This paper falls far short of fully developing such a theory. Rather, it introduces a promising line of attack that merits further development and exploration.

References

Agnew, C. E. (1976). Dynamic modeling and control of congestion-prone systems. *Operations Research*, 24(3):400–419.

Arnott, R. (2013). A bathtub model of downtown traffic congestion. *Journal of Urban Economics*, 76:110 – 121.

Arnott, R. and Inci, E. (2010). The stability of downtown parking and traffic congestion. *Journal of Urban Economics*, 68(3):260 – 276.

Arnott, R. J. and Buli, J. (2016). Solving for equilibrium in the basic bathtub model. In draft form.

Arnott, R. J., Kokoza, A., and Naji, M. (2015). A Model of Rush-Hour Traffic in an Isotropic Downtown Area. *CESifo Working Paper Series No. 5465*.

Fosgerau, M. (2015). Congestion in the bathtub. *Economics of Transportation*, 4(4):241–255.

Fosgerau, M. and Small, K. A. (2013). Hypercongestion in downtown metropolis. *Journal of Urban Economics*, 76:122 − 134.

Geroliminis, N. and Daganzo, C. F. (2008). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. *Transportation Research Part B: Methodological*, 42(9):759 − 770.

Geroliminis, N. and Levinson, D. (2009). Cordon pricing consistent with the physics of overcrowding. In Lam, W. H. K., Wong, S. C., and Lo, H. K., editors, *Transportation and Traffic Theory 2009: Golden Jubilee*, pages 219–240. Springer US.

Gibbs, W. W. (1997). Transportation's perennial problems. *Scientific American*, 277:54–57.

Gonzales, E., Chavis, C., Li, Y., and Daganzo, C. F. (2011). Multimodal transport in Nairobi, Kenya: Insights and recommendations with a macroscopic evidence-based model. In *Transportation Research Board 90th Annual Meeting*, number 11-3045.

Hardy, G., Littlewood, J., and Polya, G. (1952). *Inequalities*. Cambridge University Press, Cambridge, UK.

Lamotte, R. and Geroliminis, N. (2016). The morning commute in urban areas with heterogeneous trip lengths. Unpublished manuscript.

Small, K. A. (2015). The bottleneck model: An assessment and interpretation. *Economics of Transportation*, 4(12):110 − 117.

Small, K. A. and Chu, X. (2003). Hypercongestion. *Journal of Transport Economics and Policy*, 37(3):319–352.

Vickrey, W. S. (1969). Congestion theory and transport investment. *The American Economic Review*, 59(2):pp. 251–260.

Vickrey, W. S. (1991). Congestion in Manhattan in relation to marginal cost pricing. Memo, Columbia University.

Notational Glossary

c, $\hat{c}$	normalized, unnormalized trip cost
$\hat{c}(\hat{t})$	trip cost as a functional of departure time
$\underline{c}$	equilibrium trip cost
c_i^m, $\hat{c}_i^m$	normalized, unnormalized trip cost in mass i conditional on m masses
$c_{(m)}$	common trip cost in no-toll equilibrium with m departure masses
e	equilibrium
i	index of departure or arrival mass
k, $\hat{k}$	normalized, unnormalized density (per unit area)
m	number of masses
n_i^m, $\hat{n}_i^m$	normalized, unnormalized population density in mass i conditional on m masses
s	severity of congestion ($\equiv MCE/c$)
t, $\hat{t}$	normalized, unnormalized time
t^*	desired arrival time (set to zero in much of the paper)
v, $\hat{v}$	normalized and unnormalized velocity
v_f	unnormalized free-flow velocty (normalized free-flow velocity equals 1)
$A(m,\theta)$	constant term ($\equiv \frac{1-\theta}{\theta}\left[\frac{1-(1-\theta)^m}{(1-\theta)^m}\right]$)
$B(N)$	intermediate variable ($\equiv \sqrt{1-\theta(1-N)}$)
D	set of departure times
$D_{(m)}$, $\hat{D}_{(m)}$	normalized, unnormalized duration of rush hour with m masses
L	unnormalized trip distance (normalized trip distance equals 1)
MCE, $\hat{MCE}$	normalized, unnormalized marginal congestion externality cost
MSC, $\hat{MSC}$	normalized, unnormalized marginal social cost
N, $\hat{N}$	normalized and unnormalized population of commuters per unit area
$N_{m,m+1}$, $\hat{N}_{m,m+1}$	normalized, unnormalized population density at which switch occurs from m to $m+1$ masses
SDC, $\hat{SDC}$	normalized, unnormalized total schedule delay cost
$S(N)$	alternative measure of the severity of congestion ($\equiv \frac{MCE(N)}{c(N)-1}$)
$\hat{T}(\hat{t})$	unnormalized travel time as a function of departure time
TC, $\hat{TC}$	normalized, unnormalized total trip cost
TTC, $\hat{TTC}$	normalized, unnormalized total travel time cost
α	unnormalized unit value of travel time (normalized unit value of travel time is 1)
β	unnormalized unit value of time early (normalized unit value of time early is θ)
γ	unnormalized unit value of time late (normalized unit value of time late is ρ)
θ	$\equiv \frac{\beta}{\alpha}$
ρ	$\equiv \frac{\gamma}{\alpha}$
τ_i^m	congestion toll applied to each commuter in mass i, conditional on there being m departure masses
Δ	finite increment
Ω	unnormalized jam density (normalized jam density is 1)
*	social optimum