Abstract

In standard economic models of traffic congestion, traffic flow does not fall under heavily congested conditions. But this is counter to experience, especially in the downtown areas of major cities during rush hour. This paper analyzes a bathtub model of downtown rush-hour traffic congestion that builds on ideas put forward by William Vickrey. Water flowing into the bathtub corresponds to cars entering the traffic stream, water flowing out of the bathtub to cars exiting from it, and the height of water in the bathtub to traffic density. Velocity is negatively related to density, and outflow is proportional to the product of density and velocity. Above a critical density, outflow falls as density increases (traffic jam situations). When demand is high relative to capacity, applying an optimal time-varying toll generates benefits that may be considerably larger than those obtained from standard models and that exceed the toll revenue collected.

Keywords: Rush hour, traffic congestion, equilibrium, optimum, toll

JEL code: R41

1. Introduction

There are two standard models of traffic congestion employed by economists. In the first, which is familiar from undergraduate economics textbooks, as demand increases, equilibrium traffic flow increases. In the second, the bottleneck model (Vickrey, 1969; Arnott, de Palma, and Lindsey, 1993), which treats morning rush-hour traffic dynamics, congestion is modeled as a deterministic queue behind a bottleneck of fixed flow capacity. In neither model does traffic flow fall as demand increases. But casual experience and common sense suggest that, in the downtown areas of heavily congested cities, due to traffic jams traffic flow is lower at the peak of the rush hour than during less congested periods of the day. Only very recently have traffic engineers started to measure traffic flows at the level of downtown neighborhoods, and early results provide strong support for this phenomenon (Geroliminis and Daganzo, 2008; Daganzo, Gayah, and Gonzales, 2011).

There is abundant anecdotal evidence of downtown rush-hour traffic speeds of 2 to 5 mph in cities such as Central London before the congestion toll, and central Moscow, Shanghai, Djakarta, Istanbul, Mexico City, Cairo, and Bangkok, but no reliable documentation. It seems implausible that traffic flow can be close to its maximum (capacity flow) at such low speeds.
This phenomenon has important implications for the management of downtown traffic congestion in very congested cities. First, the time loss due to rush-hour congestion could be sharply reduced if traffic restraint policies restricting entry flow into the downtown area were implemented. Second, the benefits from applying optimal time-varying congestion tolls are substantially higher that those estimated from the standard models, and exceed the toll revenue raised.

This paper develops a bathtub model of downtown traffic congestion that captures traffic-jam situations. Think of the bathtub as being Manhattan. In the morning rush hour, cars join the traffic on Manhattan streets, entering either across the bridges and tunnels into Manhattan or from parking spaces in Manhattan. These cars correspond to the inflow of water into the bathtub. Similarly, cars leaving the traffic stream, either exiting Manhattan or entering parking spaces in Manhattan, correspond to the outflow of water from the bathtub. Traffic density corresponds to the height of water in the bathtub. Traffic velocity is assumed to be inversely proportional to traffic density. Via the fundamental identity of traffic flow, traffic flow equals traffic density times traffic velocity. And outflow from the bathtub is assumed to be proportional to traffic flow. The relationship between traffic velocity and traffic density is such that outflow is increasing in the height of water in the bathtub up to some critical height, and decreasing in the height of the water above that height.

Vickrey used the term “hypercongestion” to refer to traffic jam situations, where traffic flow is inversely related to density. In the early morning rush hour, the inflow into the bathtub exceeds the rate at which the bath drains, and the water level rises. If the water level rises much about the critical level, the bathtub takes a long time to drain. A planner regulating the inflow into the bathtub would ensure that the water level never rises above the critical height.

Vickrey’s bathtub analogy for downtown traffic congestion has the ring of truth about it. But only recently has the relationship between average traffic density and average traffic velocity at the level of downtown neighborhoods received strong empirical confirmation. Traffic engineers started to collect detailed

---

2The model of this paper was inspired by a conversation with Vickrey a few years before his death. He coined the term “bathtub model of traffic congestion”, thinking both of Manhattan and of an (imperfect) hydrodynamic analogy. A dated outline of a model and incomplete notes on it were found in his files after his death (Vickrey, 1991). I am grateful to Keith Knapp for pointing out that the term “bathtub model” is used in biology and hydrology to refer to a dynamic model (of an aquifer, for example) in which a disturbance at one location is instantaneously propagated to all other locations. That characteristic is a property of this paper’s bathtub model too.

3In the traffic engineering literature, the “capacity” of a section of road is defined to be its maximum sustainable flow. In this paper, the congestion technology has the property that flow cannot exceed the maximum sustainable flow, so that “capacity” or “capacity flow” will refer simply to maximum flow. “Capacity density” and “capacity velocity” will refer to the density and velocity corresponding to capacity flow. Here, capacity density corresponds to the critical height of water in the bathtub.

4I thank Kenneth Small for making me aware of two papers antedating Vickrey (1991) that share at least part of his conception, in particular his “macroscopic” conception of a well-defined relationship between traffic aggregates at the scale of a downtown areas. Olszewski and Suchorzewski (1987) presented a complex macroscopic model of traffic congestion in the city center of Warsaw that relates aggregate traffic flow to average speed. Ardekani and Herman (1987) estimated the parameters of Herman and Prigogine (1979)’s macroscopic two-fluid model of “town” traffic for Austin and Dallas, assuming a stable relationship between mean density, mean flow, mean velocity, and the fraction of vehicles stopped. Neither paper, however, considers the evolution of congestion over the rush hour that is an essential feature of bathtub model.
data on freeway traffic flow in the late 1970s (e.g., Hall, Allen, and Gunter, 1986). Analysis of such data (e.g., Cassidy and Bertini, 1999) suggests that freeways contain bottlenecks whose discharge rates fall only modestly as the length of the queue behind them increases. Based partly on these analyses, the prevailing wisdom in urban transportation economics is that the aggregative or macroscopic behavior of rush-hour traffic in metropolitan areas is broadly consistent with the bottleneck model, with flow being approximately constant and congestion delay taking the form of quasi-queues. Only in the last five years have comparable data been collected for downtown neighborhoods in larger cities, using a network of stationary sensors, supplemented by mobile sensors (GPS devices in taxis). Analysis of these data (e.g., Geroliminis and Daganzo, 2008) provides strong support for what the authors refer to as the existence of a stable, urban-scale macroscopic fundamental diagram (MFD) – a stable graph relating traffic flow to density at the level of a downtown neighborhood, which includes a hypercongested portion. The results of these empirical studies are broadly supported by the current generation of downtown traffic simulation models, which incorporate elements permitting flow to fall as density increases. Figure 1 reproduces Figure 3 from Gonzales, Chavis, Li, and Daganzo (2011). It plots flow versus density for a large neighborhood in Yokohama (measured), San Francisco (simulated), and Nairobi (simulated). The graph for Nairobi and San Francisco are particularly striking, showing a sharp reduction in flow with high levels of congestion.

While the paper restricts attention to traffic congestion, the bathtub model can be adopted to other congestible facilities for which heavy loading results in decreased output, such as brown-outs and black-outs in electrical systems and jammed switches in telephone circuits.
Section 2 presents a particularly simple bathtub model of the morning rush hour, in which velocity is linearly inversely related to density, commuters are identical, and the user cost function corresponds to that in the bottleneck model. Section 3 solves for equilibrium in the morning rush hour in the absence of tolls. Section 4 solves for the corresponding social optimum and for the time-varying congestion toll that supports it. Section 5 compares the no-toll equilibrium and the social optimum for a particular numerical example calibrated to correspond to a downtown area where demand is high relative to capacity. Section 6 discusses policy insights and directions for future research, and section 7 concludes.

2. The Bathtub Model

Consider an isotropic downtown area. There are $N$ identical commuters per unit area, each of whom must travel from home to work in the morning rush hour and has work start time $t^*$, and experiences travel time cost and schedule delay cost (the cost of arriving at work inconveniently early or late). Classical flow congestion is assumed, which posits a physical relationship between traffic velocity and traffic density. In particular, congestion is described by Greenshields’ Relation (Greenshields, 1935), which specifies a negative...
linear relationship between velocity at time $t$, $v(t)$, and traffic density per unit area at time $t$, $k(t)$:

$$v(t) = v(k(t)) = v_f \left(1 - \frac{k(t)}{k_j}\right),$$  \hspace{1cm} (1)

where $v_f$ is free-flow velocity and $k_j$ is jam density, which can be interpreted as the density of cars with bumper–to–bumper traffic and is taken as fixed throughout the analysis, so that $N$, which is here termed population, can be interpreted as the level of demand relative to capacity. Greenshields’ Relation implies a parabolic relationship between flow and density. Travel on the positively-sloped portion is said to be congested, and on the negatively-sloped portion, corresponding informally to traffic jam situations, is said to be hypercongested. Capacity density, the density at which flow is maximized, corresponds to the critical height of water in the bathtub.

Letting $d(t)$ be the departure rate from home at time $t$, and $a(t)$ be the arrival rate at work at time $t$, the evolution of density is given by

$$\dot{k}(t) = d(t) - a(t) .$$  \hspace{1cm} (2)

Letting $D(t)$ be cumulative departures by time $t$, $A(t)$ cumulative arrivals, $t$ the time of the first departure from home, and $\tilde{t}$ the time of the last departure,

$$D(\tilde{t}) = A(t) = 0$$

$$D(t) = N .$$  \hspace{1cm} (3)

Combining (2) and (3) yields

$$k(t) = D(t) - A(t) ,$$  \hspace{1cm} (4)

which states that, at each point in time, density equals cumulative departures minus cumulative arrivals. It would seem that the natural way to complete the specification of the model’s dynamics would be to assume that the identical commuters have the same trip length, $L$.\footnote{Indeed, with no late arrivals permitted, this is the model employed by DePalma and Arnott (2012) of morning rush hour traffic congestion along a traffic corridor of uniform width, which draws on Gordon Newell (1988). Furthermore, Arnott and Inci (2006, 2010) and Arnott and Rowse (2009) employ the steady-state version of this model.} Letting $T(t)$ denote the travel time of a commuter who departs at time $t$, cumulative departures and arrivals would then be related according to

$$D(t) = A(t + T(t)) ;$$  \hspace{1cm} (5)

a commuter who departs at time $t$ arrives at time $t + T(t)$. Unfortunately, this assumption results in intractability.\footnote{To illustrate the intractability, consider the problem of solving for the evolution of density, given the departure rate function. From (4), $k(t + T(t)) = D(t + T(t)) - A(t + T(t))$. Substituting out $A(t + T(t))$ using (5) gives $k(t + T(t)) = D(t + T(t)) - D(t)$.}
It seems that there is no elegant way of dealing with this intractability. Small and Chu (2003) tackled it by assuming that a commuter’s travel time depends on traffic density when he arrives at his destination, and interpreted this as an approximation.\(^{11}\) More recently, Fosgerau and Small (2011) obtained a tractable model without making an approximation by adapting the bottleneck model, such that the bottleneck operates at capacity up to some queue length, then drops to a fixed, reduced capacity for longer queue lengths, and finally drops to zero capacity for very long queues; unfortunately, extensions of their model are analytically unwieldy. Geroliminis and Levinson (2009) solved for the social optimum in a simpler model in which all congested travel is at free-flow speed, with only hypercongested traffic slowing down with increased density.

This paper adopts an alternative assumption, which is interpreted as an approximation:

**A-1:** When in transit, the probability that a commuter exits the traffic flow and arrives at his destination between \(t\) and \(t + dt\) is \(v(k(t))L dt\).

Assumption **A-1** is the defining assumption of the bathtub model, since it implies that the arrival rate\(^{12}\),

\[
a(t) = \frac{k(t)v(k(t))}{L},
\]

depends only on contemporaneous traffic density, which is analogous to the outflow from the bathtub depending only on the height of water in it.

The main weakness of **A-1** is that there does not appear to be a plausible set of microeconomic assumptions supporting it. It can be derived from the more primitive pair of assumptions that commuter trip lengths are negative exponentially distributed with mean \(L\) and that commuters do not know their trip lengths at the time they make their departure decisions\(^{13}\), but neither assumption is realistic. The bathtub model, for which **A-1** is an essential element, has three offsetting strengths.\(^{14}\) First, it allows the simple algebra of Poisson processes to be applied, so that the model can readily be extended in the many ways that the bottleneck model has been; second, (like Small and Chu, 2003) it builds on classical flow congestion theory, allowing its performance to be related to observations on downtown traffic flow, density, and velocity; and

\[T(t)\text{ solves } \int_t^{t+T(t)} v(k(u)) du = L\text{ (travel time equals the time to travel the distance }L). \text{ Differentiating both equations with respect to } t \text{ gives } \dot{k}(t+T(t)) = d(t + T(t)) - \frac{d(t)v(k(t+T(t)))}{v(k(t))}. \text{ This is a delay differential equation with an endogenous delay, a class of problems at the research frontier in applied mathematics.}

\(^{11}\)Small and Chu (2003) examined the transient dynamics of traffic congestion in an isotropic downtown with a generalized Greenshields’ Relation. They examined the no-toll equilibrium but not the social optimum.

\(^{12}\)Since density is measured as cars per ml\(^2\) (rather than as the typical cars per ml of road), the arrival rate is measured in units of cars/ml\(^2\)-hr.

\(^{13}\)The probability that a commuter who departs at time \(u\) is still on the road at time \(u\prime\) is \(e^{-\int_u^{u'} \frac{v(k(u'))}{L} du'}\). Expected trip length is therefore \(\int_0^\infty v(k(u))e^{-\int_0^u \frac{v(k(u'))}{L} du'} du = L\), independent of departure time.

\(^{14}\)The basic bathtub model presented here is isomorphic to a bottleneck model in which, first, the service discipline is random (de Palma and Fosgerau, 2011, extends the bottleneck model with fixed capacity to permit alternative service disciplines, including random service) and, second, bottleneck capacity depends in a particular way on queue length. The bottleneck model’s analog to the arrival rate is bottleneck capacity, \(s\), and its analog to traffic density is queue length, \(Q\). Thus, to be consistent with the bathtub model, bottleneck capacity must be related to queue length according to \(s(Q) = Qv(Q)/L\).
third, the bathtub model has intuitive appeal.

The trip cost of a commuter equals his travel time cost plus his schedule delay cost.\footnote{Here, as in most of the theoretical literature on the economics of traffic congestion, the money cost of commuting is ignored. Empirical studies documented in Small and Verhoef (2007) confirm that money cost is less than time cost, especially under congested conditions. As well, the money cost is less sensitive to congestion than the time cost. Including money cost would complicate the algebra while adding little insight.} With a value of travel time of $a$, a value of time early of $\beta$, and a value of time late of $\gamma$, the expected trip cost of a commuter who departs at time $t < t^*$ is

$$Ec(t) = a(\text{expected travel time} \mid t) + \beta(\text{expected time early} \mid t) + \gamma(\text{expected time late} \mid t) ,$$

and of a commuter who departs at $t > t^*$ is

$$Ec(t) = a(\text{expected travel time} \mid t) + \gamma(\text{expected time late} \mid t) .$$

3. The No-Toll Equilibrium

There are two equilibrium conditions. The first is Vickrey’s trip-timing equilibrium condition, that no commuter can reduce his expected trip price by altering his departure time. Since there is no toll, expected trip price equals expected trip cost, so that commuters have a common expected trip cost, $c$, over the departure interval, which is connected. The second is that all commuters must depart – that the integral of the departure rate over the departure interval equals the population. $N$ will be solved for as a function of $c$, with the inverse function relating equilibrium user cost to population. To simplify notation somewhat, time is measured relative to $t^*$, i.e. $t^* = 0$.

3.1. Early Departures

The probability that a commuter departing at time $t$ in the early morning rush hour ($t < 0$) arrives at his destination between $t$ and $t + dt$ is $\frac{v(t)}{L} dt$. If he does, his trip cost is $-\beta t$. If he does not his expected trip cost is $\alpha dt + c$; he incurs a travel time cost of $\alpha dt$ between $t$ and $t + dt$, and at time $t + dt$ faces the expected trip cost of $c$. Thus, $c = -\frac{v(t)dt}{L} \beta t + [1 - \frac{v(t)dt}{L}](\alpha dt + c)$, which simplifies to

$$c = \frac{\alpha L}{v(t)} - \beta t .$$

Rearranging yields

$$v(t) = \frac{\alpha L}{c + \beta t} .$$
which gives equilibrium velocity over the early morning rush hour. Equilibrium density over the early morning rush hour is then obtained from (9) and (1):

\[ k(t) = k_j \frac{c + \beta t - \frac{\alpha L}{v_f}}{c + \beta t} . \]  

(10)

From (6), the equilibrium arrival rate is

\[ a(t) = \frac{v(t)k(t)}{L} = \alpha k_j \frac{c + \beta t - \frac{\alpha L}{v_f}}{(c + \beta t)^2} . \]  

(11)

The equilibrium departure rate can then be derived from \( d(t) = \dot{k}(t) + a(t) \), where \( \dot{k}(t) \) is obtained from (10) and \( a(t) \) from (11):

\[ d(t) = \frac{k_j \alpha L}{v_f(c + \beta t)^2} + \frac{\alpha k_j(c + \beta t - \frac{\alpha L}{v_f})}{(c + \beta t)^2} \]

\[ = \frac{\alpha k_j[(c - \gamma t) - \frac{\alpha L}{v_f}]}{(c + \beta t)^2} . \]  

(12)

At the time of the first departure, velocity equals free-flow velocity. Thus, from (9),

\[ v_f = \frac{\alpha L}{c + \beta t} \]  

so that \( t = \frac{(\frac{\alpha L}{v_f} - c)}{\beta} \).  

(13)

3.2. Late Departures

Late departures may or may not occur. The analysis will proceed on the assumption that they do, and then obtain a condition under which indeed they do. The formulae for velocity, density, and the arrival rate as functions of time are the same as those for early departures, except that \( -\gamma t \) replaces \( \beta t \). The equilibrium departure rate equals the change in density plus the arrival rate:

\[ d(t) = -\frac{k_j \gamma \alpha L}{v_f(c - \gamma t)^2} + \frac{\alpha k_j(c - \gamma t - \frac{\alpha L}{v_f})}{(c - \gamma t)^2} \]

\[ = \frac{\alpha k_j[(c - \gamma t) - \frac{(\alpha + \gamma)L}{v_f}]}{(c - \gamma t)^2} . \]  

(14)

If there are late departures, there is a discontinuous decrease in the departure rate at \( t = 0 \), and \( \bar{t} \) solves \( d(t) = 0 \) from (14). Late departures occur if \( c > \frac{(\alpha + \gamma)L}{v_f} \) — under what are here termed heavily congested conditions but not under moderately or lightly congested conditions. The intuition is that, unless congestion is heavy, the trip-timing condition cannot be satisfied for late departures since, even with no late departures, the reduction in expected travel time cost is insufficient to offset the increase in expected time late cost.

Since the paper’s focus is on heavily congested traffic conditions, in the remainder of this section only that case will be considered. The equilibrium results under moderately and lightly congested conditions can be obtained from the author upon request.
From (14),

\[ t = \frac{c - (\alpha + \gamma)L}{\gamma} . \]  

(15)

The population corresponding to the equilibrium trip cost of \( c \) is solved for as the integral of the departure rate over the departure interval. For populations for which congestion is heavy:

\[ N = k_j \left[ \left( \frac{\alpha}{\beta} + \frac{\alpha}{\gamma} \right) \left( \ln \theta - 1 + \frac{1}{\theta} \right) + 1 + \left( \frac{\alpha}{\gamma} \right) \ln \left( \frac{\gamma}{\alpha + \gamma} \right) \right] , \]

(16)

where \( \theta = \frac{c_v f}{\alpha L} \). This equation relates \( N \) to \( \theta \) and hence to \( c_v \). Since \( N \) is monotonically increasing in \( \theta \) and hence \( c_v \), one can regard (16), more appropriately, as giving \( c_v \) as a function of \( N \), \( c_v(N) \). When population is zero and when therefore there is no congestion, uncongested user cost is \( \frac{\alpha L}{\gamma} \). Thus, \( \theta \) is the ratio of equilibrium user cost to uncongested user cost. Private congestion cost, \( pcc \), is defined to be the difference between equilibrium user cost and uncongested user cost. The ratio of private congestion cost to uncongested user cost is therefore\(^{16} \theta - 1 = \varepsilon \). \( \varepsilon \) is a measure of the severity of congestion, and equals zero under uncongested conditions. Substituting this into (16) gives the relationship between\(^{17} \varepsilon \) and \( N \), \( \varepsilon(N) \):

\[ N = k_j \left[ \left( \frac{\alpha}{\beta} + \frac{\alpha}{\gamma} \right) \left( \ln(1 + \varepsilon) - \frac{\varepsilon}{\varepsilon + 1} \right) + 1 + \left( \frac{\alpha}{\gamma} \right) \ln \left( \frac{\gamma}{\alpha + \gamma} \right) \right] . \]

(17)

Private congestion cost equals \( \varepsilon(N) \) times uncongested user cost, so that the elasticity of private congestion cost with respect to \( N \) equals the elasticity of \( \varepsilon \) with respect to \( N \) (\( E_{pcc,N} = E_{\varepsilon,N} \)). Marginal variable social cost (marginal social cost in excess of uncongested user cost), \( mvsc \), equals\(^{18} \) private congestion cost times one plus the elasticity of \( pcc \) with respect to \( N \) (\( mvsc = pcc(1 + E_{pcc,N}) \)) and congestion externality cost, \( cec \), equals private congestion cost times the elasticity of \( pcc \) with respect to \( N \) (\( cec = pcc E_{pcc,N} \)). For the parameters of the numerical example to be presented in section 5, Figure 2 plots private congestion cost and marginal variable social cost in the no-toll equilibrium, \( pcc^e(N) \) and \( mvsc^e(N) \) respectively\(^{19} \). Congestion externality cost, as function of \( N \), is given by the vertical distance between \( pcc^e(N) \) and \( mvse^e(N) \).

The elasticity of private congestion cost of the transport system with respect to the aggregate number of users, \( E_{pcc,N} \), is of central interest in the analysis of transportation policies that operate at the level of

\(^{16}\) Heavily congested conditions occur when \( \theta > \frac{\alpha + \gamma}{\alpha} \) or \( \varepsilon > \frac{\gamma}{\alpha} \). In the numerical example that will be examined in section 5, \( \frac{\gamma}{\alpha} = 2.0 \). For the inequality \( \varepsilon > \frac{\gamma}{\alpha} \) to hold, a commuter’s equilibrium trip cost must be at least three times as high as that under uncongested conditions.

\(^{17}\) Note that \( N \) does not equal zero when \( \varepsilon = 0 \) since (17) applies only under heavily congested conditions.

\(^{18}\) Private congestion cost is \( pcc(N) = c(N) - c(0) = c(0)(\frac{c(N)}{c(0)} - 1) = c(0)(\theta(N) - 1) = c(0)\varepsilon(N) \), so that \( E_{pcc,N} = \frac{E(N)}{\varepsilon(N)} \). Total variable social cost is \( N(c(N) - c(0)) = Npcc(N) = Nc(0)\varepsilon(N) \). Marginal variable social cost is therefore \( mvsc(N) = c(0)\varepsilon(N) + Ng(0)e^c(N) = pcc(N)(1 + E_{pcc,N}) \), and congestion externality cost is \( cec(N) = Ng(0)e^c(N) = pcc(N)E_{pcc,N} \).

\(^{19}\) The user cost curve equals the \( pcc^e \) curve, shifted up by uncongested user cost. Since there is no congestion toll, the supply curve relating trip price to \( N \) coincides with the user cost curve. The model can be extended to treat price-sensitive demand by adding a demand function relating trip price to \( N \). Equilibrium would then be at the point of intersection of the demand and supply curves.
the downtown or metropolitan areas. Most obviously, the elasticity provides information on how responsive congestion is to the aggregate number of users or (since congestion is almost always assumed to depend on the volume-capacity ratio) to aggregate capacity. But also this elasticity equals the ratio of congestion externality cost to private congestion cost. Combining the time a commuter loses due to congestion, which is relatively easy to measure, with this elasticity and a value of time, provides a simple way of determining the appropriate level of the average congestion toll.

In the standard, textbook model of congestion, trip cost is assumed to equal the product of travel time and the value of time. The most common form of the link travel time function is the Bureau of Public Roads function, \( T = c_0 + c_1 f^b \), where \( c_0 \), \( c_1 \), and \( b \) are constants, and \( f \) is traffic flow. The function has been extensively estimated, and the standard value of \( b \) employed in US traffic studies is 3.5. This function has the property that \( b \) is the elasticity of private congestion cost with respect to \( f \). \( f \) is often measured as traffic flow over some sub-period of the day, such as the morning rush hour, in which case it corresponds to \( N \) in our model. According to this interpretation, the conventional wisdom is that, \( E_{pcc:N} = 3.5 \). In contrast, because of its linearity, the bottleneck model gives a value of this elasticity of 1.0. The bathtub model, with Greenshields’ Relation, provides a way of reconciling these results. Eq. (17) has the properties that \( E_{pcc:N} \) starts at 0 when \( N = 0 \), increases monotonically with \( N \), and approaches infinity in the limit as \( N \) approaches infinity. Thus, in this respect, with moderate congestion the bathtub model behaves like the
bottleneck model, while with heavy (but not very heavy congestion) it behaves like the standard textbook model. It is also of historical interest that Vickrey noted fifty years ago that, consistent with the bathtub model, \( b \) rises with the level of congestion (Vickrey, 1963).

4. The Social Optimum

The social optimum can be solved for using optimal control theory, in a manner similar to Agnew (1976). This section also solves for the time-varying toll that decentralizes the social optimum. In contrast to the previous section, the analysis of this section is conducted with a general congestion function \( v(k) \) having the properties that \( v' < 0 \) and \( 2v' + kv'' < 0 \) (the second derivative of flow with respect to density is negative) for \( k \in (0, k_j) \) where \( k_j \) is jam density, rather than specifically Greenshields’ Relation.

4.1. Optimal control problem

There is one control variable, \( d(t) \), and two state variables, \( A(t) \) and \( k(t) \). The objective function is to minimize the total cost of transporting the \( N \) commuters, the sum of total travel time cost, total time early cost, and total time late cost. As with the analysis of equilibrium, time is measured relative to the common desired arrival time, \( t^* = 0 \). Total travel time cost can be written as \( \int_0^\infty \alpha k(t) \, dt \), since the total travel time between \( t \) and \( t + dt \) equals the density of cars then on the road times the increment of time. Total time early cost is \( \int_0^t \beta A(t) \, dt \) since the total time early between \( t \) and \( t + dt \) for \( t < 0 \) equals the number of commuters who have already arrived at work times the increment of time. Total time late cost is \( \int_0^\infty \gamma (N - A(t)) \, dt \) since the total time late between \( t \) and \( t + dt \) for \( t > 0 \) equals the number of commuters who have not yet arrived at work times the increment of time. The two differential equation constraints are \( \dot{k}(t) = d(t) - \frac{k(t)v(k(t))}{L} \) (which combines (2) and (6)) and \( \dot{A}(t) = \frac{k(t)v(k(t))}{L} \) (which is (6)). There is also an isoperimetric constraint, that the integral of the departure rate over the departure interval equal the population, and initial conditions on the state variables (the non-negativity constraints do not bind). The full optimal control problem is

\[
\min \left[ \int_0^L \alpha k(t) \, dt + \int_0^L \beta A(t) \, dt \right] + \left[ \int_0^\infty \alpha k(t) \, dt + \int_0^\infty \gamma (N - A(t)) \, dt \right]
\]

such that

\[
\begin{align*}
&\ i) \quad \dot{k}(t) = d(t) - \frac{k(t)v(k(t))}{L} & \lambda \\
&\ ii) \quad \dot{A}(t) = \frac{k(t)v(k(t))}{L} & \mu \\
&\ iii) \quad N = \int_0^T d(t) \, dt & \rho \\
&\ iv) \quad k(0) = A(0) = 0
\end{align*}
\]
For early departures, the Hamiltonian is

\[ H = \alpha k + \beta A + \lambda \left( \frac{d - kv(k)}{L} \right) + \mu \left( \frac{kv(k)}{L} \right) - \rho d . \]  

(19)

The first-order condition with respect to \( d \), and the co-state equations for \( k \) and \( A \) are:

\[ d : \lambda - \rho = 0 \]  

(20)

\[ k : \dot{\lambda} = - \left[ \alpha - \frac{\lambda(v + kv')(k)}{L} + \frac{\mu(v + kv')}{L} \right] \]  

(21)

\[ A : \dot{\mu} = - \beta \]  

(22)

\( \rho \) is the marginal social cost of a commuter. \( \lambda(t) \) is the shadow price of density at \( t \). Eq. (20) indicates that, over the early departure interval, the marginal social cost of a commuter is independent of when he departs, so that \( \dot{\lambda}(t) = 0 \). Also, since \( \mu(t) \) is the shadow cost of an extra arrival at \( t \), \( \mu(0) = 0 \). Thus, (21) reduces to

\[ \alpha - \frac{[\rho + \beta t](v + kv')}{L} = 0 . \]  

(23)

This equation can be explained through a perturbation argument. Increase traffic density at time \( t \) by one unit and then decrease it back to its original level a period \( dt \) later. This increases traffic density by one unit for a period of time \( dt \). The travel time cost associated with this is \( \alpha dt \). The arrival rate at time \( t \) is \( \frac{k(t)v(k(t))}{L} \), which increases by \( \frac{v(k) + kv'}{L} \) for the increment of time, and thereafter returns to its original level. Since the cost of an extra arrival at time \( t \) is \( -\beta t \), time early cost increases by \( -\beta t(v(k) + \frac{kv'}{L}) dt \). To restore density to its original level requires subtracting \( \frac{v(k) + kv'}{L} dt \) individuals at time \( t + dt \), at a saving of \( \frac{\lambda(v(k) + kv')}{L} dt \). Since the initial allocation is optimal, this perturbation has no effect on total trip cost.

Since there is no congestion at \( t = \frac{L}{v_f} \), \( (v + kv')_L = v_f \). Thus,

\[ \rho = \frac{\alpha L}{v_f} - \beta \frac{L}{v_f} . \]  

(24)

Marginal social cost equals travel time cost at free-flow speed plus the time early cost associated with arrival at \( \frac{L}{v_f} \).

Substituting \( \rho \) from (24) into (23) yields

\[ \alpha L - \left[ \frac{\alpha L}{v_f} + \beta \left( t - \frac{L}{v_f} \right) \right] (v + kv') = 0 . \]  

(25)

Since the term in square brackets is positive, \( v + kv' \) must be positive too. Since this implies that flow is increasing in density, hypercongestion does not occur in the early morning rush hour. And since the term in square brackets is increasing in \( t \), \( v + kv' \) is decreasing in \( t \), implying that velocity decreases and density increases in the early morning rush hour.
Now imagine that the shape of the function $A(t)$ has been solved for, and consider shifting the function horizontally. Shifting it to the right by an amount $dt$ decreases total time early cost by $\beta A(0) \, dt$ and increases time late cost by $\gamma (N - A(0)) \, dt$. Thus, $\mathcal{L}$ should be chosen such that

$$A(0) = \frac{\gamma N}{\beta + \gamma}. \numbertag{26}$$

It is quite possible that there are no late departures. Suppose that they do occur. The equation for late departures analogous to (25) is

$$\alpha L - \left[\frac{\alpha L}{v_f} - \beta \mathcal{L} - \gamma \mathcal{L}\right] (v + k v') = 0. \numbertag{27}$$

Since the term in square brackets is continuous in $t$ at $t = 0$, so too must be $v + k v'$, which implies that there is no departure mass at $t = 0$. Thus, traffic is congested rather than hypercongested at $t = 0^+$, where $0^+$ denotes the right-hand limit as $t$ approaches zero (and $0^-$ the left-hand limit). Since the term in square brackets is decreasing in $t$, $v + k v'$ is increasing in $t$, which implies that velocity increases and density decreases in the late morning rush hour. Thus, traffic is never hypercongested in the late morning rush hour either. Having obtained an expression for $k(t)$, $d(t)$ can be obtained as $d(t) = \hat{k}(t) + \frac{v(k(t)) k(t)}{L}$, and the time of the last departure is that $t$ for which $d(\bar{t}) = 0$. Now compare (25) and (27). Since the derivative of the term in square brackets decreases discontinuously at $t = 0$, the derivative of $v + k v'$ must increase discontinuously, which implies a discontinuous decrease in $\hat{k}$ and hence a discontinuous decrease in the departure rate.

Suppose alternatively that late departures do not occur. Then either $\bar{t} < 0$ or $\bar{t} = 0$. In the former case, $\bar{t}$ is obtained by solving $d(\bar{t}) = 0$. In the latter case, the departure rate is positive at $t = 0^-$. In both cases, traffic is congested rather than hypercongested at the end of the departure interval. Since traffic density falls monotonically after the last departure, traffic is never hypercongested. Thus, whether the last departure occurs before, at, or after the common desired arrival time, traffic is never hypercongested.

Now consider what happens as $N$ (demand relative to capacity) becomes large. Since travel is never hypercongested, a commuter’s travel time never exceeds $\frac{\alpha L}{v_c}$, where $v_c$ is the capacity velocity. The length of the departure interval, however, continues to increase without bound. As a result, the ratio of total schedule delay cost to total travel time cost increases with $N$. In the limit as $N$ approaches infinity, the social optimum departure pattern approaches that minimizing total schedule delay cost, which is achieved

---

20Late departures occur when $d(0^+) > 0$. With Greenshields’ Relation, this inequality is satisfied iff $\left(\frac{\alpha L}{v_f} - \beta \mathcal{L}\right)^2 > \alpha (\alpha + 2\gamma) \left(\frac{v_f}{v_c}\right)^2$. While this condition is not primitive, it indicates that late departures occur when demand is sufficiently high relative to capacity.
by choosing the departure rate such that density is close to capacity density over almost the entire departure interval. Thus, in the limit, total travel time cost increases linearly with $N$, while total schedule delay cost increases as the square of $N$. It can also be shown that the elasticity of marginal variable social cost and average private congestion cost, as well as the ratio of the average congestion externality cost to the average private congestion cost, increase with $N$, approaching 1.0 in the limit as $N$ approaches infinity. These results are displayed in Figure 3 for the numerical example to be presented in the next section, which assumes Greenshields’ Relation.

![Figure 3: Private congestion cost and marginal variable social cost as functions of population: Social optimum](image)

Figure 4 superimposes the graphs in Figures 2 and 3, again for the numerical example. The average congestion toll equals the vertical distance between $mvsc^{so}$ and $pcc^{so}$, while the per capita deadweight loss from not congestion tolling equals the vertical distance between $pcc^{eq}$ and $pcc^{so}$. At the level of population where the $mvsc^{so}$ and $pcc^{eq}$ curves intersect, the per capita deadweight loss from not congestion tolling equals the average congestion toll. Above this population, which is here termed a situation of high demand (relative to road capacity), imposing the optimal time-varying congestion toll would make all commuters better off even if the toll revenue were completely squandered. The Figure illustrates that, as $N$ becomes large, the ratio of congestion externality cost to private congestion cost approaches one in the social optimum but increases without limit in the no-toll equilibrium.
4.2. The optimal time-varying toll

The optimal time-varying toll, $\tau(t)$, which is applied at the beginning of a trip, equals the difference between the marginal social cost and the user cost, evaluated at the social optimum. In the decentralized optimum, expected trip price is the same for all $t$ in the departure interval. Where $P(u; t) \equiv e^{-\int_{u}^{x} \frac{v(k(x))}{L} dx}$ is the probability that a commuter is still on the road at time $u$, conditional on departure at time $t$, the expected trip price (user cost plus toll) of a commuter who departs early, at time $t$, $Ep(t)$, is

$$Ep(t) = \rho - \beta t + (\alpha - \beta) \int_{0}^{t} P(u; t) du + (\alpha + \gamma) \int_{t}^{\infty} P(u; t) du + \tau(t).$$ (28)

$(\frac{v(u)}{L}) P(u; t)$ is the probability that the commuter departing at time $t$ exits the road at time $u$. Thus, the first integral is expected travel time cost, the second integral expected time early cost, and the third integral expected time late cost.

Integration of this equation by parts yields

$$Ep(t) = \lambda = -\beta t + (\alpha - \beta) \int_{t}^{0} P(u; t) du + (\alpha + \gamma) \int_{0}^{\infty} P(u; t) du + \tau(t).$$ (28)

The minimum trip cost for the commuter who departs at time $t$ is $-\beta t$, which is the cost he incurs if he
exits the road as soon as he departs. Each extra unit of time he remains on the road in the early morning rush hour increases his trip cost by $\alpha - \beta$ (the travel time cost associated with that unit of time minus the reduction in time early cost), and each extra unit of time he remains on the road in the late morning rush hour increases his cost by $\alpha + \gamma$.

It can be shown that

$$\rho = -\beta t + (\alpha - \beta) \int_0^t \hat{P}(u; t) \, du + (\alpha + \gamma) \int_0^\infty \hat{P}(u; t) \, du .$$

(29)

where $\hat{P}(u; t) = e^{-\int_t^u \frac{v + kv'}{v} \, dx}$ is the probability that an extra commuter is still on the road at time $u$, conditional on a commuter being added at time $t$. The calculation of marginal social cost is analogous to the calculation of user cost, except that in the former what is relevant is the probability that an extra commuter is on the road at time $u$ rather than the probability that the added commuter is on the road at time $u$. By slowing them down, the added commuter increases the probability that all travelers who are either on the road when he departs or who depart between time $t$ and time $u$ are still on the road at time $u$.\footnote{In the economics textbook model of traffic congestion, total travel cost per unit time-distance is $\alpha \int_0^T f \, T'(f) \, df$, where $T'(f)$ is the time to travel a unit distance, which is regarded as a function of flow. The user cost is therefore $\alpha T(f)$, the marginal social cost is $\alpha (T(f) + fT'(f))$, and the congestion externality cost is $\alpha f T'(f)$. Now, $T(f) = \frac{1}{1 - f}$, and $\frac{\partial T}{\partial T} = \frac{-\alpha f f'}{\alpha f f'}$. Suppose, in keeping with classical flow congestion, that both $v$ and $f$ are functions of traffic density. Totally differentiating $f(k) = kv(k)$, with respect to $k$ gives $f' = v + kv''$. Interpreting $\frac{df}{df}$ as $\frac{v'}{v + kv''}$ gives $\alpha f T'(f) = \alpha (\frac{1}{1 - f} - \frac{f'}{f})$. Thus, the congestion externality cost here has similarities to the congestion externality cost of the textbook model, but differs from it due to the bathtub model’s stochasticity and its treatment of schedule delay.}

Comparing (28) and (29) gives

$$\tau(t) = (\alpha - \beta) \int_t^0 [\hat{P}(u; t) - P(u; t)] \, du + (\alpha + \gamma) \int_0^\infty [\hat{P}(u; t) - P(u; t)] \, du .$$

(30)

The corresponding expression for the toll for a commuter who departs late is

$$\tau(t) = (\alpha + \gamma) \int_t^\infty [\hat{P}(u; t) - P(u; t)] \, du .$$

(31)

The toll is positive at both the start and end of the departure interval since the first and last commuters to depart impose an expected congestion externality cost. The toll initially increases, reaching a maximum before the common work start time, and then falls.

Closed-form solutions for the various functions in the social optimum under Greenshields’ Relation can be obtained from the author upon request. Figure 5 displays the optimal toll for the numerical example of the next section, for which the common marginal social cost over the departure interval is $35.46. The maximum toll is so high because demand is so high relative to capacity.
5. Comparison of the No-toll Equilibrium and the Social Optimum

This section compares the no-toll equilibrium and the social optimum for a particular numerical example, which assumes Greenshields’ Relation. In the numerical example, free-flow velocity, $v_f$, is 20.0 mph so that (under Greenshields’ Relation) capacity velocity equals 10.0 mph; above this velocity, travel is congested, and below it, travel is hypercongested. Mean trip distance, $L$, is 5.0 miles. The units of density are normalized\(^{22}\), so that capacity flow is 1.0 car-miles per ml\(^2\)-hr, which corresponds to a maximum arrival rate of 0.2 cars per ml\(^2\)-hr and which implies a jam density, $k_j$, of 0.2 cars per ml\(^2\) and capacity density of 0.1 cars per ml\(^2\). Only to make the numerical results more intuitive, time is measured so that the common desired arrival time is 9:00 am The values of $\alpha$ (the value of in-vehicle time), $\beta$ (the value of time early), and $\gamma$ (the value of time late) are $20/\text{hr}$, $10/\text{hr}$, and $40/\text{hr}$ respectively.\(^{23}\)

$N$ is set at 0.6922. That value was chosen so that velocity at the peak of the rush hour in the no-toll equilibrium is 2.5 mph. The example describes a situation where demand is high relative to capacity, as it is during rush hour in the downtown area of almost all large cities outside Northern Europe and North America, and where therefore, in the absence of congestion tolling, rush-hour traffic is highly congested.

\(^{22}\)The congestion technology is scale invariant. What matters is the ratio of population to capacity flow.
\(^{23}\)The equilibrium and optimum solutions depend only on the ratio of these parameters. Thus, with these numbers scaled down appropriately, the example applies to poorer countries. The ratios are based loosely on Small (1982).
Figure 6 displays four panels. Panel I displays velocity as a function of time, Panel II density, Panel III the departure rate, and Panel IV the arrival rate. In each panel, the curve for the no-toll equilibrium is drawn as a solid line and that for the social optimum as a dashed line. The discussion will focus on the early morning rush hour. Consider first Panel I. In the no-toll equilibrium, the early morning rush hour starts at 5:30 am. Traffic becomes increasingly congested up to the work start time, with hypercongestion setting in (corresponding to a velocity of 10 mph) at 6:00 am, and with velocity reaching its minimum value of 2.5 mph at work start time. In the social optimum, the rush hour starts later, at 5:57 am. Traffic becomes increasingly congested over the early morning period, but never becomes hypercongested, with velocity at the peak being 11.41 mph. In the no-toll equilibrium, the common trip cost (which equals trip price since there is no toll) equals\(^{24}\) $40.00. In the social optimum, the common marginal social cost is\(^{25}\) $35.46. In

\(^{24}\)From (13), the common trip cost can be calculated as free-flow travel time (5 mls at 20 mph) times the value of time ($20/hr) plus the cost of arrival at the start of the rush hour (3.5 hrs early times the value of time early of $10/hr).

\(^{25}\)The marginal social cost of the first person to depart is given by (24), equaling free-flow travel time cost of $5.00 plus the cost of arrival at the start of the rush hour (3.0455 hrs early times the value of travel early).
the decentralized social optimum, this is also the common trip price, which equals the time-varying user cost plus the time-varying optimal toll. Thus, trip price is lower in the social optimum than in the no-toll equilibrium, implying that imposition of the optimal time-varying congestion toll would make all commuters better off, even if the toll revenues were wasted. Furthermore, in contrast to almost all previous models of the morning rush hour, imposition of the time-varying toll reduces the length of the rush hour.

Panel II displays density as a function of the toll. Since density is a negative linear function of velocity, this panel displays the same information as that in Panel I, but illustrates directly how much denser traffic is in the no-toll equilibrium than in the social optimum.

Panel III displays the departure rate as a function of time. In the no-toll equilibrium, the departure rate starts at double capacity outflow, which explains why traffic becomes hypercongested so early. The departure rate falls throughout the morning rush hour. There is a discontinuous decrease in the departure rate at \( t^* \). During the late morning rush hour, the departure rate first increases and then decreases to zero at \( \bar{T}^q \). In the social optimum, in contrast, the departure rate equals capacity outflow throughout the morning rush hour. There is a discontinuous decrease in the departure rate at \( t^* \). During the late morning rush hour, the departure rate decreases continuously, becoming zero at \( \bar{T}^{so} \).

Panel IV displays the arrival rate as a function of time. Due to the paper’s simplifying assumption \( A-I \), the arrival rate is proportional to the flow rate on the streets. In the no-toll equilibrium, the arrival rate rises rapidly between 5:30 and 6:00 am, reaching a maximum at 6:00 am when traffic flow becomes hypercongested. In the remainder of the early morning rush hour, traffic becomes increasingly hypercongested, so that the arrival rate decreases, until at work start time it is only one-half its maximum level (velocity is one-quarter capacity velocity and density is twice capacity density, so that flow is one-half capacity flow). In the late morning rush hour, density continuously decreases, with the transition from hypercongested to congested flow occurring at around 9:45. In the social optimum, travel is congested rather than hypercongested throughout the rush hour, so that flow is positively related to density. Density and hence flow increase continuously over the early morning rush hour and then decrease continuously over the late morning rush hour.

Figure 7 contrasts the no-toll equilibrium and the social optimum using a diagram that is standard in the transportation engineering literature. The curves \( D(t) \) and \( A(t) \) give cumulative departures and arrivals, respectively, as a function of time. The vertical distance between the two curves gives traffic density, while the horizontal distance gives travel time for a particular commuter. Total travel time is given by the area between the cumulative departure and cumulative arrival curves; total time early is given by the area below the cumulative arrival curve and to the left of the \( t^* \) line; and total time late is given by the area above the cumulative arrival curve, below the \( N = 0.6922 \) line, and to the right of the \( t^* \) (\( = 9.0 \)) line. Particularly striking is the much higher total travel time in the no-toll equilibrium than in the social optimum.
6. Policy Insights and Directions for Future Research

Most applied studies of congestion tolling have examined the policy using the textbook model, where the benefit from congestion tolling derives exclusively from a reduction in the number of trips taken (e.g., Anderson and Mohring, 1997). These studies typically find that the toll revenue raised is several times the efficiency gain achieved. Unless therefore the toll revenue is spent wisely tolling can reduce social surplus, and even when it raises social surplus some groups are hurt and others helped. There has therefore been much policy and academic discussion of how toll revenues can be used to benefit all major user groups (e.g. Small, 1992), or at least the super-majority needed for tolling to be politically attractive. In the basic bottleneck model of the morning rush hour in contrast, in which trip demand is inelastic, the benefits from congestion tolling derive exclusively from the reallocation of the fixed number of trips over the rush hour. The toll revenue exactly equals the efficiency gain, and so is raised with no burden. The bathtub model of downtown traffic congestion enriches the bottleneck model to allow for classical traffic engineering flow congestion, in which velocity is negatively related to density. In moderate-demand downtowns the efficiency gain from downtown congestion tolling would fall short of the revenue raised, but in high-demand downtowns could exceed it substantially.

Under ideal conditions, perfect congestion tolling is the most efficient policy. However, most cities have to date chosen not to implement congestion pricing for well-documented reasons, and those that have
employed it have implemented imperfect forms. The bathtub model can also be applied to analyze alternative congestion-relief policies. Perhaps the major insight it provides is the importance of preventing recurrent hypercongestion. One class of policies with this aim converts hypercongestion into queues. Ramp metering is a good example in the context of freeway travel. In the context of downtown travel, Carlos Daganzo and his co-authors have been examining traffic signal coordination that effectively restricts entry into zones prone to traffic jams. This policy increases congestion outside these zones, but, if well designed, reduces the overall time loss due to congestion. Another class of policies aimed at preventing recurrent hypercongestion directly regulates travel into zones during times of the day when otherwise traffic jams would recur.

The bathtub model provides a different perspective on downtown traffic congestion, that is consistent with recent theoretical and empirical developments in transportation science. Exploring this perspective further will require more extensive observation and analysis of downtown traffic, the development of traffic microsimulation models that do a better job of modeling the onset, growth, and dissipation of traffic jams, experimentation with alternative policies, as well as enrichment of the bathtub model. It should be possible to extend the bottleneck model, while maintaining analytical tractability, in many of the ways that the bottleneck model has been extended, for example to treat price-sensitive demand, heterogeneity of users, stochastic capacity and demand, and mass transit. Arnott, Inci, and Rowse have applied steady-state variants of the bathtub model to examine downtown parking policy; the stage is now set to incorporate intra-day dynamics. Other natural extensions are to provide an integrated treatment of the three-stage journey to work, suburban access, freeway travel, and downtown egress, with access and egress being modeled as bathtub congestion, and to enrich the modeling of downtown as a network of linked bathtubs, each corresponding to a neighborhood.

7. Conclusion

This paper presented a new economic model of traffic congestion – called the bathtub model – that is particularly well suited to heavily congested downtown areas. The feature of the model that distinguishes it from economists’ previous models of traffic congestion is that traffic jams cause traffic flow to fall, which results in traffic congestion being more costly than current models imply but at the same time in congestion relief policies bringing greater benefits. The bathtub model is so named since traffic density is likened to the height of water in a bathtub that has the property that the outflow is positively related to the water’s height up to a critical level, above which outflow is negatively related to the water’s height. The former region corresponds to normal (‘congested’) traffic, the latter to traffic jam situations (‘hypercongested’ traffic). According to the model, when demand is high relative to capacity, downtown congestion relief policy should focus on preventing traffic jams, which can be achieved through time-varying congestion pricing, converting
traffic jams into queues, and regulating the entry rate of cars into the downtown area.

8. Acknowledgements

I would like thank the US Department of Transportation, grant #IDTRT07-G-9000, and the California Department of Transportation, grant #65A0216, through the University of California, Transportation Center (UCTC) at UC Berkeley for financial support; Anatolii Kokoza and Huiling Zhang for excellent research assistance; and Alex Anas, Mogens Fosgerau, Nikolas Gerolominis, Eren Inci, Keith Knapp, Marvin Kraus, Robin Lindsey, Kenneth Small, Hai Yang, and participants at the Swiss Transportation Research Conference (2011), the UCI-UCLA-USC Urban Research Day Symposium (2011), and the UCR Applied Economics Workshop (2012) for useful comments. Dedicating this paper to the memory of William Vickrey is appropriate since a conversation with him was the seed from which the paper grew.
References


9. Appendix (Not for publication)

Appendix A (Not for publication)

Table 1: No-toll Equilibrium

<table>
<thead>
<tr>
<th>Case I</th>
<th>Case II (Late Departures)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Velocity</td>
<td>[v(t) = \begin{cases} \frac{c_0}{c_1 + v(t) + \beta t} = \frac{\alpha L}{\alpha L}, &amp; t \in [\bar{t}, t^<em>] \ \frac{c_0}{c_2 + \beta t} = \frac{\alpha L}{\alpha L}, &amp; t \in [t^</em>, \bar{t}] \end{cases} ]</td>
</tr>
<tr>
<td>Density</td>
<td>[k(t) = \left(1 - \frac{v(t)}{v_f}\right) k_j = \begin{cases} k_j - \frac{k_j \alpha L}{v_f (c_1 + v(t) + \beta t)}, &amp; t \in [\bar{t}, t^<em>] \ k_j - \frac{k_j \alpha L}{v_f (c_2 + \beta t)}, &amp; t \in [t^</em>, \bar{t}] \end{cases} ]</td>
</tr>
<tr>
<td>Arrival Rate</td>
<td>[a(t) = \frac{v(k)}{L} = \begin{cases} k_j - \frac{k_j c_0}{v_f (c_1 + v(t) + \beta t)} + \frac{k_j c_0 (\beta - \frac{\gamma}{L})}{v_f (c_1 + v(t) + \beta t)}, &amp; t \in [\bar{t}, t^<em>] \ k_j - \frac{k_j c_0}{v_f (c_2 + \beta t)} + \frac{k_j c_0 (\beta - \frac{\gamma}{L})}{v_f (c_2 + \beta t)}, &amp; t \in [t^</em>, \bar{t}] \end{cases} ]</td>
</tr>
</tbody>
</table>

1. These equations apply with late departures. When there are no late departures, the first equation for each of velocity, density, and the arrival rate apply for \( t \in [\bar{t}, \bar{t}] \). After \( t \), there are no departures. Over this interval, density evolves according to the differential equation \( \dot{k}(t) = -k(t) v(k(t)) \), with \( k(\bar{t}) \) determined as indicated above. Over this interval, \( v(t) \) is determined as \( v(k(t)) \) and \( a(t) \) as \( -\dot{k}(t) \).
2. Case I applies if substituting in \( t^* \) into the departure rate equation for late departures gives a negative number.
3. Greenshields’ Relation is assumed: \( v(k) = v_f (1 - \frac{k}{k_f}) \).
4. In the paper, \( t^* \) is set equal to 0.

\[ \bar{t} = \frac{c_2 - c_0 + \gamma L}{\gamma}, \quad \bar{t} = \frac{c_2 - c_0 + \gamma L}{\gamma} \]
Table 2: Social Optimum

<table>
<thead>
<tr>
<th>Table 2: Social Optimum</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Velocity</strong>^1</td>
</tr>
</tbody>
</table>
| \( v(t) = v_f \left( 1 - \frac{k}{k_j} \right) = \begin{cases} 
\frac{1}{2} v_f \left[ 1 + \frac{\alpha L}{v_f + \beta (t - \bar{t})} \right] , & t \in [\bar{t}, t^*] \\
\frac{1}{2} v_f \left[ 1 + \frac{\alpha L}{v_f^L - \beta (t - \bar{t})} \right] , & t \in [t^*, \bar{t}] 
\end{cases} \) |
| **Density**^1           |
| \( k(t) = \begin{cases} 
\frac{k_j}{2} \left[ 1 - \frac{\alpha L}{v_f + \beta (t - \bar{t})} \right] , & t \in [\bar{t}, t^*] \\
\frac{k_j}{2} \left[ 1 - \frac{\alpha L}{v_f^L - \beta (t - \bar{t})} \right] , & t \in [t^*, \bar{t}] 
\end{cases} \) |
| **Arrival Rate**^1      |
| \( a(t) = \frac{v_k}{4L} \left[ 1 - \frac{\left( \frac{v_f}{\alpha L + \beta (t - \bar{t})} \right)^2}{\left( \frac{v_f}{v_f + \alpha L} \right)^2} \right] , & t \in [t^*, \bar{t}] \) |
| **Cumulative Arrival Rate** |
| \( A(t) = \int_t^\bar{t} a(u) \, du \) |
| **Departure Rate**      |
| **Case I** (No Late Departures) |
| \( d(t) = \dot{k}(t) + a(t) = \frac{v_f k_j}{4L} + \frac{k_j \alpha L (2 \beta - \alpha)}{4L (\alpha L + \beta (t - \bar{t}))^2} , & t \in [\bar{t}, \bar{t}] \) |
| **Case II (Late Departures)** |
| \( d(t) = \dot{k}(t) + a(t) = \begin{cases} 
\frac{v_f k_j}{4L} + \frac{k_j \alpha L (2 \beta - \alpha)}{4L (\alpha L + \beta (t - \bar{t}))^2} , & t \in [\bar{t}, t^*] \\
\frac{v_f k_j}{4L} - \frac{k_j \alpha L (2 \gamma + \alpha)}{4L (\alpha L - \beta (t - \bar{t}))^2} , & t \in (t^*, \bar{t}] 
\end{cases} \) |
| **Cumulative Departure Rate** |
| \( D(t) = \int_t^{\bar{t}} d(u) \, du \) |

1. These equations apply with late departures. When there are no late departures, the first equation for each of velocity, density, and the arrival rate apply for \( t \in [\bar{t}, \bar{t}] \). After \( \bar{t} \), there are no departures. Over this interval, density evolves according to the differential equation \( \dot{k}(t) = -\frac{k_k v(k(t))}{L} \), with \( k(\bar{t}) \) determined as indicated above. Over this interval, \( v(t) \) is determined as \( v(k(t)) \) and \( a(t) \) as \( -\dot{k}(t) \).
2. Case I applies if substituting in \( t^* \) into the departure rate equation for late departures gives a negative number.
3. Greenshields’ Relation is assumed: \( v(k) = v_f (1 - \frac{k}{k_j}) \).
4. In the paper, \( t^* \) is set equal to 0.
Table 3: Optimal Toll

<table>
<thead>
<tr>
<th>Condition</th>
<th>Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t &lt; 0 )</td>
<td>( \hat{P}_1(u; t) = \frac{(2u+1-2t)^2}{(2u+1-2t)^2} e^{-2(u-t)} )</td>
</tr>
<tr>
<td></td>
<td>( P_1(u; t) = \frac{2t+1-2t}{2u+1-2t} e^{-2(u-t)} )</td>
</tr>
<tr>
<td>( 0 &lt; t &lt; t )</td>
<td>( \hat{P}_2(u; t) = \frac{(-8t+1-2t)^{1/4}}{(-8t+1-2t)^{1/4}} e^{-2(u-t)} )</td>
</tr>
<tr>
<td></td>
<td>( P_2(u; t) = \frac{(-8u+1-2t)^{1/4}}{(-8u+1-2t)^{1/4}} e^{-2(u-t)} )</td>
</tr>
<tr>
<td>( t &lt; \infty )</td>
<td>( \hat{P}_3(u; t) = \frac{(5+Ce^t)^2}{(5+C(e^t))^2} e^{4(u-t)} )</td>
</tr>
<tr>
<td></td>
<td>( P_3(u; t) = \frac{Ce^t+5}{Ce^t+5} )</td>
</tr>
</tbody>
</table>

Where \( C = \frac{e^{-2.4268}}{0.1 - \frac{0.1}{23.41}} - 5e^{-2.4268} \approx 1.1553 \)
Appendix B (Not for publication)

**Algebra for \( \dot{P}_1(u; t) \):**

The general form of \( \dot{P}_1(u; t) \) is

\[
\dot{P}_1(u; t) = e^{-\int_t^u \frac{v(k_i(x)) v'(k_i(x))}{L} \, dx} = e^{-g_i(x)}. \tag{32}
\]

Where

\[
f_i(x) = \int_t^u \frac{v(k_i(x)) + k_i(x) v'(k_i(x))}{L} \, dx = \int_t^u \frac{20 - 100k_i(x) - 100k_i(x)}{5} \, dx
\]

is dependent on the form of \( k_i(x) \). Defining \( k_1(x) \) to equal \( k(x) \) over the interval \([t, 0]\), from Table 2 (and with the normalization that \( t^* = 0 \)):

\[
k_1(x) = \frac{k_j}{2} \left( 1 - \frac{\alpha L}{\alpha} e^{\beta(x - 1)} \right) = 0.1 \left( 1 - \frac{1}{1 + 2(x - 1)} \right). \tag{33}
\]

Substituting this into the general result for \( f_i(x) \):

\[
f_1(x) = 4 \int_t^u \frac{1}{1 + 2(x - t)} \, dx = [2 \log(2x - 2t + 1)]_t^u
\]

\[
f_1(x) = 2 \log \left( \frac{2u - 2t + 1}{2t - 2t + 1} \right).
\]

Substituting this into the general form for \( \dot{P}_1(u; t) \),

\[
\dot{P}_1(u; t) = e^{-2 \log \left( \frac{2^{u-2t+1}}{2t-2+1} \right)} = \left( \frac{2t - 2t + 1}{2u - 2t + 1} \right)^2.
\]

**Algebra for \( P_1(u; t) \):**

The general form of \( P_1(u; t) \) is

\[
P_1(u; t) = e^{-\int_t^u \frac{v(k_i(x))}{L} \, dx} = e^{-g_i(x)}. \tag{34}
\]

Where

\[
g_i(x) = \int_t^u \frac{v(k_i(x))}{L} \, dx = \int_t^u \frac{20 - 100k_i(x)}{5} \, dx = \int_t^u 4 - 20k_i(x) \, dx. \tag{35}
\]

Using \( k_1(x) \) from (33):

\[
g_1(x) = 4 \int_t^u 1 - 5 \left( 0.1 - \frac{0.1}{1 + 2(x - 1)} \right) \, dx = \int_t^u 2 + \frac{2}{1 + 2(x - t)} \, dx
\]

\[
g_1(x) = [2x - \log(1 + 2(x - t))]_t^u = 2(u - t) - \log \left( \frac{1 + 2(u - t)}{1 + 2(t - t)} \right),
\]

and substituting \( g_1(x) \) into (34)

\[
P_1(u; t) = e^{-2(u-t) - \log \left( \frac{1 + 2(u - t)}{1 + 2(t - t)} \right)} = \frac{1 + 2(u - t)}{1 + 2(t - t)} e^{-2(u-t)}. \tag{36}
\]
Algebra for $\hat{P}_2(u; t)$:

$$\hat{P}_2(u; t) = e^{-f_2(x)}$$

(37)

Where

$$f_2(x) = 4 \int_t^u 1 - 10k_2(x) \, dx.$$  

(38)

Defining $k_2(x)$ to equal $k(x)$ over the interval $[0, \tilde{t}]$ from Table 2 (and with the normalization that $t^* = 0$):

$$k_2(x) = 0.1 \left( 1 - \frac{5}{5 - 10t - 40x} \right) = 0.1 \left( 1 - \frac{1}{1 - 2t - 8x} \right).$$

(39)

Substituting (39) into (38)

$$f_2(x) = 4 \int_t^u 1 - 10 \left[ 0.1 \left( 1 - \frac{1}{-8x + 1 - 2t} \right) \right] \, dx$$

$$= 4 \int_t^u 1 - 1 + \frac{1}{-8x + 1 - 2t} \, dx = \int_t^u \frac{4}{-8x + 1 - 2t} \, dx$$

$$= \left[ -\frac{1}{2} \log(-8x + 1 - 2t) \right]_t^u = -\frac{1}{2} \log \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right),$$

and substituting $f_2(x)$ into (37)

$$\hat{P}_2(u; t) = e^{\frac{1}{2} \log \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right)} = \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right)^{1/2}.$$  

(40)

Algebra for $P_2(u; t)$:

$$P_2(u; t) = e^{-g_2(x)}$$

(41)

where

$$g_2(x) = \int_t^u 4 - 20k_2(x) \, dx.$$  

(42)

Substituting $k_2(x)$ from (39) into (42),

$$g_2(x) = \int_t^u 4 - 20 \left( 0.1 - \frac{0.1}{1 - 2t - 8x} \right) \, dx = \int_t^u 2 + \frac{2}{1 - 2t - 8x} \, dx$$

$$g_2(x) = \int_t^u 2 + \frac{2}{1 - 2t - 8x} \, dx = \left[ 2x - \frac{1}{4} \log(1 - 2t - 8x) \right]_t^u$$

$$g_2(x) = 2(u - t) - \frac{1}{4} \log \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right),$$

(43)

and substituting (43) into (41), we get

$$P_2(u; t) = e^{\frac{1}{4} \log \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right)} e^{-2(u - t)} = \left( \frac{-8u + 1 - 2t}{-8t + 1 - 2t} \right)^{1/4} e^{-2(u - t)}$$

$$P_2(u; t) = \frac{-8u + 1 - 2t}{-8t + 1 - 2t}^{1/4} e^{-2(u - t)t}.$$  

(44)

Algebra for $k(t)$ for $t > \tilde{t}$
From Note 1 in Table 2:

$$\dot{k}(t) = -\frac{k(t)v(k(t))}{L}. \quad (45)$$

This Bernoulli equation can be rewritten as the linear ODE

$$k' = -\frac{k(20-100k)}{5} \Rightarrow k' + 4k = 20k^2 \quad \Rightarrow \quad 26 - v' + 4v = 20 \quad \Rightarrow \quad v' - 4v = -20 .$$

Using the integrating factor $I(t) = e^\int -4dt = e^{-4t}$,

$$\frac{d}{dt}(e^{-4t} v) = -20e^{-4t} \quad \Rightarrow \quad e^{-4t} v = \int -20e^{-4t} dt \quad \Rightarrow \quad e^{-4t} v = 5e^{-4t} + C,$$

and substituting $v = \frac{1}{k(t)}$,

$$k_3(t) = \frac{1}{5 + Ce^{4t}}. \quad (46)$$

To find $C$, solve for $k_3(t)$:

$$k_3(T) = 0.1 - \frac{0.1}{1 - 2\lambda - 8\tau} \quad \Rightarrow \quad \frac{1}{5 + Ce^{4(T)} - \frac{0.1}{1 - 2(-3.055) - 8(0.6067)}} = \frac{0.1}{1 - 2(-3.055) - 8(0.6067)} - 5$$

$$C = \frac{e^{-2.4268}}{0.1 - \frac{0.1}{2.2374}} - 5e^{-2.4268} \approx 1.1553 . \quad (47)$$

**Algebra for $\hat{P}_3(u; t)$**

$$\hat{P}_3(u; t) = e^{-f_3(x)} , \quad \text{where} \quad f_3(x) = \int_t^u 4 - 40k_3(x) \, dx . \quad (48)$$

Using $k_3(t)$ from (46),

$$f_3(x) = 4 \int_t^u 1 - \frac{10}{5 + Ce^{4x}} \, dx = \left[ 2\log(5 + Ce^{4x}) - 4x \right]_t^u$$

$$f_3(x) = 2\log \left( \frac{5 + Ce^{4x}}{5 + Ce^{4\tau}} \right) - 4(u - t) ,$$

and substituting $f_3(x)$ into (48), we get

$$\hat{P}_3(u; t) = \frac{(5 + Ce^{4\tau})^2}{(5 + Ce^{4u})^2} e^{4(u-t)} . \quad (49)$$

**Algebra for $P_3(u; t)$**

$$P_3(u; t) = e^{-g_3(x)} , \quad \text{where} \quad g_3(x) = \int_t^u 4 - 20k_3(x) \, dx .$$

---

$26v = \frac{1}{k}$ and $v' = \frac{k'}{k^2}$
Using $k_3(t)$ from (46),

$$g_3(x) = \int_t^u 4 - \frac{20}{5 + Ce^{4x}} \, dx = \left[ \log(Ce^{4x} + 5) \right]_t^u$$

$$g_3(x) = \log \left( \frac{Ce^{4u} + 5}{Ce^{4t} + 5} \right),$$

and substituting $g_3(x)$ into $P_3(u; t) = e^{-g_3(x)}$,

$$P_3(u; t) = \frac{Ce^{4t} + 5}{Ce^{4u} + 5}.$$  

(50)
Optimal toll for early departures

\[
\tau(t) = (\alpha - \beta) \int_t^0 \hat{P}_1(u; t) - P_1(u; t) \, du
+ (\alpha + \gamma) \int_t^T \hat{P}_1(0; t)\hat{P}_2(u; 0) - P_1(0; t)P_2(u; 0) \, du
+ (\alpha + \gamma) \int_T^\infty \hat{P}_1(0; t)\hat{P}_2(T; 0)\hat{P}_3(u; T) - P_1(0; t)P_2(T; 0)P_3(u; T) \, du.
\]

Substituting in the values of \( P_1(u; t) \), \( \hat{P}_1(u; t) \), \( P_2(u; t) \), \( P_3(u; t) \), and \( \hat{P}_3(u; t) \),

\[
\tau(t) = 10 \int_t^0 \frac{(2t + 1 - 2T)^2}{(2u + 1 - 2T)^2} - \frac{2t + 1 - 2T}{2u + 1 - 2T} e^{2t-2u} \, du
+ 60 \int_0^T \left[ \frac{(2t + 1 - 2T)^2}{(2u + 1 - 2T)^2} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/2}}{(1 + 2T)^{1/2}} \right]_{t=0}^{t=T} \left[ \frac{(5 + Ce^{4T})^2}{(5 + Ce^{4u}e^{4T})^2} \right]_{t=0}^{t=T} \, du
- \left[ \frac{2t + 1 - 2T}{2u + 1 - 2T} e^{2t-2u} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/4}}{(1 + 2T)^{1/4}} e^{2t-2u} \right]_{t=0}^{t=T} \, du
+ 60 \int_T^\infty \left[ \frac{(2t + 1 - 2T)^2}{(2u + 1 - 2T)^2} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/2}}{(1 + 2T)^{1/2}} \right]_{t=0}^{t=T} \left[ \frac{(5 + Ce^{4T})^2}{(5 + Ce^{4u}e^{4T})^2} \right]_{t=0}^{t=T} \, du
- \left[ \frac{2t + 1 - 2T}{2u + 1 - 2T} e^{2t-2u} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/4}}{(1 + 2T)^{1/4}} e^{2t-2u} \right]_{t=0}^{t=T} \, du
\]

and substituting the appropriate values for \( u \) and \( t \),

\[
\tau(t) = 10 \int_t^0 \frac{(2t + 1 - 2T)^2}{(2u + 1 - 2T)^2} - \frac{2t + 1 - 2T}{2u + 1 - 2T} e^{2t-2u} \, du
+ 60 \int_0^T \left[ \frac{(2t + 1 - 2T)^2}{(1 + 2T)^2} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/2}}{(1 + 2T)^{1/2}} \right]_{t=0}^{t=T} \left[ \frac{(5 + Ce^{4T})^2}{(5 + Ce^{4u}e^{4T})^2} \right]_{t=0}^{t=T} \, du
- \left[ \frac{2t + 1 - 2T}{1 + 2T} e^{2t} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/4}}{(1 + 2T)^{1/4}} e^{-2t} \right]_{t=0}^{t=T} \left[ \frac{Ce^{4T} + 5}{Ce^{4u} + 5} \right]_{t=0}^{t=T} \, du
+ 60 \int_T^\infty \left[ \frac{(2t + 1 - 2T)^2}{(1 + 2T)^2} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/2}}{(1 + 2T)^{1/2}} \right]_{t=0}^{t=T} \left[ \frac{(5 + Ce^{4T})^2}{(5 + Ce^{4u}e^{4T})^2} \right]_{t=0}^{t=T} \, du
- \left[ \frac{2t + 1 - 2T}{1 + 2T} e^{2t} \right]_{u=0}^{u=T} \left[ \frac{(-8u + 1 - 2T)^{1/4}}{(1 + 2T)^{1/4}} e^{-2t} \right]_{t=0}^{t=T} \left[ \frac{Ce^{4T} + 5}{Ce^{4u} + 5} \right]_{t=0}^{t=T} \, du
\]
where

\[ C = \frac{e^{-2.4268}}{0.1 - \frac{0.1}{2.374}} - 5e^{-2.4268} \approx 1.1553 . \]
Optimal toll for late departures

\[ \tau(t) = (\alpha + \gamma) \int_{t}^{T} P_2(u; t) - P_2(u; t) \, du \]

\[ + (\alpha + \gamma) \int_{T}^{\infty} P_2(\bar{t}; \bar{t}) \bar{P}_3(u; \bar{t}) - P_2(\bar{t}; \bar{t}) \bar{P}_3(u; \bar{t}) \, du \]

Using the expressions for \( P_i(u; t) \) and \( \bar{P}_i(u; t) \) \((i = 2, 3)\) for early departures and integrating, one obtains

\[
\tau(t) = 60 \int_{t}^{T} \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2t − 2u} \right] du
\]

\[ + 60 \int_{T}^{\infty} \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{(5 + Ce^{4t})^{2}}{(5 + Ce^{4u})^{2}} e^{4u − 4t} \right]_{t=\bar{t}} \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2(t − u) + e_{2}} du \]

\[
\tau(t) = 60 \int_{t}^{T} \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2t − 2u} \right] du
\]

\[ + 60 \int_{T}^{\infty} \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{(5 + Ce^{4t})^{2}}{(5 + Ce^{4u})^{2}} e^{4u − 4t} \right]_{t=\bar{t}} \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2t − 2t} \left( \frac{Ce^{4t} + 5}{Ce^{4u} + 5} \right) du \]

\[
\tau(t) = 60 \int_{t}^{T} \left[ \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2t − 2u} \right] \right] du
\]

\[ + 60 \int_{T}^{\infty} \frac{(−8u + 1 − 2t)^{0.5}}{(−8t + 1 − 2t)^{0.5}} \left[ \frac{(5 + Ce^{4t})^{2}}{(5 + Ce^{4u})^{2}} e^{4u − 4t} \right]_{t=\bar{t}} \frac{8u + 1 − 2t}{(−8t + 1 − 2t)^{0.25}} e^{2t − 2t} \left( \frac{Ce^{4t} + 5}{Ce^{4u} + 5} \right) du \]

where

\[ C = \frac{e^{-2.4268}}{0.1 − 0.1 \frac{1}{2374}} − 5e^{-2.4268} \approx 1.1553 . \]