Hypercongestion in downtown metropolis[*]

Mogens Fosgerau[†] Kenneth A. Small[‡]

December 13, 2012

Abstract

Engineering studies demonstrate that traffic in dense downtown areas obeys a stable functional relationship between average speed and density, including a region of ’hypercongestion’ where flow decreases with density. This situation can be described as queuing behind a bottleneck whose capacity declines when the queue is large. We combine such a variable-capacity bottleneck with Vickrey scheduling preferences for the special case where there are only two possible levels of capacity. Solving the model leads to several new insights, including that the marginal cost of adding a traveler is especially sensitive to the lowest level of capacity reached. We analyze an optimal toll, a coarse toll, and metering, showing substantial benefits from using these policies to eliminate the period of reduced capacity. Under hypercongestion, all of these policies can be designed so that travelers gain even without considering any toll revenues.

Keywords: hypercongestion; congestion; road pricing;

Forthcoming, Journal of Urban Economics

[*] Mogens Fosgerau has received support from the Danish Strategic Research Council.

[†] Technical University of Denmark, Centre for Transport Studies, Sweden and Ecole Nationale Supèrieure. mf@transport.dtu.dk

[‡] University of California, Irvine and Resources for the Future

1 Introduction

Recent studies of traffic in urban street networks carried out by Daganzo and his associates establish a number of regularities that promise to be very useful for economic modeling of urban congestion.[1] Their findings apply to neighborhoods, defined as uniformly congested parts of cities of dimensions comparable with a trip length. In such neighborhoods, regularities among aggregate variables emerge even though data from specific points appear quite chaotic.

Two such regularities are especially important for congestion modeling. First, there is a well-defined inverse-U-shaped relationship between space-averaged flow and vehicle density (the latter is proportional to occupancy, the number of cars traveling in the neighborhood). Second, the trip completion rate is proportional to the space-averaged flow.

In contrast to the microscopic level of detail involved in looking at individual streets and individual cars, we may consider neighborhoods as macroscopic entities. The observed regularities make it possible to abstract from the microscopic complexity of actual traffic networks in order to offer a remarkably simple picture of congestion at the macroscopic neighborhood level. Cars embarking on trips enter a neighborhood at some rate, adding to the network density there. Cars move ahead, passing points in the network according to a flow rate that on average depends only on network density. Trips are completed and cars leave the system at a rate that is proportional to the instantaneous flow. This last property is extremely useful, enabling us to identify flow with quantity demanded. Thus it frees us from having to take into account that trips have varying lengths and that changing the flow rate may change the distribution of trip lengths.

The relationship between flow and density identifies regions of congestion and hypercongestion, namely the rising and falling portions of the inverse-U flow-density relationship. Under congested conditions, flow increases with density but less than linearly. As more cars enter the system, flow increases at a diminishing rate until it reaches the maximum flow. Above this point, additional traffic in the system will decrease flow; this phenomenon is called hypercongestion.[2]

Conservation of vehicles dictates that at any point in time, the rate of change in occupancy is the difference between the instantaneous entry and the exit rates. Hypercongestion is thus inevitably a transient phenomenon. For hypercongestion to occur there must have been a period with entry at a rate higher than the maximum flow rate, but that rate cannot continue indefinitely as the occupancy would then increase indefinitely. More generally, it is necessary to consider intra-day dynamics in order to understand congestion and hypercongestion in a demand peak.

[1] See Daganzo (2007); Geroliminis and Daganzo (2008).

[2] In the engineering literature, these two regimes are called "uncongested" and "congested" flow, respectively.

To consider intra-day dynamics, we must take into account that travelers have preferences regarding the timing of their trips, for which we use a formulation that has become standard. We also need to determine the relationship between when trips begin and when they end. For this, we make an assumption that may appear innocuous but is actually quite strong: namely, that trips are completed in the sequence in which they are initiated. This first-in-first-out (FIFO) principle is not necessarily consistent with a microscopic picture of the evolution of individual travel speeds, but it is at least as plausible as several alternative assumptions that have been made to make analysis of hypercongestion tractable.[3] Without FIFO, analysis would become extremely complex because one would have to explicitly depict separate routes for each traveler and their interactions on a network; but general features of the model should remain because we would still have equilibrium conditions equating costs for a given traveler at different times, and the feature of capacity reduction when vehicle density becomes large.

This description of congestion in an urban neighborhood brings us close to the highly successful “bottleneck model” of Vickrey (1969). As shown by Vickrey and further elaborated by Fargier (1983), Arnott, de Palma and Lindsey (hereafter ADL) (1990,1993,1998), and others, the model depicts Nash equilibrium where travelers adjust their departure times endogenously, accounting for aversion to inconvenient schedules as well as to travel time. In the bottleneck model, there is a queue that waits behind a deterministic bottleneck. But there is no hypercongestion: the bottleneck has constant capacity, hence implies a constant trip completion rate once a queue exists. Other models have considered time-varying capacities (Zhang et al., 2010) and so can analyze exogenous temporary imbalances between inflow and capacity, but still do not depict a situation where greater inflow results in lower outflow.

In our model, by contrast, there is a pool of cars traveling, whose trip times are affected by the system’s limited processing capacity just as though they were in a queue. (Indeed, at a micro level they typically are in queues at various intersections.) But the flow rate from this "queue" declines with occupancy, producing hypercongestion. Thus we can analyze an apparently complex macroscopic system similarly to a bottleneck with variable capacity. This equivalence between the two problems is noted by Geroliminis and Levinson (2009). In comparison to their analysis, we simplify the dependency of flow on occupancy, enabling us to solve the model explicitly. Specifically, we assume the bottleneck has two capacities, the lower one being activated when occupancy reaches a certain level. Travelers

[3] Two such alternatives include that trip time depends solely on the density at the end of the trip (Small and Chu, 2003), and that trip distances are randomly distributed and unknown to the traveler (Arnott, 2011). Other papers making similar assumptions to that of Small and Chu (2003) include Henderson (1981), Mahmassani and Herman (1984), Chu (1995) and Yang and Huang (1997).

understand this and, in Nash equilibrium, account for it in their departure-time decisions.

Our results demonstrate that indeed travel costs rise much more severely with demand for travel than in the conventional bottleneck model, and thus there are correspondingly greater gains to policies such as capacity reduction and dynamic pricing that relieve congestion. Where the standard bottleneck model shows the benefits associated with reduced queueing, our model shows additional gains associated with avoiding capacity drop. The latter become especially visible in another policy analyzed here: metering, i.e. use of traffic signals to restrain the rate of inflow to certain parts of the road network.[4] The additional gains associated with avoiding capacity drop also mean that the average cost of trips will decrease. In fact, we obtain the somewhat surprising feature that if demand is not perfectly inelastic, optimally regulating hypercongestion may entail increasing traffic—it will definitely do so in the case of metering, and will do so in the case of tolling if congestion (absent pricing) is especially severe.

The nature of hypercongestion and how to deal with it in economic models has been the subject of a long discussion: its existence has been generally acknowledged but its significance debated. Partly this is because hypercongestion has been defined from microscopic data, but the interest in it arises from macroscopic phenomena. Partly it is because hypercongestion can be defined in terms of a static speed-flow relationship that entails what is effectively a backward bending supply curve (e.g. Walters, 1961). But, as we have just seen, hypercongestion is an inherently dynamic phenomenon. Our approach deals explicitly with the dynamic macroscopic properties that are the main source of interest in economic analysis of very congested systems.

We proceed by defining and analyzing the variable-capacity model just described in Sections 2 and 3. We then consider and compare a variety of policies in Section 4, namely an optimal time-dependent toll, an optimal coarse toll (with just one toll level), and metering of system inflow so as to eliminate hypercongestion. Section 5 discusses extension to the case of elastic demand. We perform numerical simulations in Section 6 and section 7 draws conclusions for policy and for research.

2 Model setup

We consider a continuum of N travelers all making trips. Each traveler must choose a departure time; the resulting arrival time is determined by the queueing

[4] When there are multiple user groups accessing the road with limited capacity at different points, metering can also be used to reduce aggregate user cost by better allocating priority among these groups (Shen and Zhang, 2010).

system and thus depends on the aggregate departure schedule. We denote cumulative departures by $R(t)$, which has derivative $\rho(t) \geq 0$ almost everywhere. All travelers depart eventually, i.e. $R(\infty) = N$. Similarly, we denote cumulative exits by $A(t)$, which is weakly increasing and satisfies $A(-\infty) = 0$ and $A(\infty) = N$. Arrivals occur later than departures, i.e. $R(t) \geq A(t)$. The number of travelers $Q(t)$ in the system at any time is the number who have departed less the number who have arrived: $Q(t) = R(t) - A(t)$.

For simplicity, we will ignore any travel time not related to congestion; adding free-flow time or cost is a trivial extension. Therefore the first time anybody departs is also the first time anybody arrives, i.e. there is no delay for the first person: formally, $\inf\{t|R(t) > 0\} = \inf\{t|A(t) > 0\}$.

The queueing system obeys the first-in-first-out (FIFO) rule. This means that the traveler departing at time t, with position $R(t)$ in the sequence of departures, has the same position in the sequence of arrivals. We denote the time of arrival as $a(t) \equiv A^{-1}[R(t)]$. (Because $A(t)$ is weakly increasing in t, this inverse exists wherever $A' > 0$, to which region we restrict attention.) The travel time for a traveler departing at time t is the horizontal distance between the functions $R(\cdot)$ and $A(\cdot)$, i.e. it is $a(t) - t$.

The FIFO assumption guarantees that the later entrants will never exit before the earlier ones. However, it does not imply a lack of effect of later travelers on earlier ones. On the contrary, and in distinction from the standard bottleneck model, they have a profound effect as now described. In short, this effect occurs because later travelers affect the length of the queue and therefore the processing rate that applies while the traveler in question is waiting in that queue, even though they are behind that traveler in position. Thus, we should not think of the queueing system as a literal queue, but rather a system where later entrants can influence earlier ones. For example, in the areawide setting described in the introduction, they might do so by blocking intersections that the earlier entrants will use to get to their exits.

The delay for each traveler is governed by a processing rate $\psi \geq 0$ for exiting the queueing system. In the Vickrey (1969) bottleneck model, ψ is a constant, so the function $A(\cdot)$ is very simple: it has derivative ψ whenever there is a queue. But here, following Geroliminis and Levinson (2009) and Gonzalez and Daganzo (2011), we assume that the processing rate at time s is a function of $Q(s)$, i.e., $\psi = \psi(Q(s))$ and that $A'(s) = \psi(Q(s))$ whenever there is a queue. Although ψ is a temporary processing rate, we refer to it as "capacity," and to our model as having variable or endogenous capacity, in much the same way that flow breakdown at real highway bottlenecks is often described as a capacity reduction.

For a traveler entering the queueing system at time t, there are $Q(t)$ earlier travelers in who must be processed before this given traveler can exit the system. They are processed at rate $\psi[Q(s)]$ over the succeeding times s. Therefore arrival time $a(t)$ is defined implicitly by

$$Q(t) = \int_{t}^{a(t)} \psi[Q(s)]\, ds. \tag{1}$$

Equation (1) shows that any traveler entering between times t and $a(t)$ can influence the travel time of the traveler entering at t, via the function $Q(s)$.

The consistency required by (1) makes the model intractable to solve in general. As a result, researchers investigating hypercongestion have replaced (1) by something simpler, for example letting $Q(s)$ inside the integral be replaced by $Q(t)$ or by $Q[a(t)]$ (see footnote 3). In this paper, we solve the problem without compromising the mutual interactions of travelers in (1) by assuming a particularly simple form for $\psi(Q)$. Namely, we assume $\psi(Q)$ is piecewise constant, taking just three values: full capacity ψ_0, reduced capacity ψ_1, or zero:

$$\psi(Q) = \begin{cases} \psi_0, & Q \leq Q_0 \\ \psi_1, & Q_0 < Q < Q_J \\ 0, & Q \geq Q_J, \end{cases}$$

where Q_0 is a critical queue size above which capacity drops and Q_J is the queue size where vehicles stop moving entirely, corresponding to the jam density in a traffic flow model. We consider only values of N for which Q never reaches jam density, since otherwise we would have infinite travel delays and the model would break down; thus we actually need deal with only two values for ψ.

Having described congestion technology, we now turn to behavior. Again, we will need simplifying assumptions in order to obtain a tractable model and results that are amenable to interpretation. First, travelers are identical. Second, travelers care about the timing of their departure and arrival with as expressed by a user cost of the α-β-γ type formulated by Vickrey (1969), estimated by Small (1982), and used by numerous authors since.[5] Specifically, the user cost associated with departing at time t and arriving at time a is

$$c(t,a) = \alpha \cdot (a-t) + \beta \cdot \max(t^* - a, 0) + \gamma \cdot \max(a - t^*, 0),$$

where α is the value of time, β is the cost of earliness, γ is the cost of lateness, and t^* is the preferred arrival time. Like most authors using this behavioral model, we require $0 < \beta < \alpha < \gamma$, assumptions supported empirically (Small, 1982), in order to produce a sensibly shaped peak period. We normalize $t^* = 0$ at no loss of generality.

[5] For example, Fargier (1983), Arnott et al. (1990, 1993). For useful reviews, see Arnott et al. (1998) or Small and Verhoef (2007).

We look for a Nash equilibrium in which the macroscopic state of the system, arising from the aggregate of individual scheduling decisions, leaves each traveler achieving the lowest possible cost given that state. Given identical travelers, this means that in equilibrium, user cost takes a constant value for all departure times for which the departure rate ρ is positive, and no lower values elsewhere.

3 Unregulated equilibrium

It is straightforward to show that in equilibrium with no toll or metering, departures and arrivals all take place during a common interval $[t_0, t_1]$. The reasoning is identical to that in the standard bottleneck model (ADL 1990). The first and last departures (at times t_0 and t_1, respectively) provide congestion-free travel, while in between the queue is always positive. The first and last departure times are related by the equal-cost condition for first and last travelers: $\gamma t_1 = -\beta t_0$.

The form of the cost function implies that the arrival time $a(t)$ and delay $[a(t) - t]$ are piecewise-linear functions of departure time, governed solely by cost parameters α, β, and γ. For early departure times, i.e. those for which $a(t) < 0$, average cost $[\alpha \cdot (a(t) - t) - \beta \cdot a(t)]$ is constant, which requires that $a(\cdot)$ be linear with slope

$$a'(t) = \alpha / (\alpha - \beta) \equiv a'_E.$$

A similar condition for departures corresponding to late arrival $a(t) > 0$ yields slope

$$a'(t) = \alpha / (\alpha + \gamma) \equiv a'_L.$$

Accounting for boundary conditions, then, the arrival rate between times t_0 and t_1 is

$$a(t) = \begin{cases} t_0 + a'_E \cdot (t - t_0), & t < t_M \\ a'_L \cdot (t - t_M), & t > t_M \end{cases} \tag{2}$$

where $t_M \equiv (\beta/\alpha) t_0$ is the departure time leading to arrival at $t^* \equiv 0$. Note that travel delay, $a(t) - t$, first grows at rate $a'_E - 1 = \beta / (\alpha - \beta)$ and then declines at rate $1 - a'_L = \gamma / (\alpha + \gamma)$.

We now derive a departure pattern consistent with this equilibrium arrival pattern, as well as with congestion technology. We can find the equilibrium departure rate $\rho(t)$ by differentiating (1), keeping in mind that the queue is $Q(t) = R(t) - A(t)$, the departure rate is $R'(t) = \rho(t)$, and the arrival rate is $A'[t] = \psi[Q(t)]$. This yields, almost everywhere, that

$$\rho(t) = a'(t) \cdot \psi[Q(a(t))]. \tag{3}$$

Table 1: Base parameters for numerical examples

Assumed Parameters:		Derived quantities	
α	1.0	δ	0.4
β	0.5	$a_E^{'}$	2.0
γ	2	a_L'	1/3
ψ_0	1	N_1^c	5.0
ψ_1	0.5	N_2^c	9.0
Q_0	2.0	N_3^c	11.0

Note that the equilibrium departure rate at time t depends on the exit rate at time $a(t)$; this is why the model with time-varying capacities is intractable in general. With our simplifying assumptions, there are only two possible values for a', as seen from (2), and two for ψ: hence there are just four possible values of ρ.

It turns out there are three distinct possible equilibrium patterns, each arising for successively larger values of N. We name them Regimes 1, 2, and 3. The math involved in determining these regimes is straightforward but tedious, and is given in the appendix. Here we provide an overview.

3.1 Regime 1

The first regime is where demand N is sufficiently low that only congestion and not hypercongestion occurs. The reduced capacity is not activated and so this is the standard bottleneck model (Arnott et al., 1993), depicted in Figure 1a using the parameters shown in Table 1. As is well known, in that case $t_0 = -(\delta/\beta) N/\psi_0$, $t_1 = -(\beta/\gamma) t_0$, and the maximum queue length is $\delta N/\alpha$, where $\delta \equiv \beta\gamma/(\beta+\gamma)$ is a measure of the strength of scheduling costs (those parts of user cost related to preferences over schedules). The condition ensuring that the low capacity is not activated is the one ensuring that maximum queue length does not reach Q_0:

$$N \leq N_1 \equiv \alpha Q_0/\delta.$$

The departure rate is defined almost everywhere by

$$\rho(t) = \begin{cases} a_E'\psi_0, & t \in (t_0, t_M) \\ a_L'\psi_0, & t \in (t_M, t_1). \end{cases} \tag{4}$$

All travelers achieve the same cost in equilibrium, equal to the cost for the first traveler departing at time t_0. Hence total cost as a function of N is $\delta N^2/\psi_0$; marginal cost (its derivative) is $2\delta N/\psi_0$, which is exactly twice the average cost.

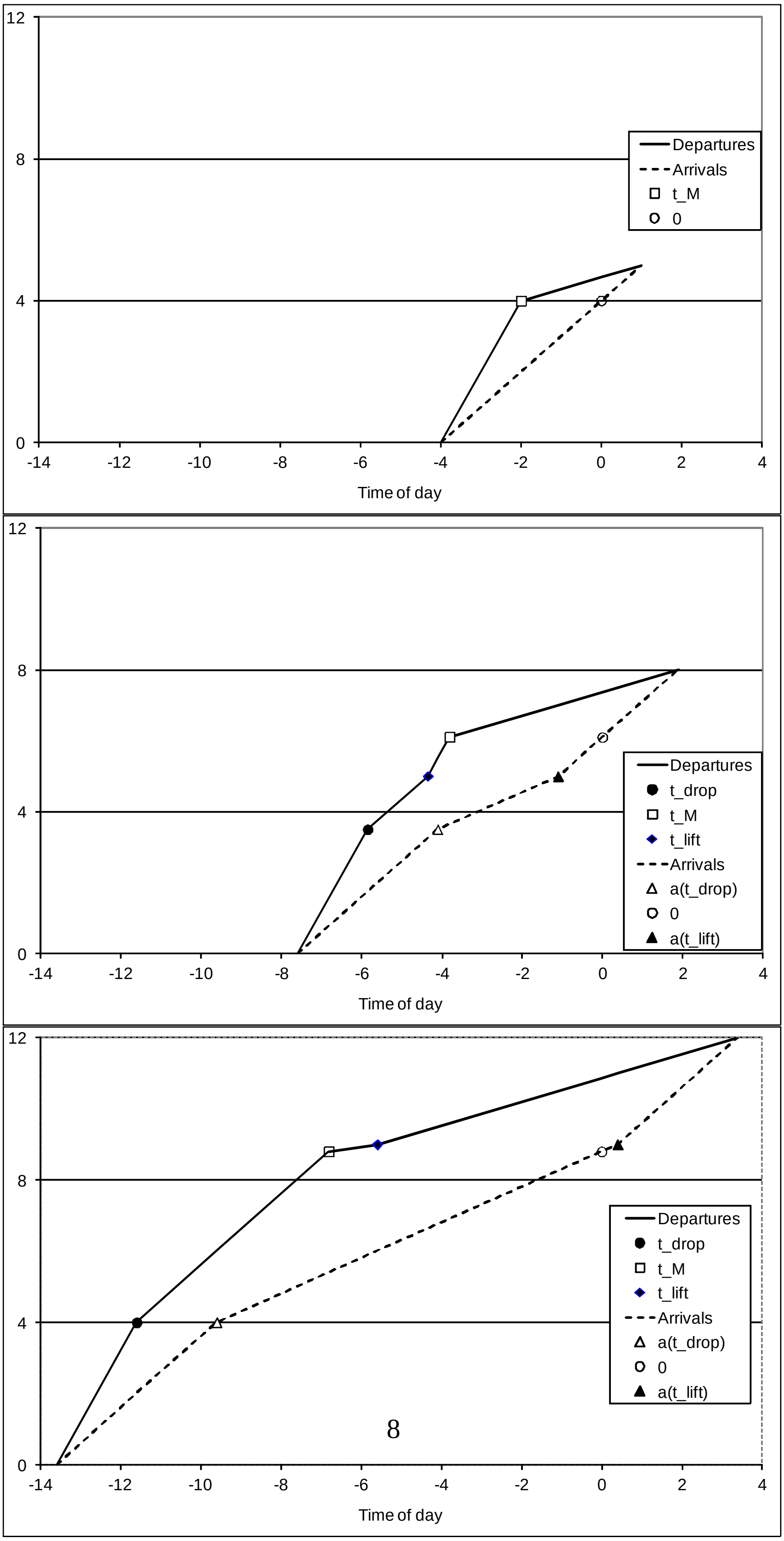

Figure 1: Cumulative departure and arrival patterns for various N. (a) Regime 1 (N=5); (b) Regime 2 (N=8); (c) Regime 3 (N=12)

3.2 Regime 2

When $N > N_1$, the lower capacity ψ_1 is activated over some non-zero time interval. Denote by $t_{drop} < t_M$ the first departure time after which a traveler will be processed at the slower rate, i.e. the first time when $Q\left[a\left(t_{drop}\right)\right] = Q_0$. Similarly, denote by t_{lift} the next departure time after that for which a departing traveler will again be processed at the faster rate, meaning that $Q\left[a\left(t_{lift}\right)\right]$ is again Q_0. If N is not too large, so that capacity drops only for a brief period, then $t_{lift} < a\left(t_{drop}\right)$; this defines Regime 2, which is depicted in Figure 1b.

Whenever $N > N_1$, some travelers now have to account for the fact that the arrival rate will become slow before they complete their trips, and hence their equilibrium condition (3) involves the lower processing rate ψ_1. (These travelers are sandwiched between others whose marginal condition involves the higher rate.) In Regime 2, all of them arrive early. Thus the full set of departure rates is

$$\rho\left(t\right) = \begin{cases} a'_E\psi_0, & t \in \left[t_0, t_{drop}\right] \\ a'_E\psi_1, & t \in \left(t_{drop}, t_{lift}\right) \\ a'_E\psi_0, & t \in \left[t_{lift}, t_M\right] \\ a'_L\psi_0, & t \in \left(t_M, t_1\right]. \end{cases} \tag{5}$$

Note that there are only three distinct rates here, since one rate occurs twice. Note also that the first and last departure rates are identical to those of Regime 1.

The arrival rate, $A'\left(t\right)$, has just two values, ψ_0 and ψ_1, the slopes of the lower curve in the figure. The first (high) value prevails until the queue builds to value Q_0 and again after it has fallen back to Q_0; the second (low) value prevails in between. The times when these kinks occur can be derived geometrically (relative to t_0) from the diagram, since we know the slopes ρ and A' as just described; these times are given in the Appendix, along with the corresponding numbers of travelers Q in the system at each time. Finally, time t_0 is determined from the condition that all travelers are accommodated, i.e. that traveler N departs at time t_1. As shown in the appendix, the result is

$$\begin{aligned} t_0 &= -\frac{\delta N}{\beta\psi_{01}} + \frac{Q_0}{\psi_0}\frac{\alpha\Delta\psi}{\left(\alpha-\beta\right)\Delta\psi + \beta\psi_1} \\ &= -\frac{\delta N}{\beta\psi_{01}} + \left(\frac{1}{\psi_{01}} - \frac{a'_e\psi_1/\Delta\psi}{1 + a'_e\psi_1/\Delta}\right)Q_0 \end{aligned} \tag{6}$$

where $\Delta\psi \equiv \psi_0 - \psi_1$ and

$$\psi_{01} \equiv \psi_0\frac{\left(\alpha-\beta\right)\Delta\psi + \beta\psi_1}{\alpha\Delta\psi + \beta\psi_1} \tag{7}$$

can be interpreted as an intermediate capacity, lying between ψ_1 and ψ_0.

It is straightforward to verify that t_0 in Regime 2 occurs earlier than in Regime 1. The marginal external cost $\delta N/\psi_{01}$ is larger than in Regime 1, and smaller than it would be if capacity were always equal to ψ_1 (namely $\delta N/\psi_1$); this observation is important in the comparison with Regime 3.

3.3 Regime 3

As N grows, the time of reduced capacity lasts longer, and for large enough N we will find that $t_{lift} > a\left(t_{drop}\right)$. This condition defines Regime 3. It implies that some travelers experience only the reduced capacity for their entire trip, since they arrive before the time when capacity goes back to its higher level. We show in the Appendix that this occurs when N exceeds the critical value

$$N_2 \equiv N_1 + \left(\frac{\alpha-\beta}{\beta}\cdot\frac{\Delta\psi}{\psi_1}+1\right)Q_0 \tag{8}$$

which is always greater than N_1. There are two possibilities: if $t_{lift} < t_M$ (Regime 3a), then (5) again applies, with just three distinct departure rates. Otherwise (Regime 3b), all four possible departure rates occur:

$$\rho\left(t\right) = \begin{cases} a'_E\psi_0, & t\in\left[t_0,t_{drop}\right] \\ a'_E\psi_1, & t\in\left(t_{drop},t_M\right] \\ a'_L\psi_1, & t\in\left(t_M,t_{lift}\right) \\ a'_L\psi_0, & t\in\left[t_{lift},t_1\right]. \end{cases} \tag{9}$$

Regime 3b is shown in Figure 1c.

The appendix also shows that in regime 3, the first departure time t_0 is

$$t_0 = -\frac{\delta N}{\beta\psi_1}\cdot\left[1-\left(\frac{\Delta\psi}{\psi_0}\frac{\alpha+\gamma}{\gamma}+\frac{\Delta\psi}{\psi_1}\frac{\alpha-\beta}{\beta}\right)\frac{Q_0}{N}\right]. \tag{10}$$

The factor in square brackets can be shown to lie between 0 and 1; thus the first departure time is again later than would be the case if capacity were ψ_1 throughout the entire period, which would be $-\left(\delta/\beta\right)N/\psi_1$.

3.4 Implications of unregulated equilibrium

There are two notable features of the unregulated equilibrium in our model.

3.4.1 Time stretching

First, the effect of activating the lower capacity is to stretch the peak period, with the beginning and ending looking just like they did before except taking place further from the desired arrival time. This is more easily seen by holding N constant while decreasing ψ_1, i.e. while exacerbating the decline in processing rate that occurs when vehicle density is high. Figure 2 shows the cumulative departure and arrival patterns for three such values of ψ_1 (with other parameters as in Table 1). As ψ_1 decreases, the delay times in the middle of the rush hour (the horizontal distance between the two curves) becomes larger, even though the queue length itself (the vertical distance) does not. To maintain equilibrium, the earliest and latest travelers suffer correspondingly higher scheduling costs.

Figure 3 zooms in on the early rush hour, namely the departures and arrivals of the first $2Q_0$ travelers. The patterns are nearly identical but displaced to earlier times as N increases. Similarly, we can see from Figure 2 that the patterns for the last $1.5Q_0$ travelers are unaffected by ψ except for being displaced to slightly later times when ψ is smaller.

3.4.2 Race to the bottom

The second notable feature is that at the margin, the cost of adding a traveler to the system is governed ultimately by the lowest capacity. By "ultimately" we mean once the transitional Regime 2 has been passed through. As seen from the above expressions for t_0, the average cost of a traveler, $ac \equiv -\beta t_0$, is a linear function of N within any given regime, whose slope increases from one regime to the next:

$$ac\left(N\right) = -\beta t_0 = \begin{cases} \frac{\delta}{\psi_0}N, & N \leq N_1 \\ ac\left(N_1\right) + \frac{\delta}{\psi_{01}}\left(N - N_1\right), & N_1 < N \leq N_2 \\ ac\left(N_2\right) + \frac{\delta}{\psi_1}\left(N - N_2\right), & N_2 \ < N. \end{cases} \tag{11}$$

(Recall we assume throughout that N is smaller than the value that would cause the queue to reach jam density.)

The marginal external congestion cost ($mecc$), i.e. marginal cost less average cost, is determined solely by the term that is linear in N,

$$mecc = N\frac{d\left[ac\left(N\right)\right]}{dN} = \frac{\delta N}{\psi_k} \tag{12}$$

where $k = 0, 01, 1$ varies by regime. This is just like the standard bottleneck model except with capacity replaced by ψ_k. We note in passing that if demand has nonzero elasticity and the only policy tool available is a uniform toll, its optimal level is $mecc$.

The average and marginal cost functions implied by 11 are illustrated in Figure 4 for the case $\psi_1 = 1/3$. The figure also shows what they would be if there were no capacity reduction. Note that average and marginal costs increase nonlinearly; average cost is convex and marginal cost is discontinuous at N_1 and N_2. This property also characterizes many static models, but not the dynamic model

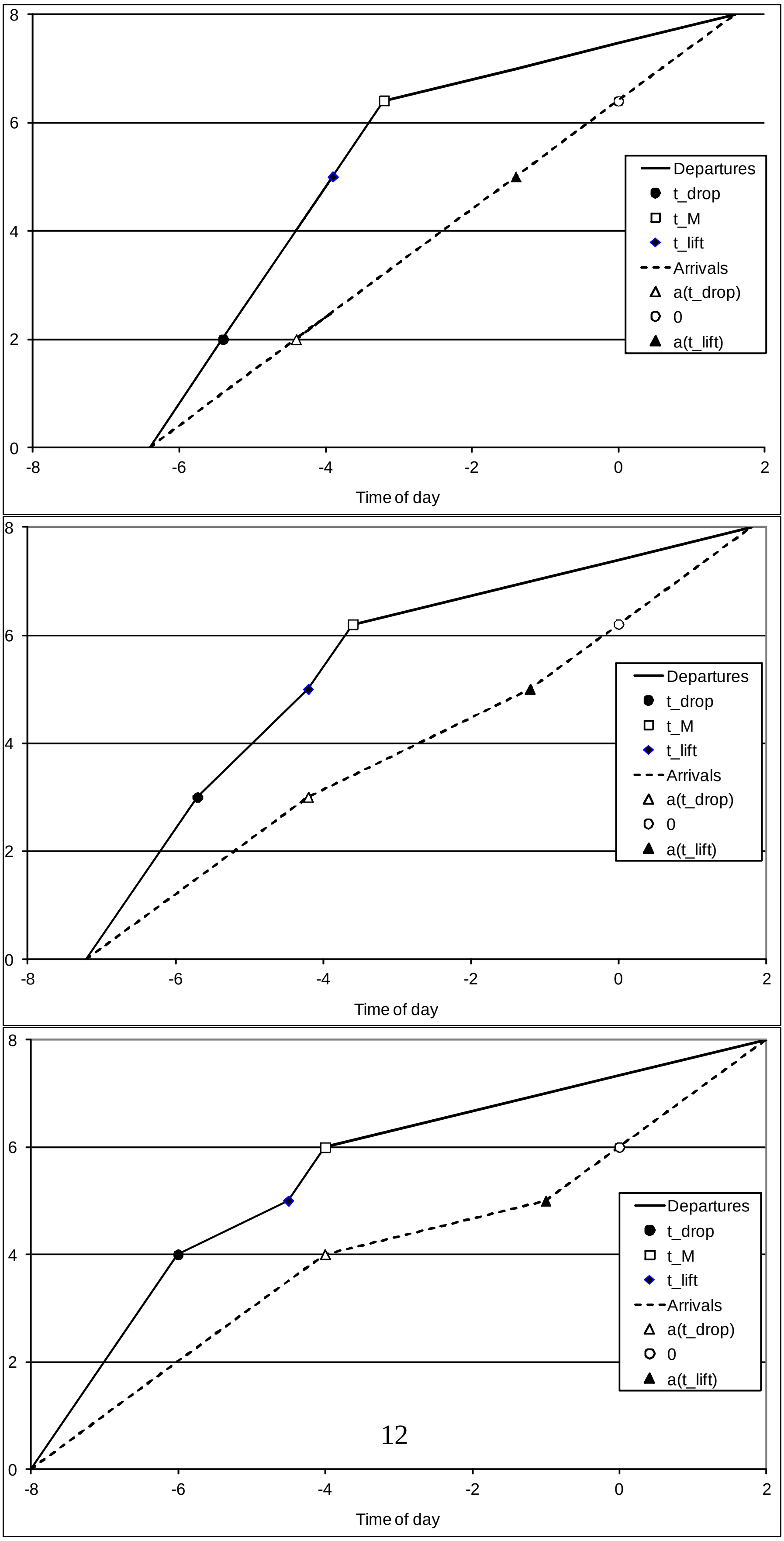

Figure 2: Cumulative departure and arrival patterns for Regime 2, $N = 8$: (a) $\psi_1 = 1$; (b) $\psi_1 = 2/3$; (c) $\psi_1 = 1/3$

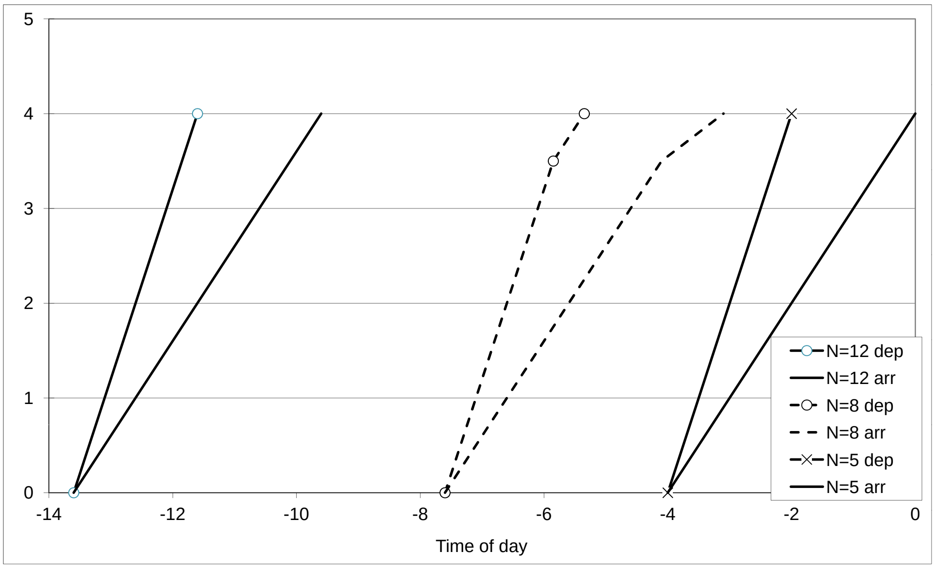

Figure 3: Cumulative departure and arrival patterns for earliest travelers at various values of N.

most commonly used in economic analysis of congestion, which is the standard bottleneck model (our Regime 1). Thus, our model gives a dynamic justification for the common assertion — typically based on a static model — that congestion becomes especially sensitive to traffic levels under highly congested conditions.

We conjecture that this conclusion is robust with respect to the inclusion of many capacity levels. More specifically, we conjecture that if the model were extended to include many possible values of successively lower capacity, then average cost would still be a convex function such that marginal external congestion cost would be increasing; the average cost would depend in a complex way on the whole range of activated capacities.

4 Policies

Hypercongestion can be reduced or eliminated by at least two types of policies. One is pricing, designed to reduce departure rates enough to keep the maximum queue length below the critical value Q_0. The other is metering, designed to move the queue outside of the region where it produces hypercongestion. For example, if the model represents areawide congestion within a central business district, vehicles might be allowed into that district at a reduced rate, with the resulting queues regulated in such a way that they do not interfere with any moving traffic.

Of the many possible policies of these types we consider the three shown in

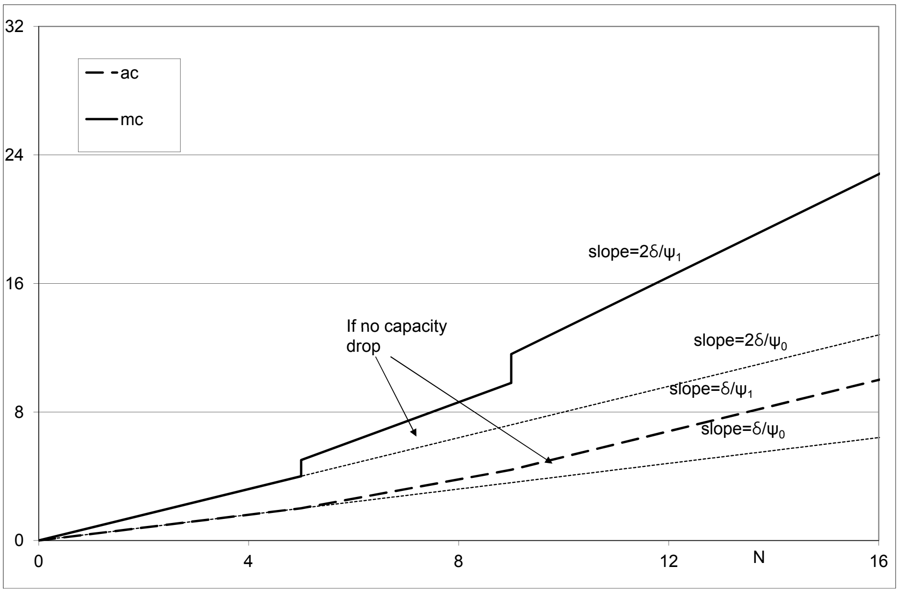

Figure 4: Average and marginal cost functions for $\psi_1 = 1/3$

Table 2: Policies considered

Policy objective	Tolling	Metering
No queue	Optimal dynamic toll	
No hypercongestion	Coarse toll to remove hypercongestion	Metering to remove hypercongestion

Table 2. The optimal dynamic toll addresses all costs of queuing while the other two (a coarse toll and metering) are aimed mainly at eliminating hypercongestion.

4.1 Optimal dynamic toll

The dynamic toll that produces the lowest average cost (and thereby maximizes welfare given perfectly inelastic demand) has already been described by ADL (1990). We know this because the same reasoning applies here as in the Vickrey model. The optimal toll must eliminate all queuing, which is an unnecessary cost to the system, while allowing travelers to fully utilize the capacity of the bottleneck in order to minimize their aggregate user costs. Since the total number of travelers is fixed, it is only the time pattern of the toll that matters: it can be increased or decreased uniformly without affecting the resulting equilibrium. For concreteness, and following convention, we assume the toll is zero for the first and last travelers. We also assume that each traveler cares only about the "inclusive price", defined as user cost plus toll.

This toll and the resulting departure and arrival patterns (which are identical) are well known. The toll is time-varying and replaces exactly the queuing cost that would have occurred in a hypothetical unregulated equilibrium with capacity ψ_0. Average revenue per traveler is $(1/2)\,\delta N/\psi_0$. Travelers have the same scheduling costs as they would in that equilibrium, but the corresponding hypothetical queuing costs are replaced by the cost (to them) of tolling. Thus, relative to that hypothetical high-capacity equilibrium, travelers perceive the same total price with or without the toll policy in place. But of course toll payments are now balanced by revenues, which are counted as a benefit.

Thus, there are two resulting sources of gain from this policy. First, capacity can be maintained at the high level ψ_0. This enables all travelers to be accommodated during a shorter time interval, having duration N/ψ_0. Second, even relative to a bottleneck with that higher capacity ψ_0, all queuing cost is eliminated. That cost is exactly half the total cost in the hypothetical unregulated equilibrium, or $(1/2)\,\delta N/\psi_0$ per traveler. The two cost savings combined are found by subtracting the average cost with optimal tolling, $ac(tolled) = (1/2)\,\delta N/\psi_0$, from that given by (11):

$$
\begin{aligned}
&ac(unregulated) - ac(tolled) \qquad (13)\\
&= \begin{cases}
\frac{\delta}{2\psi_0}N, & N \le N_1 \\
\left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N - N_1) + \frac{\delta}{2\psi_0}N, & N_1 < N < N_2 \\
\left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N_2 - N_1) + \frac{\delta\Delta\psi}{\psi_0\psi_1}(N - N_2) + \frac{\delta}{2\psi_0}N, & N_2 \le N.
\end{cases}
\end{aligned}
$$

As noted earlier, the average revenue from the optimal toll with fixed capacity ψ_0 is equal to the last term in each of these expressions. Hence, the expressions show how welfare gain is divided between travelers and the recipient of revenues. The fraction of gain accruing to travelers is zero for $N \le N_1$ and it increases as N increases. The possibility of having substantial gains accrue directly to travelers contrasts sharply with the standard bottleneck model and also with the typical static model, which assumes a sharply convex relationship between travel time and flow.

4.2 Metering to remove hypercongestion

Suppose inflow to the system can be metered to remain below a certain rate, and the resulting queue can be held where it is not part of the queue length that determines capacity. Suppose further that FIFO applies in the metered queue. Then hypercongestion can be eliminated, either by limiting the metering rate to ψ_0 or less or at least by setting it so that the resulting queue within the system (the unmetered queue) never exceeds Q_0. Assume further that travelers view waiting in either the metered or unmetered queue identically, with cost per minute α. Then the system is converted into the equivalent of a standard bottleneck model whose bottleneck capacity is ψ_0 and whose queue length is the sum of the metered and unmetered queues. In the language of areawide congestion, the travel time is the sum of time waiting in the metered queue and time spent within the congested area (the latter represented in the model by time waiting in the unmetered queue).

If storage in the metered queue is costless, then it is optimal to assure that no hypercongestion occurs since, if it did, travel times and scheduling costs could be reduced by decreasing the metering rate. Thus we already have the solution to this policy, which is the departure pattern (4) originally presented as Regime 1, but now applying for any value of N. The welfare gain per traveler, relative to the unregulated equilibrium, is given by the difference in average costs, which is easily obtained from equations (11):

$$\begin{aligned} & ac(unregulated) - ac(metered) \\ = & \begin{cases} 0, & N \leq N_1 \\ \left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N - N_1), & N_1 < N \leq N_2 \\ \left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N_2 - N_1) + \frac{\delta \Delta\psi}{\psi_0 \psi_1}(N - N_2), & N_2 < N. \end{cases} \end{aligned} \tag{14}$$

The welfare gain given in (13), from the optimal dynamic toll, can thus be decomposed into two parts. The first part is that which results from the elimination of hypercongestion; this part is the same as the welfare gain (14) from optimal metering, and it accrues directly to travelers. The second part is $\delta N/(2\psi_0)$; it is due to eliminating the queue that remains even when capacity does not drop, and it accrues to the recipient of revenues. Thus, if travelers ignore the use of revenues, they are indifferent between optimal tolling and metering.

4.3 Coarse toll to remove hypercongestion

The optimal toll varies continuously over time and so one may seek a simpler toll policy. This section considers a so-called coarse toll, which is a toll that has a single level, here denoted τ, that is applied to arrivals (i.e. exits from the bottleneck) during some interval which we denote $[t_+, t_-]$. The welfare benefits of such a toll, for the case of constant capacity, have been described elsewhere (Arnott et al., 1990; Laih, 1994, 2004; Fosgerau, 2011). Here, we analyze the case of variable capacity.

In either case, the outcome of a coarse toll depends on assumptions regarding queueing technology. The issue is how to deal with the discontinuity that exists at time t_- when the toll is lifted, which may produce an equilibrium with massed departures. Arnott et al. (1990) allow such massed departures to occur and assume they are placed randomly in the queue position. A simpler approach, adopted by Fosgerau (2011), is to assume that travelers who will arrive after time t_- queue separately while those paying the toll are being preferentially processed, even though this violates FIFO. In other words, people who don't want to pay the toll can enter the system and queue up behind each other while waiting for the toll to be removed at t_-. (There is some realism to this idea: in Stockholm, drivers have been reported waiting for the toll to be reduced before entering the city, even while others enter and pay the toll.) Fosgerau (2011) shows that in the case of constant capacity, and with the toll level set within certain limits, the departure and arrival patterns of those not paying the toll will then be exactly as in the unregulated equilibrium, while those paying the toll will encounter zero travel times at times t_+ and t_- and so adopt a departure and arrival pattern that is just like they would if they were in a Regime 1 system with starting and ending times t_+ and t_-.

To achieve this solution, we set t_+ and t_- so that $-\beta t_+ = \gamma t_-$ and $-\beta t_+ + \tau = -\beta t_0 = N\delta/\psi_0$. That is, we set

$$t_+ = \frac{1}{\beta}\left(\tau - \frac{\delta N}{\psi_0}\right), \tag{15}$$

$$t_- = \frac{1}{\gamma}\left(\frac{\delta N}{\psi_0} - \tau\right) \tag{16}$$

Then in equilibrium the number of travelers paying the toll is $\psi_0\left(t_- - t_+\right) = N - (\tau\psi_0/\delta)$; they depart and arrive during $[t_+, t_-]$ and face inclusive price $-\beta t_0$, the same as for other travelers. This is the same inclusive price that is paid by all in unregulated equilibrium, which shows that just as with an optimal toll, travelers are indifferent between a coarse toll and an unregulated equilibrium if they ignore revenues.

Figure 5 depicts the equilibrium under such a coarse toll, indicating actual cumulative departures and arrivals (solid lines) as well as the intervening departures that would occur in an untolled equilibrium (dotted lines). Note that some of the untolled group depart before some of the latest of the tolled group, yet arrive after them, since the tolled group receives preferential treatment; both contribute to the queue simultaneously for a while, as shown at the top of the figure. These two groups of travelers are shown separately as two dashed line segments. The one for late untolled travelers is placed at the top of the figure to show how it fits into the graph that applies for an untolled equilibrium (the outer triangle including the dotted line). But the actual cumulative departure rate is the solid line segment con-

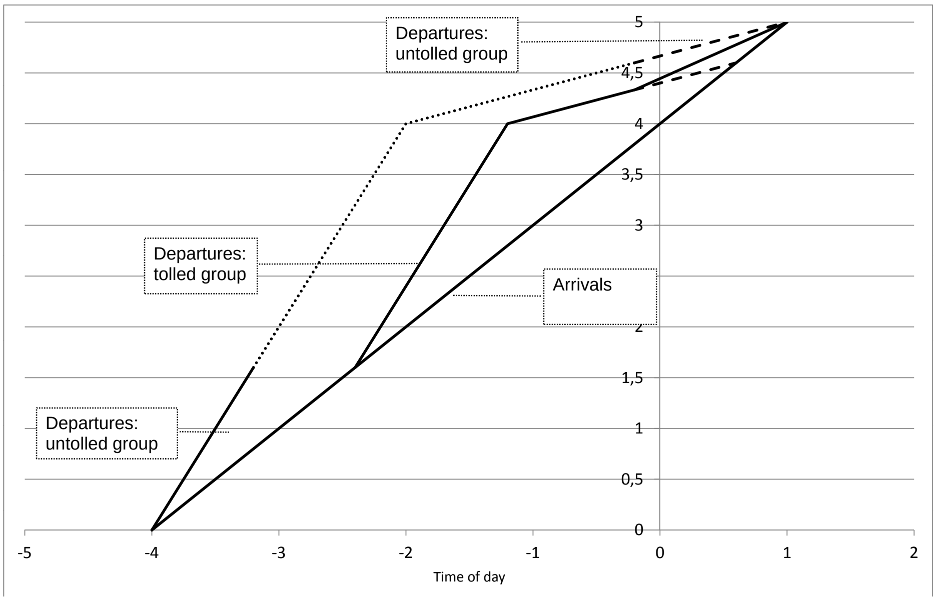

Figure 5: Coarse toll: cumulative departures and arrivals

necting these two dashed lines, which has slope equal to the sum of their slopes.

Since our solution eliminates hypercongestion, it has the same properties as the fixed-capacity coarse toll analyzed by Fosgerau (2011).[6] The toll level that maximizes welfare is also the one that maximizes revenues $\tau\psi_0\left(t_- - t_+\right) = \tau N - \tau^2\frac{\psi_0}{\delta}$, which is

$$\tau^* \equiv \frac{\delta N}{2\psi_0}. \tag{17}$$

In this case exactly half the travelers pay the toll, providing total revenue $\delta N^2/(4\psi_0)$ which is half that from an optimal toll. Average cost is $(3/4)\delta N/\psi_0$, which is midway between that from optimal tolling and that in the unregulated equilibrium. We show in the appendix that this solution is valid, in the sense that the maximal queue never exceeds Q_0, provided $N \leq 2N_1$. If $N > 2N_1$, it is not possible to eliminate hypercongestion using a coarse toll.

[6] There are other systems where capacity breakdowns may be avoided by giving some people the chance to pay for the ability to bypass service queues. One is internet service, where providers have proposed breaking "net neutrality" by giving preferred customers faster processing times. Another is electricity provision, where customers can voluntarily submit to peak pricing in lieu of being subject to blackouts. The analogy holds even if priority access is granted based on some criterion other than paying a toll: as shown by Fosgerau (2011), creating an unpriced but restricted express lane ("fast lane") can replicate the patterns of departures and queuing delays produced by a coarse toll.

The coarse toll provides a welfare gain consisting of two parts: that from removing hypercongestion, which we already calculated as (14), and that from further reduced queuing, which is the same as revenues. Thus total welfare gain per traveler is

$$ac(unregulated) - ac(coarse\ toll) \tag{18}$$
$$= \begin{cases} \frac{\delta N}{4\psi_0}, & N \leq N_1 \\ \left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N - N_1) + \frac{\delta N}{4\psi_0}, & N_1 < N \leq \min\{N_2, 2N_1\} \\ \left(\frac{\delta}{\psi_{01}} - \frac{\delta}{\psi_0}\right)(N_2 - N_1) + \frac{\delta \Delta\psi}{\psi_0\psi_1}(N - N_2) + \frac{\delta}{4\psi_0}N, & N_2 < N \leq 2N_1 \end{cases}$$

Equivalently, the welfare gain per traveler from replacing optimal metering by the optimal coarse toll is $\delta N/(4\psi_0)$, again provided $N \leq 2N_1$.

5 Elastic demand

Our analysis thus far has taken total traffic volume N to be exogenous. As a consequence, there is an indeterminacy in the optimal toll: only its pattern matters, not its absolute level. However, if demand is a less than perfectly inelastic function of inclusive price (average cost plus toll), this indeterminacy is removed as the toll level now controls the total traffic volume. We already know from Arnott et al. (1993) that the bottleneck model leads to a very neat division of the toll into a component related to timing of departures (the same as derived earlier) and one related to the total amount of traffic. This latter component is a constant equal to the marginal external cost of congestion, $mecc$, defined as the derivative of the average cost when viewed as a function of N. (Other external costs, such as from air pollution, can be added to $mecc$.)

Regardless of what policy we consider, however, a new feature arises in our model because these policies reduce average cost more than in the constant-capacity model: specifically, by (14) due to eliminating hypercongestion. This affects the inclusive price resulting from a given scenario. This will tend to increase total traffic volume, analogously to the "rebound effect" from implementing energy efficiency regulation: by making operation cheaper, increased use is attracted.

For each of the policies described here, we can calculate the conditions under which total traffic will increase due to the policy. This occurs whenever the extra monetary cost imposed by a toll level is less than (14). For the metering policy, there is no monetary cost so this condition always holds: metering will increase total traffic. For the optimal toll including $mecc = (1/2)\delta/\psi_0$, total traffic will increase if $mecc$ is less than (14).

For the coarse toll, the situation is more complicated because the toll is applied only to part of the peak period; if its level is raised in order to suppress elastic demand, the time pattern will also be distorted. Rather than solve that rather messy problem, we consider adding a uniform toll (covering all times) on top of the coarse toll. We have already seen that for this policy, $ac\,(N) = (3/4)\delta N/\psi_0$ and therefore $mecc \equiv d\,[ac\,(N)]\,/dN = (3/4)\delta/\psi_0$. Thus traffic is increased whenever this value is less than (14). Note that the greater N, the more likely traffic will be increased by either an optimal or a coarse toll, since (14) rises with N whereas $mecc$ does not.

6 Simulation study

In order to determine how our model behaves when there are more than two levels of non-zero capacity, we use numerical simulation. We have implemented a simulation model using the values $(\alpha, \beta, \gamma) = (1, 2, 4)$. Capacity is a decreasing step function of the queue length, starting at $\psi_0 = 5$ and declining by half at each queue length in the set $\{2000, 4000, 6000, ...\}$.

A simulation run begins with an arbitrarily chosen time of the first departure, t_0. At each iteration, it uses the queue pattern from the preceding iteration to compute the cumulative departures R, then recalculates the queue dynamically to determine the cumulative exit rate A at each succeeding point in time. Convergence is obtained when an iteration step does not change the queue length significantly at any time. The value of t solving $R\,(t) = A\,(t)$ is then found numerically, and the corresponding value of R is taken to be the amount of total traffic N consistent with the chosen t_0.

Figure 6 shows the cumulative departures and arrivals for a single run with departures starting at time -5000. The simulation finds the corresponding total traffic volume to be 17248. The figure clearly shows how kinks in the departure rate are related to kinks in the arrival rate at the corresponding time of arrival, as well as to the change from early to late arrival.

By carrying out such simulations runs for many values of t_0, the model produces a relationship between N and t_0. Figure 7 presents some statistics from a series of such runs with N varying from near zero to about 22800. (The first departure time ranges between zero and -8000.) The upper panel plots the maximal queue length against total traffic N. Kinks are evident, due not only to capacity thresholds but also to transitions to new regimes, such as the beginning of Regime 3 at about N=11200. The lower panel plots the resulting average cost as a function of N; this curve is convex and we conjecture that convexity is inevitable.

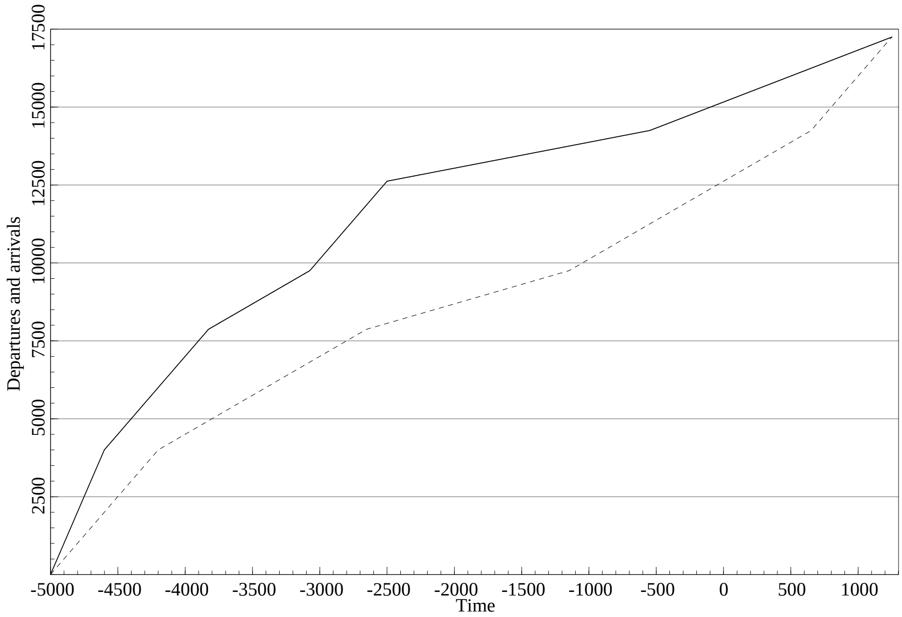

Figure 6: Simulation run: cumulative departures and arrivals given $t_0 = -5000$

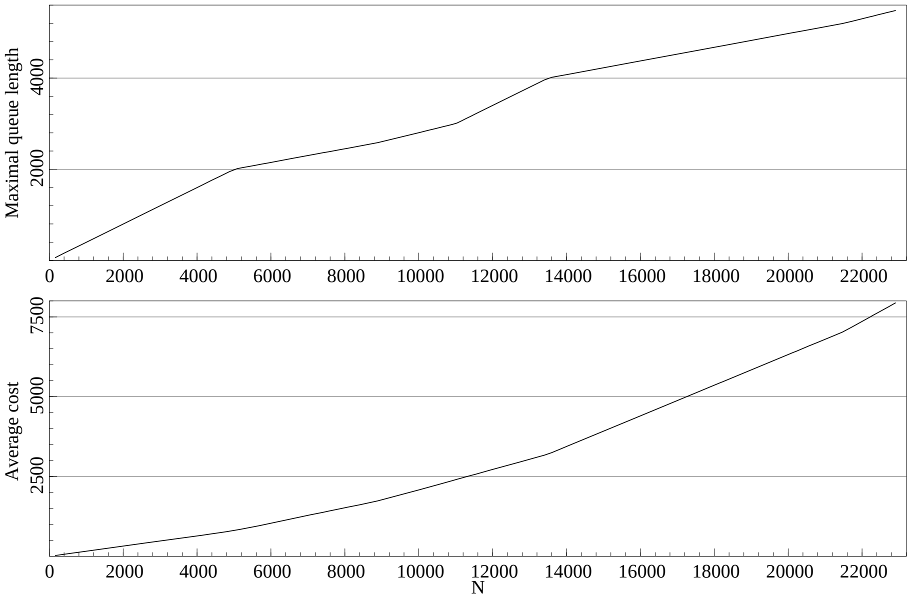

Figure 7: Simulation runs: maximal queue length (upper panel) and average cost (lower panel) as functions of N

7 Concluding remarks

In this paper we have formulated an analytically tractable model of hypercongestion that is consistent with some new results from traffic engineering. The model relies of course on some stringent assumptions. Probably the most important such assumptions are that travelers are homogeneous in preferences and travel distances; that the system is deterministic, with no uncertainty arising from accidents or other sources of unexpected congestion; and that capacity depends on network occupancy through a simple step function. Nevertheless, the model leads to several insights that appear to depend on general features that would likely remain even if those assumptions were relaxed.

First, the average cost for travelers in unregulated equilibrium is increasing and convex as a function of traffic volume. The slope of the average cost curve is related to the lowest capacity activated, but the relationship is not simple due to the forces of equilibrium that partly compensate for lower capacity by reducing the departure rate.

Second, there are two sources of benefit associated with tolling: the familiar source from the Vickrey bottleneck model, namely that queueing can be reduced without affecting travelers' inclusive price; and a new source due to eliminating or reducing the time during which capacity is reduced. This new source also applies to a metering policy, which has not previously been analyzed in the context of a single bottleneck. It can be very large if the unregulated equilibrium involves severe hypercongestion.

Third, in contrast to the Vickrey bottleneck model, this model allows analysis of metering as a policy instrument even for a single bottleneck. Metering can be used to avoid the reduction of capacity, thereby achieving the second of the benefits associated with tolling. Importantly, the benefit accrues directly to travelers; there is no intervening step in which revenue has to be collected and used beneficially.

Fourth, like the ideal time-varying toll, a coarse toll (one that varies in discrete steps) also leads to the benefits both from queue reduction and from avoided capacity reduction. Furthermore, application of such a toll enables a substantially larger traffic volume to be sustained without activating capacity reduction.

Our model opens further possibilities for analyzing new classes of policies that could not be analyzed with such explicit attention to scheduling. One, already discussed, is metering. Another consists of measures designed to change the capacity function, for example increased enforcement of regulations against blocking intersections and increased storage space on turn lanes to prevent spillbacks. These latter policies are frequently included in congestion management strategies, and our model offers a way to analyze their systemic effects by changing the queue threshold Q_0 at which the lower capacity is activated.

Tractability is the main challenge for models that deal with hypercongestion, because the travel time of one traveler is determined by the decisions of other travelers throughout the duration of the trip. Thus, generalizing to situation with more general scheduling preferences, heterogeneous travelers, travel-time uncertainty, and more realistic capacity functions will be difficult. It may be such generalizations will have to rely on simulation rather than on analytical results. It is also possible that more radical deviations from the current model framework may turn out to be fruitful. For example, one could consider equilibrium concepts other than Nash equilibrium, or abandon the first-in-first-out principle.

What seems certain is that hypercongestion is a significant phenomenon at the macroscopic level in real cities. Acknowledging and better understanding hypercongestion are fundamental to the assessment of policies that address congestion.

References

Arnott, R. (2011) A Bathtub Model of Traffic Congestion *Working Paper* .

Arnott, R., de Palma, A. and Lindsey, R. (1990) Economics of a bottleneck *Journal of Urban Economics* **27**(1), 111–130.

Arnott, R., de Palma, A. and Lindsey, R. (1993) A structural model of peak-period congestion: A traffic bottleneck with elastic demand *American Economic Review* **83**(1), 161–179.

Arnott, R., de Palma, A. and Lindsey, R. (1998) Recent developments in the bottleneck model *in* K. Button and E. T. Verhoef (eds), *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar Cheltenham, UK pp. 79–110.

Chu, X. (1995) Alternative congestion pricing schedules *Regional Science and Urban Economics* **29**(6), 697–722.

Daganzo, C. F. (2007) Urban gridlock: Macroscopic modeling and mitigation approaches *Transportation Research Part B: Methodological* **41**(1), 49–62.

Fargier, P. (1983) Effects of the choice of departure time on road traffic congestion *in* V. Hurdle, E. Hauer and G. N. Steuart (eds), *Proceedings of the Eighth International Symposium on Transportation and Traffic Theory* University of Toronto Press Toronto pp. 223–262.

Fosgerau, M. (2011) How a fast lane may replace a congestion toll *Transportation Research Part B* **45**(6), 845–851.

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. M. (2009) Cordon pricing consistent with the physics of overcrowding *Proceedings of the 18th International Symposium on Transportation and Traffic theory* .

Gonzalez, E. J. and Daganzo, C. F. (2011) Morning Commute with Competing Modes and Distributed Demand: User Equilibrium, System Optimum, and Pricing *Kuhmo Nectar conference on transportation economics* Stockholm, Sweden.

Henderson, J. (1981) The economics of staggered work hours *Journal of Urban Economics* **9**, 349–364.

Laih, C.-H. (1994) Queueing at a bottleneck with single- and multi-step tolls *Transportation Research Part A* **28**(3), 197–208.

Laih, C. H. (2004) Effects of the optimal step toll scheme on equilibrium commuter behaviour *Applied Economics* **36**(1), 59–81.

Mahmassani, H. and Herman, R. (1984) Dynamic User Equilibrium Departure Time and Route Choice on Idealized Traffic Arterials *Transportation Science* **18**(4), 362–384.

Shen, W. and Zhang, H. (2010) Pareto-improving ramp metering strategies for reducing congestion in the morning commute *Transportation Research Part A* **44**(9), 676–696.

Small, K. (1982) The scheduling of Consumer Activities: Work Trips *American Economic Review* **72**(3), 467–479.

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

Vickrey, W. (1969) Congestion theory and transport investment *American Economic Review* **59**(2), 251–261.

Walters, A. A. (1961) The Theory and Measurement of Private and Social Cost of Highway Congestion *Econometrica* **29**(4), 676–699.

Yang, H. and Huang, H. J. (1997) Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory *Transportation Research Part B: Methodological* **31**(6), 425–440.

Zhang, X., Zhang, H. M. and Li, L. (2010) Analysis of User Equilibrium Traffic Patterns on Bottlenecks with Time-Varying Capacities and their Applications *International Journal of Sustainable Transportation* **4**(1), 56–74.

A Unregulated solution: Regimes 2 and 3

As shown in the text, the queue will reach length Q_0 if $N \geq N_1$. This appendix establishes some properties of equilibrium in this case.

In these regimes, times t_M, t_{drop} and t_{lift} are defined by:

$$\begin{aligned} a\left(t_M\right) &= 0 \\ a\left(t_{drop}\right) &= \inf\left\{t | Q\left(t\right) \geq Q_0\right\} \\ a\left(t_{lift}\right) &= \inf\left\{t | t > a\left(t_{drop}\right), Q\left(t\right) < Q_0\right\}. \end{aligned}$$

We begin by establishing some general properties of equilibrium.

Lemma 1 *Suppose $N > N_1$. Then: (a) Capacity drops before and lifts after the departure that arrives just on time: i.e. $a\left(t_{drop}\right) \leq t_M \leq a\left(t_{lift}\right)$; and (b) The times when the queue exceeds the threshold Q_0 form a single continuous interval.*

Proof. (a) Considering the first inequality of part (a), suppose otherwise that there is a t such that $t_M < t < a\left(t_{drop}\right)$. The t can be chosen such that the queue is increasing at time t, since the queue reaches Q_0 for the first time at $a\left(t_{drop}\right)$. But $a'\left(t\right) < 1$ since $t > t_M$. Hence $Q'\left(t\right) = \left[a'\left(t\right) - 1\right]\psi_0 < 0$, which is a contradiction.

Considering the second inequality, suppose otherwise that there is a t such that $a\left(t_{lift}\right) < t < t_M$. This t can be chosen sufficiently close to $a\left(t_{lift}\right)$ that the queue is decreasing at time t. But $a'\left(t\right) > 1$ since $t < t_M$. Hence $Q'\left(t\right) = \left[a'\left(t\right) - 1\right]\psi_0 > 0$, a contradiction.

(b) From consumer equilibrium, the queue must be rising for arrivals at times $a\left(t\right) < t_M$, and falling for times $a\left(t\right) > t_M$. From (a), this means the queue is first rising then falling between times $a\left(t_{drop}\right)$ and $a\left(t_{lift}\right)$, so remains above Q_0 throughout this interval. Also, it is rising before $a\left(t_{drop}\right)$ and falling after $a\left(t_{lift}\right)$, so cannot be above Q_0 at any other time. ■

Note that t_{lift} may occur either before or after $a\left(t_{drop}\right)$. These possibilities distinguish regimes 2 and 3. Finding the times that characterize these two regimes requires solving some equations. We consider regime 2 first.

A.1 Regime 2

Recalling Lemma 1, regime 2 is defined by the following ordering of times.

$$t_0 < t_{drop} < t_{lift} < a\left(t_{drop}\right) < t_M < a\left(t_{lift}\right) < -\frac{\beta}{\gamma}t_0 \equiv t_1.$$

This order is sufficient to identify the changes in R and Q that occur in intervals between these points.

We can find $a\left(t_{lift}\right)$ in terms of t_0 as follows. During the interval $\left[a\left(t_{lift}\right), t_1\right]$ a number $\left[t_1 - a\left(t_{lift}\right)\right] a'_L \psi_0$ of travelers depart, during which time the queue changes by $\left[t_1 - a\left(t_{lift}\right)\right]\left(a'_L - 1\right)\psi_0$. This change must equal $-Q_0$ in order that the queue be reduced from Q_0 to zero. Therefore, using $t_1 = -\left(\beta/\gamma\right) t_0$:

$$\begin{aligned} -Q_0 &= \left[t_1 - a\left(t_{lift}\right)\right]\left(a'_L - 1\right)\psi_0 \\ \left[t_1 - a\left(t_{lift}\right)\right] &= \frac{\alpha+\gamma}{\gamma} \cdot \frac{Q_0}{\psi_0} \\ a\left(t_{lift}\right) &= -\frac{\beta}{\gamma} t_0 - \frac{\alpha+\gamma}{\gamma} \cdot \frac{Q_0}{\psi_0}. \end{aligned} \tag{19}$$

Next we may use the condition that the queue length is identical at $a\left(t_{drop}\right)$ and $a\left(t_{lift}\right)$ in order to find $a\left(t_{drop}\right)$. We do this by breaking this time interval into its two parts, each with its own departure rate and and an arrival rate equal to ψ_1, and setting the cumulative change in queue equal to zero:

$$\begin{aligned} 0 &= \left[t_M - a\left(t_{drop}\right)\right]\left(a'_E \psi_0 - \psi_1\right) + \left[a\left(t_{lift}\right) - t_M\right]\left(a'_L \psi_0 - \psi_1\right) \\ &= \left(a'_E - a'_L\right)\psi_0 \frac{\beta}{\alpha} t_0 - \left(a'_E \psi_0 - \psi_1\right) a\left(t_{drop}\right) + \left(a'_L \psi_0 - \psi_1\right) a\left(t_{lift}\right). \end{aligned}$$

Solving, substituting (19), and simplifying yields:

$$\begin{aligned} a\left(t_{drop}\right) &= \frac{a'_E - a'_L}{a'_E \psi_0 - \psi_1} \psi_0 \frac{\beta}{\alpha} t_0 + \frac{a'_L \psi_0 - \psi_1}{a'_E \psi_0 - \psi_1} a\left(t_{lift}\right) \\ &= \left[\frac{a'_E - a'_L}{a'_E \psi_0 - \psi_1} \psi_0 \frac{\beta}{\alpha} - \frac{a'_L \psi_0 - \psi_1}{a'_E \psi_0 - \psi_1} \frac{\beta}{\gamma}\right] t_0 - \frac{a'_L \psi_0 - \psi_1}{a'_E \psi_0 - \psi_1} \cdot \frac{\alpha+\gamma}{\gamma} \cdot \frac{Q_0}{\psi_0} \\ &= \left[\frac{\psi_0 - \frac{\alpha-\beta}{\alpha+\gamma}\psi_0}{\alpha\Delta\psi + \beta\psi_1} \beta - \frac{\frac{\alpha-\beta}{\alpha+\gamma}\psi_0 - \frac{\alpha-\beta}{\alpha}\psi_1}{\alpha\Delta\psi + \beta\psi_1} \frac{\alpha\beta}{\gamma}\right] t_0 - \frac{\frac{\alpha-\beta}{\alpha+\gamma}\psi_0 - \frac{\alpha-\beta}{\alpha}\psi_1}{\alpha\Delta\psi + \beta\psi_1} \cdot \frac{\alpha\left(\alpha+\gamma\right)}{\gamma} \cdot \frac{Q_0}{\psi_0} \\ &= \frac{\beta}{\gamma} \cdot \frac{\left(\beta-\alpha\right)\Delta\psi + \gamma\psi_0}{\alpha\Delta\psi + \beta\psi_1} t_0 - \frac{\alpha-\beta}{\gamma} \cdot \frac{\alpha\Delta\psi - \gamma\psi_1}{\alpha\Delta\psi + \beta\psi_1} \cdot \frac{Q_0}{\psi_0} \end{aligned} \tag{20}$$

where $\Delta\psi \equiv \psi_0 - \psi_1$.

Using these results, we can identify the relationship between N and t_0 by noting that arrival rate $A\left(t\right)$, which has kinks at $a\left(t_{drop}\right)$ and $a\left(t_{lift}\right)$, must integrate to N:

$$\begin{aligned}
N &= \left(a\left(t_{drop}\right)-t_0\right)\psi_0+\left(a\left(t_{lift}\right)-a\left(t_{drop}\right)\right)\psi_1+\left(t_1-a\left(t_{lift}\right)\right)\psi_0 \\
&= \left[a\left(t_{drop}\right)-a\left(t_{lift}\right)\right]\Delta\psi-\frac{\beta+\gamma}{\gamma}t_0\psi_0 \\
&= \frac{\beta}{\delta}\psi_0\left\{\frac{\beta\Delta\psi}{\alpha\Delta\psi+\beta\psi_1}-1\right\}t_0+\frac{\alpha\left(\beta+\gamma\right)}{\gamma}\frac{\Delta\psi}{\alpha\Delta\psi+\beta\psi_1}\cdot Q_0 \\
&= -\frac{\beta\psi_{01}}{\delta}\cdot t_0+\frac{\beta}{\delta}\frac{\alpha\Delta\psi}{\alpha\Delta\psi+\beta\psi_1}\cdot Q_0
\end{aligned}$$

This equation is solved for t_0 to yield equation (6) in the text.

Lemma 2 *The slope*

$$\psi_{01}\equiv\psi_0\frac{\left(\alpha-\beta\right)\Delta\psi+\beta\psi_1}{\alpha\Delta\psi+\beta\psi_1}$$

of average cost as a function of N in Regime 2 satisfies $\psi_1\leq\psi_{01}\leq\psi_0$.

Proof. The second inequality is obvious. The first inequality is equivalent to

$$\begin{aligned}
\psi_1 &< \psi_0\frac{\left(\alpha-\beta\right)\psi_0+\left(2\beta-\alpha\right)\psi_1}{\alpha\Delta\psi+\beta\psi_1} \\
&\Leftrightarrow \alpha\frac{\Delta\psi}{\psi_0}+\beta\frac{\psi_1}{\psi_0}<\left(\alpha-\beta\right)\frac{\psi_0}{\psi_1}+2\beta-\alpha \\
&\Leftrightarrow \alpha\left(2-\frac{\psi_1}{\psi_0}-\frac{\psi_0}{\psi_1}\right)<\beta\left(2-\frac{\psi_1}{\psi_0}-\frac{\psi_0}{\psi_1}\right)
\end{aligned}$$

But it can easily be verified that the function $f\left(x\right)=2-x-x^{-1}\leq 0$ for $x>0$ and that $f\left(x\right)=0$ only for $x=1$. If $\psi_1<\psi_0$, the first inequality is equivalent to $\alpha>\beta$, which we have assumed throughout. If $\psi_1=\psi_0$, the definition of ψ_{01} shows directly that $\psi_0=\psi_{01}$. ■

The time t_{lift} itself can be found from the fact that the number of departures during time interval $[t_{lift},t_1]$ is equal to the number of arrivals during $[a(t_{lift}),t_1]$. Breaking the first of these intervals into its two parts with different departure rates, this equality is written:

$$a_E'\psi_0\cdot\left[t_M-t_{lift}\right]+a_L'\psi_0\cdot\left[t_1-t_M\right]=\psi_0\cdot\left[t_1-a\left(t_{lift}\right)\right]$$

from which, using (19) along with earlier results for t_M and t_1:

$$\begin{aligned} t_{lift} &= t_M + \frac{a'_L}{a'_E} \cdot [t_1 - t_M] - \frac{1}{a'_E} \left[\frac{\alpha+\gamma}{\gamma} \cdot \frac{Q_0}{\psi_0} \right] \\ &= \frac{\beta}{\alpha} \left(1 - \frac{\alpha-\beta}{\gamma} \right) t_0 - \frac{(\alpha-\beta)(\alpha+\gamma)}{\alpha\gamma} \cdot \frac{Q_0}{\psi_0}. \end{aligned} \tag{21}$$

Note that this derivation does not depend on the relative positions of $a(t_{drop})$ and t_{lift}, so remains valid in Regime 3.

We can now demonstrate that the times $a(t_{drop})$ and t_{lift} approach each other as N increases, and equal each other at the critical value given in Section 3.3 of the text. Combining (21) and (20), we find from the coefficients of t_0 that:

$$\begin{aligned} \frac{\alpha\delta}{\beta^2} \cdot \frac{d}{dt_0} [t_{lift} - a(t_{drop})] &= 1 - \frac{\alpha\psi_0}{\alpha\Delta\psi + \beta\psi_1} \\ &= \frac{(\beta-\alpha)\psi_1}{\alpha\Delta\psi + \beta\psi_1} \\ &< 0. \end{aligned}$$

Given then $dt_0/dN < 0$ from (6), we therefore know that $d[t_{lift} - a(t_{drop})]/dN > 0$ during Regime 2.

We can furthermore find the threshold N_2 for which $t_{lift} = a(t_{drop})$, marking the boundary between Regimes 2 and 3, as the value for which $t_{lift} = a(t_{drop})$. Equating (21) and (20), and using (6) to eliminate t_0, leads to:

$$\frac{N_2 - N_1}{Q_0} = 1 + \frac{\alpha-\beta}{\beta} \frac{\Delta\psi}{\psi_1}$$

as stated in (8) in Section 3.3 of the text. The derivation is tedious but straightforward, so we did not include it here but it is available on request.

A.2 Regime 3

Regime 3 is characterized by the conditions that

$$\begin{aligned} t_0 &< t_{drop} < a(t_{drop}) < t_M < a(t_{lift}) < -\frac{\beta}{\gamma} t_0 \equiv t_1 \\ a(t_{drop}) &< t_{lift} \end{aligned}$$

The times t_{drop} and $a(t_{drop})$ are simpler to derive than in regime 2. They are determined by two conditions. First, the number of departures during interval

$[t_0, t_{drop}]$ (which occur at rate $a'_E\psi_0$) is equal to the number of arrivals during interval $[t_0, a(t_{drop})]$ (which occur at rate ψ_0). This yields:

$$a(t_{drop}) - t_0 = \frac{\alpha}{\alpha - \beta}(t_{drop} - t_0). \tag{22}$$

Second, the critical queue length Q_0, which by definition occurs at time $a(t_{drop})$, is equal to the number of departures between t_{drop} and $a(t_{drop})$. (This is because everyone already in the queue at time t_{drop} passes through it by time $a(t_{drop})$, by definition of the latter.) These departures occur at rate $a'_E\psi_1$. Therefore:

$$Q_0 = \frac{\alpha}{\alpha - \beta} \cdot [a(t_{drop}) - t_{drop}]\,\psi_1.$$

Combining with (22), we obtain

$$t_{drop} - t_0 = \frac{Q_0}{\psi_1}\frac{(\alpha - \beta)^2}{\alpha\beta} \tag{23}$$

and hence

$$a(t_{drop}) - t_0 = \frac{Q_0}{\psi_1}\frac{(\alpha - \beta)}{\beta}.$$

This enables us to see how $[t_{lift} - a(t_{drop})]$ varies with N for this regime, just as we did for Regime 2. Combining the above equation with (21), which as noted remains valid in this regime, we see from the coefficients of t_0 that:

$$\begin{aligned}\frac{d}{dt_0}[t_{lift} - a(t_{drop})] &= \frac{\beta}{\alpha}\left(1 - \frac{\alpha - \beta}{\gamma}\right) - 1 \\ &= \frac{(\beta + \gamma)(\beta - \alpha)}{\alpha\gamma} \\ &< 0.\end{aligned}$$

As with Regime 2, this implies that $[t_{lift} - a(t_{drop})]$ is increasing in N; since it begins at zero for $N = N_2$, it must be greater than zero for $N > N_2$, as stated in the text.

To determine the relationship between N and t_0, we combine (19) with (22) to obtain:

$$\begin{aligned}a(t_{lift}) - a(t_{drop}) &= (t_1 - t_0) - \frac{\alpha + \gamma}{\gamma} \cdot \frac{Q_0}{\psi_0} - \frac{Q_0}{\psi_1}\frac{(\alpha - \beta)}{\beta} \\ &= -\frac{\beta}{\delta}t_0 - \left(\frac{1}{\psi_0} \cdot \frac{\alpha + \gamma}{\gamma} + \frac{1}{\psi_1}\frac{(\alpha - \beta)}{\beta}\right)Q_0 \\ &= -\frac{\beta}{\delta}t_0 - \left(\frac{\alpha + \gamma}{\gamma} + \frac{\psi_0}{\psi_1}\frac{(\alpha - \beta)}{\beta}\right)\frac{Q_0}{\psi_0}\end{aligned}$$

We then integrate A' from t_0 to t_1, making use of the fact that (19) from Regime 2 applies also in this regime since it was derived from the arrival rate aftertime $a\left(t_{lift}\right)$. The result is:

$$\begin{aligned}
N &= \psi_0 \cdot \left[a\left(t_{drop}\right) - t_0\right] + \psi_1 \cdot \left[a\left(t_{lift}\right) - a\left(t_{drop}\right)\right] + \psi_0 \cdot \left[t_1 - a\left(t_{lift}\right)\right] \\
&= \frac{\psi_0}{\psi_1}\frac{(\alpha-\beta)}{\beta}Q_0 - \frac{\beta}{\delta}\psi_1 t_0 - \left(\frac{\alpha+\gamma}{\gamma}\frac{\psi_1}{\psi_0} + \frac{(\alpha-\beta)}{\beta}\right)Q_0 + \frac{\alpha+\gamma}{\gamma}Q_0 \\
&= -\frac{\beta}{\delta}\psi_1 t_0 + \left[\left(\frac{\psi_0}{\psi_1} - 1\right)\frac{\alpha-\beta}{\beta} + \left(1 - \frac{\psi_1}{\psi_0}\right)\frac{\alpha+\gamma}{\gamma}\right]Q_0 \\
&= -\frac{\beta}{\delta}\psi_1 t_0 + \frac{\Delta\psi}{\psi_0\psi_1}\left[\frac{(\alpha-\beta)}{\beta}\psi_0 + \frac{\alpha+\gamma}{\gamma}\psi_1\right]Q_0.
\end{aligned}$$

Solving for t_0 yields (10).

B Coarse toll

We need to determine when the system with coarse toll (17), applied during the optimal interval $[t_+, t_-]$ given by (15) and (16), satisfies the requirement that the queue never exceed Q_0.

First, consider the queue faced by travelers paying the toll. Travelers in the tolled group arriving exactly at the preferred arrival time experience the maximum queue in this group. To be in equilibrium with non-toll payers, whose average cost is $N\delta/\psi_0$, the time spent queueing must be worth $N\delta/\psi_0 - \tau^*$; that is, the queue duration must be $\left(N\delta/\psi_0 - \tau^*\right)/\alpha$. With a processing rate of ψ_0, the maximum queue length in the tolled group is then $Q_{toll}^{\max}\left(\tau^*\right) = \left(N\delta - \psi_0\tau^*\right)/\alpha = N\delta/(2\alpha)$.

Next, consider the early group of non-tolled travelers. They depart starting at time t_0 at a rate greater than capacity. Their queue is steadily increasing and reaches its maximum as they approach their last departure time, which is $a^{-1}\left(t_+\right)$, i.e., the departure time allowing them to arrive just at time t_+ when the toll kicks. To find $a^{-1}\left(t_+\right)$, we make use of three conditions:

1. Their departure rate is $\rho_E = a'_E\psi_0$, from (4), with $a'_E = \alpha/\left(\alpha-\beta\right)$;

2. The earliest departure is $t_0 = -N\delta/\left(\beta\psi_0\right)$, from the condition that the first traveler has cost $N\delta/\psi_0$, as stated above (15);

3. The total number of departures in the early group, $\rho_E \cdot \left[\left(a^{-1}\left(t_+\right) - t_0\right)\right]$, equals the total number of arrivals from this group, $\psi_0 \cdot \left(t_+ - t_0\right)$.

Using these three conditions and substituting (15) for t_+ and (17) for τ^*, we obtain a departure interval of duration

$$\Delta t_E \equiv a^{-1}\left(t_+\right) - t_0 = \frac{1}{a'_E \psi_0} \cdot \frac{\delta N}{2\beta}$$

Over the interval of departures, then, cumulative departures are

$$\rho_E \Delta t_E = \frac{\delta N}{2\beta}$$

while cumulative arrivals during that same time interval are

$$\psi_0 \Delta t_E = \frac{\delta N}{2\beta} \cdot \frac{1}{a'_E}.$$

The difference is the maximum queue:

$$Q^{\max}_{early}\left(\tau^*\right) = \frac{N\delta}{2\alpha},$$

showing that $Q^{\max}_{early}\left(\tau^*\right) = Q^{\max}_{toll}\left(\tau^*\right)$.

Consider now late travelers who pay no toll. They begin departures at some time t_0^{late} with corresponding arrival at time t_-. Both tolled and late untolled travelers depart at rate $a'_L \psi_0 \equiv \left[\alpha/\left(\alpha+\gamma\right)\right]\psi_0$ during this period. Thus the combined departure rate for tolled and late travelers is $\rho_L = \left[2\alpha/\left(\alpha+\gamma\right)\right]\psi_0$, hence their queue changes at rate $(\rho_L - \psi_0)/\psi_0 = \left[2\alpha/\left(\alpha+\gamma\right) - 1\right]\psi_0$. But $\alpha < \gamma$ by assumption, so this rate is negative, that is, the queue is dissipating.

So with the optimal coarse toll and capacity at ψ_0, the maximal queue is $N\delta/(2\alpha) = (1/2)Q_0\left(N/N_1\right)$, and the maximum is attained both by early and tolled travelers. Therefore the system remains in Regime 1, and hence solution (17) is valid, provided $N \leq 2N_1$. To put the result another way, the optimal coarse toll makes it possible to accommodate twice as many travelers without activating the lower capacity as is the case with an unregulated equilibrium. This completes the needed proof for Section 4.3.

C List of Symbols

A	Cumulative arrivals
$a(t)$	Arrival time for departure at time t
a'_E	$\beta / (\alpha - \beta)$
a'_L	$\gamma / (\alpha + \gamma)$
N	Total number of travelers
Q	Queue length (in number of travelers)
Q_0	Value of Q at which capacity drops
R	Cumulative departures
t^*	Preferred arrival time: normalized to zero
t_0	Time of first departure
t_1	Time of last departure [$= -(\beta/\gamma)\, t_0$]
t_{drop}	First departure time for which lower capacity is encountered by end of trip (Regimes 2 and 3 only)
t_{lift}	Last departure time for which lower capacity is encountered by end of trips in (Regimes 2 and 3 only)
t_M	Departure time leading to arrival at time t^* [i.e. $a^{-1}(0) = (\beta/\alpha)\, t_0$]
α	Utility loss per unit of travel time
β	Utility loss per unit of early arrival
γ	Utility loss per unit of late arrival
δ	$\beta\gamma / (\beta + \gamma)$
ρ	Departure rate (travelers per unit time)
$\psi(Q)$	Bottleneck capacity as function of Q
ψ_0, ψ_1	Higher, lower values of ψ, respectively
$\Delta\psi$	$\psi_0 - \psi_1$