Congestion in the bathtub

Fosgerau, Mogens

Technical University of Denmark, Royal Institute of Technology, Sweden

2015

Congestion in the bathtub

Mogens Fosgerau[*]

March 20, 2015

Abstract

This paper presents a model of urban traffic congestion that allows for hypercongestion. Hypercongestion has fundamental importance for the costs of congestion and the effect of policies such as road pricing, transit provision and traffic management, treated in the paper. In the simplest version of the model, the unregulated Nash equilibrium is also the social optimum among a wide range of potential outcomes and any reasonable road pricing scheme will be welfare decreasing. Large welfare gains can be achieved through road pricing when there is hypercongestion and travelers are heterogeneous.

Anybody living in a major city will appreciate that congestion is a significant issue for economic policy. For the US, for example, it is estimated that urban road congestion in 2011 caused a total of 5.5 billion hours of delay (Schrank et al., 2012). Congestion is not only costly. It also has impacts on the local economy, it affects the functioning of labor markets, and it is an offsetting force balancing urban agglomeration effects.[1] It is therefore important for a range of economic issues to understand the nature of urban traffic congestion.

[*] Technical University of Denmark, Denmark, and Royal Institute of Technology, Sweden, mf@transport.dtu.dk. I have benefited from comments by Ken Small, Jan Brueckner, Per Olsson, Jos van Ommeren, Antonio Russo, conference audiences in Atlanta, Copenhagen, Toulouse and Stockholm, as well as support from the Danish Strategic Research Council. I am particularly happy to have received many insightful comments from Richard Arnott.

[1] See, e.g., Duranton and Puga (2004), Rosenthal and Strange (2004), Moretti (2011).

Traffic congestion is essentially dynamic: the traffic system has memory and conditions at one point in time affects conditions later on the same day. Therefore the timing of trips is fundamental and must be taken into account by economic analysis. The dynamic aspect of traffic congestion matters also from a spatial economic point of view due to the connection between the timing and the length of commutes (Fosgerau and de Palma, 2012).

The seminal Vickrey (1969) bottleneck model has shaped our intuition about urban congestion dynamics.[2] That model describes queueing before a bottleneck, for example located at the entrance to the central business district of a city where drivers enter during the morning commute. The bottleneck allows cars to enter only at a certain maximal rate. Drivers have similar preferences regarding when they would like to arrive at work and therefore a queue first builds up before the bottleneck and then dissipates every morning. In equilibrium, drivers trade off the inconvenience of deviating from their preferred schedule against the time lost queueing. The bottleneck model allows the inconvenience of the timing of trips as well as the dynamics of congestion to be accounted for in the economic analysis of congestion.

The defining property of bottleneck congestion is the constant capacity of the bottleneck, which implies that delaying arrivals at the bottleneck can reduce delays; nobody will arrive later, provided the bottleneck capacity remains fully utilized. This feature of bottleneck congestion implies that a time varying toll can be designed to induce drivers in the middle of the peak to delay their departures, such that revenue is raised, queueing is reduced and no driver is made worse off. Arnott et al. (1993) exhibit a stylized case in which an efficiency gain can be harvested through the imposition of a time-varying toll that eliminates queueing and the efficiency gain is equal to half the congestion cost that the bottleneck imposes on drivers in unregulated equilibrium. Drivers will be indifferent between the tolled and the untolled equilibrium, all of the efficiency gain will be captured as toll revenue.

Leaving the bottleneck, we shall now discuss flow congestion. It is well established that the instantaneous speed at a single point on a road is a decreasing

[2] Vickrey's paper is extensively cited and has spawned a lively literature on regulating congestion dynamics, see de Palma and Fosgerau (2011).

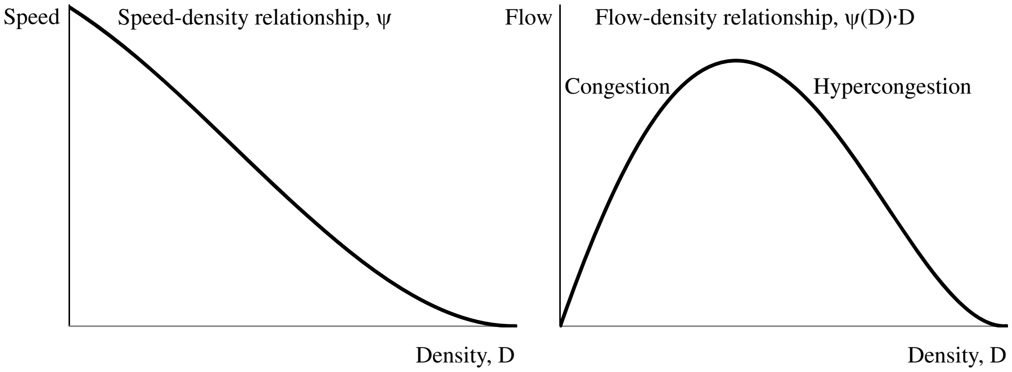

Figure 1: Fundamental diagram of traffic flow

function of the instantaneous density of cars at that point (Greenshields, 1935). The fundamental identity of traffic flow holds that flow, i.e. the number of cars passing the point per time unit, equals speed times density, where density is the number of cars per distance unit. Flow is then (with appropriate shape restrictions on the speed-density relationship) an inverse u-shaped function of density. On the upward sloping part we talk about congestion, as higher density leads to higher flow but reduced speed. On the downward sloping part we talk about hypercongestion. Here higher density is associated with both lower flow and reduced speed. The combined relationships between speed, density and flow are called the fundamental diagram of traffic flow. This is illustrated in Figure 1.

A recent range of contributions have shown that such a fundamental diagram of traffic flow also applies at the level of an urban neighborhood meeting certain conditions (Daganzo, 2007; Geroliminis and Daganzo, 2008; Daganzo et al., 2011).[3] The underlying mechanism is that drivers continuously adapt their route choices to avoid more congested parts of the road network. This adaptation process tends to equalize congestion across space. A stable relationship emerges between the density of cars in the network and the space-averaged speed. This is a very important finding, since it allows urban congestion to be analyzed in an aggregate manner, without having to refer to specific road networks. I shall refer

[3] There is a second macroscopic relation named the network exit function (Gonzales and Daganzo, 2012), which relates the rate at which trips are completed to density. I do not employ this relationship as the model presented here produces the time at which trips are completed as a function of departure times and trip lengths using only the macroscopic speed-density relationship.

to this type of congestion as *bathtub congestion*. Just like the water level is the same everywhere in a bathtub, the level of congestion and hence the speed is the same everywhere in an urban area subject to bathtub congestion.[4] Figuratively speaking, we can think of a car that drives a trip of a certain length in a bathtub: It does not matter where it begins and ends it trip, its effect on the speed of other cars depends only on *when* it is present in the bathtub.

The bottleneck model does not describe bathtub congestion well, since the inverse-u relationship between flow and density does not occur in the bottleneck model. Flow out of the bottleneck does increase with density before the bottleneck until the point where the capacity flow is reached. At higher densities, however, the flow does not decrease but stays constant. Thus the bottleneck does not generate hypercongestion.

Bottleneck congestion may be considered appropriate as a description of urban congestion for example concerning commuting flows towards a city centre, where congestion is concentrated near the entrance to the centre. Given the now existing empirical evidence, bottleneck congestion can no longer be considered appropriate as a description of congestion at the urban level. For homogeneously congested downtown urban areas, we now have empirical evidence that bathtub congestion is an appropriate description.[5]

This paper presents a model that I call the bathtub model. The bathtub model is similar to the bottleneck model in describing a fixed mass of homogeneous drivers who care about the timing of their trips. The main difference is the congestion technology embodied in the model. Where the bottleneck model builds on bottleneck congestion, the bathtub model (unsurprisingly) builds on bathtub congestion. Thus it incorporates hypercongestion., allowing increases in flow to be associated with increases in speed. In this paper I show that the bathtub model can

[4] Richard Arnott has pointed out that I use the term "bathtub model" in the sense of hydrology and he prefers calling it an isotropic model. Vickrey worked on what he also called a bathtub model of congestion, which was based on the intuition that now materializes in Daganzo's work. Vickrey never completed this work but a note has been preserved (Vickrey, 1991). He used as fundamental the idea that outflow from the bathtub is proportional to the height of the water in the bathtub. This is similar to the second macroscopic relationship mentioned in the previous footnote whereby the rate of trip completion depends on density. My model simply computes the times when trips are completed as a function of departure time and speed.

[5] Ji and Geroliminis (2012) consider partitioning a road network into a small number of uniformly congested subnetworks (bathtubs). This paper considers just a single bathtub.

be used to give a unified treatment of a range of issues related to urban congestion and hypercongestion, as discussed in the following.

The bathtub model leads to conclusions that are radically different from those of the bottleneck model. The bottleneck conclusions depend on the property of bottleneck congestion that it is possible to delay departure times without affecting arrival times. In the bathtub, such trip retiming has small or even no effect on travel times under some circumstances. The underlying principle is illustrated in Figure 2, which shows a short trip and a long trip in an urban area subject to bathtub congestion. Each trip has some fixed length and a duration that depends on the average speed obtained. The short trip is carried out within the duration of the long trip; I call this *regular sorting*. The speed is low when both trips are ongoing and high when only one trip is in progress.

Notice first that the duration of the short trip does not depend on the timing of that trip. Under regular sorting, the speed for the short trip is always low. Notice next that, still under regular sorting, the duration of the long trip is also independent of the relative timing of the two trips: the long trip covers the same distance as the short trip during the interval when both are ongoing; the remaining distance is covered at the high speed, which is the same before and after the short trip.

Section 2 generalizes this simple example to the case where there is a continuum of drivers with a distribution of trip lengths and shows that travel times for all trip lengths are completely determined under regular sorting. Thus the specific departure and arrival time profiles do not matter at all for travel times. Under a regularity assumption, it is shown that Nash equilibrium in the timing of trips is in fact regularly sorted. Moreover, taking the travel time as given, each driver travels at his optimal time. Since travel times cannot be reduced as long as regular sorting is maintained, this implies that the Nash equilibrium is also the social optimum among regularly sorted outcomes. Hence any policy that changes the departure schedule can only make drivers worse off, if regular sorting is maintained.

Section 3 allows demand for car travel to be elastic in two different ways. First, by introducing an alternative mode of travel. I call it "transit" for concreteness, but the defining characteristic is just that it provides a speed that is attractive when car speed drops due to congestion. Travelers have no specific preferences

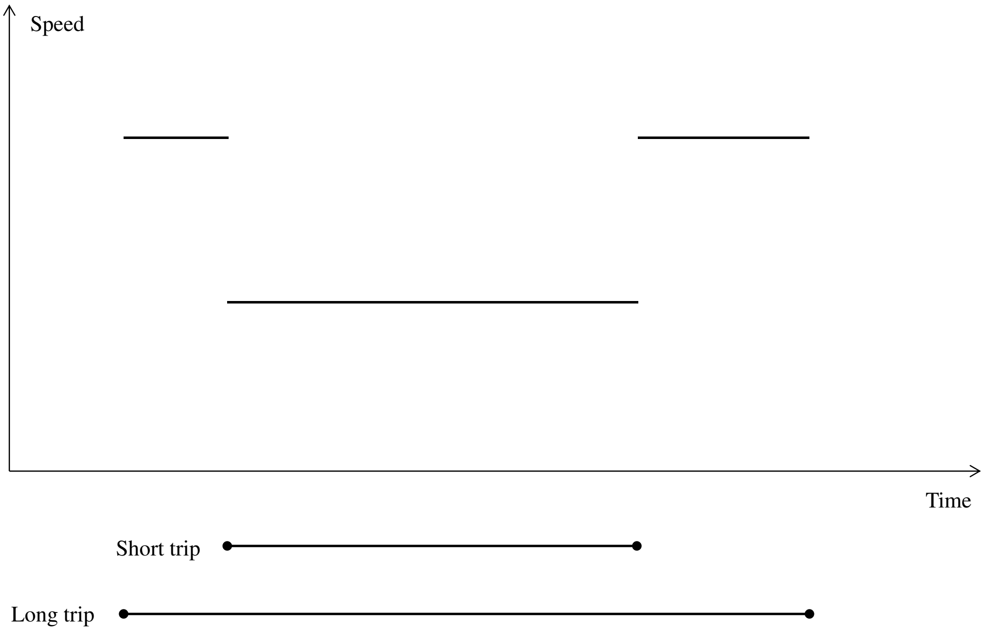

Figure 2: A short and a long trip with regular sorting. Durations are independent of timing.

regarding mode of travel. In equilibrium, the availability of transit allows travelers with short trips to escape from the lowest car speeds during the height of the congested peak and instead travel at the higher transit speed. The remaining car drivers are those with trip lengths above some threshold and they gain a speed increase from the absence of the transit users. This mechanism is similar to that described in Anderson (2014), who argue that transit users are those whose road alternative is most congested and hence that their impact on road congestion is disproportionately large. In contrast to previous transportation and urban economics literature, Anderson (2014) conclude that the congestion relief benefits are large enough to justify investment in transit infrastructure. This paper finds that it is welfare increasing to induce a higher level of transit use than in equilibrium. A tolling/subsidy scheme that achieves this can be very simple, it suffices to induce more travelers to use the alternative mode and this can be attained by a constant toll on car trips and/or a subsidy to transit use. If the transit speed is higher than the critical speed below which there is hypercongestion, then hypercongestion does not occur in equilibrium and the scope for achieving efficiency gains from time-varying road pricing is then limited.

The second way of allowing for elastic demand is to introduce an outside option with a utility that is the same for all. Then a fixed charge per trip or a toll that is charged at a flat rate will induce those travelers with the longest trip to abandon their car trips; the remaining car drivers will experience higher speeds. It is welfare increasing to have a positive charge or flat rate toll. This reproduces the standard conclusion from static models of congestion that welfare can be improved through pricing.

Section 4 extends the analysis in Section 2 to allow also for heterogeneity of the trip timing preferences. Preferences are shifted in time by a constant, which has some distribution in the population of drivers. Again under a regularity condition, drivers in equilibrium will sort regularly according to trip length within each time shift group and they will sort according to the shift in preferences. There is now a scope for achieving efficiency gains by changing departure patterns, which arises due to the effect that travelers with different time shifts of their preferences has on the speed of each other.

Figure 2 shows a short and a long early trip and a short and a long late trip.

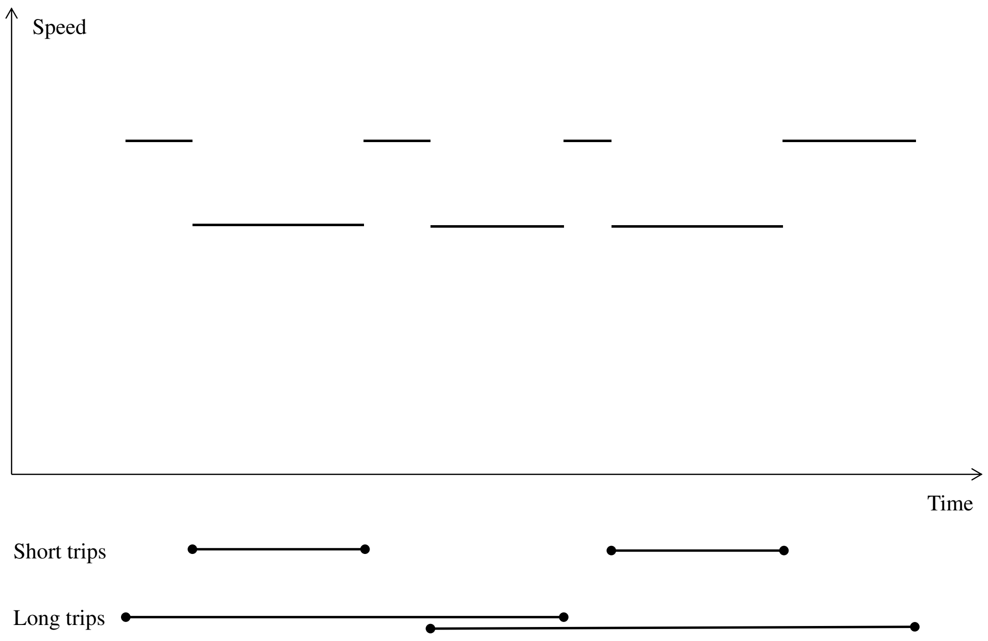

Figure 3: Two groups of a short and a long trip. Regular sorting within groups, but not overall. Long trips overlap and durations depend on timing.

There is regular sorting within each group. Still, the travel time depends on the departure times in this case, depending on how much the long trips overlap in time. Thus heterogeneity with respect to trip timing preferences makes trip timing relevant for travel times, even if there is regular sorting within each group of travelers.

The scope for achieving efficiency gains by affecting departure schedules is investigated through a series of simulations of a congested demand peak at varying levels of congestion. The simulations show that the scope for efficiency gains from road pricing depends dramatically on the level of congestion, I examine three cases.

In the least congested case, there is no hypercongestion. Road pricing achieves a small welfare gain, but at the price of raising a revenue that is much larger than the welfare gain. Thus car drivers will lose substantially from road pricing if road pricing revenues are not returned to them. The intermediate case just reaches hypercongestion at the middle of the peak when road pricing is not in place. In this case, road pricing can produce a larger welfare gain but the revenue from pricing is still much larger than the welfare gain. The flow profile is not much affected by road pricing in this case, which means that simply counting traffic will not easily reveal that road pricing has had any effect. However, the flow without pricing results with a high density of cars traveling at a low speed, while the same flow in the presence of pricing results with a low density of cars traveling at a high speed. In the most congested case, the capacity of the bathtub is only slightly smaller than in the intermediate case, but the congestion outcomes are quite different. Without road pricing, traffic flow has two maxima, one at the beginning of the peak and one at the end. In between, flow drops noticeably due to hypercongestion. Road pricing maintains flow above capacity and thus produces a large welfare gain. The revenue from pricing is a little larger than the welfare gain, which means that drivers would still lose if revenues were not returned to them. Also in this case, pricing has remarkable effects on the timing of trips. Travelers who ex ante travel early in the peak are induced by pricing to travel such that they complete their trips earlier than otherwise. Thereby they free capacity for travelers in the middle of the peak, who then gain from higher speed. The increase in speed is so large that capacity is also freed for the early travelers, some of whom can depart later than they would without pricing, even if they arrive earlier. The same, seemingly paradoxical phenomenon emerges for travelers at the end of the peak who, in the presence of pricing, will depart later but still arrive earlier than they would have without pricing.

Section 5 discusses traffic management measures that are increasingly used to avoid hypercongestion. Hypercongestion occurs when a higher density of cars in the traffic network blocks road capacity, thereby causing flow, the total distance driven per time unit, to decrease. Traffic management measures dealing with this seek to place the high density of cars at places where it does not block road capacity: they turn hypercongestion into mere queues. This can be described in the bathtub model and a simulation shows that the welfare gain from traffic management may be close to that of road pricing.

I now turn to discussing some related literature. Arnott (2013) provides a recent and comprehensive review of the bathtub literature. Here I review just the part of the literature that is most closely related to the current paper.

Distance plays a very important role in the bathtub. Distance is, however, essentially ignored in the basic Vickrey bottleneck model. Vickrey (1969) employed so-called α-β-γ trip timing preferences, represented by a utility function that is linear in travel time and separable in travel time and arrival time. Then driver heterogeneity with respect to distance to the bottleneck can be ignored. Arnott (1998) combined a model of urban spatial structure with the α-β-γ bottleneck model; optimal tolling does not change transport costs for travelers so when the revenues are not returned, optimal tolling will have no effect on urban structure. Fosgerau and de Palma (2012) presented a model with more general scheduling preferences where distance from the home to the bottleneck does matter and used this model to analyze commuting in a city where workers live at various distances from the CBD. In the bathtub, each driver has some distance to cover in the urban area and there is an aggregate distribution of trip lengths, which matters for outcomes. In this model, optimal tolling will have an effect on urban structure.

Arnott (2013) develops a version of Vickrey's bathtub model, which is similar to the present model in several ways. Arnott (2013) uses a linear speed-density relationship to describe congestion technology while the present paper uses a general speed-density relationship. The present paper uses general trip timing preferences that comprise the α-β-γ trip timing preferences of Arnott/Vickrey as a limiting case. The main difference concerns the treatment of trip distance. Arnott uses Vickrey's assumption that outflow from the bathtub is proportional to density.[6] This requires that remaining trip lengths at any time in the bathtub have an exponential (i.e. memoryless) distribution. This assumption can be interpreted as saying that each driver's trip length is random with a certain distribution and that drivers do not know their trip lengths at the time they make their departure decisions, which is somewhat awkward. This paper merely assumes that drivers choose departure time optimally knowing their trip length and the speed at which they will travel and the physics of the model keeps track of the distance driven for each driver. Arnott (2013) finds that the efficiency gain from congestion tolling might be smaller than the toll revenue when congestion is light and larger, even much larger, when congestion is severe.

Fosgerau and Small (2013) also analyze a bathtub type model. In their model there is a bottleneck with a capacity that depends on the number of cars queueing; tractability is achieved by simplifying the bottleneck capacity function to a step function and there is no congestion outside the bottleneck. Unfortunately, increasing the number of steps in the capacity function leads to a proliferation of cases and Fosgerau and Small only analyze a two-step capacity. The present model does not resort to such ad hoc devices and allows for a general distribution of trip lengths.

Verhoef (2003) considers a model in which car travel a road that has a bottleneck in the middle. This leads to hypercongestion on the part of the road upstream of the bottleneck. Hypercongestion does, however, not matter economically in this model: the flow rate and speed inside the hypercongested queue do not matter for the total trip time. In contrast, this paper is concerned with a situation where hypercongestion is of first order economic importance.

Section 1 introduces the trip timing preferences, which are common to the bottleneck and the bathtub models. Section 2 introduces the bathtub model. Section 3 allows demand to be elastic through the introduction of a transit mode and an outside option. Section 4 introduces heterogeneity in trip timing preferences. Sec-

[6] Vickrey's assumption has empirical support. It corresponds to the network exit function discussed in footnote 3.

tion 5 considers traffic management in the bathtub. Section 6 concludes. Proofs omitted in the text are in Appendix B.

1 Driver preferences

We begin by formulating the travelers' preferences for the timing of a trip. Let h, w be real functions satisfying $h, w > 0$, $h' < 0 < w'$ and $h(0) = w(0)$ as well as the technical conditions given in the footnote.[7]

The function h describes utility accumulated at the origin of the trip from some initial time, set to zero at no loss of generality, until departure time a, this amounts to $\int_0^a h(s)\,ds$. Similarly, $\int_b^0 w(s)\,ds$ is the utility accumulated at the destination from the arrival time until some arbitrary time, also set to zero at no loss of generality. Let τ be a toll payment and define utility as

$$u(a, b, \tau) = \int_0^a h(s)\,ds + \int_b^0 w(s)\,ds - \tau. \tag{1}$$

The marginal utilities of later departure and earlier arrival are positive and decreasing. There is a continuum of N travelers and they have identical trip timing preferences represented by u. We are concerned with the interaction of congestion dynamics with the timing of departures and regard N as fixed.

To talk about welfare, I assume a social welfare function that depends only on the average utility of travelers gross of toll payment, that is, it depends only on the average of $u + \tau$. This welfare measure will also apply when demand is allowed to be elastic, since then there will be a fixed population of potential travelers who will have an outside option with a fixed utility associated.

A range of different pricing schemes may be cast as modifications of the utility rates. This is useful, since then the analysis with general utility rates applies as well under these pricing schemes. A charge may be defined up to a constant as $\tau_1(a) = \int_a^0 \pi_1(s)\,ds$, which is a charge at the origin of a trip, as $\tau_2(b) = \int_0^b \pi_2(s)\,ds$, which is a charge at the destination, or as $\tau_3(a, b) = \int_a^b \pi_3(s)\,ds$, which is a charge while the trip is ongoing. Insertion of these charges in the

[7] The rates of change $\dot{h}(\cdot) \equiv \partial \ln h(\cdot)/\partial t$ and $\dot{w}(\cdot)$ are bounded. The function $(a, b) \to \left(\frac{h'(b)}{h'(a)-w'(b)}, \frac{w'(b)}{h'(a)-w'(b)}\right)$ is bounded, continuous and Lipschitz.

definition of utility in equation (1) may be cast as a modification of the utility rates as follows.

$$\begin{aligned} u\left(a, b, \tau_1\left(a\right)\right) &= \int_0^a \left(h\left(s\right) + \pi_1\left(s\right)\right) ds + \int_b^0 w\left(s\right) ds, \\ u\left(a, b, \tau_2\left(a\right)\right) &= \int_0^a h\left(s\right) ds + \int_b^0 \left(w\left(s\right) + \pi_2\left(s\right)\right) ds, \\ u\left(a, b, \tau_3\left(a, b\right)\right) &= \int_0^a \left(h\left(s\right) + \pi_3\left(s\right)\right) ds + \int_b^0 \left(w\left(s\right) + \pi_3\left(s\right)\right) ds. \end{aligned} \tag{2}$$

2 The bathtub model

The bathtub model considers car trips that take place on the roads in an urban area where a uniform but time-varying speed S prevails, with speed at any time t satisfying $S\left(t\right) > 0$. The part of trips outside the urban area is uncongested with constant distance which is normalized to zero at no loss of generality. For a given departure time a and trip length l, the arrival time $b\left(a, l\right)$ is given implicitly by

$$l = \int_a^{b(a,l)} S\left(t\right) dt. \tag{3}$$

Denote by $D\left(t\right)$ the number of drivers on the road at time t; in a real city this quantity is proportional to the density of cars on the streets when the road network is held constant. The speed-density relationship is ψ, where $\psi > 0, \psi' < 0$, and it relates the instantaneous speed to instantaneous density by

$$S\left(t\right) = \psi\left(D\left(t\right)\right).$$

The speed is measured as distance per time unit. Density is the number of cars driving in the area. The total distance driven per time unit is the speed times the density, which is equal to the flow. The flow is thus the rate at which the total distance driven increases.

There is a distribution of trip lengths l in the population of drivers. Let demand $\Phi\left(l\right)$ be the number of drivers with trip length of at least l. This distribution is absolutely continuous with c.d.f. $1 - \Phi$ and density $\phi = -\Phi'$.

The arrival time $b\left(a, l\right)$ for a driver depends both on the departure time a and the trip length l. We take for granted that departure and arrival schedules a and b are differentiable functions of trip length. It follows immediately from equation (3) that the partial derivatives of $b = b(a, l)$ are

$$\frac{\partial b(a, l)}{\partial a} = \frac{S(a)}{S(b)}, \frac{\partial b(a, l)}{\partial l} = \frac{1}{S(b)}.$$

Drivers are indexed by their trip length. If $a(l)$ is the departure time for drivers with trip length l then, denoting $b(l) = b(a(l), l)$,

$$b'(l) = \frac{1}{S(b(l))} + \frac{S(a(l))}{S(b(l))} a'(l). \tag{4}$$

Define for convenience the functions

$$H(a) = \frac{h(a)}{S(a)}, W(b) = \frac{w(b)}{S(b)},$$

expressing the utility rates in terms of utility per distance unit rather than utility per time unit[8], and note that $\dot{H}(a) = \dot{h}(a) - \dot{S}(a), \dot{W}(b) = \dot{w}(b) - \dot{S}(b)$.

A main feature of the bathtub model is the heterogeneity with respect to trip length. Through the first-order condition for the choice of departure time, it leads to simple differential equations for the departure and arrival times, which will be used throughout the paper. The first-order condition for the choice of departure time for a driver with trip length l is

$$0 = \frac{\partial u(a, b(a, l))}{\partial a} = h(a) - w(b) \frac{\partial b(a, l)}{\partial a} = S(a)(H(a) - W(b(a, l))) \tag{5}$$

The second-order condition is discussed in this footnote.[9] The first-order condi-

[8] At the time of departure, the utility rate is $h(a) = \frac{du}{da}$. The inverse of speed is the time per distance $S(a)^{-1} = \frac{da}{dl}$, such that $\frac{h(a)}{S(a)} = \frac{du}{dl}$ is the rate at which utility is achieved per distance (not) traveled.

[9] Omitting some function arguments for the sake of clarity, the corresponding second-order condition is

$$0 \geq S(a)\left(H'(a) - W'(b)\frac{S(a)}{S(b)}\right) = S(a) H(a)\left(\dot{H}(a) - \dot{W}(b)\frac{S(a)}{S(b)}\right).$$

This constrains how quickly the speed can decrease at the time of departure or increase at the time of arrival.

tion holds for all l and is equivalent to $H\left(a\left(l\right)\right)=W\left(b\left(a\left(l\right),l\right)\right)$. Differentiating the first-order condition with respect to l shows that

$$0=H'\left(a\left(l\right)\right)a'\left(l\right)-W'\left(b\left(a,l\right)\right)\left(\frac{1}{S\left(b\left(l\right)\right)}+\frac{S\left(a\left(l\right)\right)}{S\left(b\left(l\right)\right)}a'\left(l\right)\right).$$

If the second-order condition holds with strict inequality, then this can be solved to yield the following differential equations for a and b.

$$\begin{aligned} a'\left(l\right) &= \tfrac{\dot{W}(b(a,l))}{S(b(l))\dot{H}(a(l))-S(a(l))\dot{W}(b(a,l))}, \\ b'\left(l\right) &= \tfrac{\dot{H}(a(l))}{S(b(l))\dot{H}(a(l))-S(a(l))\dot{W}(b(a,l))}. \end{aligned} \tag{6}$$

Together with the boundary condition $a\left(0\right)=b\left(0\right)=0$, these differential equations determine the equilibrium departure and arrival schedules.

If $a'<0<b'$ then shorter trips are carried out entirely within the duration of longer trips. As mentioned in the Introduction, this is called *regular sorting*. It is illustrated in Figure 4. Regular sorting is an analytically useful property. Under regular sorting, $D\left(a\left(l\right)\right)=D\left(b\left(l\right)\right)=\Phi\left(l\right)$,which says that those on the road when a trip of length l begins or ends are those with trips longer than l. Then, when a trip of length l begins or ends, the speed is $S\left(a\left(l\right)\right)=S\left(b\left(l\right)\right)=\psi\left(\Phi\left(l\right)\right)$. By equation (4), the travel time $b\left(l\right)-a\left(l\right)$ then depends on trip length according to $b'\left(l\right)-a'\left(l\right)=\frac{1}{S(\Phi(l))}$. The driver with trip length $l=0$ has a travel time of zero and then the travel time profile is completely determined. This argument establishes the following result.

Theorem 1 *Under regular sorting and the specifications in this section, the travel times for all drivers are completely determined by the speed-density relationship ψ and by the distribution of trip lengths Φ.*

This result means that travel times do not depend on the departure schedule as long as there is regular sorting. In the bottleneck model, in contrast, travel times depend strongly on the departure schedule.

We shall use the following regularity assumption.

Assumption 1 *In equilibrium, $\dot{H}\left(a\right)<0,\dot{W}\left(b\right)>0$ for all times a,b.*

This assumption is trivially satisfied when the speed is constant. It remains satisfied if speed does not drop too quickly at times of departure or rise too quickly at times of arrival. The assumption is not a primitive condition, as would have been desirable, but depends on endogenous variables. The numerical results presented below specifies h and w as exponential functions: in this case, Proposition 1 in Appendix C translates the assumption into a straightforward bound on the slope coefficients of h and w which is satisfied provided the slope coefficients are sufficiently large. By construction, Assumption 1 is satisfied in Nash equilibrium in the numerical example below. As stated in the following theorem, the assumption implies regular sorting.

Theorem 2 *Under Assumption 1, Nash equilibrium exists uniquely in the bathtub model with heterogeneous trip lengths. Drivers sort regularly such that $a'(l) < 0 < b'(l)$ for all l. Specifically,*

$$\begin{array}{l} a'(l) = \frac{1}{\psi(\Phi(l))} \frac{w'(b(l))}{h'(a(l)) - w'(b(l))} < 0, \\ b'(l) = \frac{1}{\psi(\Phi(l))} \frac{h'(a(l))}{h'(a(l)) - w'(b(l))} > 0. \end{array} \tag{7}$$

The Nash equilibrium is socially optimal among departure schedules that maintain regular sorting.

Theorem 2 is proved in Appendix B except for the last statement which is established here. Under regular sorting, the first-order condition for the choice of departure time (5) reduces to

$$h(a(l)) = w(b(l)). \tag{8}$$

This is also the first-order condition for the choice of departure time for a driver facing a constant travel time $b(l) - a(l)$; hence all drivers depart at the time that would be optimal if their travel time was fixed at the equilibrium value. By Theorem 1, the travel time cannot changed given regular sorting. This implies that the Nash equilibrium is the social optimum among the cases with regular sorting.

Before presenting the first simulation example, it is useful to make a few observations regarding the relationships between speed, flow and density under bathtub (and flow) congestion. I use a linear speed-density relationship $S = 1 - \gamma D$, which has a maximal speed of 1 and reaches zero speed at the jam density $1/\gamma$. A linear speed-density relationship is roughly in accordance with the empirical evidence in Geroliminis and Daganzo (2008). The flow-density relationship $F = SD = D(1 - \gamma D)$ has maximum at $D^* = 1/2\gamma$; the corresponding maximal flow is $F^* = 1/4\gamma$ and the corresponding speed is $S^* = 1/2$. In the congested region, at densities less than the critical density D^* and speeds above 0.5, a higher flow is associated with a lower speed. In the hypercongested region, at densities above the critical density and speeds below 0.5, a higher flow is associated with higher speed. We may write flow as a function of speed as $F = S(1 - S)/\gamma$, which is a parabola with maximum at $S = S^*$ and which attains the value zero when $S = 0$ and when $S = 1$. Any flow less than the maximal flow may occur either at a speed smaller than S^* or at a speed higher than S^*.

Figure 4 shows departure and arrival schedules in a simulation of Nash equilibrium; details of the simulation are given in Appendix C, which also shows that the simulation is consistent with regular sorting. The speed drops by 70% at the height of the peak when all drivers are on the road, which means that it reaches hypercongestion.

The first panel of the figure shows the departure and arrival schedules, they are regularly sorted. The second panel shows the speed profile. There is hypercongestion when the speed is below 0.5. The flow profile on the third panel drops in the middle of the peak, which is again evidence of hypercongestion.

The simulation uses a uniform distribution of trip lengths. Then the slope of the arrival rate schedule a' is proportional to the rate at which trips are completed. Observe that at the end of the peak, the slope of the arrival rate schedule increases: this is a straightforward consequence of the decreasing density at the end of the peak. This is inconsistent with the properties of the network exit function in, e.g., Gonzales and Daganzo (2012), which would say that the rate at which trips are completed should tend to zero as density tends to zero. This is a reason why I deviate from the engineering literature in not assuming a network exit function that relates the rate of trip completions to density. The present model tracks individual trips from beginning to end.

Through simulation it has been verified that social optimum may not be regularly sorted. When social optimum is not sorted, then finding the social optimum

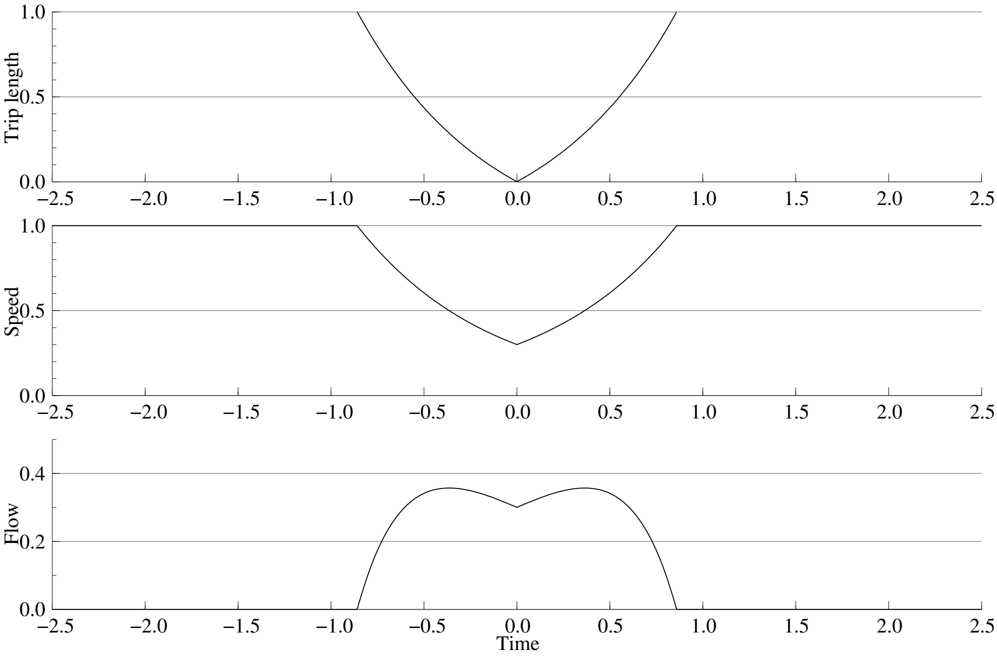

Figure 4: Simulation with regular sorting and hypercongestion

analytically is extremely difficult and probably impossible as it involves very complex differential equations.[10] A comparison of the social optimum to the Nash equilibrium in the simulation example shows that the average duration of trips decreases by 0.24 time units from 1.03, but the average utility increases only 0.14 utility units corresponding to about 0.04 time units.[11] Thus most of the travel time gain is offset by worse timing of trips in this case. In contrast to equilibrium, the social optimum does not have the first departure at time 0. Hence it is not possible to implement the social optimum via a toll of the type τ_3 in (2).

[10] Equations (7) is a pair of ordinary differential equations. Without regular sorting the equivalent differential equations would be a kind of delay differential equations (Kuang, 1993), but where the increments $a'(l), b'(l)$ depend on a, b at other values of l as well as on the inverses of these functions.

[11] The marginal utility of travel time is computed in Appendix D. This is used to translate changes in utility into changes in travel time. Computing the marginal utility of travel time requires a little consideration since it depends on when the trip takes place and on how departure and arrival times are supposed to adjust. The appendix computes the marginal utility change for a trip with specific beginning and end times with the change corresponding to an increase in travel time that maintains the ratio between the marginal utility of distance at the beginning and the end of the trip. When the trip is optimally timed by the driver then this corresponds to the optimal change in the timing of the trip.

3 Elastic demand

We shall now extend the basic model of the previous section to allow demand to be elastic. We shall do this in two ways. First, we introduce an idealized alternative mode of transportation. For simplicity, we call it "transit". The alternative mode can be thought to represent public transport of some kind, in Copenhagen or Amsterdam it could be walking or cycling, in Asian cities the alternative mode could be moped. The main property is that it provides a speed that is better than the speed achieved by car when there is most congestion. We shall see that the presence of such a mode induces travelers with short trips to escape from the car, thereby relieving road congestion during the middle of the peak which allows the peak to be shorter.

The second extension we consider is to allow travelers to have an outside option with some constant utility. Then travelers with long trips will be first to choose the outside option. This relieves congestion for all remaining car drivers and the peak will again be shorter.

3.1 Transit

Travelers are still assumed to have the preferences specified by the utility in equation (1). Thus they do not have specific preferences regarding the transportation mode but care only about the duration and timing of trips. For a given departure time a traveler simply prefers the mode that provides the faster trip. The transit mode is uncongested, providing a constant speed S_T regardless of trip length. The transit speed is lower than the free flow speed of cars but higher than the car speed that results when all cars are on the road, i.e. $\psi(N) < S_T < \psi(0)$. This assumption ensures that some but not all travelers will choose the transit mode. We have the following result.

Theorem 3 *Assume that transit is available. Let $\tau \geq 0$ be a fixed charge per car trip. Under Assumption 1, Nash equilibrium exists uniquely. Both the group of car drivers and the group of transit users sort regularly and the shortest trip length of car drivers is not shorter than the longest trip length of transit users.*

Clearly, all travelers gain from the availability of an uncongested transit mode. Without transit, transit users would have been driving at speeds less than the transit speed. The same is true of car drivers, their speed does not drop below the transit speed when transit is available. They cover the distance of the shortest car trip faster than they would have if transit was not available and hence depart later and arrive earlier than otherwise.

The same equilibrium emerges if the charge τ is replaced by a subsidy of the same size to transit trips. A charge may also be just partially replaced by a subsidy and it is thus always possible to combine charge and subsidy to achieve revenue neutrality.

The availability of transit also makes it relevant to consider again the potential for welfare gains from charging road use. Consider the last car drivers to depart, just before transit use begins. They have trip lengths just longer than l and their average speed is slightly above S_T, hence they do not want to change mode. If they were somehow induced to change to transit then they would lose a bit of time, but all the remaining car drivers would gain some time due to increased speed. This argument suggests that it would be socially optimal to increase the use of transit compared to equilibrium. The next theorem shows that this is indeed the case.

Theorem 4 *Assume that transit is available. Under Assumption 1, increasing transit use from the equilibrium level is welfare improving. The regular welfare optimum may be implemented via a fixed charge applied to car drivers.*

The social optimum may be implemented in a variety of ways. As has been noted, it leads to identical outcomes (except for monetary payments) if a charge to car drivers is replaced by a subsidy to transit users. It is similarly equivalent to charge car drivers at a flat rate during the time of the shortest car trip or to subsidize transit trips at a flat rate during the time of the longest transit trip. It is even feasible to allow charges or subsidies to vary over time during these intervals.

The results concerning transit are now illustrated through some examples. A range of simulations have been carried out, summarized in Table 1. The setup of the simulations is the same as in Section 2, except now the transit mode is included. The simulations are also illustrated in Figure 5.

Table 1: Simulation results with transit

Sim. no.	Transit speed S_T	Charge on car trips τ	Transit share	Last arrival	Utility min.	Utility mean	Avg. duration	Min. car speed
1	0	0	0	0.76	-4.61	-2.63	0.88	0.40
2	0.5	0	0.17	0.74	-4.43	-2.53	0.84	0.50
3	0.5	0.8	0.53	0.64	-3.63	-2.28	0.77	0.72

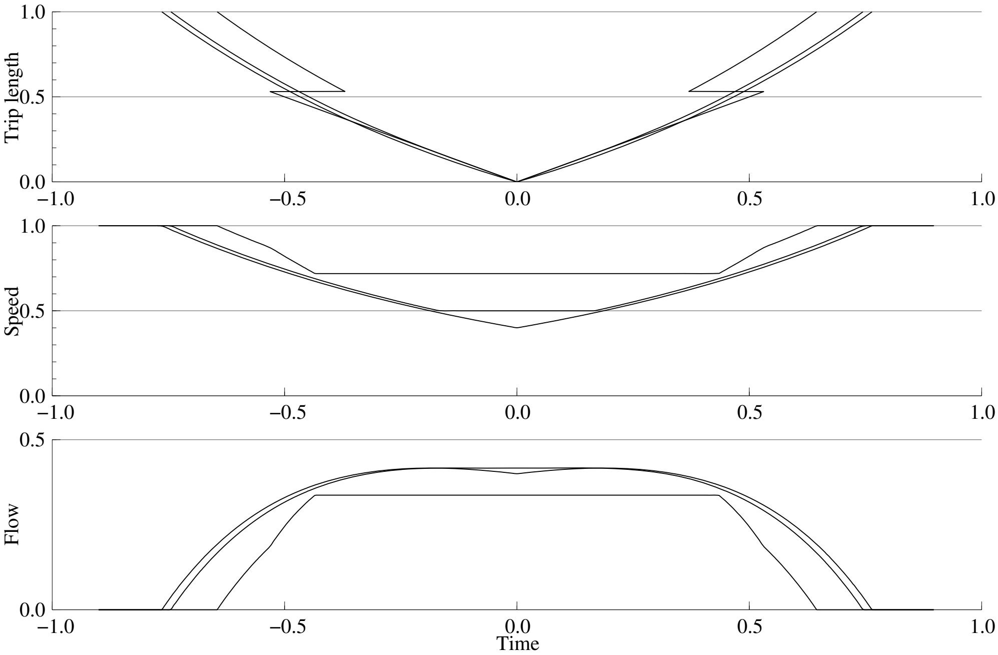

Figure 5: Simulations with transit

In the first simulation, the transit speed is low, making the transit mode share zero. The car speed reaches a minimum of 0.4, which means that there is hypercongestion, reflected in the drop in the flow at the middle of the peak. Simulation 2 sets a transit speed $S_T = 0.5$, which is higher than the previous minimum car speed and equal to the critical speed below which there is hypercongestion. No charge or subsidy is applied. This leads to a transit mode share of 0.17 and a minimum car speed that is equal to the transit speed. All remaining car drivers achieve higher speeds than before. The third simulation sets the welfare maximizing charge on car drivers of 0.8, it is equivalent to a subsidy to transit users. This increases the transit share to 0.53 and the minimum car speed increases to 0.72.

3.2 An outside option

Assume now that all travelers have an outside option yielding a constant utility u_0 that is the same for all. Maximizing utility, they will only drive if their utility (1) associated with driving is greater than u_0. This setup has the immediate consequence that it is the drivers with the longest trips who are at the margin. Under regular sorting and a constant charge τ or a toll charged at a constant rate π_3 during the trip, the toll payment is higher for drivers with longer trips. Hence the marginal driver is the driver with the longest trip also under these kinds of pricing. In the model with homogeneous drivers, under regular sorting, there is no efficiency gain available from affecting the timing of trips, but only from pricing some of the longest trips off the road. This can be achieved by a constant toll charged of all drivers. We obtain the following result.

Theorem 5 *Assume that an outside option is available. Under Assumption 1, it is welfare improving to impose a fixed charge τ or a toll that is charged at a fixed rate π_3 during trips. The welfare optimum is achieved by such a toll.*

A constant toll $\pi_3 > 0$ drives the travelers with the longest trips to the outside option. Hence speeds improve throughout the peak for all who remain on the road and the peak will be shorter.

If both transit and an outside option are available, then it is welfare improving to affect the margins for both long and short trips. A flat charge on driving would however affect both margins. Therefore welfare optimum requires that transit can be subsidized (or taxed) independently.

4 The bathtub model with heterogeneous preferences

This section extends the bathtub model in Section 2 to allow for heterogeneous time shifts of trip timing preferences[12], while maintaining heterogeneity with respect to trip length. Recall that the utility rates are defined with $h(0) = w(0)$ and

[12] This is analogous to allowing the preferred arrival time to be heterogenous under standard $\alpha - \beta - \gamma$ trip timing preferences (Vickrey, 1969).

that this implies that a driver with zero trip length will depart and arrive at time 0. We now introduce a time shift c and define

$$u\left(a,b,\tau|c\right)=u\left(a-c,b-c,\tau\right),$$

such that a driver with zero trip length and time shift c will prefer to depart and arrive at time c. The following theorem uses a generalized version of Assumption 1 to allow for heterogeneous time shifts and shows that, in Nash equilibrium, drivers sort on time shift c while regular sorting on trip length holds for each time shift c.

Theorem 6 *Assume that Nash equilibrium exists and satisfies $\dot{h}\left(a-c\right)-\dot{S}\left(a\right)<0<\dot{w}\left(b-c\right)-\dot{S}\left(b\right)$ for all a,b,c. Then, for fixed c, drivers sort regularly in equilibrium and, for fixed l, the departure time is increasing as a function of c.*

Think of drivers as being grouped according to their time shift c and restrict attention to cases with regular sorting. With just one group we have seen that it is not possible to reduce travel times by changing departure times without going outside regular sorting. With more time shift groups having trips that overlap in time, departure times can be changed to reduce the amount of overlap between groups while maintaining regular sorting. Hence there is a scope for achieving a welfare gain from tolling in this case if drivers with small c (those who prefer early trips) can be induced to depart earlier, and conversely drivers with large c can be induced to depart later, since then speed will increase while those in the middle are traveling.

We shall investigate this using simulation, now including a uniform distribution of time shifts c on the interval $[-1,1]$. If we interpret the time unit as hours then the simulation could describe a commuting peak that in the absence of congestion would last from 6.30am to 9.30am with the shortest trips taking place during 7-9 am. Congestion increases the duration of the peak. The speed-density relationship is again linear. The slope is set to produce varying degrees of congestion. This is equivalent to changing the size of the population. We shall look at three sets of simulations, distinguished just by the slope of the speed-density relationship.

The simulations are carried out by first computing the departure and arrival times for each c using the differential equations (6) derived from the first-order condition, taking a speed profile S as given. The corresponding second-order condition is verified. Then the speed profile is updated using the resulting departure and arrival schedules to compute a new density profile, which then leads to a new speed profile. These steps are repeated until the speed profile is constant to a high degree of precision. The simulation allows for modification of the utility rates by tolling as discussed in Section 1. Regular sorting is not imposed on the simulations but results anyway for most travelers.

Each set of simulations comprises two scenarios, a base scenario and a tolled scenario. The base scenario is equilibrium without any regulation. This is compared the tolled scenario, which is equilibrium in the presence of a toll τ_3 charged during trips. The toll τ_3 is computed to approximate the welfare optimal toll.[13]

Figure 6 shows the first set of simulations with light congestion corresponding to a maximum flow of 0.25. The first panel shows departure and arrival schedules for the travelers with $c = -1, 0$ and 1, the second panel shows the speed profiles and the third shows the flow profiles. The flow at a point in time is the rate of increase of the total distance driven by all vehicles in the system and hence the area under the flow profile is the total distance driven, which is constant in these simulations.

In the base scenario, the speed reaches a minimum of around 0.7 so there is no hypercongestion. The flow profile is unimodal. Introducing the toll causes early travelers to depart and arrive earlier and late travelers to depart and arrive later. The flow profile becomes flatter and more spread out. The speed is decreased on the shoulders of the peak and increased in the middle. Table 2 summarizes the welfare consequences of the charge for the three sets of simulations. The first column shows the critical flow F^*, which is set by design for each set of simulations and which is the only parameter that distinguishes between them. The second column shows the minimum speed that is reached in each simulation. We

[13] The toll τ_3 is continuous and piecewise linear with eight equidistant separation points in the interval $[-2, 2]$. It is is zero outside this interval. The toll is restricted to be symmetric since the trip timing preferences are symmetric in time in the simulation; this implies that the toll is described by four parameters for the values of τ_3 at the separation points. The toll was identified in a search in the welfare increasing direction over a four-dimensional grid of parameter values.

Table 2: Simulation results

Sim.	Max. flow	Min. speed	Gross util. change (2)	Toll rev. (3)	Ratio (3)/(2)
1 base	0.250	0.68			
1 toll	0.250	0.72	0.031	0.716	23.4
2 base	0.167	0.38			
2 toll	0.167	0.60	0.845	2.415	2.86
3 base	0.156	0.24			
3 toll	0.156	0.59	2.883	3.061	1.06

recall that the limit between congestion and hypercongestion is at the critical speed 0.5. The third column shows the average utility gross of toll, which is also the welfare measure. The fourth column shows the toll revenue per driver. Utility is money-metric such that these figures are comparable. The last column shows the ratio of toll revenue to the change in gross utility.

In the first simulation, the charge leads to a welfare gain of 0.03 but the toll revenue is 23 times larger. Thus car drivers pay a high price for a relatively small welfare gain in this situation.

The second set of simulations in Figure 7 concern a setting with a moderate level of congestion with a maximum flow of 0.17. The slope of the speed-density relationship is set such that the speed drops to about 0.4 in the base scenario, which means that hypercongestion occurs. The flow profile is a little bit bimodal, reflecting that flow drops during the middle of the peak due to hypercongestion. As under light congestion, the toll causes early travelers to complete their trips even earlier and conversely for late travelers. This reduces the density such that hypercongestion does not occur in the tolled scenario and speeds are increased substantially during the middle of the peak. The total duration of the peak is hardly affected by the toll - traffic is redistributed within the peak but it does not become much longer: early travelers depart a little bit earlier but arrive much earlier, conversely for later travelers, travelers in the middle travel much faster. The flow profile, however, changes only a little. This reflects the fact that the same flow can occur both at a high speed and at a low speed. The welfare gain from tolling is 0.85, much higher than under light congestion, and the toll revenue is 2.86 times the welfare gain, so that the price paid by car drivers is not so extreme

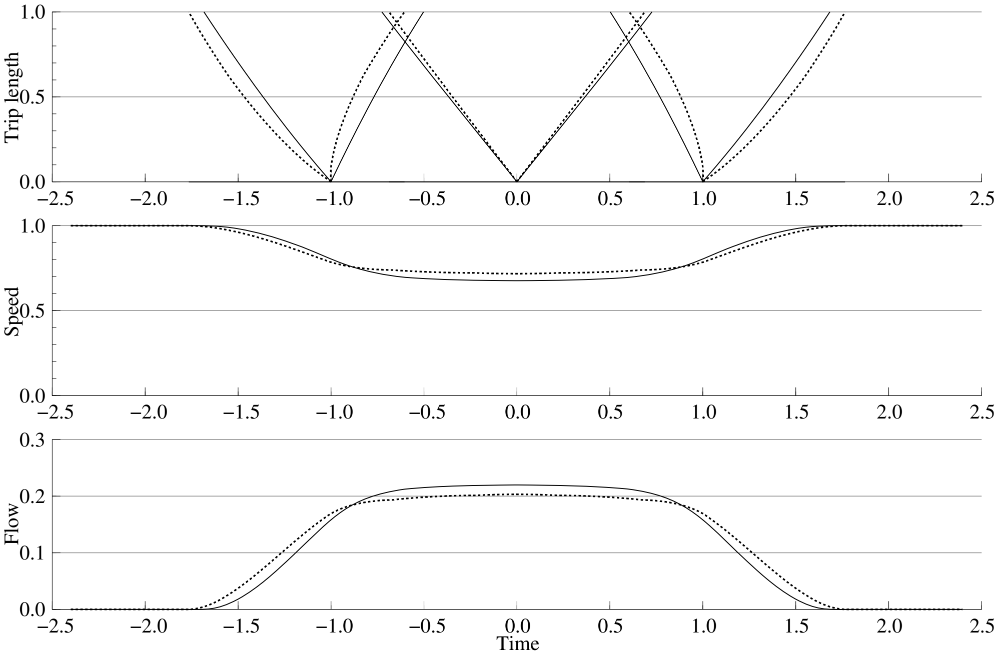

Figure 6: Simulation with heterogeneous time shifts and light congestion

relative to the welfare gain, but still higher.

The third set of simulations in Figure 8 concern a setting with a high level of congestion. The maximum flow of 0.16 is actually only slightly smaller than in the previous simulation, but the dynamics of congestion cause speed to drop substantially compared to the previous simulation. Speed now drops to around 0.25 in the base scenario and the bimodality of the flow profile is quite pronounced. With congestion this strong, the effect of the toll is quite remarkable. Early travelers arrive earlier and late travelers arrive later. This leaves room for the travelers in the middle who achieve much higher speeds and hypercongestion is removed. The travelers in the middle now occupy capacity for a much shorter time and hence they interact less with travelers on the shoulders of the peak. The speed is then increased so much for the travelers on the shoulders that some early travelers can depart *later* in the tolled scenario than in the base, even if they arrive earlier. The net result is a peak that is shorter than in the base scenario, even if the speed increase is achieved by inducing travelers to spread out in time. The welfare gain is 2.9, which is much higher than before. The toll revenue is 1.1 times the welfare

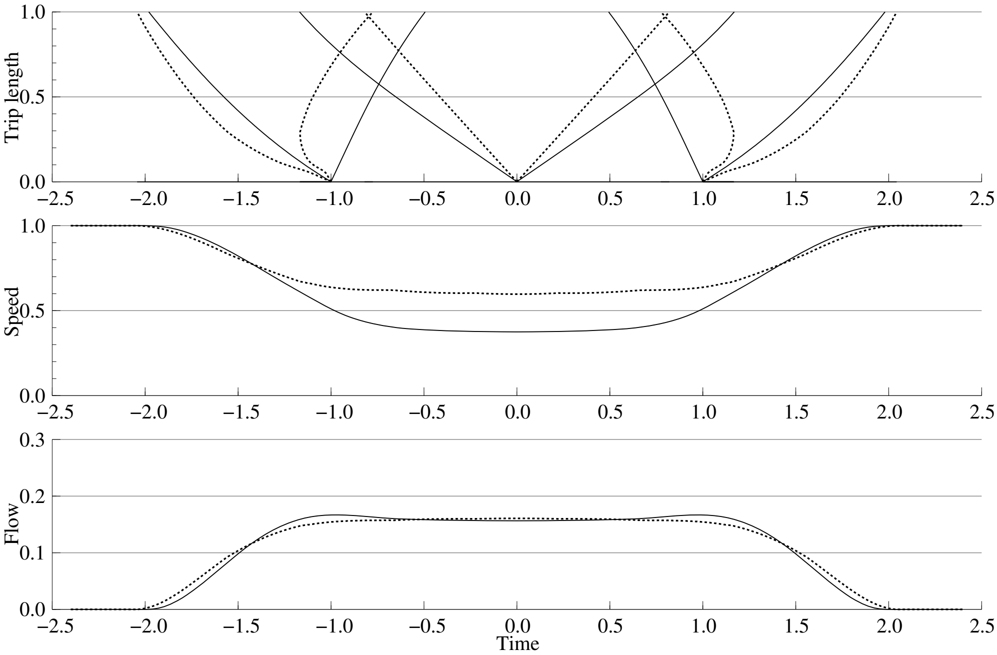

Figure 7: Simulation with heterogeneous time shifts and moderate congestion

gain so travelers would experience a net loss from tolling if toll revenues were not returned.

5 Turning hypercongestion into queues

As we have seen, congestion becomes very costly when it reaches the level of hypercongestion. Hypercongestion occurs when the density of cars becomes so large that it reduces the flow of traffic. In concrete physical terms this can happen for example at a turn lane on a freeway that leads to some bottleneck such as a traffic light. If the flow into the turn lane exceeds the capacity at the bottleneck then a queue will build. Eventually the queue will spill back on the freeway and reduce flow for traffic that does not use the turn lane. Traffic management measures for such a situation include lengthening the turn lane to avoid the queue reaching the freeway or increasing the bottleneck capacity, perhaps by increasing the green time for traffic leaving the freeway.

Another type of situation is capacity drop at a merge on a freeway (Cassidy

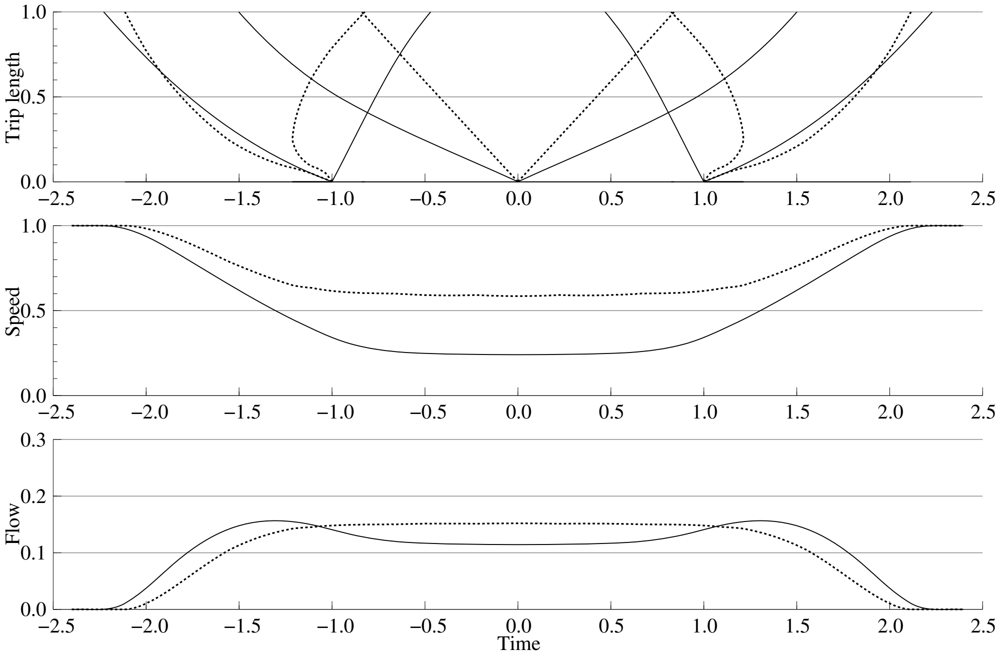

Figure 8: Simulation with heterogeneous time shifts and heavy congestion

and Rudjanakanoknad, 2005). A high level of merging traffic leads to a merge queue. This queue induces lane changing behavior, which causes the queue to spread laterally into the freeway and flow out of the merge is reduced. A traffic management strategy here is ramp metering, whereby the volume of merging traffic is metered to avoid the capacity drop. Metering moves queueing upstream of the merge and thereby avoids hypercongestion at the merge.

In urban street networks, spill backs can lead to hypercongestion when queues become so long that flows through intersections are reduced. Traffic management may reduce such problems by, e.g., using traffic lights that adapt to the length of the queues on the different roads that lead into intersections (Gershenson, 2005; de Gier et al., 2011; Gershenson and Rosenblueth, 2012). Then priority can be given flows such that intersection capacity is better utilized while spillbacks can be minimized.

These examples have in common that excessive densities of cars at critical locations impede traffic flow. The traffic management strategies to deal with such situations involve essentially directing the high densities to places where they do

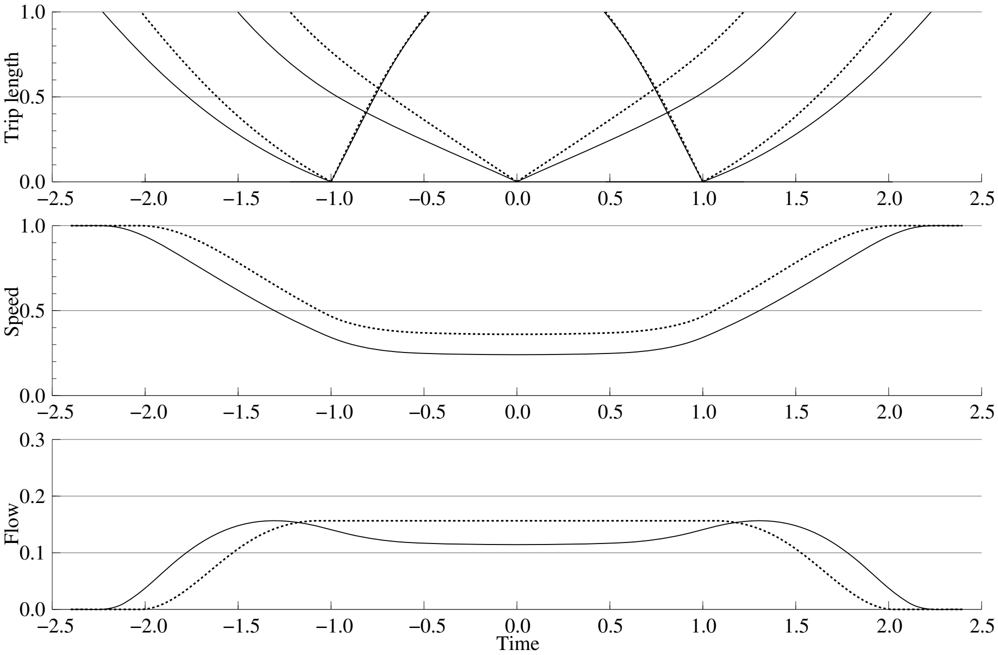

Figure 9: Simulation with heterogeneous time shifts, with and without hypercongestion

not impede traffic flow. Thus hypercongestion is converted into mere queues.

The effect of traffic management can be illustrated using the simulation model of the previous section. I will compare the most congested example to a situation in which flow does not decrease at all at high densities. This is a stark contrast that shows the potential if hypercongestion can be completely avoided through traffic management.

The simulation in this section compares two scenarios. The first is the same as shown in Figure 8 with toll applied. The second is the same except it has a modified relationship between flow and density. The relationship is the same as before at densities below the critical density. At higher densities, the relationship is modified such that flow remains equal to the maximum flow. This is achieved by modifying the speed-density relationship into

$$\psi\left(D\right)=\begin{cases}1-\gamma D, & D\leq\frac{1}{2\gamma}\\ \frac{1}{4\gamma D}, & D>\frac{1}{2\gamma}.\end{cases}$$

Then speed decreases as density increases but flow remains constant at $1/4\gamma$ at high densities.

Figure 9 shows the simulation results. Removing hypercongestion causes flow not to drop, even if speed still drops below the critical speed that would otherwise be associated with the transition to hypercongestion. Speeds improve such that the duration of the peak becomes shorter. This enables all drivers to depart later and/or arrive earlier. The welfare gain associated with the removal of hypercongestion is 1.98, which can be compared to the welfare gain from road pricing of 2.88 obtained in the third simulation above. Thus most of the welfare gain that was obtained from pricing can be obtained from traffic management given that traffic management is able to remove hypercongestion.

A possible concern with this simulation is that density might increase to a level that would be infeasible due to space constraints. It turns out in this case that removing hypercongestion actually reduces the equilibrium density at all times during the peak. There are two opposing forces in action. One force is that the speed increase induces trips to concentrate more in time, i.e. early departures will occur later and conversely for late arrivals. This force tends to increase density in the middle of the peak. The opposing force is that trips are completed faster under the higher speed, so that each trips contributes to the density for a shorter duration. The latter effect dominates in this simulation.

6 Conclusion

This paper has contributed a tractable bathtub model, employing a physically realistic description of congestion dynamics in a downtown area where a macroscopic relationship between speed and density prevails. The bathtub model allows for hypercongestion and can be used as a unified framework to analyze a range of issues. The model leads to a number of conclusions that do not arise in the bottleneck model and demonstrates the importance of taking hypercongestion into account.

There are a number of unresolved questions that may be asked of the model presented here. These questions are not easy and their answers require new insights beyond what has been established in this paper. In the model with homogeneous drivers (Section 2), we may ask what happens if Condition 1 that leads to regular sorting is violated. Does equilibrium still exist in this case? Another question, raised by Richard Arnott, is whether hypercongestion can exist in social optimum. Assuming for a moment that social optimum involves hypercongestion, then if the distribution of trip lengths allows it, it seems plausible that it is possible to rearrange the departure schedule to avoid exceeding the critical density that leads to hypercongestion while maintaining the same flow but at higher speeds and lower densities. This possibility would contradict that a departure schedule leading to hypercongestion could be socially optimal. More generally, we may ask for conclusions regarding existence and uniqueness of Nash equilibrium and social optimum. These are difficult questions due to the complexity of delay differential equations.

There are several directions in which it is relevant and perhaps feasible to extend the current model. One direction is to introduce some spatial differentiation, perhaps by allowing multiple connected bathtubs. Another is to connect to urban economics models such as the monocentric city model. There is the issue of uncertainty, since random variability is a major aspect of congestion. Finally, of course, more may be done to allow for traveler heterogeneity. I am sure these directions do not exhaust the possibilities for extension and application of the bathtub model.

References

Anderson, M. L. (2014) Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion *American Economic Review* **104**(9), 2763–2796.

Arnott, R. A. (1998) Congestion Tolling and Urban Spatial Structure *Journal of regional science* **38**(3), 495–504.

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

Arnott, R. A., 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.

Cassidy, M. J. and Rudjanakanoknad, J. (2005) Increasing the capacity of an isolated merge by metering its on-ramp *Transportation Research Part B: Methodological* **39**(10), 896–913.

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

Daganzo, C. F., Gayah, V. V. and Gonzales, E. J. (2011) Macroscopic relations of urban traffic variables: Bifurcations, multivaluedness and instability *Transportation Research Part B: Methodological* **45**(1), 278–288.

de Gier, J., Garoni, T. M. and Rojas, O. (2011) Traffic flow on realistic road networks with adaptive traffic lights *Journal of Statistical Mechanics: Theory and Experiment* **2011**(04), P04008.

de Palma, A. and Fosgerau, M. (2011) Dynamic Traffic Modeling *in* A. de Palma, R. Lindsey, E. Quinet and R. Vickerman (eds), *A Handbook of Transport Economics* Edward Elgar.

Duranton, G. and Puga, D. (2004) Micro-foundations of urban agglomeration economies *in* J. Henderson and J.-F. Thisse (eds), *Handbook of Regional and Urban Economics* Vol. Volume 4 of *Cities and Geography* Elsevier pp. 2063–2117.

Fosgerau, M. and de Palma, A. (2012) Congestion in a city with a central bottleneck *Journal of Urban Economics* **71**(3), 269–277.

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

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

Gershenson, C. (2005) Self-Organizing Traffic Lights *Complex Systems* **16**.

Gershenson and Rosenblueth, D. A. (2012) Self-organizing traffic lights at multiple-street intersections *Complexity* **17**.

Gonzales, E. J. and Daganzo, C. F. (2012) Morning commute with competing modes and distributed demand: User equilibrium, system optimum, and pricing *Transportation Research Part B: Methodological* **46**(10), 1519–1534.

Greenshields, B. D. (1935) A Study of Traffic Capacity *Proceedings Highway Research Record* **14**, 448–477.

Ji, Y. and Geroliminis, N. (2012) On the spatial partitioning of urban transportation networks *Transportation Research Part B: Methodological* **46**(10), 1639–1656.

Kuang, Y. (1993) *Delay Differential Equations: With Applications in Population Dynamics* Academic Press.

Moretti, E. (2011) Chapter 14 - Local Labor Markets *in* David Card and Orley Ashenfelter (ed.), *Handbook of Labor Economics* Vol. Volume 4, Part B Elsevier pp. 1237–1313.

Rosenthal, S. S. and Strange, W. C. (2004) Evidence on the nature and sources of agglomeration economies *in* J. V. Henderson and J.-F. Thisse (eds), *Cities and Geography* Vol. 4 Elsevier pp. 2119–2171.

Schrank, D., Eisele, B. and Lomax, T. (2012) TTI's 2012 Urban Mobility Report *Technical report* Texas Transportation Institute, Texas A&M University College Station.

Taylor, M. E. (2011) *Introduction to Differential Equations* American Mathematical Society.

Verhoef, E. T. (2003) Inside the queue:: hypercongestion and road pricing in aÂă-continuous timeâĂŞcontinuous place model of traffic congestion *Journal of Urban Economics* **54**(3), 531–565.

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

Vickrey, W. S. (1991) Congestion in Midtown Manhattan in Relation to Marginal Cost Pricing.

A Notation

Symbol	Definition
s, t	Points in time
$S(t)$	Speed at time t
$D(t)$	Density of cars at time t
$\psi(D)$	Speed-density relationship in general form, linear form $S = 1 - \gamma D$ used in simulations
$h(t), w(t)$	Utility rate at the origin and at the destination
$\alpha_0, \alpha_1, \beta_0, \beta_1$	Parameters in specific form, $h(s) = \exp(\alpha_0 - \alpha_1 s)$, $w(s) = \exp(\beta_0 + \beta_1 s)$
$H(t) = \frac{h(t)}{S(t)}, W(t) = \frac{w(t)}{S(t)}$	Utility rates in distance terms
u	Utility
τ	Toll
π	Toll rate
N	Number of drivers
l	Trip length
Φ, $(\phi = -\Phi')$	Survivor functions for trip length in the model with heterogeneous trip lengths
$\dot{x}(t) = \frac{x'(t)}{x(t)}$	Dot above a function of time denotes rate of change
a, b	Departure and arrival times

B Proofs

Proof of Theorem 2 . The first-order condition for utility maximization for a driver with trip length l taking the speed profile S as given is (5) and by Assumption 1 the corresponding second-order condition (footnote 9) is satisfied with strict inequality.

The first-order condition holds for all l and is equivalent to $H(a(l)) = W(b(a(l), l))$. From (6), Assumption 1 implies that $a'(l) < 0 < b'(l)$ for all l such that regular sorting applies. Using equation (8), this can be simplified to obtain the expression in equations (7) stated in the theorem. Finally, equations (7) is a pair of differential solutions with initial condition $a(0) = b(0) = 0$. This system has a unique solution given our requirements on h, w (Taylor, 2011). ■

Proof of Theorem 6. Consider some fixed trip length l. The first-order condition for the choice of departure time $a\left(l|c\right)$ is

$$0=\frac{h\left(a\left(l|c\right)-c\right)}{S\left(a\left(l|c\right)\right)}-\frac{w\left(b\left(a\left(l|c\right),l\right)-c\right)}{S\left(b\left(a\left(l|c\right),l\right)\right)}.$$

For visual clarity, write a for $a\left(l|c\right)$ and b for $b\left(a\left(l|c\right),l\right)$. The second-order condition is (using the first-order condition)

$$SOC\equiv\frac{h\left(a-c\right)}{S\left(a\right)}\left(\dot{h}\left(a-c\right)-\dot{S}\left(a\right)-\frac{S\left(a\right)}{S\left(b\right)}\left(\dot{w}\left(b-c\right)-\dot{S}\left(b\right)\right)\right)<0.$$

Differentiate the first-order condition with respect to c to find

$$0=SOC\frac{\partial a}{\partial c}-\frac{h\left(a-c\right)}{S\left(a\right)}\left(\dot{h}\left(a-c\right)-\dot{w}\left(b-c\right)\right),$$

which yields

$$\frac{\partial a}{\partial c}=\frac{\dot{h}\left(a-c\right)-\dot{w}\left(b-c\right)}{\dot{h}\left(a-c\right)-\dot{S}\left(a\right)-\frac{S(a)}{S(b)}\left(\dot{w}\left(b-c\right)-\dot{S}\left(b\right)\right)},$$

which is strictly positive as required.

Now fix c, differentiate the first-order condition with respect to l and rearrange to find

$$\begin{aligned}\frac{\partial a}{\partial l}&=\frac{\dot{w}\left(b-c\right)-\dot{S}\left(b\right)}{S\left(b\right)\left(\dot{h}\left(a-c\right)-\dot{S}\left(a\right)\right)-S\left(a\right)\left(\dot{w}\left(b-c\right)-\dot{S}\left(b\right)\right)}\\ \frac{\partial b}{\partial l}&=\frac{\dot{h}\left(a-c\right)-\dot{S}\left(a\right)}{S\left(b\right)\left(\dot{h}\left(a-c\right)-\dot{S}\left(a\right)\right)-S\left(a\right)\left(\dot{w}\left(b-c\right)-\dot{S}\left(b\right)\right)}.\end{aligned}$$

Then there is regular sorting for each c under the assumptions of the Theorem. ■

Prof of Theorem 3. The number of car drivers is strictly between 0 and N, since otherwise somebody could change mode and gain. The car drivers do not interact with the transit users, and so Theorem 1 and Theorem 2 apply to them. Then the road speed profile S is U-shaped and reaches a minimum for the duration of the shortest car trip where the density is equal to the mass of car drivers. The minimum car driver speed is greater than the transit speed. Due to sorting, a driver with a longer trip experiences higher average speed than the shortest car trip. Hence all travelers with longer trips than the minimum car trip length will go by car. and $\psi^{-1}\left(S_T\right)=\Phi\left(l_T\right)$. ■

Proof of Theorem 4. Consider equilibrium with transit where all travelers with trip length less than l_* use transit and all others use car. Denote by $D_* = \Phi(l_*)$ the number of car drivers. We shall investigate the effect of changing the threshold l_*.

Let a_C, b_C denote departure and arrival schedules for car drivers and let a_T, b_T be those for transit users. Aggregate welfare is

$$W = \int_0^{l_*} u\left(a_T(l), b_T(l)\right)\phi(l)\,dl + \int_{l_*}^{\infty} u\left(a_C(l), b_C(l)\right)\phi(l)\,dl.$$

Departures are optimally chosen so we may appeal to enveloping and ignore the effect of l_* on departures. Arrivals are given by physics. A marginal increase in l_* of dl_* shifts a mass of $\phi(l_*)\,dl_*$ out of car into transit. The speed for the remaining car drivers increases by $-\psi'(\Phi(l_*))\phi(l_*)\,dl_*$ during the interval $[a_C(l_*), b_C(l_*)]$, such that they cover the distance l_* earlier by $-\psi'(\Phi(l_*))\phi(l_*)(b_C(l_*) - a_C(l_*))\,dl_*$ time units. Hence,

$$\frac{\partial b_C(l_*)}{\partial l_*} = \psi'(\Phi(l_*))\phi(l_*)(b_C(l_*) - a_C(l_*)).$$

Denote $\Delta u = u(a_T(l_*), b_T(l_*)) - u(a_C(l_*), b_C(l_*))$ and differentiate the welfare measure with respect to l_* to find

$$\begin{aligned}\frac{\partial W}{\partial l_*} &= \phi(l_*)\Delta u + \int_0^{l_*}\frac{\partial}{\partial l_*}\left(\int_{b_T(l)}^{0} w(s)\,ds\right)\phi(l)\,dl + \int_{l_*}^{\infty}\frac{\partial}{\partial l_*}\left(\int_{b_C(l)}^{0} w(s)\,ds\right)\phi(l)\,dl \\ &= \phi(l_*)\Delta u - \int_0^{l_*} w(b_T(l))\frac{\partial b_T(l)}{\partial l_*}\phi(l)\,dl - \int_{l_*}^{\infty} w(b_C(l))\frac{\partial b_C(l)}{\partial l_*}\phi(l)\,dl\end{aligned}$$

The travel time for transit users is unaffected, so this reduces to

$$\frac{\partial W}{\partial l_*} = \phi(l_*)\Delta u - \int_{l_*}^{\infty} w(b_C(l))\frac{\partial b_C(l)}{\partial l_*}\phi(l)\,dl.$$

Inserting the expression for $\frac{\partial b_C(l)}{\partial l_*}$ leads to

$$\frac{\partial W}{\partial l_*} = \phi(l_*)\Delta u - \psi'(\Phi(l_*))\phi(l_*)(b_C(l_*) - a_C(l_*))\int_{l_*}^{\infty} w(b_C(l))\phi(l)\,dl.$$

When $l_* = l_T$, then $\Delta u = 0$ and so $\frac{\partial W}{\partial l_*} \geq 0$. If also $\psi'(\Phi(l_T)) < 0$ then $\frac{\partial W}{\partial l_*} > 0$.

It remains to be shown that the regular welfare optimum may be implemented through a fixed charge on car drivers. So consider the welfare optimum. Car drivers and transit users do not interact, so the group of car drivers must also be in welfare optimum when considered in isolation. By Theorem 2, car drivers must be in Nash equilibrium. A similar argument shows that transit users must also be in Nash equilibrium. It must also be the case that car drivers have longer trips than transit users. Then it only remains to see that a fixed charge τ on car users may be set to achieve to optimum number of car drivers and transit users. ■

Proof of Theorem 5. We shall consider the case of a constant rate toll π_3, the case of a fixed charge τ is completely analogous. Utility is given by (2). Regular sorting holds among car drivers as before and Theorem 2 remains valid, since $\pi_3' = 0$. Equilibrium utility is then decreasing in trip length. This implies that there is a threshold trip length l_* that separates car drivers from those who choose the outside option. The threshold decreases as π_3 increases.

A marginal increase in l_* of dl_* shifts $\phi(l_*)\,dl_*$ drivers onto the road which affects all drivers with shorter trips. Note that

$$\begin{aligned}
b'(l) - a'(l) &= \frac{1}{\psi(\Phi(l_*) - \Phi(l))} \\
b(l) - a(l) &= \int_0^l \frac{1}{\psi(\Phi(l_*) - \Phi(l'))} dl' \\
\frac{\partial(b(l) - a(l))}{\partial l_*} &= \phi(l_*) \int_0^l \frac{\psi'(\Phi(l_*) - \Phi(l'))}{\psi(\Phi(l_*) - \Phi(l'))^2} dl'
\end{aligned}$$

and this is negative when $\psi'(\Phi(l_*) - \Phi(l')) < 0$.

Welfare is

$$W = \int_0^{l_*} u(a(l), b(l))\,\phi(l)\,dl + \Phi(l_*)\,u_0.$$

Differentiate welfare to get

$$\frac{\partial W}{\partial l_*} = [u(a(l_*), b(l_*)) - u_0]\,\phi(l_*) - \int_0^{l_*} h(a(l))\,\frac{\partial(b(l) - a(l))}{\partial l_*}\phi(l)\,dl. \quad (9)$$

Evaluate at the no toll equilibrium point where $u(a(l_*), b(l_*)) = u_0$. The first term in (9) is then zero such that

$$\frac{\partial W}{\partial l_*}\Big|_{u(a(l_*),b(l_*))=u_0} < 0.$$

This establishes that it is welfare improving to price some drivers off the road compared to no-toll equilibrium. The conclusion that welfare optimum is achieved for some toll rate π_3 follows from 2. ■

C Simulation details

The simulations specify h and w as exponential functions

$$h(s) = \exp(\alpha_0 - \alpha_1 s), w(s) = \exp(\alpha_0 + \beta_1 s), \tag{10}$$

where $\alpha_1, \beta_1 > 0$, then $\dot{h}(s) = h'(s)/h(s) = -\alpha_1$ and $\dot{w}(s) = w'(s)/w(s) = \beta_1$. Thus this specification satisfies the requirements made on h and w.

The simulations use $N = 1$, $\alpha_0 = 0$, $\alpha_1 = \beta_1 = 2$, trip lengths that are uniformly distributed on $[0, 1]$ and the speed density relationship $\psi(D) = 1 - \gamma D$. The speed-density relationship leads to a maximal speed drop of γ.

Proposition 1 *Assume that*

$$\frac{\alpha_1 \beta_1}{\alpha_1 + \beta_1} > -\psi'(\Phi(l))\phi(l). \tag{11}$$

The Nash equilibrium in the bathtub model with homogeneous drivers and the specification just stated is given by

$$\begin{aligned} a(0) &= b(0) = 0, \\ a'(l) &= -\frac{1}{\psi(\Phi(l))}\frac{\beta_1}{\alpha_1 + \beta_1}, b'(l) = \frac{1}{\psi(\Phi(l))}\frac{\alpha_1}{\alpha_1 + \beta_1} \end{aligned}$$

Proof. There will be sorting by construction and hence the speed is $S(a(l)) = S(b(l)) = \psi(\Phi(l))$. It is straightforward that $l = \int_{a(l)}^{b(l)} S(t)\,dt$. The first-order condition for utility maximization, $h(a(l)) = w(b(l))$ is equivalent to $0 = \alpha_1 a(l) + \beta_1 b(l)$, which holds by the definition of $a'(l)$ and $b'(l)$.

Note that (omitting some function arguments)

$$\begin{aligned} S(a(l)) &= S(b(l)) = \psi(\Phi(l)), \\ S'(a(l)) &= -\frac{\psi'(\Phi(l))\phi(l)}{a'(l)}, \\ S'(b(l)) &= -\frac{\psi'(\Phi(l))\phi(l)}{b'(l)}. \end{aligned}$$

Then Assumption 1 is satisfied since

$$\begin{array}{rcl} 0 & > & \dot{H}\left(a\left(l\right)\right) \\ & \Updownarrow & \\ \dfrac{\alpha_1\beta_1}{\alpha_1+\beta_1} & > & -\psi'\left(\Phi\left(l\right)\right)\phi\left(l\right) \\ & \Updownarrow & \\ 0 & < & \dot{W}\left(b\right) \end{array}$$

and these inequalities hold by the assumption in equation (11). The second-order condition for utility maximization is implied by Assumption 1. ■

The simulation in Section 2 uses this proposition to compute the Nash equilibrium. The social optimum is computed using an iterative procedure, where the derivative of the departure schedule $a'\left(l\right)$ is given by a sixth-order polynomial and parameters of this polynomial are found to maximize the average utility of drivers. Note that regular sorting is not imposed in the computation of social optimum. This simulation has a maximum speed drop of 60%, which satisfies assumption (11) and leads to a situation in which the social optimum is not regularly sorted.

The simulation in Section 3 is based on the simulation in Section 2 with a few modifications. First, the threshold trip length separating car drivers and transit users is calculated numerically. This is feasible due to the sorting of departures within car drivers and transit users. Departure and arrival schedules for transit users are computed directly. The speed for the marginal car driver is calculated appealing to sorting. The departure and arrival schedules for car drivers with longer trip lengths are computed by integrating the differential equations for a' and b' in Proposition 1.

In Section 4, the location of scheduling preferences is heterogeneous with a uniform distribution on the interval $[-1, 1]$. The slope of the speed-density relationship is varied across simulations to produce different levels of congestion. Given a speed profile, the departure schedule for drivers is computed using the first-order condition for utility maximization. Then the speed profile is updated given the departure schedule and this is iterated until convergence. The second-order condition is verified at convergence for all simulations. Then the simulations do not rely on regular sorting and they are hence valid even if regular sorting does not result.

D The marginal utility of travel time

In order to compare changes in travel times to changes in utility, it is useful to have an expression for the marginal utility of travel time. So we would like to consider the marginal utility associated with an increase in travel time for a trip beginning at time a and ending at time b. However, both the departure time and the arrival time are flexible, so we need to decide how they both change as travel time increases.

First, travel time is $b - a$ such that a and b are related by $b' - a' = 1$ as travel time increases.

The first-order condition for utility maximization, $h\left(a\right)/S\left(a\right) = w\left(b\right)/S\left(b\right)$, states that the marginal utility associated with increasing distance is the same at the beginning and the end of the trip. Differentiating this and solving leads to

$$a' = \frac{\left(\dot{w} - \dot{S}\left(b\right)\right)}{\left(\dot{h} - \dot{S}\left(a\right)\right) - \left(\dot{w} - \dot{S}\left(b\right)\right)}.$$

We use this expression also for trips that are non-optimally timed, since this maintains the difference between the marginal utilities associated with increasing distance. Differentiating gross utility, $u\left(a, b, 0\right) = \int_0^a h\left(s\right) ds + \int_b^0 w\left(s\right) ds$, under these conditions leads to the following marginal utility of increased travel time:

$$\begin{aligned} u'\left(a, b, 0\right) &= \left(h\left(a\right) - w\left(b\right)\right) a' - w\left(b\right) \\ &= \frac{h\left(a\right)\left(\dot{w} - \dot{S}\left(b\right)\right) - w\left(b\right)\left(\dot{h} - \dot{S}\left(a\right)\right)}{\left(\dot{h} - \dot{S}\left(a\right)\right) - \left(\dot{w} - \dot{S}\left(b\right)\right)}. \end{aligned}$$

Specialize this to exponential utility rates (10) to find that the marginal utility of increased travel time is

$$u'\left(a, b, 0\right) = -e^{a_0} \frac{e^{-\alpha_1 a}\left(\beta_1 - \dot{S}\left(b\right)\right) + e^{\beta_1 b}\left(\alpha_1 + \dot{S}\left(a\right)\right)}{\left(\beta_1 - \dot{S}\left(b\right)\right) + \left(\alpha_1 + \dot{S}\left(a\right)\right)}.$$