TI 2002-062/3
Tinbergen Institute Discussion Paper

Inside the Queue

Erik T. Verhoef

Dept of Spatial Economics and Business Administration, Vrije Universiteit Amsterdam, and Tinbergen Institute.

Tinbergen Institute
The Tinbergen Institute is the institute for economic research of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam, and Vrije Universiteit Amsterdam.

Tinbergen Institute Amsterdam
Roetersstraat 31
1018 WB Amsterdam
The Netherlands
Tel.: +31(0)20 551 3500
Fax: +31(0)20 551 3555

Tinbergen Institute Rotterdam
Burg. Oudlaan 50
3062 PA Amsterdam
The Netherlands
Tel.: +31(0)10 408 8900
Fax: +31(0)10 408 9031

Please send questions and/or remarks of non-scientific nature to wdriessen@few.eur.nl.
Most TI discussion papers can be downloaded at http://www.tinbergen.nl.

INSIDE THE QUEUE

Hypercongestion and Road Pricing in a Continuous Time – Continuous Place Model of Traffic Congestion*

Erik T. Verhoef**
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
E-mail: everhoef@econ.vu.nl
http://www.econ.vu.nl/medewerkers/everhoef/et.html

This version: 27/05/03

Key words: congestion, road pricing, networks
JEL codes: R41, R48, D62

Abstract
This paper develops a continuous-time – continuous-place economic model of road traffic congestion with a bottleneck, based on car-following theory. The model integrates two archetype congestion technologies used in the economics literature: 'static flow congestion', originating in the works of Pigou, and 'dynamic bottleneck congestion', pioneered by Vickrey. Because a closed-form analytical solution of the formal model does not exist, its behaviour is explored using a simulation model. In a setting with endogenous departure time choice and with a bottleneck along the route, it is shown that 'hypercongestion' can arise as a dynamic – transitional and local – equilibrium phenomenon. Also dynamic toll schedules are explored. It is found that a toll rule based on an intuitive dynamic and space-varying generalization of the standard Pigouvian tax rule can hardly be improved upon. A naïve application of a toll schedule based on Vickrey's bottleneck model, in contrast, appears to perform much worse and actually even reduces welfare in the numerical model.

*The author would like to thank Jan Brueckner, Robin Lindsey, Se-Il Mun, Ken Small, and two anonymous referees for helpful comments on an earlier draft. Any remaining deficiencies, of course, are the author's responsibility alone.

**The author is affiliated to the Tinbergen Institute, Roetersstraat 31, 1018 WB Amsterdam. This research has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

1. Introduction

The consistent and realistic economic modelling of road traffic congestion is a challenging task. Many different modelling approaches have been proposed (see Lindsey and Verhoef, 2000, for an overview). One reason for this variety is that different (policy) questions often require different types of models; compare for example the design of optimal signalling at a complicated junction with questions involving long-run spatial planning. Another reason is that traffic congestion in reality is a complex phenomenon, involving complicated temporal and spatial dynamics. Especially economists, however, often seek manageable analytical formulations to characterize the problem studied and to derive (optimal) policy rules. As different economic models make different simplifications from reality, different and sometimes contrasting insights and policy recommendations may result. This even holds for what is probably the simplest set-up, also considered in this paper, where identical users – typically commuters – use a single road to get from a single origin to a single destination.

This paper aims to make two contributions to the literature on traffic congestion in this simple setting. A first goal is to identify the optimal toll schedule for a road on which traffic congestion is a temporarily and spatially differentiated non-stationary state phenomenon, as it typically is in reality. For that purpose, a continuous-time – continuous-place congestion technology is considered, where drivers' speeds may vary continuously over time and place depending on traffic conditions, even when being in a jam. Although no closed-form solutions for the model and for its optimal prices can be derived, a simple pricing rule is proposed that appears to be 'nearly' optimal according to a heuristic analysis of alternative toll schedules. This toll rule is an intuitive dynamic and space-varying generalization of the optimal (Pigouvian) tax rule for static models. An important further issue addressed concerns the extent to which insights on optimal tolling from earlier models, employing simpler assumptions on congestion technology, can be used for designing congestion toll schedules in this more elaborate setting. Surprisingly, a naïve application of a toll schedule based on Vickrey's (1969) bottleneck model appears to perform much worse than the simple pricing rule developed, and actually even reduces welfare in the numerical model used.

A second goal is to contribute to the understanding of 'hypercongestion' (explained in more detail below), which is an important topic of debate in the economic literature on traffic congestion. The paper will demonstrate that hypercongestion will arise as a dynamic – transitional and local – equilibrium phenomenon on a road with a downstream bottleneck of sufficiently small capacity, provided demand is sufficiently high. The standard static economic model of traffic congestion in contrast typically cannot explain whether or not hypercongestion will occur in equilibrium – that is, whenever the inverse demand function intersects the backward-bending average cost function both in its normally congested and in its hypercongested segment. The dynamic extension of the standard static model presented here therefore does not share this disturbing indeterminacy with the original static model.

The paper starts with a brief literature review, and proceeds with a description of the model in Section 3. Section 4 considers the free-market ('no-toll') equilibrium, while Section 5 is concerned with congestion pricing. Section 6 concludes.

2. Prior literature

Two main strands of economic[1] modelling approaches can be distinguished for studying traffic congestion on a single road and identical users, the set-up also considered in this paper: static and dynamic models.

Static stationary-state economic models have in common that traffic speeds, flows, and densities are assumed to be constant along the road and – as it were – over time, in the sense that they do not change during a driver's trip, nor differ between drivers. Following Pigou (1920), these models are typically implicitly or explicitly based on what engineers call the 'fundamental diagram of road traffic congestion', which shows how stationary state speed falls monotonously with traffic density. The models proposed differ particularly in terms of the argument deployed in the specification of cost and inverse demand functions used for the characterization of free-market and efficient (toll-supported) market equilibria. Proposed arguments include traffic flow (*e.g.* Walters, 1961), the number of trips (*e.g.* Hills, 1993), traffic density (*e.g.* Evans, 1992; and recently Ohta, 2001) and the rate of trips started (Verhoef, 1999). The different approaches followed for instance give different answers to the question of whether 'hypercongestion' can arise as a stationary state equilibrium phenomenon on a single road of constant capacity, and even on the question of whether hypercongestion can possibly be economically efficient.[2] A hypercongested equilibrium in a stationary state model is one in which the traffic density is so high, and consequently the speed is so low, that the traffic flow (the product of speed and density in a stationary state) is below the maximum possible flow for the road (its capacity), and where the speed is below the maximum possible speed at that traffic flow. Figure 1-II in Section 2.2 below will illustrate this. This phenomenon is still under heavy debate, and also this paper will pay attention to hypercongestion, albeit in the context of a dynamic non-stationary state model.

Dynamic economic models typically have in common that road users have a desired arrival time (at work), deviations from which imply that schedule delay costs will be incurred. With identical users, a dynamic equilibrium in terms of departure time choices must then entail constancy of total trip costs during the peak; *i.e.*, the sum of travel delay costs, schedule delay costs and tolls (if levied) must be equal for all times at which departures occur, and higher otherwise (the essence of dynamic equilibrium does not change when user heterogeneity is introduced, but the analytics may become complicated; see Lindsey, 2001).

Dynamic models in particular differ with respect to the 'congestion technology' considered. Probably the most widely used is what will be called 'pure bottleneck congestion'

---

[1] The term 'economic model' is here used loosely to identify (static or dynamic) equilibrium models of traffic congestion designed to analyze the efficiency of congested road use and the impacts of policies (typically pricing) upon it.

[2] See also Small and Chu (1997) and Verhoef (1999). In brief, a hypercongested equilibrium appears feasible on a single link of constant capacity when density or flow are used as arguments. However, it appears dynamically unstable (there are no equilibrium paths towards such an equilibrium from any other feasible initial equilibrium), and hence irrelevant as a candidate stationary-state market equilibrium, when the rate of trips started is used as the argument in a dynamic extension of the flow-based formulation, where flows are determined endogenously as a result of the history of the rate of trips started (Verhoef, 1999, 2001). Hypercongestion can never be efficient in a flow-based formulation (Small and Chu, 1997), but it has been claimed to be possibly efficient in a density-based formulation (Ohta, 2001).

in the sequel. This involves a bottleneck with a 'kinked' performance function: for arrival rates of vehicles not exceeding its capacity, and in absence of a queue, the bottleneck's outflow is equal to its inflow and no congestion occurs. In all other instances, a queue will grow or shrink at a rate equal to the difference in the arrival rate at the back of the queue and the bottleneck's capacity. The queue in such models is typically 'vertical': it takes up no road space, and vehicles exit the queue on a first-in-first-out basis. Vehicles exit the queue at a constant rate equal to the bottleneck's capacity, speed is not defined in the spaceless queue, and there is no interaction between the queue and upstream traffic approaching it. Originally proposed by Vickrey (1969), this congestion technology has been made popular especially by the work of Arnott, De Palma and Lindsey (1993, 1998), Braid (1989), and Small (1992).

An alternative dynamic congestion technology is based on flow congestion, and determines a vehicle's speed – assumed constant during the trip – as a function alone of the flow at the road's entrance at the instant the trip is started (Henderson, 1974) or at the road's exit at the instant the trip is ended (Chu, 1995). This approach was called a 'no-propagation' model in Lindsey and Verhoef (2000), to reflect that it does not consider possible interactions between vehicles that start their trips at even slightly different instants, even if the distance between them is changing during the trip. There is therefore no propagation of shock-waves. Chu (1995) compared the pure bottleneck and no-propagation flow-based dynamic model for identical users with linear schedule delay cost functions, and identified four important differences: with flow-congestion, (i) in the free-market equilibrium, the ratio of total travel delay costs to total schedule delay costs is unequal to unity; (ii) optimal tolls save a smaller fraction of total variable costs (*i.e.*, less than 50%); (iii) the period of arrivals becomes longer with tolling compared to no-tolling; and (iv) the private costs including tolls is higher in the optimum than in the no-toll equilibrium (Chu, 1995, p. 340-341).

The other extreme of 'instantaneous propagation' was considered by Agnew (1977) and Mahmassani and Herman (1984). This entails the assumption that densities and speeds are uniform along the roadway at every instant. An implication is that an increased inflow at road's entrance would immediately slow down traffic along the entire road, even near its exit.

Even within the two families of static and dynamic economic models of road traffic congestion for identical users on a single road, important differences thus exist between them in terms of modelling characteristics and economic policy prescriptions. Between these families, evidently, even bigger differences may exist.[3] The sharp differences in insights obtained with models that basically attempt to describe the same phenomenon are due to the fact that different models make different simplifications from reality. It is therefore natural to direct research efforts to the construction of models that simultaneously capture more of the complexities of congestion. This enables investigation of the extent to which certain conclusions from certain models are robust, as opposed to dependent on specific simplifying assumptions. Although the dynamic dimension of traffic congestion has received increasing

[3] For instance, a static model is by definition incapable of showing how optimal tolls should vary over time unless demands and travel times in all time periods are assumed independent, and predicts zero efficiency gains when demand is perfectly inelastic. Optimal pricing in for instance the pure bottleneck model in contrast requires a continuously varying toll, which yields efficiency gains completely independent of the prevailing demand elasticity: trip prices with and without optimal tolling are equal, and hence overall demand (over the entire peak) will not change with optimal tolling (Vickrey, 1969, assumed perfectly inelastic demand).

attention in the economic literature over the last decades, the spatial dimension has been largely neglected – not insofar as network aspects of congestion and congestion pricing are concerned (*e.g.* Verhoef, 2002), but in particular insofar as spatial interactions on the links are concerned. This neglect even led Mun (1999) to state that: "The traffic jam has not been successfully treated in the literature of transport economics" [p. 323]. The following section presents a model that aims to simultaneously capture temporal and spatial dynamics in an economic model of traffic congestion. The traffic jams occurring in this model are no 'black box' queues as in previous studies. Instead, the model explicitly describes how drivers behave when being in a jam and when approaching it, and investigates the implications for the efficiency of congested traffic and road pricing.

3. The model

The model developed in this paper is based on a simple form of car-following theory, explored earlier in Verhoef (2001). The 'network' is now extended from a single constant-capacity road, with possibly a vertical queue before its entrance, to one containing a bottleneck that is due to a decrease in the number of lanes. Hence, the queue possibly arising before the bottleneck is now modelled explicitly. Furthermore, departure time decisions are endogenized, based on trade-offs between schedule delay costs and travel delay costs, as in the other dynamic models just discussed. This section presents the details of the model. First the demand side is discussed, then the cost side in Section 3.2, followed by equilibrium issues in Section 3.3, and finally a brief comparison with other economic models in Section 3.4.

3.1. Demand side

The model considers identical users that wish to travel from a single origin to a single destination. In contrast to many other models, these individuals and their vehicles are not treated as a continuum, but instead as discrete entities. Otherwise, the demand side is modelled in exactly the same way as in most prior dynamic economic models (see for instance Arnott *et al.*, 1998; Chu, 1995; Mun, 1999). Individuals exhibit price-taking behaviour. Although elasticity of demand could be considered, it is assumed that demand is perfectly inelastic so that the number of individuals, N, is given (Table 1 summarizes notation), and the only decision that individuals make is their departure time. For the numerical model uses N=2500 (5000 for some exercises). The travel costs c incurred by an individual user consist of two components: the travel time costs c^{tt} associated with the time spent on the road, and the schedule delay costs c^{sd} associated with arrival at the destination at a time different from the preferred arrival time t^*. In addition to these 'real' costs, a toll τ may apply. The trip price p considered by an individual is defined as the sum of c and τ.

As customary, constant shadow prices for travel time, time early and time late are assumed to apply, denoted α, β and γ, respectively. Empirical evidence and/or technical, equilibrium-related considerations have led most analysts to set $\beta<\alpha<\gamma$, an assumption that here too will be made. Maintaining the 1-2-4 ratio between these three parameters approximately found by Small (1982), and setting α equal to the Dutch average of 7.5, leads to β=3.75 and γ=15 (all prices in this paper are in €). For notational convenience, the desired arrival time t^* is set at t=0.

Symbol	Description	Value in numerical model
Latin		
c^{sd}	Schedule delay cost (on a per-user basis)	
c^{tt}	Travel time costs (on a per-user basis)	
c	Travel costs (on a per-user basis): $c = c^{tt}+c^{sd}$	
D	Traffic density	
$D^{\#}$	Density consistent with the maximum flow F_{max} for stationary states	0.055 veh./m
F	Traffic flow	
F_{max}	Maximum flow for stationary states	0.965 veh./s
i	Index for individuals (order of departure)	
N	Number of travellers during the peak	2500 (in some exercises 5000)
p	Trip price (on a per-user basis): $p = c+\tau = c^{tt}+c^{sd}+\tau$	
S	Speed	
S^*	Free-flow (maximum) speed	33 ⅓ m/s
$S^{\#}$	Speed consistent with the maximum flow F_{max} for stationary states	17.551 m/s
t	Clock time	
t_A	Arrival time	
t_D	Departure time	
t_1	Instant x_1 is passed	
t_2	Instant x_2 is passed	
t_q	Instant at which a driver obtains his minimum speed	
t^*	Most preferred arrival time	0
tt	Travel time	
x	Position along the road	
X	Length of the road	30 000 m
x_1	Beginning of the bottleneck	9 000 m
x_2	Ending of the bottleneck	11 000 m
Greek		
α	Per-unit-of-time cost of travel time	7.5 €/hr
β	Per-unit-of-time cost of early arrival	3.75 €/hr
γ	Per-unit-of-time cost of late arrival	15 €/hr
δ	Distance from the leading vehicle	
δ_{min}	Minimum distance between cars, for which speed falls to 0	5 m
δ^*	Minimum distance between vehicles consistent with a speed S^*	100 m
$\delta^{\#}$	Distance between vehicles consistent with the maximum flow F_{max} for stationary states	18.195 m
ρ_D	Departure rate	
ρ_A	Arrival rate	
τ	Toll	

Table 1. Notation of key variables

The trip price as a function of arrival time t_A can then be written as:

$$p(t_A) = c(t_A) + \tau(t_A) = c^{tt}(t_A) + c^{sd}(t_A) + \tau(t_A)$$
$$= \begin{cases} \alpha \cdot tt(t_A) - \beta \cdot t_A + \tau(t_A) & \text{for } t_A \le t^* = 0 \\ \alpha \cdot tt(t_A) + \gamma \cdot t_A + \tau(t_A) & \text{for } t_A > t^* = 0 \end{cases} \quad (1)$$

3.2. Supply side

3.2.1. Car-following congestion technology

The congestion technology used in the present model is based on that developed in Verhoef (2001), who presented a first-order car-following model as the simplest plausible dynamic extension of the standard static model of road congestion. This standard static model considers stationary state equilibria only, and characterizes these in terms of traffic flow F, density D and speed S, which are all presupposed to be constant over space (along the road) and time (during a trip and between trips). Drivers' behaviour in this standard model is represented by the fundamental diagram of traffic congestion, which shows how stationary state speed falls with stationary state density, thus giving rise to the well-known backward-bending speed-flow function. The dynamic extension of the model in Verhoef (2001) was aimed at describing transitional phases, in order to investigate the dynamic stability of the stationary state equilibria identified in the standard model.[4] This required the model to be dynamic, and to allow for traffic densities, speeds and flows that vary over time and along the road. This was realized by first transforming the 'density-speed' relation used in the standard static model into an equivalent 'distance-speed' relation, where the distance δ concerns a driver's distance from the car in front of him; his 'leader' in car-following terminology. For stationary states, $\delta \equiv 1/D$; hence equivalent functions $S(D)$ and $S(\delta)$ are easily established. By assuming that a driver uses the same function $S(\delta)$ to choose a speed also in transitional phases, a continuous-time – continuous-place dynamic model of road traffic congestion results that is equivalent to the standard static model for stationary states, but that can also be used for non-stationary state analysis. Verhoef (2001) considered stability analysis as one application of such non-stationary state analysis; the present paper considers the dynamic equilibrium modelling of peak congestion.

The basic single-lane model thus uses a very simple car-following (first-order differential) equation of the type:

$$S_i \equiv \dot{x}_i = S(\delta_i) \equiv S(x_{i-1} - x_i) \tag{2}$$

where a dot denotes a time-derivative, subscript i refers to individuals indexed by order of departure, and x denotes place (along the road). The function $S(\cdot)$ is assumed to be continuous; $0 \le S_i \le S^*$, $S_i = 0$ for $\delta_i \le \delta_{min}$ and $S_i > 0$ otherwise; $0 \le S' < \infty$, and $S'(\cdot) = 0$ for $S(\cdot) = S^*$ (S^* represents the maximum speed; δ_{min} the minimum possible distance between cars' fronts; and a prime a derivative). Equation (2) is different from standard car-following models, for which acceleration rather than speed is – admittedly more realistically – assumed to be the variable under instantaneous control of the driver, and that have a more complicated structure allowing for instance for reaction time lags (see for instance Lindsey and Verhoef, 2000). Although the numerical model used could in principle handle more complicated and higher-order types of car-following equations, it was decided to nevertheless employ a formulation based on (2).

[4] It was found that hypercongested equilibria are dynamically unstable (in contrast to normally congested equilibria) in the following sense: starting from any stationary state equilibrium (hypercongested or normally congested) different from the hypercongested equilibrium under investigation, a change of the rate of trips started at the entrance of the road to a level equal to the flow in the hypercongested equilibrium investigated will *not* result in that hypercongested equilibrium being approached (also not asymptotically).

Most importantly, this means that the model can still be seen as a direct continuous-time – continuous-place extension of the standard static model, so that for instance conclusions to be drawn with respect to hypercongestion and optimal tolling can be attributed to these two extensions, without having to worry about the possible implications of reaction time lags and/or the specific assumptions made on the acceleration and deceleration capabilities of vehicles. In further defence of (2), observe that just as in the classical car-following model, also with (2) the acceleration or deceleration of a driver depends on the differences in distance and speed between him and his leader. For instance, if a driver is driving slower than his leader, the distance between them is increasing, and therefore this driver must be accelerating. Finally, note that equation (2) does not seem to preclude the physical impossibility of infinite deceleration or acceleration. However, as long as the very first driver modelled will never drive at infinite speed, none of the following drivers will ever perform infinite acceleration, and the said possible drawback of formulation (2) will not become manifest.

One technical problem with this type of formulation concerns the determination of a driver's speed during the last few moments of his trip, when his leader has already completed the trip and therefore his location is strictly speaking undefined. As in Verhoef (2001), this problem will be dealt with by calculating drivers' speeds and locations also beyond the road's exit, so after they have completed their actual trip. When it is assumed that the implied imaginary part of the road, after the exit, has exactly the same characteristics and capacity as the road itself, the model will generate trips (in terms of speed as a function of clock-time) for successive drivers that are continuous and smooth, and that appear regular also for the relatively small segment just before the road's exit. It is certainly an artificial assumption, but probably the only one possible that will not introduce an additional bottleneck – or more generally: an additional source of shock-waves – into the model.

Another technical issue concerns the fact that a model that uses a car-following equation as in (2) to determine speeds for a sequence of users will not have a manageable closed-form solution. This even holds for a simple road of constant capacity and with exogenous departure rates as in Verhoef (2001), and will also be the case for the more complex case considered here. As analytical investigation of the model is not possible, numerical methods will be used instead. For the numerical model, the same distance-speed function $S(\delta)$ is used as in Verhoef (2001), depicted in Figure 1-I (the units are meters m for distance and seconds s for time):

$$\begin{aligned} &S(\delta) = 0 \quad \text{if } \delta \le 5 \\ &S(\delta) = 33\tfrac{1}{3} - \tfrac{33\frac{1}{3}}{(100-5)^5} \cdot (100-\delta)^5 \quad \text{if } 5 < \delta \le 100 \\ &S(\delta) = 33\tfrac{1}{3} \quad \text{otherwise} \end{aligned} \tag{3}$$

It is thus assumed that the minimum possible distance between cars, for which speed falls to zero, is 5 meters (approximately the length of a car), and that a maximum free-flow speed of $33\frac{1}{3}$ m/s (120 km/hr) is obtained if $\delta \ge 100$ meters. The same free-flow speed is chosen if no leader is present. For intermediate values of δ, an arbitrary polynomial function is used, that secures $S(\delta)$ to be continuous at $\delta=5$ and $\delta=100$, and smooth at $\delta=100$. The implied speed-

flow curve for constant speed stationary states[5] (for a single lane) is shown in Figure 1-II, where flow is calculated as $F=S(\delta)/\delta$. The maximum flow of F_{max}=0.965 veh./s is consistent with a speed $S^{\#}$=17.551 m/s (= 63.18 km/h) and a distance between cars of $\delta^{\#}$=18.195 meters ($D^{\#}$=0.055 vehicles per meter). This maximum flow is appreciably higher than the usually empirically measured maxima of around 2500 vehicles per hour per lane (*e.g.* Small 1992, Figure 3.4), but this deviation is not expected to affect the qualitative properties of the model.

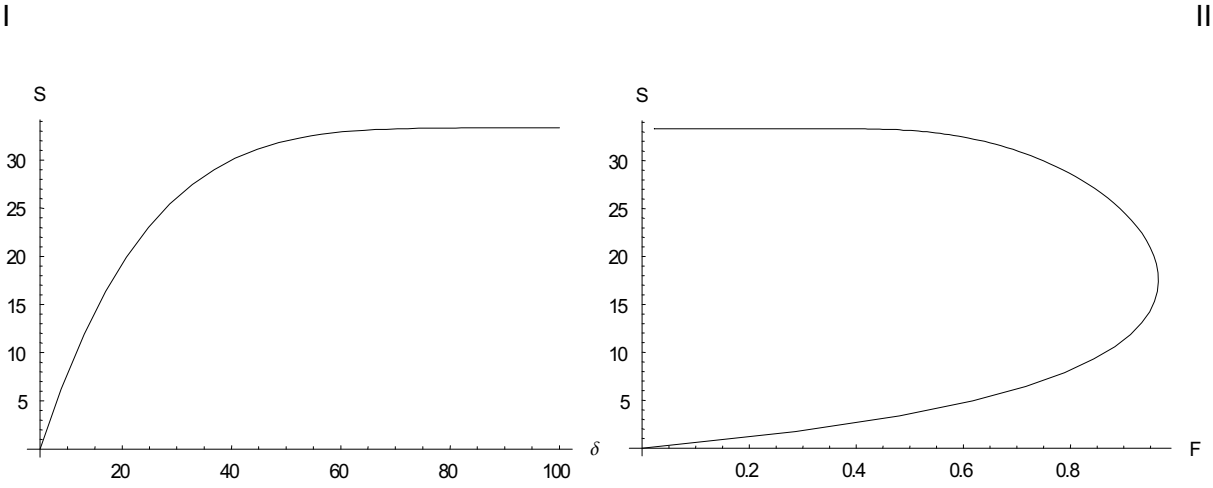


Figure 1. The distance-speed function (I) and the implied speed-flow function for stationary states (II) for the numerical simulation model

Figure 1-II clearly exhibits the well-known backward bending shape mentioned earlier. The standard practice is to derive a backward-bending average cost function (with flow as the argument) from this speed-flow function, by calculating average costs as $\alpha \cdot X/S$ (X is the length of the road). The upper segment of this backward-bending average cost function, and hence the lower segment of the speed-flow function in Figure 1-II, corresponds to hypercongestion. Verhoef (2001) showed that for a single road of constant capacity, the hypercongested equilibria suggested by Figure 1-II and the average cost curve that can be derived from it are in fact dynamically unstable, and can therefore not arise as a stationary state equilibrium for the entire road, following any feasible pattern of departure rates over time. This would falsify the standard discussion of hypercongestion, which is typically given on the basis of single-road models. However, it was hypothesized in that same paper that hypercongestion could arise as soon as the capacity of the road is non-constant. The network set-up in the present paper, discussed immediately below, allows verification of this claim.

[5] A stationary state for the dynamic model is defined as a situation where the flow is constant over time for every point along the road. This implies that the flow must be constant along the road. Verhoef (2001) demonstrates that speed and density (or its inverse, the distance between cars) need not be constant along the road in a stationary state; *i.e.*, acceleration or deceleration during trips is possible in a stationary state. Interestingly, the standard definitional relation $F=S \cdot D$ or its equivalent $F=S/\delta$ do not apply when a stationary state involves acceleration or deceleration (Verhoef, 2001).

3.2.2. The spatial lay-out of the road network and the implied bottleneck

In order to model the behaviour of vehicles in the traffic jam (or queue) explicitly, a simple road network is used, depicted in Figure 2. Between two exogenously determined points x_1 and x_2, the number of (unidirectional) lanes reduces from 2 to 1, which – with a sufficiently high traffic flow – implies that a bottleneck is present as the capacity of the road reduces and traffic has to merge. As explained in Appendix 1, travel time minimization and hence cost minimization implies that in a dynamic equilibrium, successive drivers will alternately choose the left and right lane when starting their trips, and hence, will have another leader after merging than before merging – as indicated by the black and white vehicles in Figure 2.

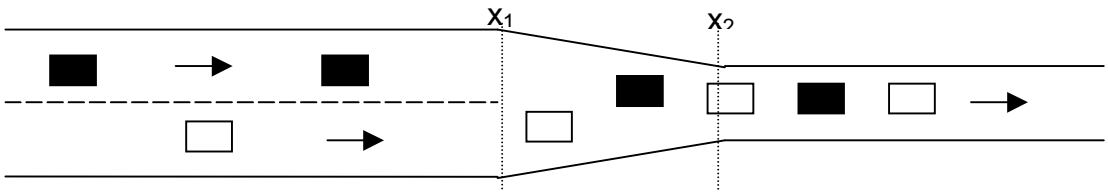


Figure 2. The spatial lay-out of the road network and the implied bottleneck

To prevent discrete changes in speed, it is assumed that the merging of traffic is a smooth process in the sense that a given driver *i* gradually switches from considering driver *i–2* as his leader to considering driver *i–1* as driver *i–1* proceeds from x_1 to x_2 and thus to an increasing degree moves in front of driver *i*, 'pushing' driver *i–2* increasingly out of direct sight. This is accomplished in the numerical model by defining a driver's δ_i^t (for use in (3) to determine a driver's speed at time *t* when a leader is present) as follows:

$$\delta_i^t = \begin{cases} x_{i-2}^t - x_i^t & \text{if } x_{i-1}^t < x_1 \\ w(t)\cdot\left(x_{i-2}^t - x_i^t\right) + (1-w(t))\cdot\left(x_{i-1}^t - x_i^t\right) & \text{if } x_1 \le x_{i-1}^t \le x_2 \\ x_{i-1}^t - x_i^t & \text{if } x_{i-1}^t > x_2 \end{cases} \tag{4}$$

$$\text{with:} \quad w(t) = 1 + 2\cdot\left(\frac{t - t_{i-1,1}}{t_{i-1,2} - t_{i-1,1}}\right)^3 - 3\cdot\left(\frac{t - t_{i-1,1}}{t_{i-1,2} - t_{i-1,1}}\right)^2$$

where subscripts denote drivers, superscripts *t* time, and $t_{i-1,1}$ ($t_{i-1,2}$) denotes the instant at which driver *i–1* passes point x_1 (x_2). The function *w* thus defines the weights attached to drivers *i–2* and *i–1*, which sum up to unity. An (otherwise arbitrary) functional specification for *w* was chosen that secures that the weight for driver *i–2* falls continuously over time from 1 to 0 as driver *i–1* proceeds from x_1 to x_2, and that *w* has a zero time derivative at the instants driver *i–1* passes x_1 and x_2. Note that the specification of (4) uses the equilibrium property derived in Appendix 1, that successive drivers will alternately choose the left and right lane when starting their trips. Otherwise, the car-following equation and its parametrization as given in (3) applies throughout the trip. The full model thus consists of a set of *N* first-order differential equations as in (3), with (4) used as the argument in the function $S(\cdot)$.

The three segments of the road thus distinguished will be referred to as the upstream segment ($x<x_1$), the bottleneck ($x_1 \leq x \leq x_2$), and the downstream segment ($x>x_2$). The numerical model considers a road of 30 000 meters, with the bottleneck located between x_1 = 9 000 and x_2 = 11 000. This implies that a trip at a free-flow speed of $S^* = 33\frac{1}{3}$ would take 900 seconds or 15 minutes.

3.3. Dynamic equilibrium conditions

The dynamic equilibrium condition employed states that the equilibrium trip price as given in equation (1) should be equal for all users, and should not be lower for arrivals before the first or after the last equilibrium arrival times. This condition may at first sight seem identical to the deterministic Nash (dynamic) equilibrium condition used in other dynamic models with continua of users (see Section 2), requiring that in equilibrium no user can reduce his trip price by unilaterally changing his departure time. However, the present condition in fact does not reflect a deterministic Nash equilibrium. Due to the discreteness of drivers in the present model, a deterministic dynamic Nash equilibrium does not exist, for the same reasons as identified by Bernstein (1994) in the context of the bottleneck model: given the departure times of drivers *i–1* and *i+1*, driver *i* would then choose a departure time only marginally earlier than driver *i+1*, which however would induce the latter to also adjust his departure time to an instant marginally earlier than another driver, *etc*.

The dynamic equilibrium condition employed thus cannot be justified as representing a pure deterministic Nash equilibrium. It is, however, the condition that approaches this appealing equilibrium concept as closely as possible in a set-up with discrete users, and could be defended as an intuitive approximation of the equilibrium that would arise once uncertainty about other drivers' exact departure times were introduced, and individuals were assumed to play a mixed strategy when choosing departure times. The symmetric equilibrium probability density function of departure times then too would be such that the equilibrium expected trip price is constant between the first and last arrival, and higher otherwise. As little insight is expected from introducing this type of uncertainty explicitly, however, the present deterministic dynamic equilibrium condition seems a plausible and acceptable approximation for its more realistic, but computationally more demanding, stochastic counterpart.

Appendix 2 derives that under some plausible technical conditions, a dynamic equilibrium in terms of *N* departure times would exist and would be unique. In the numerical model, this dynamic equilibrium is approached up to the level of accuracy allowed by employing a step size of 0.01 seconds in calculating equilibrium departure times. For the base case of the model this results, with an average trip price of € 4.343906, in a maximum of € 4.343896 and a minimum of € 4.343917, the standard deviation being equal to $6 \cdot 10^{-6}$.

3.4. A brief comparison with prior models

Having described the model in detail, the congestion technology will now be compared briefly with the technologies assumed in other dynamic economic models.

Three main differences with the pure bottleneck model (Vickrey, 1969; Arnott, De Palma and Lindsey, 1998) stand out. The first is that in the present model, a traffic queue will not be 'vertical' and spaceless, but instead does take up road space, and that drivers'

behaviour while being in a queue is modelled explicitly. Secondly, and related to the first, a queue in the present model does affect upstream traffic: drivers do slow down when approaching the queue. Both differences imply that the physics of the traffic queue – which can be considered as the key manifestation of traffic congestion in reality – is integrated in greater detail and more elaborately in the present model than in the pure bottleneck model. A third difference is that the present model lacks the rather unrealistic kinked performance function of the bottleneck model, and that drivers instead can – and will – choose speed as a continuous variable, depending on the traffic situation.

Compared to the 'zero-propagation' models of Henderson (1974) and Chu (1995), the present model has the property that drivers' equilibrium speeds can and will vary while driving, and that the congestion encountered depends on the (recent and less recent) history of departure rates, before the driver has departed himself, rather than on an instantaneous departure or arrival rate alone. A main difference with the 'instantaneous-propagation' model of Agnew (1977) is that speed and density are not assumed to be constant along the road at each instant, and that therefore for instance a worsening in upstream traffic congestion would not slow down downstream drivers. Compared to these two groups of models, the present formulation has the further feature that traffic jams (or queues) of a time-varying length can arise endogenously in the model. This, however, is also due to the assumed network structure (Figure 2), and not only to the assumed congestion technology *per se*.

The model that probably is closest to the one proposed here is that of Mun (1999). Mun (1999) considers a road comparable to that depicted in Figure 2. His model, however, assumes that as long as the departure rate of vehicles is less than the capacity of the downstream road segment, speed and local density are dependent on the inflow rate at the time of departure alone, which entails a 'zero-propagation' type of congestion technology. Otherwise, a queue will be building up before the bottleneck, the speed and density in which are determined by the capacity of the bottleneck alone. The back of the queue may then propagate as a shock wave along the road's upstream (higher-capacity) segment, and thus does not at all affect upstream traffic until the shock wave is trespassed. The propagation of the queue's tail is the only form of propagation present in the model. Speed is adjusted in a discrete step when trespassing the shock wave. The present model, in contrast, does not have 'zero-propagation', determines the speed inside the queue on the basis of driver's behaviour, and models explicitly how drivers slow down when approaching the queue.

Finally, for the comparison with the standard static economic model of traffic congestion, it is more useful to focus on the similarities than on the differences. The single main similarity is that the congestion technology defined is such that for stationary states, with traffic flows constant over time and place, the dynamic model could produce exactly the same equilibria as the standard static model (ignoring the details of traffic merging, and ignoring that stationary state traffic will in fact not occur in a dynamic equilibrium). The model thus is a straightforward dynamic extension of the standard static model, making its insights for stationary states useful for a better understanding of that static model.

It appears reasonable to say that the present model is more realistic than the alternatives to which it was just compared, on the grounds mentioned.[6] Note that this improved realism is here meant to concern the continuous-time – continuous-space car-following formulation as such; not necessarily the specific car-following equation employed in the numerical model. This increase in realism comes at the price of a lack of analytical tractability, and an increased computational complexity. These drawbacks are acceptable for the two main purposes of the present paper – the analysis of congestion tolls with time- and space-varying congestion; and the further investigation of hypercongestion as a dynamic equilibrium phenomenon – but may be considered prohibitive for other research questions.

4. 'No-toll' equilibrium: hypercongestion as a dynamic equilibrium phenomenon

The free-market no-toll equilibrium for the base case of the numerical model was found by numerically searching the departure pattern that leads to constant trip prices as defined in equation (1) under zero taxes, such that 2500 drivers are accommodated and nobody wants to travel outside the peak. It entails departures between t_D=–3270.16 and t_D=–304.85, implying a time span of nearly 50 minutes. Arrivals occur over a period of the same duration, but 900 seconds later. 1952 drivers (78%) arrive early, and 548 (22%) late. The longest travel time is 2085 seconds for the driver arriving nearest to t^*, 2.3 times as long as the free-flow travel time of 900 seconds; the average is 1522 seconds, 1.7 times the free-flow travel time. The upper panel in Figure 3 shows the departure and arrival rates – for ease of comparison both plotted against the drivers' arrival times – which are calculated as the inverse of the time lags between the discrete vehicles. The irregularities in the departure rate are therefore due to the step-size of 0.01 seconds used in determining equilibrium departure times, which translates into departure rates that for instance can take on only 13 values in the interval [1.6, 2]. Note that this does not cause any significant irregularities in equilibrium trip prices (middle panel of Figure 3), for which the minimum and maximum differ by only € $2.1 \cdot 10^{-5}$.

The general shape of the departure rate function is comparable to those shown in Chu (1995) and Mun (1999); the arrival rate function is somewhat different as in the present model, the arrival rate will not gradually approach zero towards the end of the peak, but instead rises continuously over time (with the exception of the final driver), approaching the downstream segment's capacity from below. This is related to one of the model's properties to be discussed below, namely that near the end of the peak, the drop in speeds caused by the bottleneck will be moving downstream towards the road's exit as a 'footloose queue'. As a result, successive drivers' 'finishing speeds' are decreasing over time (see also Figure 4-I), which is intuitively consistent with a rising arrival rate with speeds above $S^{\#}$. The departure and arrival rate functions are quite different from those applying in the pure bottleneck model's equilibrium, which are piecewise constant for departures, and constant for arrivals.

---

[6] The discussion focused on what was called 'economic models' before. Compared to the hydrodynamic LWR model of Lighthill and Whitham (1955) and Richards (1956), important differences include that the present model (i) treats vehicles as discrete entities and not as forming a continuum; (ii) does therefore not use a strictly local measure of density but instead the (positive) distance from a driver's leader as the argument for speed choice; and (iii) does not assume that the stationary-state relation $F=S \cdot D$ would also hold for accelerating or decelerating vehicles (which in fact turns out not to be the case; see footnote 5). Newell (1988) has shown that the LWR model, as the present model, can generate hypercongestion only on non-uniform roads.

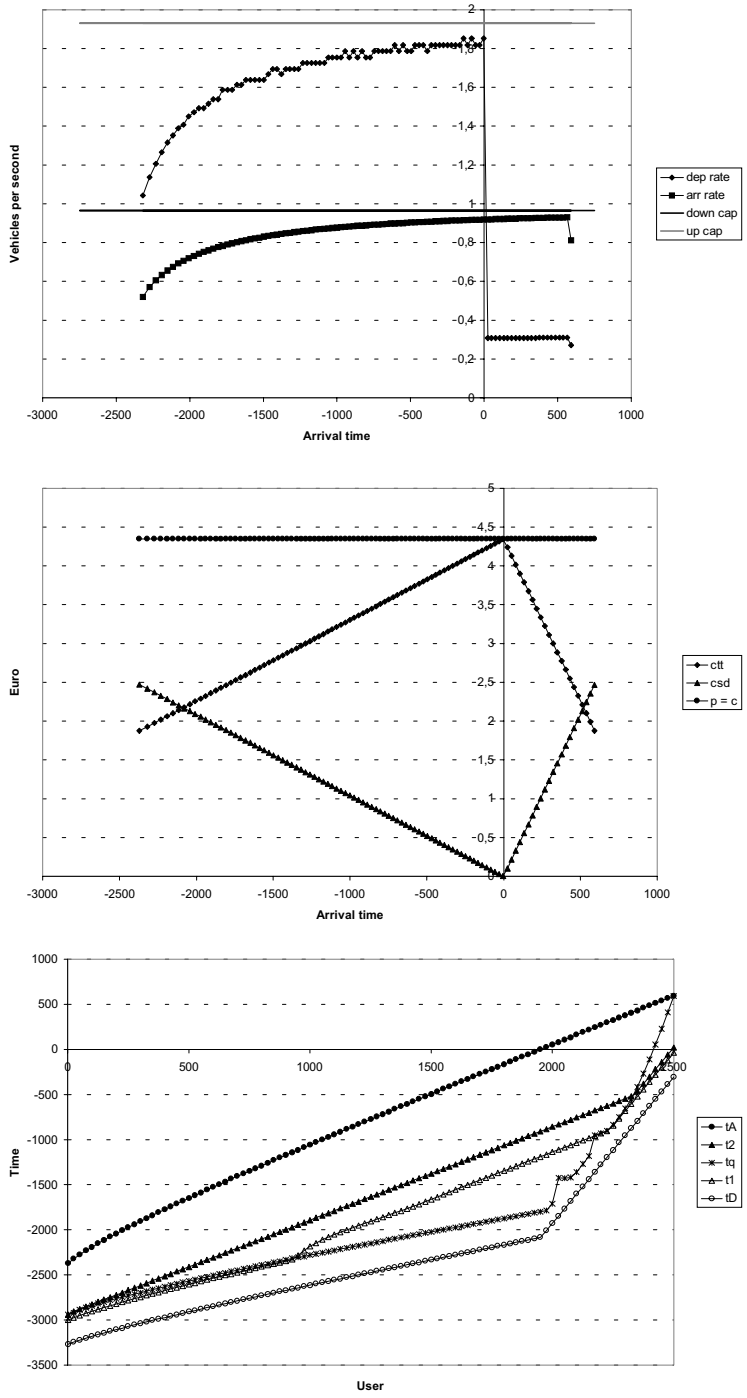


Figure 3. Arrival and departure rates by arrival time (top), cost components by arrival time (middle) and key clock-times by user (bottom) in the no-toll equilibrium (N=2500) (plotted points are the values for drivers i=1,25,50,…,2500)

The fact that the departure rate exceeds the arrival rate for early arrivals, and reversely for late arrivals, is consistent with equilibrium travel times rising and falling for early and late arrivals, respectively. This also explains the sharp drop in the departure rate at the desired arrival time t^*: travel times should there suddenly change from being rising, to falling over

time. The middle panel in Figure 3 shows how the constancy of the equilibrium trip price and the piecewise linearity of the schedule delay cost function together imply that also travel time costs c^{tt} as a function of arrival time t_A must be piecewise linear (in absence of tolling).

The bottom panel in Figure 3 shows some key clock-times for successive drivers. The bottom and upper line, showing t_D and t_A, respectively, follow a pattern consistent with c^{tt} in the middle panel. The three other curves show the instants t_1 and t_2 at which a driver passes x_1 and x_2, respectively; and t_q which is defined as the instant at which a driver obtains the minimum speed applying during his trip. The latter shows how the first few drivers experience the lowest speed very near x_2, the end of the bottleneck. Gradually, however, the location of minimum speeds propagates upstream, and passes x_1 for around the 950th driver. Interestingly, the discrepancy between t_1 and t_2 starts decreasing immediately afterwards, which in combination with the fact that travel times still increase indicates that the most severe congestion (in terms of speeds) propagates upstream during this phase of the peak. When the departure rate drops sharply after t_A=0, however, t_q starts rising more steeply in the diagram, intersects the lines indicating t_1 and t_2, and equals t_A for the last driver. In other words, the drop in speeds due to the bottleneck then propagates downstream until it reaches the exit. (The irregularities in the t_q function for this phase of the peak are due to the fact that the clock-time – speed functions of these drivers have a relatively long nearly flat segment at a low speed, which however exhibits two local minima; see also Figure 4-I).

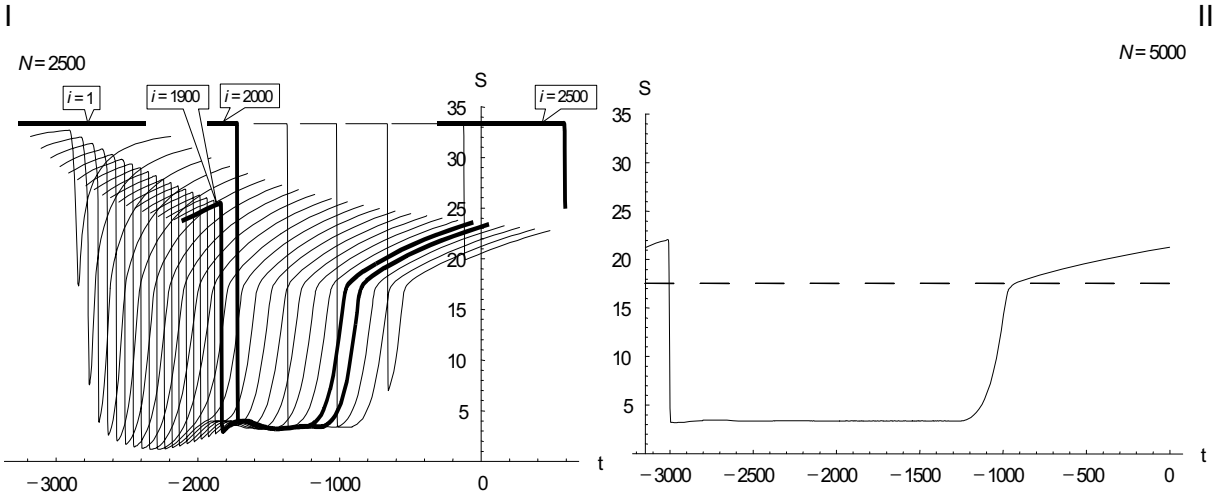


Figure 4. Clock-time – speed functions in the no-toll equilibrium (N=2500) for drivers i=1,100,200…2500 (I), and for driver i=3935 in the N=5000 no-toll equilibrium (II)

These findings are consistent with the plots of successive drivers' clock-time – speed functions shown in Figure 4-I. Three types of trips are present in this diagram. The first and last driver have a trip at a constant maximum speed S^* – the last driver approximately so (his slowing down just before completing the trip, shown in the graph, costs around half a second, and departing later might save some of this but would lead to higher schedule delay costs). Other drivers arriving early first accelerate, then slow down abruptly and after a longer or shorter period of a low – typically hypercongested – speed, they accelerate again. The last

driver with this type of trip is driver 1900, plotted in bold (driver 1952 arrives closest to t^*). For late arrivals, the reduced departure rate implies that the first part of the trip can be traveled at the maximum speed S^*, after which a similar pattern as for early arrivals applies. The first driver plotted with this type of trip is driver 2000 (also in bold).

The model thus produces, in its dynamic equilibrium, a phenomenon that can be characterized loosely as a 'footloose queue', for instance if we define the queue as that part of the trip where a driver drives below $S^\#$ (this pragmatic definition would mean that in the equilibrium, all drivers between approximately i=100 and i=2350 experience queuing).[7] In the last phase of the peak, the queue propagates downstream until it reaches the road's exit. This equilibrium property contradicts the assumption of a given location of the queue's head, often made in dynamic economic models. It may mean that in the empirical estimation (or calibration) of dynamic models on the basis of traffic data, one should allow for the possibility that the object of study (the queue caused by a bottleneck) need not always be there where it would be expected to be (just before the bottleneck). A comparable pattern was found by May (1990, p.209), who too considered a reduction in the number of lanes, a comparable but exogenous time pattern of departure rates, and an LWR (Lighthill and Whitham, 1955; Richards, 1956) hydrodynamic congestion technology.

The clock-time – speed functions in Figure 4 illustrate that the model generates hypercongestion as a dynamic equilibrium phenomenon on the upstream segment. Drivers arriving near t^* obtain an approximately constant speed significantly below $S^\#$ (=17.551) over a substantial 'upstream' part of their trips. Figure 4-II shows this for the last driver arriving before t^* (i=3935) in the equilibrium with the larger demand of N=5000. The associated speed is the hypercongested speed for which the flow on the upstream segment has become equal to downstream capacity (the reader may verify in Figure 1-I that a speed of 3.38 m/s is indeed consistent with a hypercongested flow equal to half the maximum flow). In contrast to the standard (flow-based) static model, which in general cannot explain whether or not hypercongestion will occur in equilibrium,[8] it appears that the present model will always generate hypercongestion as an equilibrium phenomenon on the upstream link, provided demand (N) is large enough and the capacity on the downstream link is sufficiently small

[7] Clearly, the dynamic equilibrium patterns generated in the continuous-time – continuous-place formulation prevent an exact and unambiguous definition of 'the queue'. For the present purposes, it is also not necessary to provide such a definition (one could think of formulations that are based on second or third derivatives of the clock-time – speed functions). But it is noteworthy that a model that aims to describe the queue in more detail produces outcomes that makes the exact definition of 'the queue' far less straightforward than in simpler models.

[8] To be precise, whenever the inverse demand function intersects the backward-bending average cost function both in its normally congested and in its hypercongested segment, multiple candidate equilibria can be identified in the static model. Even local stability analysis, within the limits of a static approach implying that only stationary states can be considered, appears inconclusive in the most general case where the inverse demand function (with flow as its argument) intersects the normally congested segment of the average cost function once, and the hypercongested segment at least twice, including at least once from below and once from above. The intersection with the normally congested segment is locally stable both for price and quantity perturbations. An intersection with the backward-bending hypercongested segment is locally stable for quantity perturbations and locally unstable for price perturbations if the inverse demand function cuts the average cost function from above, and reversely if from below. Whichever view is taken on the appropriateness of either type of perturbation for local stability analysis, therefore, at least two candidate equilibria – one normally congested, one hypercongested – will still result for this most general case (for a graphical exposition, see *e.g.* Verhoef, 1999).

compared to that of the upstream link (a certain combination of these two conditions is required). Furthermore, it appears that no 'severe' hypercongestion will occur on the downstream segment according to the model. That is: speeds below $S^{\#}$ may apply locally and temporarily during the last phase of the peak as the queue propagates downstream while dissolving (*e.g.*, t_q for driver i=2300 in Figure 4-I involves $S<S^{\#}$ and occurs on the downstream segment), but the flow will asymptotically approach the downstream segment's capacity during the first phase of the peak.

A formal proof of these properties of the dynamic equilibrium will not be provided, but a verbal sketch of a proof will now be given that makes use of some of the results obtained by Verhoef (2001). In doing this, it is helpful to consider Figure 4-II as a useful illustration for the 'attractor' for early arrivals. The 'attractor' is defined as the situation just before t^* that would be approached asymptotically with an increasing N, and hence with an increasing time period over which early arrivals occur. Why would this attractor involve hypercongestion on the upstream segment and a maximum flow on the downstream segment?

First observe that the dynamic equilibrium condition requires travel times to increase linearly with arrival time during the first phase of the peak. This implies that the departure rate should exceed the capacity of the downstream road segment, denoted $F_{\max}^{ds}$, from some moment onwards (nearly immediately so in the numerical model; see Figure 3). If not, the travel time would have an upper limit implied by the non-hypercongested speeds consistent with $F_{\max}^{ds}$ for the upstream segment and the bottleneck, and $S^{\#}$ for the downstream road segment (compare also Proposition 2 in Verhoef, 2001). Such an upper limit is inconsistent with the dynamic equilibrium condition if N becomes sufficiently large.

As a result, the flow at every point along the upstream segment and along the bottleneck will exceed $F_{\max}^{ds}$ from some instant onwards, too, and so will the rate at which users arrive near the entrance of the downstream segment. Proposition 5 in Verhoef (2001) derived for a single constant-capacity road, now applied to the current situation, then implies that the flow on the downstream segment will asymptotically approach $F_{\max}^{ds}$ from below. Figures 4-I and 4-II indeed illustrate that the speeds over this segment approach a constant value $S^{\#}$ (indicated by the dashed line in Figure 4-II) more closely as the peak's first phase extends. The upper panel in Figure 3 further confirms that the arrival rate at the road's exit approaches $F_{\max}^{ds}$ (=0.965) from below.

Because the downstream flow is smaller than the upstream flow, the average distance between vehicles on the upstream segment and hence their speeds will be decreasing over time. The closer the flow at a certain point on the upstream segment approaches $F_{\max}^{ds}$ from above, however, the less rapidly would distances between cars and hence their speeds decrease between that point and the downstream segment. The changes in flow, speed and density will thus be such that a stationary state with a flow equal to $F_{\max}^{ds}$ will be approached asymptotically from above (undershooting is implausible).

Two reasons can be given why it should be a hypercongested stationary state that will be approached on the upstream segment. First, starting with a flow exceeding $F_{\max}^{ds}$, consistent downward adjustments in speeds cannot bring the system towards a non-hypercongested stationary state with a flow $F_{\max}^{ds}$, but only to a hypercongested one – compare the speed-flow function in Figure 1-II for intuition. Secondly, since the close approximation of a stationary

state with a flow equal to $F_{\max}^{ds}$ requires the inflow at the road's entrance (at x=0) to obtain that value too, and hence to fall below the departure rate which exceeds $F_{\max}^{ds}$, the stationary state that will be approached must involve a growing queue before the upstream segment's entrance. (The explicit modeling of the implied queue would of course require the addition of yet another upstream link to the model, which however would yield no further insights and is thus avoided.) But a growing queue before the entrance cannot be consistent with a persistent non-hypercongested stationary state with a flow $F_{\max}^{ds}$ on the upstream segment. Instead, the inflow into the road would then remain equal to the departure rate, which exceeds $F_{\max}^{ds}$, as long as the speed at x=0 remains positive – so that the process of decreasing speeds as described above would again occur at least as long as the speed at x=0 is positive.[9] Figure 4-II confirms that it is indeed a hypercongested stationary state with a flow equal to $F_{\max}^{ds}$ that is approached asymptotically. A flow equal to $F_{\max}^{ds}$, under hypercongested conditions, is thus the attractor for the upstream segment for early arrivals.

In summary, the model thus predicts, for a sufficiently large N, that the arrival rate for early arrivals will exceed $F_{\max}^{ds}$ from some point onwards, that the flow along the entire road will asymptotically approach $F_{\max}^{ds}$, and that this flow will be achieved under hypercongested conditions on the upstream road segment. The latter represents queuing in the present model. Paradoxically, therefore, the lowest dynamic equilibrium speeds will apply where capacity is highest (see Figure 4-II) if the peak lasts long enough. This indicates that cost-benefit analyses of road expansion that naïvely use (dynamic) equilibrium speeds as an indicator for the potential benefits may lead to a ranking of projects that would be the exact reverse of the 'true' ranking. In the present model, expanding the upper segment to three lanes would lead to an ever lower (approximate) dynamic equilibrium speed for drivers arriving near t^*, namely the hypercongested speed consistent with ⅓ of $F_{\max}^{ds}$ according to Figure 1-II. Substantial benefits, in contrast, might result from expanding the downstream segment. This underlines the importance of considering network effects seriously when performing cost-benefit analyses.

The present model thus avoids the ambiguity of the standard static model, which in general cannot explain whether or not hypercongestion will occur, and instead shows that hypercongestion will occur in a dynamic equilibrium in a queue caused by a downstream bottleneck provided N is sufficiently large (relative to the reduction in capacity due to the bottleneck), whereas no severe and 'structural' hypercongestion on the lowest capacity segments of the road need be expected, but maximum flows will be approached instead.

[9] This observation means that if a vertical queue were allowed to arise before the road's entrance, a sufficiently large N would result in what Verhoef (2001) called a 'variable-speed' hypercongested stationary state on the upstream road segment, as opposed to the 'single-speed' hypercongested stationary states as depicted in Figure 1-II. This distinction between 'single-speed' and 'variable speed' stationary states as such is important and interesting for stationary state analyses as in Verhoef (2002), but ignored here as it does not affect the main finding that hypercongestion will occur on the road's upstream segment.

5. Equilibria with tolling: (near-)optimal toll schedules for a continuous-time – continuous-place congestion technology

The dynamic equilibrium as identified in Section 4 will not be efficient. In general one can expect inefficiencies resulting from the external costs associated with congested road use; in the present model's equilibrium, such inefficiencies are for instance clearly exemplified by the occurrence of hypercongestion. The question thus arises whether the dynamic optimum for the model can be identified or approximated, and if so, what its properties are compared to those in the free-market ('no-toll') equilibrium discussed above, and which type of toll schedule should be used to achieve a decentralized optimum as a dynamic equilibrium.

The optimality conditions for the general model can be found by minimizing the sum of travel time costs c^{tt} and schedule delay c^{sd} over all N users by setting their departure times t_D optimally. The structure of the full model, consisting of N interdependent first-order differential equations $S(\delta_i^t) \equiv \dot{x}(\delta_i^t)$ such as defined by (3) and (4) for the numerical model, implies that a driver's travel time and hence his arrival time given his departure time, and as a consequence his travel time costs and schedule delay costs, will depend on the departure times of all prior users, in addition to his own departure time. Similarly, a driver's departure time will affect all later drivers' travel time costs and schedule delay costs, given their own departure times. The formal optimization problem can thus be written as:

$$\underset{t_D^1,\ldots,t_D^N}{\text{Min}} \sum_{i=1}^{N} c_i^{tt}\left(t_D^1,\ldots,t_D^i\right) + c_i^{sd}\left(t_D^1,\ldots,t_D^i\right) \tag{5}$$

where t_D^i is driver i's departure time, implying necessary first-order conditions:

$$\frac{\partial C}{\partial t_D^i} = \sum_{j=i}^{N} \frac{\partial c_j^{tt}(\cdot)}{\partial t_D^i} + \sum_{j=i}^{N} \frac{\partial c_j^{sd}(\cdot)}{\partial t_D^i} = 0 \quad \text{for all } i \tag{6}$$

where C indicates total costs (summed over all users). Equations (6) are of little or no practical use for finding the optimum in a fully specified model. The reason is that no closed-form expressions for the functions $c_i^{tt}(\cdot)$ and $c_i^{sd}(\cdot)$ will generally exist. Moreover, the sheer number of first-order conditions will further complicate their use in an applied context, even if closed-form expressions would exist.

Unfortunately, neither the general car-following equation, nor the one used in the numerical model, appeared to have any properties that allow useful substitutions or other solution strategies circumventing the complexities just mentioned. Also reformulations of (6) with arrival times, departure time lags or arrival time lags as its arguments did not appear to yield any manageable results. As a result, only an 'approximate optimal toll' schedule can be presented. This section proceeds by first presenting the derivation of its underlying tax rule in Section 5.1. Section 5.2 compares its efficiency impacts with that of alternative schedules, with the aim to assess the 'degree of optimality' of the toll schedule presented.

5.1. Deriving an approximate optimal tax rule

The main two difficulties preventing an (easy) determination of the dynamic optimum for the full model are the fact that traffic will generally be non-stationary throughout the optimum so that for every driver the speed will vary continuously over time and place, and the large number of arguments in the objective function (5): all N departure times. The approximate toll

schedule is derived using a simple approximation of the original model, that avoids these two complexities. Specifically, it is assumed that at every instant and at every point along the road, the optimality of traffic conditions can be analyzed as if a stationary state with a constant flow and a constant speed would apply locally at that instant. Next, a continuum of drivers are now considered, so that the objective can be rewritten as the minimization of total costs by choosing an optimal time profile of arrival rates at the road's exit, $\rho_A(t_A)$. Finally, a simplifying approximation is used for the assessment of the marginal impact of a marginal change in $\rho_A(t_A)$ upon speeds at earlier instants at upstream locations. The following notation, definitions and assumptions are further used:

t_A^x denotes the instant that a driver arriving at t_A passes point x.

ρ_t^x denotes the instantaneous traffic flow at point x at instant t.

$\rho_{t_A^x}^x$ therefore denotes the instantaneous traffic flow at point x at instant t_A^x. The important simplifying assumption that will be made is that for the derivation of optimality conditions (*i.e.*, when evaluating marginal modifications of the optimal arrival rate pattern), $\rho_{t_A^x}^x$ can be considered as a function of $\rho_A(t_A)$ alone. Specifically, it is assumed that a marginal change in $\rho_A(t_A)$ leaves ρ_t^x unaffected for $t \neq t_A^x$, but leads to an equally large marginal change in ρ_t^x for $t = t_A^x$: $\mathrm{d}\rho_{t_A^x}^x / \mathrm{d}\rho_A(t_A) = 1$. This can be motivated loosely by noting that having one additional user arriving during the time unit centered by t_A implies, under the simplifying assumptions, that one additional user will pass any point x during the time unit centered by t_A^x. It is an inexact approximation only because in general, under non-stationary travel conditions, $\rho_{t_A^x}^x \neq \rho_A(t_A)$, which contradicts that $\mathrm{d}\rho_{t_A^x}^x / \mathrm{d}\rho_A(t_A) = 1$ should always hold.

$k_x^{tt}(\rho^x)$ denotes the per-user – per-unit-of-distance travel time cost function for point x, which is assumed to be a function of the instantaneous local flow level alone. This function can be derived directly from the static speed-flow function by multiplying the inverse of speed by the value of travel time.

$-T$ and T denote lower and upper arrival time limits used in the specification of the objective function, which are presupposed to be non-binding.

Under these simplifying assumptions, the problem of minimizing social costs can be represented by the following Kuhn-Tucker specification:

$$\Lambda = \int_0^X \int_{-T}^{T} \rho_t^x \cdot k_x^{tt}(\rho_t^x)\,\mathrm{d}t\,\mathrm{d}x + \int_{-T}^{T} \rho_A(t_A) \cdot c^{sd}(t_A)\,\mathrm{d}t_A + \lambda \cdot \left(N - \int_{-T}^{T} \rho_A(t_A)\,\mathrm{d}t_A \right) \qquad (7)$$

s.t. $\rho_A(t_A) \geq 0$

The first term represents total travel time costs, the second total schedule delay costs, and the third the constraint that N drivers be accommodated during the peak (the Lagrangian multiplier λ will, in the optimum, therefore be equal to marginal costs). The following Kuhn-Tucker conditions with respect to $\rho_A(t_A)$ can be determined after substituting t_A^x for t in the first term in (7) for every t_A evaluated , and using the assumption that $\mathrm{d}\rho_{t_A^x}^x / \mathrm{d}\rho_A(t_A) = 1$:

$$\int_0^X k_x^{tt}(\rho_{t_A^x}^x)\,\mathrm{d}x + \int_0^X \rho_{t_A^x}^x \cdot \frac{\mathrm{d}k_x^{tt}}{\mathrm{d}\rho_{t_A^x}^x}\,\mathrm{d}x + c^{sd}(t_A) - \lambda \geq 0 \quad \forall t_A;$$

$$\rho_A(t_A) \geq 0 \quad \forall t_A; \tag{8}$$

$$(\rho_A(t_A)) \cdot \left(\int_0^X k_x^{tt}(\rho_{t_A^x}^x)\,\mathrm{d}x + \int_0^X \rho_{t_A^x}^x \cdot \frac{\mathrm{d}k_x^{tt}}{\mathrm{d}\rho_{t_A^x}^x}\,\mathrm{d}x + c^{sd}(t_A) - \lambda \right) = 0 \quad \forall t_A$$

The third condition shows that whenever a positive arrival rate applies, the first condition should apply as an equality. The constant trip price condition for positive arrivals implies that over the period with positive arrivals, the following should hold:

$$\int_0^X k_x^{tt}(\rho_{t_A^x}^x)\,\mathrm{d}x + \tau(t_A) + c^{sd}(t_A) = p^* \quad \text{if } \rho_A(t_A) > 0 \tag{9}$$

where p^* denotes the optimal (constant) trip price. Setting the constant in the optimal toll schedule equal to zero (with inelastic demand, any constant can be added to a time-varying toll schedule without affecting its impacts), (8) and (9) together imply the following optimal toll schedule for the simplified model:

$$\tau(t_A) = \int_0^X \rho_{t_A^x}^x \cdot \frac{\mathrm{d}k_x^{tt}}{\mathrm{d}\rho_{t_A^x}^x}\,\mathrm{d}x \quad \text{if } \rho_A(t_A) > 0 \tag{10}$$

First observe that with a zero constant in the toll schedule, an intuitive equality of trip price p^* and marginal costs λ is established in the optimum, that the constancy of λ reflects the intuitive property that marginal costs should be constant over time throughout the peak (as long as arrivals occur), and that marginal costs would exceed the equilibrium trip price for arrivals before the first or after the last one in the optimum. Next, note that equation (10) gives a straightforward time- and space-varying generalization of the standard Pigouvian congestion tax applying in static models (this is consistent with the tax rule found by Chu, 1995, who however presented a time-varying generalization alone). Given the intuitive task of an optimal toll to internalize the congestion externality, equation (10) implies that this externality involves travel time costs only, not schedule delay costs. The intuition behind this is that the schedule delay costs associated with the 'consumption' of an arrival at a certain time do not depend on the behaviour of other drivers, whereas the travel time costs do (under congested conditions). An externality is therefore imposed only via the latter cost category.

The immediate advantage of using the 'approximate optimal tax rule' – defined as a tax rule based on (10) also for non-stationary traffic – in the original model, rather than the model's true optimality conditions as implicit in (6), is that (10) can be used to calculate taxes on the basis of instantaneous and local traffic conditions.[10] Any numerical procedures based

[10] This is in the first place a computational advantage. For calculating tolls in practice, it may be advisable for safety reasons to prevent drivers from having the opportunity to save on instantaneous tolls by speeding up (a tolling scheme proposed for Cambridge, England, was rejected on these grounds). One way of doing so is to charge arrival time dependent, 'full-trip' equilibrium tolls, as shown in Figure 6 – independent of actual speeds encountered. This assumes that the regulator can correctly predict the optimum. Alternatively, drivers can be charged instantaneous local tolls based on moving averages for, say, a 5 minute period. This presumably takes away most incentives to drive dangerously in order to save tolls.

on (6), instead, would involve a re-evaluation of the entire peak after the adjustment of one (or some) departure times.

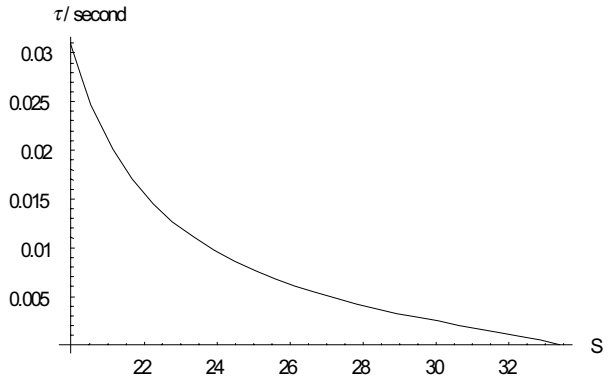


Figure 5. The approximate tax rule for the numerical model: the per unit-of-time toll level as a function of instantaneous speed

For the numerical model, the approximate tax rule (10) was calculated by integrating a per-unit-of-time toll over t, rather than a per-unit-of-distance toll over x, as in (10). The equivalence between these two alternative specifications is easily established (provided speeds are positive) by having an instantaneous toll of y per unit of time corresponding to a local toll of y/S per unit of distance, with S denoting the instantaneous speed. Figure 5 shows the resulting approximate tax rule for the numerical model, which was derived from the upper segment of the speed-flow curve in Figure 1-II in the standard way. Note that, by defining the instantaneous toll as a function of S rather than flow, a tax rule is obtained that can be applied to an individual driver's trip: no instantaneous measure of local traffic flow has to be calculated. Furthermore, the same tax rule can be applied to the road's upstream and downstream segment: whereas equal stationary state speeds would imply equal per-unit-of-time taxes, equal traffic flows would not, due to the difference in capacity (the same tax rule as shown in Figure 5 is applied as a driver passes the bottleneck). Finally, the per-unit-of-time tax approaches infinity as the instantaneous speed approaches $S^{\#}$. No effort was made to calculate taxes for candidate optima involving hypercongestion, on the grounds that any such candidate optimum could be improved upon by an alternative configuration where the same instantaneous flow is realized under non-hypercongested conditions. The existence and uniqueness of an approximate non-hypercongested optimum (*i.e.*, an equilibrium with the approximate optimal tax rule applying) can be made plausible in a way comparable to what is done for the no-toll equilibrium in Appendix 2. Specifically, a speed-dependent toll as shown in Figure 5 could simply be added up to c^{tt} without causing any fundamental differences in the further argumentation as long as, indeed, only speeds above $S^{\#}$ are considered.

It can be expected that the approximate optimal toll will perform better in the full model, the less strongly its resulting equilibrium traffic conditions deviate from the factually incorrect assumption of stationary state conditions applying at every instant and every

location, underlying the simplified model. The next sub-section presents the results of applying the approximate optimal toll in the model, and will shed some light on this issue.

5.2. Applying the approximate optimal toll schedule

The upper panel in Figure 6 depicts the equilibrium departure and arrival rate for the original model using the approximate tax rule proposed above. This equilibrium was found by numerically searching that departure pattern which leads to constant trip prices as defined in equation (1), with the tax rule from equation (10) and Figure 5 applied, such that 2500 drivers are accommodated and nobody wants to travel outside the peak.

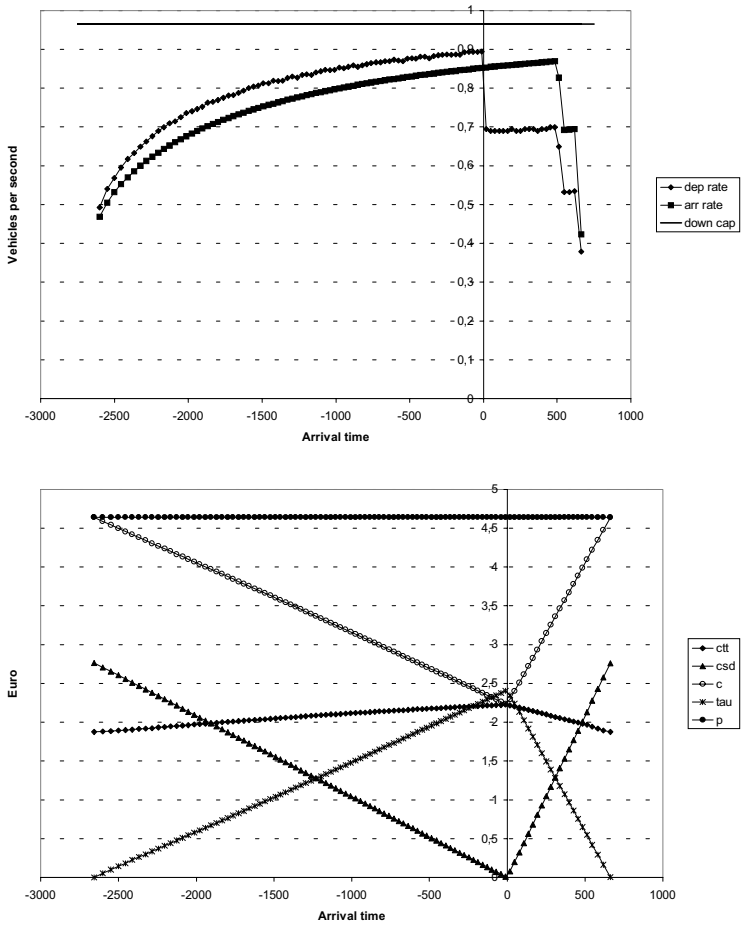


Figure 6. Arrival and departure rates by arrival time (top) and cost components by arrival time (bottom) in the approximate optimum (N=2500) (plotted points are the values for drivers i=1,25,50,...,2500)

Compared with the no-toll equilibrium (see Figure 3), the departure rate for early arrivals has decreased significantly, approximately by a factor 2, whereas the arrival rate has fallen far less strongly. This is consistent with the significant increase in speeds realized, leading to an elimination of hypercongestion as expected; see also the clock-time – speed functions in

Figure 7.[11] Consistent with the elimination of severe hypercongestion due to tolling, the departure and arrival time intervals have only increased relatively mildly, by 12.0%. As the first and final driver face a zero toll, drive at S^* throughout their trips and have equal c^{sd}, it should not come as a surprise that the trip price net of free-flow travel time costs in the approximate optimum is also 12.0% higher than in the no-toll equilibrium. The full trip price has risen by 6.8%.

Other key (aggregate) indicators have changed as follows: total costs have fallen by 22.6%, total variable costs (excluding travel time costs associated with free-flow speeds) by 39.8%; total travel time costs have fallen by 34.6% and total variable travel time costs by 84.6%; and total (variable) schedule delay costs have increased by 9.8%. The model thus produces the same qualitative differences compared to Vickrey's (1969) model of pure bottleneck congestion as found by Chu (1995) and listed towards the ending of Section 2 above. The lower panel in Figure 6 shows the various cost components by arrival time. The two main differences with the patterns shown in Figure 3 are that c^{tt} rises and falls far less steeply, which is replaced by a rise and fall in the toll level. Interestingly, despite the highly non-linear character of the model, the approximate optimal toll schedule looks nearly (piecewise) linear, as in the pure bottleneck model. An explanation is not easily given, especially not because it will be shown below that the most optimal toll schedule will not be exactly piecewise linear, and because a closer inspection of the approximate toll schedule has revealed that its slope does vary around the average over both phases of the peak – without any clear trend, though.

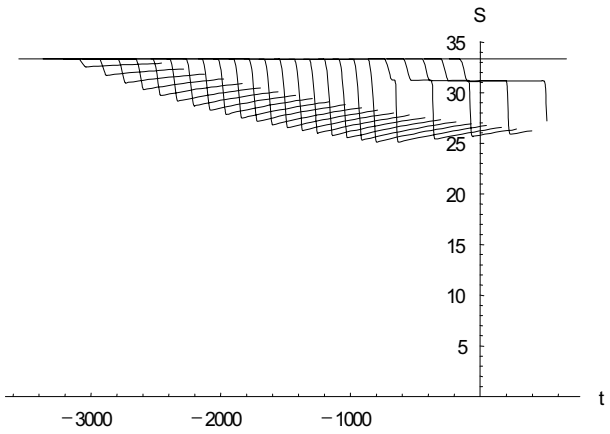


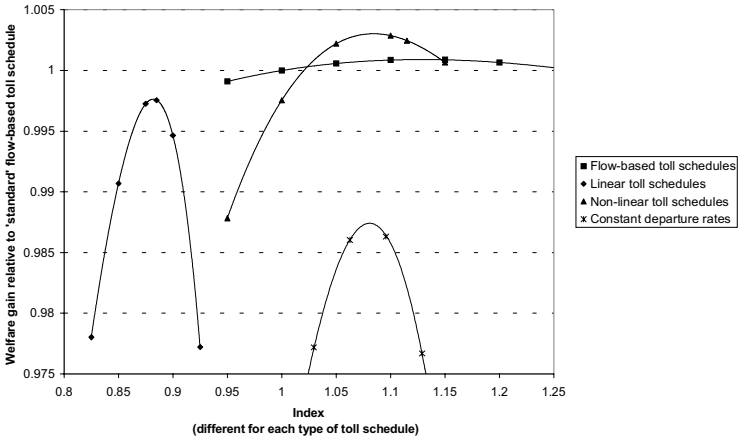
Figure 7. Clock-time – speed functions in the approximate optimum (N=2500) for drivers i=1,100,200…2500

The substantial relative cost savings achieved with the approximate toll schedule (nearly 40% of total variable costs and 85% of variable travel time costs) already suggest that it will be difficult to improve significantly upon it in terms of efficiency. By means of comparison: the

[11] The downward steps in the range $t > 0$ in Figure 6 occur when the kinks in the clock-time – speed functions, in Figure 7, pass the road's exit.

pure bottleneck model has a congestion technology that is probably more optimistic than the present one in terms of the relative cost savings that can be achieved with optimal tolling, because reductions in travel time costs in the pure bottleneck model do not require increases in schedule delay costs. That model achieves a 50% reduction of total variable costs and 100% of variable travel time costs. For the present model, in contrast, it is unlikely that any significant further reductions in travel time costs can be achieved without having the simultaneous increase in schedule delay costs outweighing these gains. Indeed, one of the experiments discussed below involves the highest departure rate allowing free-flow speeds throughout the peak (ρ_D=1/3), with the departure interval timed optimally, which however leads to a cost increase (27% of total variable costs) compared to the no toll equilibrium, rather than a cost decrease as with the approximate optimal toll schedule.

Despite this rather convincing performance of the approximate optimal toll schedule, it of course remains important to investigate its degree of optimality in some more detail, if only because the underlying tax rule was derived for a simple approximation of the original model. A rather extensive test was carried out, the results of which are depicted in Figure 8 in terms of the 'index of relative welfare improvement' ω: the cost reduction for a scheme, as a proportion of the cost reduction achieved with the standard approximate optimal toll schedule just discussed. Four 'families' of alternative toll schedules were investigated.



Note: Indices are defined as follows:

1. Flow-based toll schedules: the proportion of the approximate per-unit-of-time tax rate actually charged.
2. Linear toll schedules: slope of the toll schedule as a proportion of β (for early arrivals) and $-\gamma$ (for late arrivals.
3. Non-linear toll schedules: starting from the slopes of the best among the tested linear toll schedules (0.885 times α and γ, respectively), the non-linear toll schedules maintain this average slope for both phases of the peak, but allow this slope to vary at a constant rate over time. The index gives the slope as a proportion of the average slope at t^*. A value greater than unity thus indicates two convex segments, and a value smaller than unity two concave segments.
4. Constant departure rates: the constant departure rate as a proportion of the average departure rate applying in the equilibrium with the standard approximate tax rate.

The solid lines represent 4th-order polynomial fits for the 6 or more plot-points, as estimated by the spread-sheet computer programme used.

Figure 8. Index of relative welfare improvement (relative to the welfare gain due to the standard approximate optimal toll schedule) for various alternative toll schedules (N=2500)

The first concerns 'flow-based toll schedules', one of which is the standard approximate toll schedule itself. The variants simply multiply the taxes calculated with the standard rule with a fixed proportion (the index used in Figure 8), which, when equal to unity, therefore

reproduces the approximate optimum discussed above. If the approximate tax rate were exact, any value different from unity would lead to a lower welfare level. Clearly, this is not the case: a multiplication by a value between 1.1 and 1.15 leads to a welfare gain that is nearly 0.09% higher than for the standard tax. The difference is of course negligible in practical terms, but indicative for the approximate character of the tax rule proposed.

The second family of 'linear toll schedules' involve piecewise linear toll schedules, with a maximum for an arrival at t^* and slopes equal to a fixed proportion of β and $-\gamma$ for early and late arrivals, respectively (experiments where the proportion differ for early and late arrivals did not produce welfare gains compared to the equal proportions cases presented here). The highest efficiency for this family applies when this proportion is around 0.885, very near the value of 0.875 that reproduces the average slopes applying in the standard approximate optimal toll schedule. The maximum welfare gain remains some 0.2% below the base gain, indicating that a strictly piecewise linear toll is not optimal for the model. A toll schedule with slopes equal to β and $-\gamma$ for early and late arrivals, respectively, resulting in the equilibrium mentioned above with free-flow speeds throughout the peak, has an index of relative welfare improvement ω of –0.669.

The third family of 'non-linear toll schedules' encompasses non-linear variants of the best among the linear toll schedules. These schedules maintain an average slope of 0.885 times β and $-\gamma$ for the two phases of the peak, but allow these slopes to vary at a constant rate over time in such a way that either two concave or two convex segments result. The index used in Figure 8 gives the slope of the toll schedule at t^* as a proportion of this average slope. A value greater than 1 thus indicates two convex segments, and below 1 concave segments. The results further confirm that a piecewise linear toll schedule is not optimal, as the highest efficiency in Figure 7 – with ω 0.3% above unity – occurs for two convex segments at a 'degree of convexity' of 1.1. This is the highest efficiency found in the experiments.

Finally, a fourth family of toll schedules considered are those that support a constant departure rate as an equilibrium (these tolls were calculated such that p remains constant over the peak, with a zero toll for the first user). The results show that this type of scheme performs less well than the standard approximate toll schedule, the welfare gains remaining some 1.4% smaller at best for a departure rate ρ_D=0.825. The results (again) revealed that among the constant departure rates neither one equal to 1/3 (allowing free-flow speeds throughout, with ω=–0.669), nor one near F_{max} for the downstream road segment (ρ_D=0.965, with ω=0.741) would be the best choice. Both ω's fall well below the range plotted in Figure 8.

Two main conclusions can be drawn from these exercises. The first is that the near-optimality of the proposed approximate tax rule seems to be confirmed, for a number of reasons. First, the greatest improvement in efficiency found involved only a very modest 0.3% extra cost reduction. Next, both linear and non-linear (*i.e.* more convex and more concave) deviations from the approximate toll schedule were considered, suggesting that the 'vicinity' of the approximate toll schedule is effectively scanned. These tests generally show a progressively decreasing ω as the toll schedules considered deviate more strongly from the approximate toll schedule (for schedules beyond its very near vicinity). Combined with the relatively flatness of the three upper curves in Figure 8 and the occurrence of nearly equally-valued local maxima for each of them, this suggests that the experiments presented in Figure

8 would be near the true optimum for the model. Furthermore, the clock-time – speed functions depicted in Figure 7 show that the assumption of instantaneous and local stationary state conditions underlying the derivation of the approximate tax rule may not be violated to too large an extent. And finally, there is the consideration given before discussing the experiments, namely that the substantial relative cost savings achieved with the approximate toll schedule (nearly 40% of total variable costs and 85% of variable travel time costs) suggest that it will be difficult to improve significantly upon it in terms of efficiency.

A second main conclusion emerging concerns the limited applicability of insights from other dynamic economic equilibrium models for the current continuous-time – continuous-place congestion technology. In particular, a regulator in control of the road considered here might be tempted to base a toll schedule on those identified as optimal in one of these prior models. The question then arises what the efficiency impacts would be.

The results presented above in the first place warn against a naïve application of insights from the pure bottleneck model. Specifically, it was shown that if the regulator would simply copy the pure bottleneck model's optimal toll schedule, which is piecewise linear with slopes equal to β and $-\gamma$ for the two phases of the peak, a welfare loss instead of a welfare increase will in fact result due to tolling (ω=–0.669), while – as predicted also by the bottleneck model – no travel delays would remain existent. A much better rule of thumb based on the pure bottleneck model would use the property that the departure rate throughout the optimum should be equal to the bottleneck's capacity (ω=0.74) (note that an arrival rate constant and equal to the bottleneck's capacity is physically impossible in the present model if the first driver travels at a free-flow speed). But even then, an appreciable further welfare gain can be realized by basing the tolls on the insights from another classic model, namely the standard static economic model of traffic congestion (the approximate toll schedule, with ω=1), rather than on those of the pure bottleneck model, even though the current model's no toll equilibrium at first glance might suggest a much closer correspondence with the latter. Moreover, lower 'target' constant departure rates perform better than ω=0.74 – see Figure 8.

An important question in this context is whether the performance of naïve applications of insights from the pure bottleneck model improves as N increases, for instance because the no-toll equilibrium will then become increasingly similar to that of the pure bottleneck model in the sense that the downstream road segment will operate near capacity over a larger share of the peak period. To investigate this question, four regimes were evaluated also for N=5000[12], namely the no-toll equilibrium, the approximate optimal toll schedule, and constant departure rates equal to ⅓ and 0.965. For the former (a rate of ⅓), ω has fallen even further to –0.7995, whereas for the latter (a rate of 0.965), it has increased further to 0.8758. This suggests that for a larger peak, the 'good' rule-of-thumb from the pure bottleneck model (seeking flows equal to the bottleneck's capacity) will perform better, while the 'bad' one (eliminating all travel time losses) performs worse.[13] Consequently, especially as N becomes

[12] The longest travel time in the no-toll equilibrium now is 3145 seconds (compared to 2085 for N=2500), 3.5 times as long as the free-flow travel time of 900 seconds; the average is 2066 seconds (1522 for N=2500), 2.3 times the free-flow travel time; and the trip price equals € 6.55 (€ 4.34 for N=2500).

[13] This is somewhat surprising in the light of the fact that the ratio between the slopes of the approximate optimal toll and β (for early arrivals) or $-\gamma$ (for late arrivals) appears to increase and slowly approach unity as N grows.

larger, a very simple first check for the degree of optimality of a proposed toll schedule would be to see whether the departure rate during the peak approaches the capacity of the bottleneck sufficiently close (but does not exceed it to avoid wasteful hypercongestion).

In a similar way, the applicability of insights from the models by Chu (1995) and Mun (1999) for the current continuous-time – continuous-place congestion technology can be assessed. To start with Chu's (1995) model, it is interesting to note that the approximate tax rule proposed here would remain optimal if the congestion technology in reality would – contrary to expectation – replicate the zero-propagation technology assumed in Chu (1995). That is, the optimal tax rule in Chu (1995) equates the optimal congestion tax to the derivative of the travel time costs for the entire trip with respect to the arrival rate, which under the zero-propagation assumption implies that the result should be identical to a tax calculated using equation (10). The reverse, however, is certainly not necessarily true. For instance, Chu's (1995) tax rule, when naïvely applied to the current model, would produce taxes that are still increasing after the preferred arrival time when evaluated in the approximate optimum depicted in Figure 6, as the arrival rate also in the approximate optimum increases for a substantial period after t^* due to the downward propagation of the drop in speeds. It should be noted that this inapplicability of Chu's (1995) tax rule becomes especially manifest when indeed a bottleneck exists along the route, that causes arrival rates to increase also after the preferred arrival time.

Finally, Mun's (1999) optimal toll schedule involves convex segments during the shoulders of the peak, and linear segments with slopes equal to β and $-\gamma$ during the central part of the peak. This central part will extend both in an absolute and in a relative sense as demand becomes larger. This, combined with the findings reported above for the pure bottleneck model, suggests that also a naïve copying of Mun's (1999) optimal toll schedule to the present context would be non-optimal, and might even lead to welfare losses if the central part of the peak becomes big enough in a relative sense.[14]

6. Conclusion

This paper presented an economic dynamic equilibrium model of traffic congestion on a single road with a bottleneck, with identical users. The congestion technology proposed is based on car-following theory, and probably provides the simplest plausible dynamic extension of the standard static economic model of traffic congestion. The implied continuous-time – continuous-place congestion technology can be considered as a major extension compared to the technologies considered in earlier economic models of traffic congestion in terms of realism. The specification allows for the fact that congestion in reality typically is a non-stationary state phenomenon, in the sense that drivers' speeds will vary

---

Whereas this ratio for $N = 2500$ is around 0.875, it falls to around 0.86 for $N = 500$, and rises to around 0.898 for $N = 5000$.

[14] A further contrast with Mun's (1999) model is that the approximate toll rule presented here suggests that its application would always lead to a peak of a longer duration, and hence an optimal trip price that always exceeds the price in the no-toll equilibrium. Mun (1999) presented examples where the opposite holds, a result which did, however, not occur in a revised version of his original model (Mun, 2002). For reasons of space, this issue will not be investigated any further in this paper.

continuously over time and place during their trips under congested conditions, and that these patterns will vary over drivers when considering the entire peak period.

An important objective was to investigate the extent to which insights from models with simpler congestion technologies can be expected to be applicable in more realistic cases, where congestion indeed varies continuously over time and place. A counter-intuitive finding is that a toll schedule based on the insights of the standard static economic model of traffic congestion (coined the 'approximate optimal toll schedule' – a straightforward dynamic and space-varying generalization of the standard Pigouvian tax rule), outperformed toll schedules based on Vickrey's (1969) pure bottleneck model, even though the model's no-toll equilibrium would suggest a much closer correspondence with the latter than with the former model. A simplistic copying of the optimal toll schedule applying in the pure bottleneck model was even shown to lead to a welfare loss in the numerical version of the model, the relative size of which increases with the level of demand during the peak. The qualitative properties of the (approximate) optimum, compared to the no-toll equilibrium, were shown to be similar to those found by Chu (1995) for a 'no-propagation' dynamic flow-based congestion technology, but again it was argued that a naïve copying of the tax rule in Chu (1995) would not be optimal for the present model. The same holds for the optimal toll schedule found by Mun (1999). These findings confirm the relevance of using a fully specified continuous-time – continuous-place formulation for the design of optimal toll schedules.

The results suggest that the approximate toll schedule, although not truly optimal, is sufficiently close to optimal to justify its use in the current setting. This means that the task of designing a (near-)optimal toll schedule for a road on which congestion takes the form as assumed here may be easier than perhaps suggested by the prohibitive difficulties encountered in attempting to find the model's true optimum. The evaluation of the approximate tax rule requires knowledge only of the value of time and the car-following equation. Clearly, the optimum schedule can be predicted only if demand responses can be predicted correctly, which would require knowledge of schedule delay cost functions, but the same consideration would apply for other dynamic models.

A second main objective was to provide further insight into the phenomenon of hypercongestion, which is a much debated issue in the economic literature on road traffic congestion. The present model avoids the ambiguity of the standard static model, which in general cannot explain whether or not hypercongestion will occur, and instead shows that hypercongestion will occur in a dynamic equilibrium in a queue caused by a downstream bottleneck provided the level of demand is sufficiently large (relative to the reduction in capacity due to the bottleneck). Perhaps surprisingly, hypercongestion was shown to occur where capacity is relatively large. No severe and 'structural' hypercongestion on the lowest capacity segments of the road need be expected, but maximum flows will be approached instead. Hypercongestion is thus generated as a – transitional and local – dynamic equilibrium phenomenon in the present model.

References

Agnew, C.E. (1977) "The theory of congestion tolls" *Journal of Regional Science* **17** (3) 381-393.

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

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

Bernstein, D. (1994) "Non-existence of Nash equilibria for the deterministic departure time choice problem", manuscript.

Braid, R.M. (1989) "Uniform versus peak-load pricing of a bottleneck with elastic demand" *Journal of Urban Economics* **26** 320-327.

Chu, X. (1995) "Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach" *Journal of Urban Economics* **37** 324-343.

Evans, Alan W. (1992) "Road congestion: the diagrammatic analysis" *Journal of Political Economy* **100** (1) 211-217.

Henderson J.V. (1974) "Road congestion: a reconsideration of pricing theory" *Journal of Urban Economics* **1** 346-365.

Hills, P. (1993) "Road congestion pricing: when is it a good policy?: a comment" *Journal of Transport Economics and Policy* **27** 91-99.

Lighthill, M.J. and G.B. Whitham (1955) "On kinematic waves, II A theory of traffic flow on long crowded roads" *Proceedings of the Royal Society (London)*, **229A**, 317-345.

Lindsey, C.R. (2001) "Existence, uniqueness and trip cost function properties of user equilibrium in the bottleneck model with multiple user classes". Working paper, Department of Economics, University of Alberta.

Lindsey, C.R. and E.T. Verhoef (2000) "Congestion modelling". In: D.A. Hensher and K.J. Button (eds.) (2000) *Handbook of Transport Modelling, Handbooks in Transport 1* Elsevier / Pergamon, Amsterdam, pp. 353-373.

Mahmassani, H. and R. Herman (1984) "Dynamic user equilibrium departure time and route choice on idealized traffic arterials" *Transportation Science* **18** 362-384.

May, A.D. (1990) *Traffic Flow Fundamentals* Prentice Hall, Englewood Cliffs, New Jersey.

Mun, S.-I. (1999) "Peak-load pricing of a bottleneck with traffic jam" *Journal of Urban Economics* **46** 323-349.

Mun, S.-I. (2002) "Bottleneck congestion with traffic jam: a reformulation and correction of earlier result". Working paper, Graduate School of Economics, Kyoto University.

Newell, G.F. (1988) "Traffic flow for the morning commute" *Transportation Science* **22** (1) 47-58.

Ohta, H. (2001) "Probing a traffic congestion controversy: density and flow scrutinized" *Journal of Regional Science* **41** 659-680.

Pigou, A.C. (1920). *Wealth and Welfare*. Macmillan, London.

Richards, P.I. (1956) "Shock waves on the highway" *Operations Research*, **4**, 42-51.

Small, K.A. (1982) "The scheduling of consumer activities: work trips" *American Economic Review* **72** 467-479.

Small, K.A. (1992) *Urban Transportation Economics* Fundamentals of Pure and Applied Economics 51, Harwood, Chur.

Small, K.A. and X. Chu (1997) "Hypercongestion" Paper prepared for the meeting of the American Real Estate and Urban Economics Association, New Orleans, Jan. 1997.

Verhoef, E.T. (1999) "Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing" *Regional Science and Urban Economics* **29** 341-369.

Verhoef, E.T. (2001) "An integrated dynamic model of road traffic congestion based on simple car-following theory: exploring hypercongestion" *Journal of Urban Economics* **49** 505-542.

Verhoef, E.T. (2002) "Second-best congestion pricing in general networks: algorithms for finding second-best optimal toll levels and toll points" *Transportation Research* **36B** 707-729.

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

Walters, A.A. (1961) "The theory and measurement of private and social cost of highway congestion" *Econometrica*, **29**, 676-697.

Appendix 1. Alternating initial lane choice in a dynamic equilibrium

In Section 3.2.2 it was mentioned that in a dynamic equilibrium, drivers will alternately choose the left and right lane when starting their trips. This appendix will provide a proof for this assertion. Specifically, it will be shown that given the departure time for driver *i* that would equate his trip price to that of his leader *i–1* when starting on the other lane than driver *i–1* did, his travel time and hence trip costs would be higher when making the other choice, if the dynamic equilibrium condition of equal trip prices is fulfilled for all earlier drivers. (In a similar fashion, the counterpart could be proven that given the departure time for driver *i* that would equate his trip price to that of his leader *i–1* when starting on the same lane as driver *i–1* did, his travel time and hence trip costs would be lower when making the other choice.)

First observe that in a dynamic equilibrium overtaking cannot occur, as it would violate the constant trip price condition (beyond x_2, overtaking is physically impossible anyway with a car-following equation for which S falls to 0 for a positive δ_{min}). This means that driver *i–2* will be ahead of *i–1*, and driver *i–1* ahead of *i*, throughout their trips. Driver *i* will therefore certainly follow *i–1* during the final segment of the trip. But having followed driver *i–2* as long as possible (call this trip 2), rather than following driver *i–1* throughout (trip 1), means that he will be at x_2 sooner: because driver *i–2* is ahead of driver *i–1* throughout, following him when possible allows a higher speed for every x; see also (4). And being at x_2 sooner means that *i* will also be at X sooner. This can be understood by observing that at the instant driver *i* would be at x_2 with trip 1, he is already at a certain $x>x_2$ with trip 2 (consider positive speeds only). Hence, at that instant t, we have $\delta_1^t>\delta_2^t$ (subscripts now identify the two trips considered). From that instant onwards, trip 1 will 'catch up' on trip 2 according to:

$$\frac{\mathrm{d}\left(\delta_1^{\,t}-\delta_2^{\,t}\right)}{\mathrm{d}\,t}=S(\delta_2^{\,t})-S(\delta_1^{\,t}) \qquad \text{(A.1.1)}$$

According to this equation, the speed for trip 1 will approach that for trip 2 asymptotically from above: as long as $\delta_1^t>\delta_2^t$, $\delta_1^t-\delta_2^t$ is decreasing over time (recall that $S'<\infty$). This implies that δ_1^t will asymptotically approach δ_2^t from above, and hence that trip 2 is completed earlier than trip 1. (This proof closely follows that of Lemma 1 in Verhoef, 2001). Choosing the same lane as driver *i–1* and thus trip 1 would therefore indeed lead to a higher trip price.

Appendix 2 Existence and uniqueness of a dynamic equilibrium

An important question is whether a dynamic equilibrium as defined in Section 3.3 exists, and if so, whether it is unique. That is: is there one set of N departure times that implies an equal trip price for all drivers, and a higher trip price for arrivals before the first or after the last driver – ignoring trivial non-uniqueness resulting from the discreteness of vehicles (*i.e.*, a very small shift in the first driver's departure time will typically allow another equilibrium for N drivers)? This question will be considered below. In particular, it will be proven that for a given departure time $t_D^1 << 0$ for the first driver i=1, such that driver 1 arrives before t = 0, there will be a unique set of $N(t_D^1)$ departure times of the following drivers that equate their trip prices to the target equilibrium price p^* applying for driver 1, provided some plausible technical assumptions are satisfied. It is considered unnecessary to prove that the equilibrium number of drivers N will be a step-wise decreasing function of t_D^1, where the steps result from the discreteness of the vehicles. In what follows, it is assumed that t_D^1 is chosen to be consistent with N according to this step-function.

First observe that the first driver (i=1) will drive at S^* throughout his trip by definition as he has no leader; and that in equilibrium, the last driver (i=N) must be driving (approximately) at S^* throughout his trip and arrive at $t_A^N \approx -(\beta \cdot t_A^1)/\gamma$ (t_A^i denotes driver i's arrival time). If arrivals would have terminated well before t_A^N thus defined, driver 1 could benefit from rescheduling and arriving, after a trip at a constant speed S^*, approximately δ^*/S^* time units after the original last driver N, and thus save on c^{sd} without incurring higher c^{tt}. And if an arrival at t_A^N would involve a travel time significantly above X/S^*, the driver arriving at that instant could benefit from rescheduling and arriving, after a trip at speed S^*, just before the original first driver, and thus save on c^{tt} without incurring higher c^{sd}. Likewise, any driver arriving after t_A^N could save on c^{sd} and possibly c^{tt} by making a similar adjustment. The near-equality of c^{tt} for drivers 1 and N thus implies that in equilibrium, also near-equality of c^{sd} must hold; *i.e.*, $-\beta \cdot t_A^1 \approx \gamma \cdot t_A^N$. We assume in the remainder that t_D^1 is chosen in the allowable range for N users such that driver N departs exactly at $t_A^N = (\beta \cdot t_A^1)/\gamma$ and therefore must travel at S^* to obtain the equilibrium trip price p^* experienced by driver 1.

The next question we consider is whether for every other driver i, with 1<i<N, there exists one single departure time t_D^i that implies a trip price $p^i(t_D^i)$ equal to p^*, given the departure times of all drivers 1,…,*i–1* and N, given that these departure times imply trip prices for these other drivers equal to p^*, and under the assumption of a 'temporary absence' of drivers *i+1…N–1*. If this can be proven for every i, the same equilibrium price p^* would then be proven to apply for all drivers only under a unique set of departure times. We consider the properties of the function $p^i(t_D^i)$ over the relevant open domain $\left(t_D^{i-1}, t_A^N - X/S^*\right)$.

First, if driver i departs immediately after driver i–1, $p^i(t_D^{i-1}+\varepsilon) > p^{i-1}(t_D^{i-1}) = p^*$ will hold (with ε being a very small positive number). As argued in Appendix 1, driver i–1 has selected the most preferable leader on the first segment, and will therefore arrive earlier than driver i. With a positive 'finishing speed', driver i cannot arrive immediately after driver i–1 (an absolute minimum time span is approximately equal to $1/F_{max} > \varepsilon$). Because of his longer travel time, driver i will therefore have a higher trip price than driver i–1 (who had p^* by construction), both for early arrivals and late arrivals (recall that $\gamma > \alpha > \beta > 0$). Call this result 1.

Next, consider departure times immediately (ε seconds) before $t_A^N - X/S^*$; that is, just before that of the final driver *N*. Because driver *N* drives at a free-flow speed S^* and because the highest *i* we can consider in our thought-experiment is driver *N–1*, the average speed for these departure times will be (approximately) constant and equal to S^*, so that $p^i(t_D^i)$ will be increasing with a slope (approximately) equal to γ (a marginal change in departure time hardly affects *tt*, and therefore leads to an equally sized change in the arrival time), approaching p^* from below as t_D^i approaches $t_A^N - X/S^*$ from below. Call this result 2.

Given results 1 and 2, there will be only one departure time t_D^i that implies a trip price $p^i(t_D^i)$ equal to p^* provided the function $p^i(t_D^i)$ has at most one local minimum over the open interval $\left(t_D^{i-1}, t_A^N - X/S^*\right)$ considered. Rewriting equation (1) in the main text in terms of departure times yields in absence of tolls:

$$p(t_D^i) = \begin{cases} \alpha \cdot tt(t_D^i) - \beta \cdot \left(t_D^i + tt(t_D^i)\right) & \text{for } t_D^i + tt(t_D^i) \le t^* = 0 \\ \alpha \cdot tt(t_D^i) + \gamma \cdot \left(t_D^i + tt(t_D^i)\right) & \text{for } t_D^i + tt(t_D^i) > t^* = 0 \end{cases} \tag{A.2.1}$$

from which we derive the following partial derivatives:

$$\frac{\mathrm{d}p(t_D^i)}{\mathrm{d}t_D^i} = \begin{cases} (\alpha - \beta) \cdot \dfrac{\mathrm{d}tt(t_D^i)}{\mathrm{d}t_D^i} - \beta & \text{for } t_D^i + tt(t_D^i) \le t^* = 0 \\ (\alpha + \gamma) \cdot \dfrac{\mathrm{d}tt(t_D^i)}{\mathrm{d}t_D^i} + \gamma & \text{for } t_D^i + tt(t_D^i) > t^* = 0 \end{cases} \tag{A.2.2}$$

We make the plausible technical assumption that $tt(t_D^i)$ is continuous and decreases in a convex fashion in t_D^i for departures close after t_D^{i-1}, until it possibly reaches a constant value for later departures when driver *i* would drive at S^* throughout his trip. Because $\gamma > \alpha > \beta > 0$, the derivative $\mathrm{d}p^i(t_D^i)/\mathrm{d}t_D^i$ will then be negative throughout for early arrivals; will be positive for departures near $t_A^N - X/S^*$ where $\mathrm{d}tt(t_D^i)/\mathrm{d}t_D^i \approx 0$, and will be an increasing function of t_D^i and will therefore be equal to zero at most once for late arrivals. This means that the function $p^i(t_D^i)$ indeed has at most one local minimum over the open interval $\left(t_D^{i-1}, t_A^N - X/S^*\right)$ considered, which next implies that there is indeed one unique dynamic equilibrium. The plausible technical assumptions we make without providing a formal proof therefore concern the continuity and convexity of the function $tt(t_D^i)$.