TIME-VARYING TOLLS IN A DYNAMIC MODEL OF ROAD TRAFFIC CONGESTION WITH ELASTIC DEMAND

Erik Verhoef
Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
Email: everhoef@econ.vu.nl

Abstract

In this paper, a dynamic model of road traffic congestion is presented, with an elastic overall demand for morning peak road usage, and with the congestion technology used being 'flow congestion'. It is demonstrated that in such a case, the optimal time-varying toll should include a 'flat', time-invariant component when road users share the same desired arrival time. This has important consequences for the design of optimal toll schemes in reality, because it implies that optimal tolls cannot be set if the regulator has no information on the road users' desired arrival times.

1. Introduction

Road traffic congestion and road pricing have been studied by economists for a long time already, with seminal contributions provided by, for instance, Pigou (1920), Knight (1924), Walters (1961) and Vickrey (1969). Various approaches in modelling traffic congestion have been taken, most of which are static in nature (see, for instance, Else, 1981, Evans, 1992, Verhoef *et al.*, 1995ab, 1996ab). However, it has often been emphasized that road traffic congestion in reality is an essentially dynamic process, the modelling of which would therefore require a dynamic approach. Two families of such dynamic approaches can be distinguished. The first one is the so-called 'bottleneck approach', originally developed by Vickrey (1969), and later on refined end extended in various directions by Arnott *et al., 1*990ab, 1991ab, 1992, 1993, 1994), Braid (1989, 1996), and others (see Arnott *et al.*, 1997, for a comprehensive overview). In this approach, commuters jointly using a road network wish to be at work at the same time, but due to a limited capacity somewhere in the network (the bottleneck), this is physically impossible, and schedule delay costs are therefore unavoidable. In the unregulated equilibrium, a queue builds up and decays during the morning peak, whereas in the optimum, travel delays – which are a pure loss in this approach – are completely avoided and only schedule delay costs remain (at least, on a one-link network). Although the bottleneck approach may offer a realistic representation of congestion for some situations (e.g., at bridges or tunnels), it is not necessarily the best modelling approach in all cases; in particular because it is rather rigid in assuming that drivers either drive at the free-flow speed, or have a zero speed when waiting in the queue. The second approach, originally proposed by Henderson (1974, 1981) uses 'flow congestion', and is capable of dealing with situations in which speeds decrease to still positive values due to congestion. In this model, travel delays to some extent remain existent also in the social optimum (see Chu, 1995, for an interesting comparison between these two approaches).

In both models, the distribution of travel delays and scheduling costs over the peak and the duration of the peak in the unregulated equilibrium and the social optimum are determined endogenously. An important equilibrating principle in both approaches is that private costs of road usage, consisting of travel time costs, schedule delay costs, and the toll in case one is levied, should not vary over the peak. Otherwise, some drivers could benefit from choosing a different departure time, and the system could not be in equilibrium. Furthermore, both models have in common that the optimal toll is time-dependent, reaching its peak for the drivers arriving at the desired arrival time.

Probably because of the complexities resulting from a dynamic modelling approach, these models typically assume an inelastic demand for road usage, with the exceptions of Braid (1989) and Arnott *et al.* (1993). It is therefore not surprising that the optimal time-varying tolls typically start at zero in the beginning of the peak and return to zero at the end. Any additional flat toll component would merely imply a lump-sum tax, with no impact on travel behaviour.

In this paper, a dynamic model of road traffic congestion is presented, with the congestion technology used being flow congestion, and with an elastic overall demand for morning peak road usage. Such elasticity of demand could result, *inter alia*, from the availability of alternative transport modes. It will be demonstrated that in such a case, the optimal time-varying toll should include a 'flat', time-invariant component when road users share the same desired arrival time. This has important consequences for the design of optimal toll schemes in reality, because it implies that optimal tolls cannot be set if the regulator has no information on the road users' desired arrival times. In particular, a non-zero optimal toll is still in order when 'observable' congestion, in terms of travel time delays, has gone to zero. The reason that this flat toll has not been derived in the literature so far is the typical practice to concentrate only on 'instantaneous' external costs of travel time delays, caused by direct interaction of road users, and to ignore the 'inter-temporal' externality that users impose on all other users throughout the peak, which is caused by the impact of the total level of road usage, over the entire peak, on equilibrium travel times at each instant during this peak. Furthermore, the flat toll component is relevant only with elastic demand, and the study of road traffic congestion with elastic demand has so far been restricted to exercises with the bottleneck model only. As is hypothesized in the appendix, the flat toll component may vanish for the case of bottleneck congestion. Clearly, now that electronic road pricing, allowing time-varying tolling, is likely to be introduced soon at various places, this conclusion is not only of academic relevance, but may have great practical importance as well.

The plan of this paper is as follows. Section 2 introduces the demand side of the model, which is well in line with earlier dynamic models of peak-hour congestion. Section 3 solves the model for the case of inelastic demand. The congestion technology used is 'flow congestion', as in the Henderson approach. Section 4 considers elastic demand; and Section 5 presents a simple numerical illustration, showing the non-triviality of the flat toll. Finally, Section 6 concludes.

2. The demand side of the model and the implications for equilibria

Suppose that a number of commuters consider using the same one-link road network of length L to get from home to work. In what follows, a slightly different definition for the various relevant values of time than is common in the literature will be used. In particular, instead of using a value of travel time as such, it is assumed that this value arises because, when travelling, people are not able to spend time at home or at work. The value of time at home is constant and is denoted α. Letting t denote arrival time, the value of time at work before the desired arrival time t^* is constant and denoted β, and after t^* it is also constant and denoted γ. In order to make t^* indeed the desired arrival time, it has to be assumed that $\beta<\alpha<\gamma$. The parameters α, β, γ, and t^* are equal across individuals. The free-flow speed on the road is denoted S^*, and when driving alone, a user would spend the minimum possible travel time $T^*=L/S^*$ on the road. The minimum possible total travel costs (including travel

time and schedule delay costs) would arise if there were only one user, who is able to drive at speed S^* and would choose her arrival time t so as the maximize the total value of time enjoyed. This occurs if the arrival time is exactly at t^*: travelling earlier and arriving at t would imply an additional loss of $(\alpha-\beta)(t^*-t)$, and arriving later implies an additional loss of $(\gamma-\alpha)(t-t^*)$. This single user would incur the minimum possible total travel costs of $\alpha \cdot L/S^*$, compared to the hypothetical situation in which L=0 and travelling is not necessary.

In general, assuming that users do not consider speed variations during the trip as an additional source of disutility in itself, but only care about the total time spend travelling as preventing them from being either at home or at work, and of course care about actual arrival times, an individual's travel cost for arriving at t, after a trip with average speed S, can be written as:

$$c_t = \alpha \cdot \left(t^* - t + L / S_t\right) + (t - t^*) \cdot \left(\delta \cdot \beta + (1-\delta) \cdot \gamma\right) \tag{1}$$

where $L/S_t=T_t$ gives the travel time, which may vary according to the time of arrival t due to changes in average speeds S_t experienced, and δ is a dummy taking on the value of 1 in case of early arrival ($t<t^*$) and the value of 0 in case of late arrival ($t>t^*$). Equation (1) can most easily be verified by considering deviations from the minimum possible travel cost implied by driving at speed S^* and arriving at t^*. Arriving at t instead of t^* implies additional scheduling costs of $(\alpha-\beta)(t^*-t)$ for $t<t^*$ and $(\gamma-\alpha)(t-t^*)$ for $t>t^*$. Furthermore, driving at a lower average speed than S^* implies that one should depart earlier to arrive at t, yielding additional costs valued at α.

In the unregulated no-toll equilibrium, these travel costs should not vary over time, at least not as long as the road is actually used, because otherwise some drivers could benefit from unilaterally changing their arrival time. Denoting the change over time with a dot, this implies that in the no-toll equilibrium:

$$\dot{c}_t = -\alpha + \alpha \cdot L \cdot \frac{-\dot{S}_t}{S_t^{\,2}} + \delta \cdot \beta + (1-\delta) \cdot \gamma = 0 \tag{2}$$

Solving the implied differential equation for S_t gives the following equilibrium pattern of average speeds, experienced by drivers arriving at t, over the peak:

$$S_t = \frac{1}{t \cdot \dfrac{\alpha - \delta \cdot \beta - (1-\delta) \cdot \gamma}{\alpha \cdot L} + \delta \cdot A_1 + (1-\delta) \cdot A_0} \tag{3}$$

where A_i give constants to be determined. To determine these, it is observed that the very first (and very last) driver should have speed S^*, because otherwise they could make a non-marginal gain by departing earlier (later), drive at S^* and arriving only marginally earlier (later) than they would have by choosing to drive under congestion. Denoting the arrival times of the first and last driver(s) as t_F and t_L respectively, (3) can be solved as:

$$S_t = \frac{1}{(t - t_F) \cdot \frac{\alpha - \beta}{\alpha \cdot L} + \frac{1}{S^*}} \quad \textit{for } t_F \le t \le t^* \tag{4a}$$

$$S_t = \frac{1}{(t_L - t) \cdot \frac{\gamma - \alpha}{\alpha \cdot L} + \frac{1}{S^*}} \quad \textit{for } t^* \le t \le t_L \tag{4b}$$

It is clear from (4ab) that in equilibrium, average speeds experienced start at a value of S^* for the driver(s) arriving at t_F, subsequently decrease and reach their minimum at t^*, and then increase until they reach S^* again at t_L. Furthermore, since only one speed can prevail at t^*, the entire duration of the peak can be split into two periods of generally unequal lengths according to:

$$\frac{t^* - t_F}{t_L - t^*} = \frac{\gamma - \alpha}{\alpha - \beta} \tag{5}$$

Equations (4ab) can be substituted into (1) to obtain the following expressions for the travel costs over time:

$$c_t = \alpha \cdot \left(t^* - t + L \cdot \left((t - t_F) \cdot \frac{\alpha - \beta}{\alpha \cdot L} + \frac{1}{S^*} \right) \right) + \beta \cdot (t - t^*) \quad \textit{for } t_F \le t \le t^* \tag{6a}$$

$$c_t = \alpha \cdot \left(t^* - t + L \cdot \left((t_L - t) \cdot \frac{\gamma - \alpha}{\alpha \cdot L} + \frac{1}{S^*} \right) \right) + \gamma \cdot (t - t^*) \quad \textit{for } t^* \le t \le t_L \tag{6b}$$

It is easily checked that the costs implied by (6ab) are constant over time, as required, and are equal to:

$$c = (\alpha - \beta) \cdot (t^* - t_F) + \frac{\alpha \cdot L}{S^*} = (\gamma - \alpha) \cdot (t_L - t^*) + \frac{\alpha \cdot L}{S^*} \tag{7}$$

From (7), it follows that ‘it takes a peak to have congestion’: if the peak has no positive duration so that $t_F = t^* = t_L$, all drivers have the minimum possible total travel cost of $\alpha \cdot L / S^*$. Likewise, higher travel costs occur only if $t_F < t^* < t_L$.

The same steps can be taken to find the equilibrium pattern of speeds in case tolling, possibly time-varying, is applied. Denoting the total toll incurred during the trip when arriving at time t as τ_t, the private cost including the toll can be written as $k_t = c_t + \tau_t$:

$$k_t = \alpha \cdot \left(t^* - t + L / S_t \right) + (t - t^*) \cdot \left(\delta \cdot \beta + (1 - \delta) \cdot \gamma \right) + \tau_t \tag{8}$$

In the toll equilibrium, these private costs including the toll should be constant:

$$\dot{k}_t = -\alpha + \alpha \cdot L \cdot \frac{-\dot{S}_t}{S_t^{\,2}} + \delta \cdot \beta + (1 - \delta) \cdot \gamma + \dot{\tau} = 0 \tag{9}$$

As long as the tolls are not decreasing in time at t_F and increasing in time at t_L, also now the first and last driver should drive at speed S^*, for the same reason as before. This information is used in finding the solution to the differential equation (9), which is solved in the same way as (2). The following solution is obtained:

$$S_t = \frac{1}{(t-t_F)\cdot\frac{\alpha-\beta}{\alpha\cdot L}-\frac{1}{\alpha\cdot L}\cdot(\tau_t-\tau_F)+\frac{1}{S^*}} \quad \textit{for } t_F \le t \le t^* \tag{10a}$$

$$S_t = \frac{1}{(t_L-t)\cdot\frac{\gamma-\alpha}{\alpha\cdot L}-\frac{1}{\alpha\cdot L}\cdot(\tau_t-\tau_L)+\frac{1}{S^*}} \quad \textit{for } t^* \le t \le t_L \tag{10b}$$

where τ_F and τ_L denote the tolls at t_F and t_L, respectively. Substitution of (10ab) in (8) yields the following expressions for the travel costs, including the toll, over time:

$$k_t = \alpha\cdot\left(t^*-t+L\cdot\left((t-t_F)\cdot\frac{\alpha-\beta}{\alpha\cdot L}-\frac{1}{\alpha\cdot L}\cdot(\tau_t-\tau_F)+\frac{1}{S^*}\right)\right)+\beta\cdot(t-t^*)+\tau_t \tag{11a}$$

$$\textit{for } t_F \le t \le t^*$$

$$k_t = \alpha\cdot\left(t^*-t+L\cdot\left((t_L-t)\cdot\frac{\gamma-\alpha}{\alpha\cdot L}-\frac{1}{\alpha\cdot L}\cdot(\tau_t-\tau_L)+\frac{1}{S^*}\right)\right)+\gamma\cdot(t-t^*)+\tau_t \tag{11b}$$

$$\textit{for } t^* \le t \le t_L$$

It is again easily checked that the travel costs including the toll, implied by (11ab), are indeed constant over time and are equal to:

$$k = (\alpha-\beta)\cdot(t^*-t_F)+\tau_F+\frac{\alpha\cdot L}{S^*} = (\gamma-\alpha)\cdot(t_L-t^*)+\tau_L+\frac{\alpha\cdot L}{S^*} \tag{12}$$

It is worth emphasising that the term τ_t has dropped out of (12), which is a direct consequence of the equilibrium property of private costs including the toll being constant over time.

Before switching to the supply side of the model, it is worthwhile considering the question of whether overtaking is possible in the no-toll and time-varying toll equilibria. Only the latter case is considered, since the no-toll equilibrium is simply a special case of the time-varying toll equilibrium where the toll is constant at a zero level. Overtaking in this model means that the driver being overtaken has a later arrival time and an earlier departure time than someone else. With t still denoting arrival times, let $t_{D,t}$ denote the departure time of a driver arriving at t. The relation between t and $t_{D,t}$ is: $t_{D,t}=t-L/S_t$. Substituting (10ab) into this expression, it can be derived that, in the time-varying toll equilibrium, the implied change in $t_{D,t}$ due to a change in the arrival time t is given by:

$$\dot{t}_{D,t} = \frac{\beta}{\alpha}+\frac{1}{\alpha}\cdot\dot{\tau}_t \quad \textit{for } t_F \le t \le t^* \tag{13a}$$

$$\dot{t}_{D,t} = \frac{\gamma}{\alpha}+\frac{1}{\alpha}\cdot\dot{\tau}_t \quad \textit{for } t^* \le t \le t_L \tag{13a}$$

For the no-toll equilibrium, the second terms drop out, and since the remaining terms are positive, it is clear that overtaking does not occur in the no-toll equilibrium. The fact that the remaining term is smaller than 1 before t^*, and greater than 1 after t^*, reflects the decrease and increase in average speeds before and after t^*. Finally, it turns out that overtaking could in principle be induced by time varying tolls if they decrease sufficiently rapidly in time. Before t^*, this is unlikely in any case, as one would expect increasing tolls. After t^*, overtaking could be induced if the tolls decrease so rapidly that it becomes worthwhile to postpone the arrival time infinitely because the savings in toll exceed the value of time at work after t^*. This, of course, is also an unlikely event in any optimal tolling scheme. Most important, in either case, overtaking can only be induced if tolls decrease so rapidly over time that all trips are postponed infinitely, and overtaking is therefore inconsistent with positive road usage.

3. Optimal time-varying tolls with inelastic demand

The analysis in the previous section applies to any specific form of congestion technology. The users only care about arrival times and total travel times and tolls incurred over the entire trip, regardless of whether actual speeds and toll rates vary during the trip. Such variations during a trip could be modelled with the framework presented above. In order to find the optimal pattern of speeds, tolls and road usage over time, under the restriction of the behavioural responses to tolls as outlined in the previous section, the supply side of the model should be made explicit. Various approaches can be taken here, even if the choice set is restricted to flow congestion, as opposed to bottleneck congestion (Small, 1992 (pp. 61-74), gives an overview of both dynamic and static representations of congestion technology). Apart from the hydro-dynamic fluid model, which is difficult to apply economically and to solve analytically (see Newell, 1988), two extreme variants would be (1) a model in which 'group velocity' is infinite and all vehicles on the link have the same speed at each instant, depending on the total number of vehicles present on the entire link, regardless of their exact position on the link; and (2) a model with zero 'group velocity', in which the speed of a vehicle depends only on the number of vehicles which depart and arrive at the same instants as the vehicle itself, so that different speeds can apply simultaneously at different places along the link. The second approach is taken by Henderson (1974, 1981) and Chu (1995), and will also be followed here. It is of course difficult to say which of the approaches is more realistic; taking the latter approach has as an advantage that the analysis is at least comparable with previous efforts in the dynamic modelling of road traffic congestion.

Because, in this approach, the speed of a certain vehicle is constant during the trip, one can economize on mathematical complexity by using a congestion function of the type $T=T(n)$ $(\partial T/\partial n>0)$, giving the total travel time as a function of the number of vehicles (n) departing and arriving jointly, rather than working with the traditional type of function $S=S(n)$ $(\partial S/\partial n<0)$, showing how average speeds decrease with increasing numbers of vehicles travelling jointly. Since $T(n)=L/S(n)$ by definition, it is evident that there is a one-

to-one mapping between the two types of functions, and the choice of working with T(n) can be made without loss of generality.

For reasons of exposition, first the optimal tolls are derived for a given total number of road users N; so, under the assumption of inelastic demand. This will provide information on how a given number of users is to be distributed most efficiently over time, which is a matter of social cost minimization. The tolling principle derived will subsequently be used in the next section when considering the case of elastic demand. Whichever value of N turns out to be optimal there, it will be clear that also with elastic demand, efficiency requires the optimally determined number of road users to be distributed optimally over time; that is, against minimum social costs. The optimization problem of minimizing the social costs of having N users completing their trips is:

$$\begin{aligned} MAX \quad & \int_{t_F}^{t^*} -n_t \cdot \left((\alpha-\beta)\cdot(t^*-t)+\alpha\cdot T(n_t)\right) dt \quad + \\ & \int_{t^*}^{t_L} -n_t \cdot \left((\gamma-\alpha)\cdot(t-t^*)+\alpha\cdot T(n_t)\right) dt \end{aligned} \tag{14}$$

$$s.t. \quad \int_{t_F}^{t_L} -n_t \; dt = -N; \qquad t_F, t_L \quad free; \qquad T(n_F) = T(n_L) = T^*$$

This is a degenerate isoperimetric dynamic optimization problem with fixed endpoints $T(n_F)=T(n_L)=T^*=L/S^*$: the first and last driver should have the minimum travel time T^*, which will prove to be valid below because the optimal time-varying tolls are found to be increasing (decreasing) in time at t_F (t_L); and free initial and terminal times t_F and t_L, which will have to be determined optimally in solving the problem. The integral constraint secures that total usage is indeed equal to N, and is specified in the negative so that the Lagrangian multiplier will be positive and can therefore be readily interpreted as a shadow price. The Lagrange integrand F for this problem is:

$$\begin{aligned} F = {} & \delta \cdot \left[-n_t \cdot \left((\alpha-\beta)\cdot(t^*-t)+\alpha\cdot T(n_t)\right)\right] + \\ & (1-\delta)\cdot\left[-n_t \cdot \left((\gamma-\alpha)\cdot(t-t^*)+\alpha\cdot T(n_t)\right)\right] + \lambda\cdot n_t \end{aligned} \tag{15}$$

and the Euler-Lagrange equation for this degenerate problem is:

$$\begin{aligned} \frac{\partial F}{\partial n_t} = {} & \delta\cdot\left[-\left((\alpha-\beta)\cdot(t^*-t)+\alpha\cdot T(n_t)\right)-\alpha\cdot n_t\cdot\frac{\partial T}{\partial n_t}+\lambda\right]+ \\ & (1-\delta)\cdot\left[-\left((\gamma-\alpha)\cdot(t-t^*)+\alpha\cdot T(n_t)\right)-\alpha\cdot n_t\cdot\frac{\partial T}{\partial n_t}+\lambda\right]=0 \end{aligned} \tag{16}$$

Because (14) is an isoperimetric problem, the Lagrangian multiplier λ will be constant. This, in combination with (16), implies that in order to minimize the social costs, the 'marginal instantaneous social travel costs' should be constant over time. In other words: the total travel costs cannot be reduced by moving one of the users in time. The fixed

endpoint transversality conditions state that the Lagrange integrand should take on the value of zero at the initial and terminal time, and thus provide the following solution for the multiplier λ:

$$\lambda = (\alpha - \beta) \cdot (t^* - t_F) + \alpha \cdot T^* = (\gamma - \alpha) \cdot (t_L - t^*) + \alpha \cdot T^* \tag{17}$$

Equation (17) fixes the optimal level of the marginal instantaneous social travel costs at the level that also applies for the first and last driver(s). Inserting (17) into (16) yields:

$$\alpha \cdot \left(T(n_t) - T^*\right) = (\alpha - \beta) \cdot (t - t_F) - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t_F \le t \le t^* \tag{18a}$$

$$\alpha \cdot \left(T(n_t) - T^*\right) = (\gamma - \alpha) \cdot (t_L - t) - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t^* \le t \le t_L \tag{18b}$$

At the same time, rewriting (10ab) in terms of T_t yields:

$$\alpha \cdot \left(T(n_t) - T^*\right) = (\alpha - \beta) \cdot (t - t_F) - (\tau_t - \tau_F) \quad \textit{for } t_F \le t \le t^* \tag{19a}$$

$$\alpha \cdot \left(T(n_t) - T^*\right) = (\gamma - \alpha) \cdot (t_L - t) - (\tau_t - \tau_L) \quad \textit{for } t^* \le t \le t_L \tag{19b}$$

so that the following result is obtained for the optimal pattern of tolls over time:

$$\tau_t - \tau_F = \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t_F \le t \le t^* \tag{20a}$$

$$\tau_t - \tau_L = \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t^* \le t \le t_L \tag{20b}$$

So, differences in tolls over time reflect differences in marginal external instantaneous travel delay costs. This secures that the marginal instantaneous social travel costs are constant over time, as is required by (16). Note that the implication of (20ab) that marginal external instantaneous travel delay costs be zero at t_F and t_L is consistent with the boundary conditions that traffic should have the free-flow speed at these instants.

The average social travel costs c_t under optimal time-varying tolling can be found by substitution of (20ab), and of the identity that $k_t=c_t+\tau_t$ into (12):

$$\begin{aligned} c_t &= (\alpha - \beta) \cdot (t^* - t_F) + (\tau_F - \tau_t) + \alpha \cdot T^* = \\ &(\alpha - \beta) \cdot (t^* - t_F) - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} + \alpha \cdot T^* = \\ &\lambda - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t_F \le t \le t^* \end{aligned} \tag{21a}$$

$$
\begin{aligned}
c_t &= (\gamma - \alpha) \cdot (t_L - t^*) + (\tau_L - \tau_t) + \alpha \cdot T^* = \\
&(\gamma - \alpha) \cdot (t_L - t^*) - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} + \alpha \cdot T^* = \\
&\lambda - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \quad \textit{for } t^* \le t \le t_L
\end{aligned}
\tag{21b}
$$

Hence, although the marginal instantaneous social travel costs are constant over time, average social costs are not.

Having derived the optimal pattern of the time-varying toll for any total number of users N, the stage is now set to solve for the optimal N, as well as the optimal t_F, t_L, τ_F, and τ_L in the next section.

4. Optimal time-varying tolls with elastic demand

Earlier attempts to model traffic congestion in a dynamic framework with elastic demand are, to the best of my knowledge, restricted to exercises with the bottleneck model. Both Braid (1989) and Arnott *et al.* (1993) conclude that the optimal tolls at the very beginning and end of the peak – τ_F and τ_L in the above notation – should be zero. Chu (1995), although using a model with inelastic demand, strongly suggests that this should also hold for models with flow congestion, when stating "Since there is no travel delay at i′ [t_F in the present notation] or u′ [t_L], the optimal toll must be zero there" (p. 331). Also in the present model, there are no travel delays at t_F and t_L, suggesting that Chu's reasoning would also apply here.

However, such reasoning focuses on instantaneous externalities through direct interactions on the road only, and ignores the inter-temporal externality that road users pose upon each other. In particular, one could argue that taking away one user from the network would lead to a new optimal pattern of usage and social costs over the entire peak, implying that not only those users who are driving at exactly the same time as the one that is taken away will benefit, but, after the new equilibrium has established, all drivers will. In order to see whether this is the case, one should solve the model presented in the previous section for the optimal N, t_F, t_L, τ_F, and τ_L, knowing that the optimal distribution of users and tolls over time should also in this case be consistent with (16) and (20ab). When switching to elastic demand, N is of course no longer fixed. This implies that also the Lagrangian multiplier λ is no longer fixed, but becomes an endogenous variable $\lambda(N)$. Furthermore, an inverse demand relation D(N) should be introduced, where total benefits can be found as the area under the demand curve up to N: $\int_0^N D(N)\, dn$. Using the expressions for average social costs obtained in (21ab), the social optimization problem – given the condition that optimal time varying tolls according to (20ab) are used – becomes:

$$
MAX \quad \int_0^N D(N)\, dn \quad - \quad \left(N \cdot \lambda(N) - \int_{t_F}^{t_L} n_t \cdot \left(\alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \right) dt \right) \tag{22}
$$

The first-order condition is:

$$D(N) - \lambda(N) - N \cdot \frac{\partial \lambda}{\partial N} + \frac{\partial \int_{t_F}^{t_L} n_t \cdot \left(\alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \right) dt}{\partial N} = 0 \tag{23}$$

which can be rewritten as:

$$D(N) - \lambda(N) - \frac{1}{\partial N} \cdot \left(\int_{t_F}^{t_L} n_t \cdot \partial \lambda \; dt - \partial \int_{t_F}^{t_L} n_t \cdot \left(\alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \right) dt \right) = 0 \tag{24}$$

and, because $\alpha \cdot n_t \cdot \partial T / \partial n_t$ is equal to zero at t_F and t_L, and ∂n_t goes to zero more quickly than $\partial(\alpha \cdot n_t \cdot \partial T / \partial n_t)$ also as:

$$D(N) - \lambda(N) - \frac{1}{\partial N} \cdot \left(\int_{t_F}^{t_L} n_t \cdot \partial \left(\lambda - \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \right) dt \right) = 0 \tag{25}$$

Using (16), (25) can finally be written as:

$$D(N) - \lambda(N) - \frac{1}{\partial N} \cdot \left(\int_{t_F}^{t_L} n_t \cdot \partial \left(\alpha \cdot T(n_t) \right) dt \right) = 0 \tag{26}$$

At the same time, (12) shows that, under time-varying tolling, the private costs of road usage including the toll, is constant over time and is equal to (after substitution of (17)):

$$k(N) = \lambda(N) + \tau_F = \lambda(N) + \tau_L \tag{27}$$

Knowing that road users will keep on entering the road up to the point where D(N)=k(N), the following optimal tax rules for τ_F and τ_L can therefore be derived:

$$\tau_F = \tau_L = \frac{1}{\partial N} \cdot \left(\int_{t_F}^{t_L} n_t \cdot \partial \left(\alpha \cdot T(n_t) \right) dt \right) \tag{28}$$

(see also the Appendix for a graphical illustration of the crucial result that marginal social costs exceed $\lambda(N)$). Inserting (28) in (20), the following expression for the optimal time-varying toll can finally be derived:

$$\tau_t = \frac{1}{\partial N} \cdot \left(\int_{t_F}^{t_L} n_t \cdot \partial \left(\alpha \cdot T(n_t) \right) dt \right) + \alpha \cdot n_t \cdot \frac{\partial T}{\partial n_t} \tag{29}$$

Clearly then, in addition to the time-varying component in the toll as derived in Section 3, covering the instantaneous marginal external travel delay costs, also a ‘flat’, time-invariant toll component should be present, covering the inter-temporal externality which is equal to the marginal external costs that a user imposes on all other users, because of the impact of the total usage N on equilibrium travel times, and hence instantaneous travel costs, at all instants.

This conclusion has important implications. In particular, it implies that optimal tolls not only depend on the actual situation observed on the network: at t_F and t_L, where no observable congestion occurs, a positive toll is still in order. Instead, apart from observing what is going on the road, the regulator needs to have information on the distribution of desired arrival times before being able to set tolls optimally, because the underlying reason of the fixed component in the optimal time-varying toll in (29) is the assumed equality of desired arrival times. This can most easily be understood by considering the extreme example in which the actual arrival rates obtained by using (20) with $\tau_F=\tau_L=0$ would happen to coincide with the distribution of desired arrival times. In that case, (20) indeed produces the optimal pattern of road usage over time, whereas at the same time we know that the optimum is an equilibrium because no user has the incentive to change arrival times. Any additional flat toll would then reduce overall welfare.

5. A simple numerical illustration

This section presents an admittedly simple numerical illustration of the flat component of the optimal time-varying toll. It should be emphasized that the purpose of the simulation model presented is not to give any realistic representation of the dynamics of road traffic congestion; the only purpose is to present a system that contains the relevant elements to the model discussed above, and that can reproduce the flat toll component in an analytically tractable mathematical formulation.

	No-tolls	Time-varying tolls
$\dot{n}_t$	$-\frac{\nu}{\alpha \cdot k}$ *for* $n_t > 1$	$-\frac{\nu}{2 \cdot \alpha \cdot k}$ *for* $n_t > 1$
n_t	$(t_E - t) \cdot \frac{\nu}{\alpha \cdot k} + 1$	$(t_E - t) \cdot \frac{\nu}{2 \cdot \alpha \cdot k} + 1$
n_0	$t_E \cdot \frac{\nu}{\alpha \cdot k} + 1$	$t_E \cdot \frac{\nu}{2 \cdot \alpha \cdot k} + 1$
$N = \frac{1}{2} \cdot t_E \cdot (n_0 - 1) + t_E \cdot 1$	$\frac{\nu}{2 \cdot \alpha \cdot k} \cdot {t_E}^2 + t_E$	$\frac{\nu}{4 \cdot \alpha \cdot k} \cdot {t_E}^2 + t_E$
t_E	$\frac{-1 + \sqrt{1 + 2 \cdot \frac{\nu}{\alpha \cdot k} \cdot N}}{\frac{\nu}{\alpha \cdot k}}$	$\frac{-1 + \sqrt{1 + \frac{\nu}{\alpha \cdot k} \cdot N}}{\frac{\nu}{2 \cdot \alpha \cdot k}}$
c_t	$\nu \cdot t_E + \alpha \cdot T^*$	$\frac{1}{2} \cdot \nu \cdot t + \frac{1}{2} \cdot \nu \cdot t_E + \alpha \cdot T^*$
$TC = \int_0^{t_E} n_t \cdot c_t$	$N \cdot (\nu \cdot t_E + \alpha \cdot T^*)$	$N \cdot \alpha \cdot T^* + \frac{\nu^2}{6 \cdot \alpha \cdot k} \cdot {t_E}^3 + \frac{3}{4} \cdot \nu \cdot {t_E}^2$

Table 1. Some equilibrium values of the simulation model

In the first instance, the inverse demand is assumed to be linear and given by:

$$D = d_1 - 0.5 \cdot d_2 \cdot N \tag{30}$$

and the function giving travel times as a function of n_t is piecewise linear:

$$T_t = T^* + k \cdot (n_t - 1) \tag{31}$$

where the kink is needed to avoid zero usage at t_F and t_L. By setting $\alpha-\beta=\gamma-\alpha=v$, a symmetric system is created that allows us to study 'half the peak' only, knowing that the other half is an exact mirror image. This allows us to set $t^*=0$ and consider positive t only; to avoid confusion, the single terminal time is denoted t_E. The demand curve for this half of the peak becomes:

$$D = d_1 - d_2 \cdot N \tag{32}$$

The average cost function can then be written as:

$$\begin{aligned} c_t &= v \cdot t + \alpha \cdot \left(T^* + k \cdot (n_t - 1)\right) \quad for\ n_t > 1 \\ c_t &= v \cdot t + \alpha \cdot T^* \quad for\ n_t \le 1 \end{aligned} \tag{33}$$

Using that average costs should be constant over time in the no-toll equilibrium, and that marginal instantaneous social travel costs should be constant over time with optimal time-varying tolls, the expressions presented in Table 1 can be derived (where TC denotes total costs). These expressions were used in the simulation model, which generated the curves in Figures 1-3.

In the numerical example, the following parameter values are chosen: $d_1=1000$; $d_2=1$; $\alpha=10$; $v=10$; $k=10$; and $T=10$. A first observation that can be made is that the cost functions are endogenous; that is, dependent on the question of whether time-varying tolling is applied. If so, total, marginal and average social costs are lower for a given number of users. This is shown in Figure 1, where the total costs (TC) with and without time-varying tolling are depicted for the parameter values given above. In this and the following figures, curves representing the case with time-varying tolling are marked with an asterisk and those representing the case without time-varying tolling with an apostrophe.

Figure 2 shows the various market outcomes that can arise in the numerical example. In this figure, the demand curve D(N), average costs under no-tolling (AC'), marginal social costs under no-tolling and time-varying tolls (MC' and MC^*), and the curve $\lambda(N)$ is given, which is calculated according to (17) with t_E calculated according to the expression in the right column in Table 1. The no-toll equilibrium is given by the intersection of D and AC', and implies an overall equilibrium demand of N_0. The overall optimum N^* can be achieved by a combination of optimal time-varying tolling with a fixed component *e–d*. This flat toll serves to charge for the difference between MC^* and λ in the

optimum. The total welfare gain, compared to the no-toll equilibrium, is given by *aef*+*ejkf*, being the sum of the cost reduction resulting from distributing N^* optimally over time, and the net welfare gain of having N_0–N^* removed. Without the flat toll, road users would keep on entering up to the point where $D=\lambda$, which is at N_1. The welfare gain compared with the no-toll equilibrium is now *ahi*+*gjki*, and thus falls short of the maximum achievable welfare gain with *egh*.

Of course, this may seem a rather modest welfare loss due to the absence of a flat toll. On the one hand, this could be caused by the particular model assumed and the parameters chosen. More fundamentally, however, this is also partly caused by a phenomenon that could loosely be referred to as 'diminishing benefits of regulation'. Should one start by considering

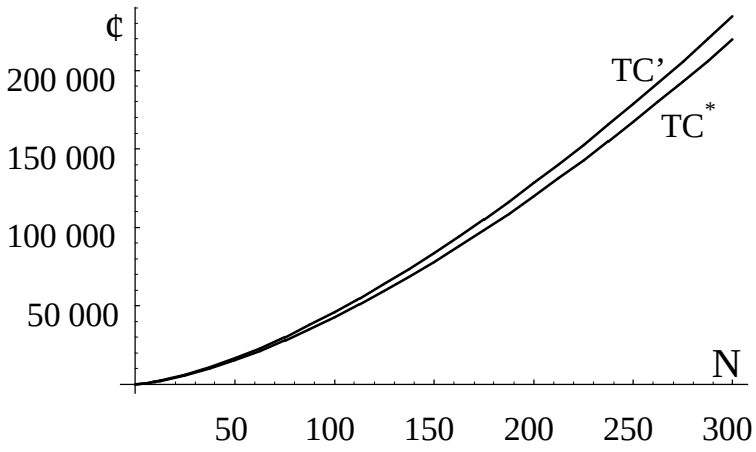

Figure 1. Total costs with (TC^*) and without (TC') time-varying tolling

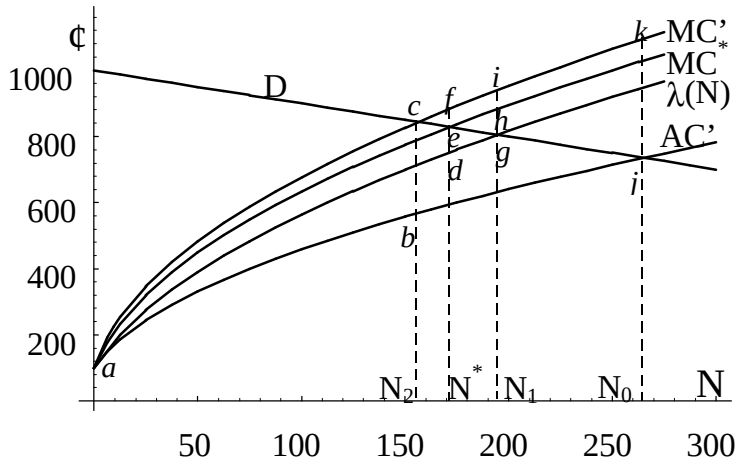

Figure 2. Various market equilibria

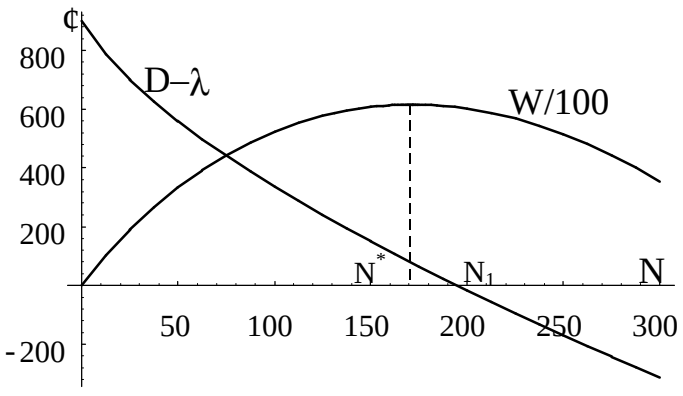

Figure 3. Optimal usage, and usage without a flat toll component

the welfare gain that can be achieved with flat tolling only, the second-best optimum N_2, to be achieved with a flat toll

c–b would be found, with a welfare gain, compared with the no-toll equilibrium, of *cjk*. The switch towards optimal time-varying tolling then 'only' yields the additional welfare gain of *aec*.

Finally, Figure 3, showing the course of D–λ and total welfare W (divided by 100), where total welfare is total benefits minus total costs under time-varying tolling, is included to demonstrate that indeed, maximum welfare does not occur where D equals λ, which would be the market outcome with time-varying tolling without a flat component.

6. Conclusion

In this paper, a dynamic model of road traffic congestion was presented, with the congestion technology used being 'flow congestion' (as opposed to 'bottleneck congestion'), and an elastic overall demand for morning peak road usage. Such elasticity of demand could result, *inter alia*, from the availability of alternative transport modes. It was demonstrated that in such a case, the optimal time-varying toll should include a 'flat', time-invariant component when road users share the same desired arrival time. This has important implications for the design of optimal toll schemes in reality, because it implies that optimal tolls cannot be set if the regulator has no information on the desired arrival times of the road users. In particular, a non-zero optimal toll is still in place when 'observable' congestion, in terms of instantaneous travel time delays, has gone to zero. Therefore, apart from observing what is going on the road, the regulator needs to have information on the distribution of desired arrival times before being able to set tolls optimally, because the underlying reason of the flat component in the optimal time-varying toll is the assumed equality of desired arrival times. Clearly, now that electronic road pricing, allowing time-varying tolling, is likely to be introduced soon at various places, this conclusion is not only of academic relevance, but has great practical importance as well.

An important reason that this flat toll has not been derived in the literature so far is the typical practice to concentrate only on 'instantaneous' external costs of travel time delays, caused by direct interaction of road users, and to ignore the 'inter-temporal' externality that users impose on all other users throughout the peak, caused by the impact of the total level of road usage, over the entire peak, on equilibrium travel times at each instant during this peak. Furthermore, the flat toll component is relevant only with elastic demand, and the study of road traffic congestion with elastic demand has so far been restricted to exercises with the bottleneck model only. As is hypothesized in the appendix, the flat toll component may vanish for the case of bottleneck congestion.

Future research may in particular be directed to alternative formulations of congestion technology. In particular, the assumption of 'zero group velocity', although not crucial to the result obtained, may be considered as an oversimplification. Such exercises may provide further information on the relative importance of a flat toll component in realistic situations.

References

Arnott, R., A. de Palma and R. Lindsey (1990a) "Departure time and route choice for the morning commute" *Transportation Research* **24B** (3) 209-228.

Arnott, R., A. de Palma and R. Lindsey (1990b) "Economics of a bottleneck" *Journal of Urban Economics* **27** 11-30.

Arnott, R., A. de Palma and R. Lindsey (1991a) "A temporal and spatial equilibrium analysis of commuter parking" *Journal of Public Economics* **45** 301-335.

Arnott, R., A. de Palma and R. Lindsey (1991b) "Does providing information to drivers reduce traffic congestion?" *Transportation Research* **25A** (5) 309-318.

Arnott, R., A. de Palma and R. Lindsey (1992) "Route choice with heterogeneous drivers and group-specific congestion costs" *Regional Science and Urban Economics* **22** 71-102.

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 (1994) "The welfare effects of congestion tolls with heterogeneous commuters" *Journal of Transport Economics and Policy* **28** 139-161.

Arnott, R., A. de Palma and R. Lindsey (1997) "Recent developments in the bottleneck model". In: K.J. Button and E.T. Verhoef (1997) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (forthcoming).

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

Braid, R.M. (1996) "Peak-load pricing of a transportation route with an unpriced substitute" *Journal of Urban Economics* **40** (179-197).

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

Else, P.K. (1981) "A reformulation of the theory of optimal congestion taxes" *Journal of Transport Economics and Policy* **15** 217-232.

Evans, A.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.

Henderson J.V. (1981) "The economics of staggered work hours" *Journal of Urban Economics* **9** 349-364.

Knight, F.H. (1924) "Some fallacies in the interpretation of social cost" *Quarterly Journal of Economics* **38** 582-606.

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

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

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

Verhoef, E.T., P. Nijkamp and P. Rietveld (1995a) "Second-best regulation of road transport externalities" *Journal of Transport Economics and Policy* **29** 147-167.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1995b) "The economics of regulatory parking policies" *Transportation Research* **29A** (2) 141-156.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1996a) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.

Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1996b) "Information provision, flat- and fine congestion tolling and the efficiency of road usage" *Regional Science and Urban Economics* **26** 505-529.

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

Walters, A.A. (1961) "The theory and measurement of private and social cost of highway congestion" *Econometrica* **29** (4) 676-697.

Appendix

This appendix gives a brief diagrammatic sketch of the conclusion that marginal social costs under time-varying tolling are larger than λ, which is the private costs including the time-varying toll that users would face under time-varying tolling without a flat toll component; see (21ab) and (20ab). Figure 4 shows the course of λ as a function of the initial and terminal times t_F and t_L according to equation (17). The intersection with the vertical axis is therefore at $\alpha \cdot T^*$.

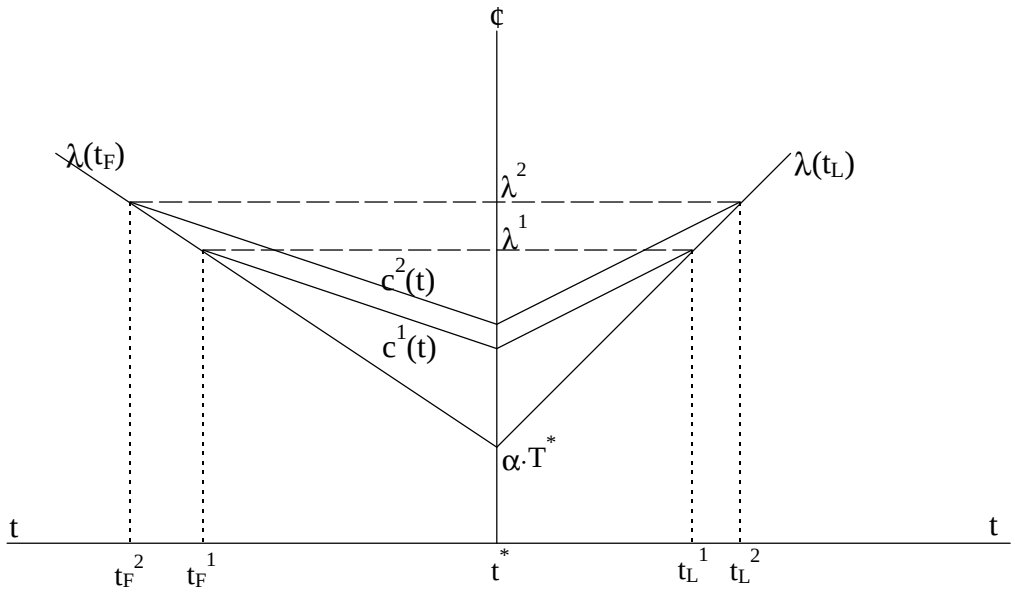

Figure 4. Average social costs and λ as a function of t, t_F, and t_L

Under time-varying tolling without a flat toll component, each pair of values of the initial and terminal times t_F and t_L with equal λ corresponds to one particular equilibrium and one particular value of N, and the duration of the peak is given by the difference between these two extreme time values. The c-curves plot for two such pairs of initial and terminal times (denoted with superscripts 1 and 2) the average social costs as a function of the arrival time t under time-varying tolling. Due to the optimality principle that marginal instantaneous travel costs should be constant over time and equal to λ (see (16)), these average cost social curves lie below the value of λ for that particular pair of t_F and t_L, which is given by the dashed horizontal lines. At the same time, these c-curves should intersect with the λ-curve under absence of a flat toll component, and should lie above this λ-curve between the intersections, as travel costs at $t_F<t<t_L$ under congestion are higher than they would have been if t were t_F or t_L – which is indicated by the λ-curve. The vertical difference

between the dashed horizontal lines and these average cost curves gives the time-varying toll. Due to the constancy of the values of time α, β and γ, the λ-function is linear in t_F and t_L also in a more general model where the congestion function itself is not necessarily linear. The steeper slope on the right hand side reflects that, in the sketched case, it is assumed that $\gamma-\alpha>\alpha-\beta$. The linearity of the c-curves, chosen for ease of diagrammatic representation, is consistent with a linear congestion technology as described in (31). However, for each flow congestion technology chosen, the c-curves would in any case intersect with the λ-curve at the relevant t_F and t_L, and would lie between the λ-curve and the equilibrium value of λ given by the horizontal dashed lines.

It is now easy to see that marginal social costs are higher than λ. The reason is that for every marginal increase in N, λ increases and therewith induces a marginal upward shift in the c-curve, and therewith in average social costs for the entire peak. Without a flat toll component, the marginal users would only consider λ as the private costs (including the time-varying toll) of using the road, ignoring the marginal impact on average social costs at all other instants in the peak than the instant at which she chooses to travel – which is in fact immaterial, because private costs including the toll will be constant in equilibrium. The optimal flat toll given in (28) charges exactly for this effect.

Only if the c-curve would have the same slopes, on both sides of t^*, as the λ-curve, and would therefore overlap with this λ-curve, this effect would vanish and λ would exactly represent marginal social costs. No flat toll component is necessary in that case. Presumably, in the bottleneck model, where it is optimal to avoid all travel delays, this is actually the case. This explains that in the bottleneck model with elastic demand, no flat component is found for the optimal time-varying toll (Braid, 1989; Arnott *et al.*, 1993). With flow congestion, in contrast, it is always optimal to maintain some travel delays. Therefore, the c-curve will always lie above the λ-curve, and a positive flat toll component is therefore always in order with optimal time-varying tolling.