Verhoef, Erik

Working Paper

Optimal Congestion Pricing with Diverging Long-run and Short-run Scheduling Preferences

Tinbergen Institute Discussion Paper, No. 17-077/VIII

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Verhoef, Erik (2017) : Optimal Congestion Pricing with Diverging Long-run and Short-run Scheduling Preferences, Tinbergen Institute Discussion Paper, No. 17-077/VIII, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/177645

TI 2017-077/VIII
Tinbergen Institute Discussion Paper

Optimal Congestion Pricing with Diverging Long-run and Short-run Scheduling Preferences

Erik (E.T.) Verhoef[1]

1: VU Amsterdam; Tinbergen Institute, The Netherlands

OPTIMAL CONGESTION PRICING WITH DIVERGING LONG-RUN AND SHORT-RUN SCHEDULING PREFERENCES*

Erik T. Verhoef**
Department of Spatial Economics, VU University Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands
Tel.:+31-20-5986094, Email: e.t.verhoef@vu.nl

Visiting Professor, Department of Economics, University of Gothenburg

This version: 18/08/17

Key words: Congestion pricing, dynamic congestion, scheduling
JEL codes: R41, R48, D62

Abstract

Recent empirical work has suggested that there is an important distinction between short-run versus long-run scheduling behaviour of commuters, reflected in differences in values of time and schedule delays, as well as in preferred arrival moments, for the short-run versus the long-run problem. Peer et al. (2015) for example find that the average value of time when consumers form their routines in the long-run problem may exceed by a factor 6 the short-run value that governs departure time choice given these routines. For values of schedule delay, in contrast, the short-run value exceeds the long-run value, by a factor 2. And, when forming routines, consumers in fact choose a most preferred arrival time that may deviate from the value they would choose in absence of congestion because a change in routines may mean that shorter delays will be encountered. This paper investigates whether this distinction between short-run and long-run scheduling decisions affect optimal pricing of a congestible facility. Using a stochastic dynamic model of flow congestion for describing short-run equilibria and integrating it with a dynamic model of routine formation, it is found that consistent application of short-run first-best optimal congestion pricing does not optimally decentralize the optimal formation of routines in the long-run problem. A separate instrument, next to road pricing, is therefore needed to optimize routine formation.

* Financial support from ERC (AdG Grant #246969 OPTION) is gratefully acknowledged.

** Affiliated to the Tinbergen Institute, Gustav Mahlerplein 117, 1082 MS Amsterdam.

1. Introduction

The values that travellers attach to travel time losses (values of time: VoT's) and to deviations from most preferred departure and arrival moments (values of schedule delay early or late: VSDE, VSDL) are key concepts in transport economics. These values are, for example, essential for the economic assessment of travel time gains from investments in infrastructure, or for the specification of optimal pricing when congestion affects travel times. Nevertheless, and perhaps surprisingly after so many years of research into the valuation of time, it is still a field of rapid conceptual and methodological development; see, for example, Small (2012) for a recent review.

One topic of interest concerns what Small (2012) refers to as the "time horizon" in scheduling decisions. The general idea is that last-minute changes to an intended trip plan may bring considerably higher schedule delay costs than deviations that are known more in advance. This reflects the intuitive notion that the traveller herself, as well as persons she possibly has to coordinate with, will generally have more flexibility to adjust plans when doing so more in advance, as changes to schedules will then typically be less difficult and hence less costly to achieve. The value of unexpected time gains, in contrast, may actually be much smaller than that of gains that can better be anticipated. The latter can be exploited more effectively (or at least by definition not less so), and be put to better use, by optimizing daily plans – which becomes more difficult, the shorter the notice.

Peer *et al.* (2015) test these hypotheses empirically, and indeed report substantial differences between what they call short-run and long-run values of time and schedule delay. They find that the long-run VOT exceeds the short-run value by a factor 6, whereas the short-run VSD's exceed the short-run values by a factor larger than 2. Moreover, they identify empirically the intuitive distinction between the long-run preferred arrival time (LRPAT) and its short-run counterpart (SRPAT). The LRPAT represents what would be the most desirable moment of arriving at work if there were no congestion, ever, for sure. This is in fact the interpretation of the preferred arrival time (PAT) in dynamic models of congestion and congestion pricing, where dynamic equilibrium conditions typically specify that identical travellers are indifferent between travel moments that are actually used, and the PAT is, in equilibrium, therefore not any more preferable than any other arrival moment that is being used. The SRPAT, in contrast, is the most desired arrival moment after travellers have adjusted their daily routines so as to optimally account for recurrent patterns of congestion. It is therefore an endogenous variable, that travellers can choose optimally. Again, the empirical patterns reported by Peer *et al.* (2015) confirm what one would expect: LRPAT's are more strongly clustered in time than SRPAT's, reflecting that drivers adjust their routines, and choose SRPAT's, so as to more easily avoid the most severe congestion.

The distinction between short-run and long-run valuations and preferred arrival times will have notable implications for empirical work. For example, consider a traveller with a LRPAT of 9:00, who has chosen an SRPAT of 7:00 to avoid the most severe congestion. If, on a given day, she arrives at 8:30 because of unforeseen congestion and therefore misses the start of a meeting scheduled at 8:00, she will incur schedule delay late, not early. A choice

model to be estimated should recognize that as such, but it will not if the LRPAT is used. Furthermore, it naturally becomes more important in empirical modelling to differentiate the coefficients attached to short-run versus long-run travel delays and schedule delays, the more strongly the time horizon at which these delays become known to the traveller affects the travellers' valuations of such delays.

The distinction can also be expected to have policy implications. For example, a measure that reduces the unanticipated part of expected travel time losses – such as incident management policies – would have to be evaluated using the short-run valuations, whereas a measure such as a capacity expansion, to the extent that it brings structural and predictable changes in expected travel times irrespective of whether these are stochastic, should be assessed on the basis of long-run values.

Another pressing question raised is whether, and if so, how, the distinction between short-run and long-run scheduling decisions affect optimal pricing of a congestible facility. This is the question addressed in the present paper. More specifically, we will study optimal dynamic congestion pricing when long-run and short-run scheduling preferences diverge in the way just described: with differences in preferred arrival times and in unit values of time and schedule delays. The short run will refer to the driver's departure time decision, taking as given the actual traffic conditions on that day, the short-run valuations, and the fact that the SRPAT has been chosen and cannot be changed in the short run. This short-run departure time choice thus involves a trade-off between schedule delay cost, relative to the SRPAT, and travel delay cost; both weighted using the short-run valuations. The long run refers to the individual's choice of the SRPAT, which will also be referred to as the formation of 'routines' in what follows. Choosing the optimal SRPAT involves making trade-offs between expected values of short-run travel cost components versus long-run costs. Expected short-run costs will vary with SRPAT, because this SRPAT affects the optimal short-run choices of departure times and hence short-run travel time and schedule delay costs. Differences in expected travel times will furthermore affect the long-run valuation of these travel times. An additional long-run cost component is the long-run schedule delay costs, reflecting the disutility associated with the endogenous SRPAT deviating from the exogenous LRPAT. The model to be developed below captures both these short-run and long-run decisions under simultaneous satisfaction of short-run and long-run equilibrium conditions, taking into account how congestion effects will make individuals' short-run and long-run decisions interdependent.

An important conceptual question studied in this paper is this: does the distinction between short-run and long-run scheduling behaviour necessitate the use of an additional 'long-run' congestion tax, aimed at affecting this long-run choice of SRPAT's or routines, in addition to 'short-run' tolls that are set to optimize departure time decisions given the choice of SRPATS? Intuition might suggest this would not be needed: consistent application of socially optimal short-run prices usually provides the optimal incentives to also optimize long-run choices from a social perspective. To make this more concrete, Verhoef (2000) presents two examples where optimal Pigouvian road pricing, with tolls equal to marginal external costs, not only affects the short-run mobility, but also some long-run choice. One example is the choice of residential location, and the other one is the choice of vehicle

technology. In both examples, these long-run choices are optimized also from the social perspective if travellers correctly anticipate the short-run Pigouvian taxes they will face given the long-run choices they make. No separate policy concerning residential or technology choice is therefore needed, if the only market failure is an already optimally priced externality from transport. By the same logic, one might conjecture that a consistent application of optimal short-run tolls in the current context would not only lead drivers to choose socially optimal departure times in the short run given the SRPAT's chosen, but would also perfectly decentralize the socially optimal choice of SRPAT's in the long-run.

We will see that this is not the case. If short-run and long-run scheduling preferences diverge, optimal tolling to affect departure times in the short run does not provide the incentives to choose socially optimal SRPAT's in the long run. A separate long-run tax is therefore needed to achieve the overall first-best outcome. In the analytical framework, it will simply be assumed that such a (long-run) tax is available, in addition to regular (short-run) tolls for the use of the roads, and its theoretically optimal time pattern will be determined. In practice, it may not be easy to implement such a tax, since the choice of SRPAT's cannot be observed. The results indicate that in such cases, second-best instruments that would affect the choice of SRPAT's, such as subsidies on staggered work-hours arrangements, may be welfare enhancing even if short-run optimal road pricing is in place. Or, indeed, that there might be a case for a second-best correction on short-run optimal road pricing, in order to approximate the long-run tax that would otherwise have been implemented. The aim is to show that short-run optimal road tolls alone will not achieve the overall first-best optimum, and the derivation of a non-zero long-run optimal tax is an analytically transparent way of doing that.

The behavioural model will assume that unit values of travel time and schedule delays differ between the short-run and long-run problem in the way that was established empirically and discussed above: the VoT is higher in the long run than in the short run; the VSD's are lower. Drivers are homogeneous with respect to their valuations and their LRPAT, and only differ through their endogenous equilibrium choices of SRPAT's, and their actual departure times. Besides analytical tractability, this has the conceptual advantage that all heterogeneity observed in equilibrium is endogenous, and not the result of possibly arbitrary assumptions.

There is stochasticity in travel times, stemming from variations in road capacity between successive days. Without such stochasticity, successive working days will be exact replicas and the equilibria in the full model would become degenerate in the sense that drivers would either set the SRPAT equal to the single identical arrival time that is chosen every identical day, or would all set the SRPAT equal to the LRPAT. It is the variation in actual arrival times between days that makes the choice of SRPAT an interesting problem. A simple form of stochasticity will be employed: road capacity can take on two values, where with a certain probability π it is relatively low – for example due to an incident – while with probability $(1-\pi)$ it is high.

These assumptions match those in Peer and Verhoef (2013), who also study optimal congestion pricing with diverging long-run and short-run scheduling preferences. The difference between their paper and the present one is in the specification of the congestion cost function, and their results justify making this effort. Peer and Verhoef (2013) employ the

conventional Vickrey (1969) bottleneck model. There is an indeterminacy in their results, even if restrictions on parameter values are imposed in order to avoid corner solutions. In particular, they show that SRPAT's will be chosen such that the density of SRPAT's is lower than the high capacity of the bottleneck, in the 'good state'. In the short-run equilibrium with high capacity, every driver will therefore choose to arrive at her SRPAT without incurring any queuing: the capacity exceeds instantaneous demand throughout the peak. Peer and Verhoef (2013) subsequently show that with optimal short-run congestion pricing, which entails a classic "triangular" toll schedule in the low-capacity state and a zero toll in the queue-free high-capacity state, there is a remaining inefficiency in the long-run problem of choosing SRPAT's which justifies the corrective use of a long-run toll. However, because there is no queuing in the high-capacity state, the regulator can apply, within limits, short-run tolls in that state without affecting the equilibrium. And it turns out that these limits are sufficiently wide to allow the regulator to specify a toll schedule in that state that does not upset the short-run no-queuing optimum, while through its expected value it affects SRPAT decisions in the same way as the deterministic optimal long-run toll would. In other words: also with only short-run tolls can the short-run and long-run optima simultaneously be decentralized.

This result for the bottleneck model is interesting in its own right, but it leaves unanswered the conceptual question of whether or not short-run tolls alone will in general be sufficient to decentralize the full optimum when short-run and long-run scheduling preferences diverge. The present paper takes up that question, by considering a different congestion technology, and in particular one that does not have the property that there is a short-run state in which optimal tolls are not uniquely defined. No longer, therefore, is there a short-run toll that can be shaped to fulfil, in expected terms, the role of the long-run toll without causing efficiency losses in the short-run. It should therefore become unambiguously clear whether or not a separate long-run toll is needed in order to achieve the full optimum.

The congestion technology employed is the one proposed by Chu (1995), based on earlier work by Henderson (1974, 1981). The model considers travel on a single homogeneous road, and assumes that a traveller's speed on that road is constant over time during the trip and depends only on the arrival rate at the road's exit at the moment the trip is completed (Chu), or the departure rate at the road's entrance at the moment the trip is started (Henderson).[1] The model thus simplifies the modelling of dynamic congestion by ignoring that travellers who have departed at different instants may interfere in terms of causing mutual congestion delays during their trips. Lindsey and Verhoef (2000) refer to this assumption as "no propagation", to distinguish it from finite-speed propagation of shock waves that for example kinematic models of traffic flows have, and from the other extreme of "instantaneous propagation" as assumed by Agnew (1977). The latter assumes that at any given moment, all vehicles that are on the road have the same speed irrespective of their location, where any given driver's speed may thus vary continuously over time as aggregate averaged density on the road varies.

[1] A problem with the latter formulation is that it could imply that, in equilibrium, an individual driver might benefit from unilaterally rescheduling the departure time, and overtake drivers who departed earlier (Chu, 1995).

The no-propagation model is definitely a simplification from reality. The reason for adopting it here is that it is arguably the simplest dynamic congestion technology that avoids the non-uniqueness of optimal toll schedules, and that therefore lends itself for answering the main question. If we find that the long-run toll is needed to achieve efficiency in this setting, it will also be needed in more sophisticated models, that may for example feature combinations of flow congestion and bottleneck queuing on the same network (such as the model proposed by Mun, 1994).

The plan of the paper is as follows. Section 2 will give a compact review of the Chu (1995) model, and will present the extensions made for this paper. Next, Section 3 derives analytically the no-toll equilibria, the short-run optima, and the long-run optima, and demonstrates that consistent first-best optimal short-run congestion pricing does not optimally decentralize the choice of short-run preferred arrival times in the long-run problem. Section 4 gives a numerical illustration, and Section 5 concludes the paper.

2. A dynamic model of flow congestion

2.1. The original Chu (1995) model

Chu's (1995) dynamic model of flow congestion considers N identical travellers who use a single road for their trip during the morning commute. They have perfectly inelastic demand, a common desired arrival time t^*, a value of time α, and values of schedule delay β for early arrivals and γ for late ones. As is customary, we define $\delta \equiv (\beta \cdot \gamma)/(\beta+\gamma)$ as a composite schedule delay cost coefficient. We denote a traveller's arrival time as t'. In the basic model, the capacity of the road is given, and is denoted K (there are some changes in notation here compared to the exposition by Chu, 1995), and the travel time $T(t')$ associated with an arrival at t' depends on both K and on the instantaneous arrival rate $r(t')$. To obtain closed-form solutions, a functional form for the travel time function $T(r(t');K)$ needs to be specified, and like Chu, a power-law or BPR (Bureau of Public Roads) function will be used:

$$T(r(t');K) = T_f + \left(\frac{r(t')}{K}\right)^{\chi} \tag{1}$$

where χ is the parameter that determines the curvature of $T(\cdot)$. Note that the regular BPR function has a second parameter, pre-multiplying the second term in (1), but this one can be dropped without loss of generality because the units of K can be chosen freely. With $\tau(t')$ denoting a time-varying toll when levied, the generalized price for an arrival at t' can be written as the sum of $\tau(t')$, the travel time cost $c_T(t')$, and the schedule delay cost $c_{SD}(t')$:

$$p(t') = \tau(t') + c_T(t') + c_{SD}(t') = \tau(t') + \alpha \cdot T(r(t');K) + \begin{cases} \beta \cdot \left(t^* - t'\right) & \text{if } t' \le t^* \\ \gamma \cdot \left(t' - t^*\right) & \text{if } t' > t^* \end{cases} \tag{2}$$

As for Vickrey's (1969) bottleneck model (see, for instance, Arnott, de Palma and Lindsey, 1993), the dynamic equilibrium can be found by determining the conditions for arrival rates that keep $p(t')$ constant for early (before t^*) and late (after t^*) arrivals. These should then be combined with the condition that the schedule delay cost for the very first driver arriving at t_q and the very last driver arriving at $t_{q'}$ should be the same, and the condition that N drivers

should arrive between t_q and $t_{q'}$. A complication compared to the bottleneck model is that for the present congestion technology, arrival rates vary continuously over time both in the no-toll equilibrium and in the first-best optimum. For the no-toll equilibrium, this follows directly from (2) by setting $\tau(t')$ equal to zero, substituting (1) for $T(\cdot)$, and then determining the arrival pattern that makes its time-derivative equal to zero – as it should be in a dynamic equilibrium. For the optimum, the solution follows from the same sequence of steps, but now with the optimal toll rule substituted in for $\tau(t')$. Chu (1995) has shown that this optimal toll rule involves a direct, time-varying application of the Pigouvian marginal external cost rule:

$$\tau(r(t')) = r(t') \cdot \frac{\partial c_T(r(t'))}{\partial r(t')} \tag{3}$$

This form of the toll secures that every driver faces the full marginal social cost for an arrival at t': the self-incurred travel time and schedule delay costs, as well as the travel delay costs imposed on others. When applied for a BPR function such as the one in (1), this toll adds a multiple χ of the delay cost experienced by the user to her generalized price (*e.g.* Small and Verhoef, 2007). This constant multiple χ causes the analytical closed-form expressions for the optimum to be not too different from those for the no-toll equilibrium.

The instantaneous arrival rates thus derived for the no-toll equilibrium and for the optimum have a simple polynomial form, and their integrals over time have convenient expressions as well. This allows the derivation of manageable analytical closed-form expressions for the model's equilibrium and optimum. These are given in Table 1. The expressions are entirely consistent with those in Chu (1995). To allow for the extension of the model with two possible realizations of capacity, an index i for the state of nature is attached to capacity K_i, and therefore also to the short-hand variable Ψ_i.

Table 1 shows that the analytical expressions are only slightly less transparent than those for the basic bottleneck model (Arnott, De Palma and Lindsey, 1993). Furthermore, because Ψ_i is proportional to $N^{\chi/(1+\chi)}$, the average cost and the generalized price increase less than proportionally with N. For intuition behind this result, one can imagine the marginal N^{th} user to be added at t^*, with all other drivers shifting a bit earlier for early arrivals, and a bit later for late arrivals, otherwise maintaining exactly the same departure pattern to maintain equilibrium under linear schedule costs. These required shifts decrease with N, as the arrival rate at t^* increases with N. Therefore, also the associated increases in schedule delay costs for the very first and last drivers decrease with N; and hence – by equality of generalized prices in equilibrium – so do the increases in equilibrium levels of average cost and generalized price.

What is not immediately clear from Table 1 is that the equilibrium travel delays will display a linear pattern over arrival time. For the no-toll equilibrium, this is to compensate for the assumed linear pattern of schedule delay cost. For the first-best optimum, both the toll and the travel delays – which, as said, maintain a constant ratio of χ between them – change (piecewise) linearly over time, to compensate for changes in schedule delay cost. The non-linearity of the travel delay function (if $\chi \neq 1$) therefore translates into non-linear arrival rates over time; not in non-linear time patterns of travel delays or tolls.

	No-toll equilibrium	First-best optimum
Short-hand Ψ_i	$\Psi_i = \left(\frac{N}{K_i} \cdot \frac{1+\chi}{\chi} \cdot \frac{\delta}{\alpha} \right)^{\frac{\chi}{1+\chi}}$	
Arrival rate $r(t')$ early ($t' \le t^*$)	$r(t') = K_i \cdot \left(\frac{\beta}{\alpha} \cdot \left(t' - t_q \right) \right)^{\frac{1}{\chi}}$	$r(t') = K_i \cdot \left(\frac{1}{1+\chi} \cdot \frac{\beta}{\alpha} \cdot \left(t' - t_q \right) \right)^{\frac{1}{\chi}}$
Arrival rate $r(t')$ late ($t' > t^*$)	$r(t') = K_i \cdot \left(\frac{\gamma}{\alpha} \cdot \left(t_{q'} - t' \right) \right)^{\frac{1}{\chi}}$	$r(t') = K_i \cdot \left(\frac{1}{1+\chi} \cdot \frac{\gamma}{\alpha} \cdot \left(t_{q'} - t' \right) \right)^{\frac{1}{\chi}}$
Early interval: $t^* - t_q$	$t^* - t_q = \Psi_i \cdot \frac{\alpha}{\beta}$	$t^* - t_q = (1+\chi)^{\frac{1}{1+\chi}} \cdot \Psi_i \cdot \frac{\alpha}{\beta}$
Late interval: $t_{q'} - t^*$	$t_{q'} - t^* = \Psi_i \cdot \frac{\alpha}{\gamma}$	$t_{q'} - t^* = (1+\chi)^{\frac{1}{1+\chi}} \cdot \Psi_i \cdot \frac{\alpha}{\gamma}$
Generalized price p	$p = \Psi_i \cdot \alpha$	$p = (1+\chi)^{\frac{1}{1+\chi}} \cdot \Psi_i \cdot \alpha$
Average generalized cost $\overline{c}$	$\overline{c} = \Psi_i \cdot \alpha$	$\overline{c} = \frac{(1+\chi)^{\frac{2+\chi}{1+\chi}}}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha$
Total travel delay cost TDC	$TDC = \frac{1+\chi}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$	$TDC = \frac{(1+\chi)^{\frac{1}{1+\chi}}}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$
Total schedule delay cost SDC	$SDC = \frac{\chi}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$	$SDC = \frac{\chi \cdot (1+\chi)^{\frac{1}{1+\chi}}}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$
Total toll revenue TR	$TR = 0$	$TR = \frac{\chi \cdot (1+\chi)^{\frac{1}{1+\chi}}}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$
Total social cost C	$C = \Psi_i \cdot \alpha \cdot N$	$C = \frac{(1+\chi)^{\frac{2+\chi}{1+\chi}}}{1+2\cdot\chi} \cdot \Psi_i \cdot \alpha \cdot N$
Toll $\tau(t')$	$\tau(t') = 0$	$\tau(t') = \alpha \cdot \chi \cdot \left(\frac{r(t')}{K_i} \right)^{\chi}$ $= \alpha \cdot \chi \cdot \left(T - T_f \right)$ $= \frac{\chi}{1+\chi} \cdot \begin{cases} \beta \cdot \left(t' - t_q \right) & \text{if } t' \le t^* \\ \gamma \cdot \left(t_{q'} - t' \right) & \text{if } t' > t^* \end{cases}$

Note 1: Besides some changes in notation, these expressions are equivalent to those in Chu (1995).
Note 2: Costs and prices are net of free-flow travel times T_f. Inclusion would require adding $\alpha \cdot T_f$ for average cost and generalized price measures, and $N \cdot \alpha \cdot T_f$ for inclusion in total costs measures.

Table 1. Equilibrium and first-best optimum

What is also not immediately evident, but can be found after integrating the equilibrium arrival rates over the relevant time intervals, is that the proportions of early and late drivers are always such that a fraction $\gamma/(\beta+\gamma)$ of the N drivers will arrive early, and a fraction $\beta/(\beta+\gamma)$ late. This is true both in the no-toll equilibrium and in the first-best optimum. This means that cumulative arrivals will always be equal to $N\cdot\gamma/(\beta+\gamma)$ at t^*, independent of the realization of capacity, and whether or not optimal tolls are in place.

Unlike what is found in the conventional bottleneck model with the same preference structure, the imposition of first-best pricing will lengthen the duration of the peak through a wider dispersion of arrival times, will increase total schedule delay cost, and will increase the generalized price of travelling. Chu (1995) discusses the differences between these two models in much greater detail.

Finally, the expressions in Table 1 imply what may have been expected intuitively: a higher capacity will shorten the peak and will increase average arrival rates, as well as the arrival rate at t^*, as well as the arrival rate at any given amount of time after the start of the peak for early arrivals, or before the end of the peak for late arrivals.

2.2. Introducing diverging long-run and short-run scheduling preferences

The distinction between long-run and short-run scheduling preferences brings a number of changes to the model. The short-run model remains close to the original model of Chu just presented, with the exception that instead of a single exogenous desired arrival time t^*, there will now be a distribution of desired arrival times. Using $t^\#$ to denote the SRPAT, the cumulative distribution will be denoted $Z(t^\#)$, while $z(t^\#) \equiv \dot{Z}(t^\#)$ gives the density of desired arrival times at $t^\#$ (the dot above $\dot{Z}(t^\#)$ denotes a time derivate). The density $z(t^\#)$ is given, and so is each individual's $t^\#$, for the short-run problem of choosing the optimal arrival time t'. But $z(t^\#)$ is of course endogenous in the long-run problem, where travellers choose their $t^\#$.

As explained, to make the distinction between the short and long run meaningful, there has to be some form of variability between successive short-run states, to avoid successive peaks becoming exact replicas for which the short-run and long-run problems would basically coincide. For reasons of analytical transparency, a simple form of stochasticity is introduced. The road's capacity can take on two possible values K_i: a high capacity K_0 in state 0 that occurs with a probability $(1-\pi)$, and a low capacity K_1 in state 1 that occurs with probability π – one might think of "an incident" such as bad weather. A single capacity level applies throughout a peak; *i.e.*, the realization of K_i materializes before the first traveller departs, and is assumed to be known by all drivers. As a result, short-run equilibria are fully deterministic.

While the endogenous SRPAT's $t^\#$ may and will differ between drivers, they have the same LRPAT t^*, an assumption that secures that the equilibrium distribution of $z(t^\#)$ is entirely the result of endogenous scheduling behaviour, not of any *ad hoc* assumed differences in t^*. The short-run values of travel time components will be denoted α, β and γ, following conventions. The long-run schedule delay costs associated with $t^\#$ deviating from t^* will have unit values that are lower than in the short-run problem, as was found empirically by Peer *et al.* (2015). For convenience, the same ratio g (with $0\leq g\leq 1$) between long-run and short-run

values is assumed to apply for early and late deviations. The VOSD's in the long-run problem thus become $g \cdot \beta$ for $t^{\#}$ earlier than t^{*}, and $g \cdot \gamma$ for $t^{\#}$ later than t^{*}. Expected values of travel delays are valued higher in the long-run than in the short run; again consistent with empirical evidence in Peer *et al.* (2015). The relative premium is a (with $a \geq 0$), so that the long-run value of expected travel times becomes $(1+a) \cdot \alpha$. One further parameter restriction is imposed: $g > \pi$. Otherwise, long-run schedule delay cost are so low that SRPAT's will be chosen in such a dispersed manner that no traveller would ever incur short-run schedule delay costs.

The assumptions $0 \leq g \leq 1$ and $a \geq 0$ thus match empirical results in Peer *et al.* (2015). At the same time, these assumptions are consistent with what one would expect on basis of theoretical considerations. For g, the assumption reflects that the per-day schedule delay cost of changing the routine to a certain short-run preferred arrival time and then arriving at that moment cannot exceed, and is typically lower than, the schedule delay cost that would be incurred when choosing that arrival time without having optimized routines for it. Hence, if travellers can better prepare for a certain arrival moment when they know it further in advance, and if a better preparation at least is not counter-productive in that it would make the total schedule delay cost higher, g cannot exceed 1 and is typically smaller. For a, the assumption reflects that the per-day value of a structural and certain time gain, following a shift of the individual's travel time distribution so that the expected travel time decreases, is at least as high as and typically higher than the value of an incidental, uncertain travel time gain on a random day. The motivation is that a traveller can always employ a certain time gain in the same way as an uncertain gain would be used, which is why a is not negative, but usually use it better because it is anticipated – and that makes a positive. Therefore, the assumptions on g and a are by no means *ad hoc*, and have a solid empirical as well as theoretical basis.

With the desired arrival times dispersed over drivers in the short-run problem, it is no longer necessarily true that short-run equilibria and optima as described in Section 2.1 will emerge. If the density of desired arrival times is high enough relative to the capacity, they will. Such a case will be referred to as a "condensed peak". The density of SRPAT's is so high that it is not an equilibrium to have all drivers arrive at their SRPAT's: drivers would then prefer to reschedule, away from the peak. In equilibrium, to make each driver indifferent with respect to marginal changes in her actual arrival time in that marginal changes in schedule delay cost exactly compensate for marginal changes in travel time cost, actual arrival rates should then still follow the patterns identified in Table 1, and early drivers generally arrive before and late drivers after their SRPAT.

For this case, it will be assumed that drivers arrive in order of their SRPAT's. This is one of many possible pure-strategy Nash equilibria for the short-run choice of departure moment. In particular, a driver with an early $t^{\#}$ (between t_q and t^{*}) would face minimum but constant short-run costs for any arrival moment between t_q and $t^{\#}$, and higher short-run costs on any other moment; and similarly for a driver with a late $t^{\#}$ (between t^{*} and $t_{q'}$) for any arrival moment between $t^{\#}$ and $t_{q'}$.

It is important to emphasize that while an assumption that drivers arrive in order of their PAT's is innocent in a conventional short-run model with an exogenous distribution of PAT's, it has implications in the present set-up. That is, although an early driver is indeed

indifferent in the short run between any arrival moment between t_q and $t^{\#}$, there is an implication for the long-run problem through the premium a on the valuation of expected travel times. (A symmetric reasoning applies to late drivers, and is suppressed for the sake of compactness of the argument.) The assumption of "keeping the order" applies this premium a to the actual travel delay that the driver incurs in the condensed state if she indeed sticks to the order of drivers. But she could also be assumed to consider the possibility of driving earlier in that condensed state, given the indifference she has between all arrival moments between t_q and $t^{\#}$. An alternative assumption reflecting this would be that in the long-run problem, an early driver would assign uniform probabilities to her choosing any arrival time between t_q and $t^{\#}$ in a condensed peak. Due to the linearity of travel delays over the peak, the expected delay is then half the travel delay at $t^{\#}$. Since, however, we will see below that in an equilibrium with a condensed state and under the assumption of drivers arriving in order of SPRAT there is a fixed proportion (over drivers) between a driver's actual arrival time t' and her SRPAT $t^{\#}$, and given the linear pattern of travel delays over the peak, there is also a fixed proportion between the travel delay at t' and half the travel delay at $t^{\#}$. In other words, the model would be consistent also with this alternative assumption, and would produce identical results, after an appropriate adjustment of the constant term a. For that reason, the assumption that drivers "keep order" seems convenient but innocent.[2]

Besides the condensed peak just discussed, it is also possible that the distribution of desired arrival times $z(t^{\#})$ is so dispersed, relative to capacity, that everybody arrives at his or her desired arrival time. The resulting time pattern of short-run travel time costs should then be so flat that no-one has an incentive to deviate from this equilibrium because the resulting increase in short-run schedule delay costs would outweigh the value of travel time gains achieved. In such cases, we will speak of a "dispersed peak". Obviously, drivers also now arrive in order of their SRPAT's, simply because everyone arrives at their SRPAT. In principle, it could be possible that a short-run peak is partly "condensed" and partly "dispersed", for a sufficiently irregular pattern of $z(t^{\#})$. It will be shown below that with an endogenous distribution of $z(t^{\#})$, only entirely dispersed or entirely condensed peaks will occur.

In order to more easily characterize equilibria in terms of the results for the original model summarized in Table 1, a number of definitions is introduced. The first is that of the "reference arrival rate" for the prevailing capacity i, which is defined as the arrival rate that would occur with that capacity if all desired arrival times were equal to t^* – as they are in the original model. These arrival rates thus follow the patterns defined in Table 1, and will be indicated with a subscript for the state, and if needed to avoid confusion a superscript *NT* for the no-toll equilibrium (r_i^{NT}) and *FB* for the short-run first-best equilibrium (r_i^{FB}).

[2] The paper's main result will also hold if, instead of assigning uniform probabilities to arrivals between t_q and $t^{\#}$ in a condensed peak, an early driver would assign probabilities that follow the aggregate equilibrium arrival rate. Essential for the main result is that (expected) travel delay increases with $t^{\#}$ for early drivers, and decrease with $t^{\#}$ for late drivers, so that also a driver's (expected) marginal external cost varies with $t^{\#}$ in the same manner. This would also be true under third possible assumption; however, modelling it this way complicates the analytics considerably.

Next, we use superscripts *SR* to denote price and cost components that are variable in the short-run problem of choosing the arrival time t'. The summation of these cost components is represented by c^{SR}; the summation of price components p^{SR}, where the road toll τ^{SR}, when levied, can make the difference between them. Hence, to describe departure time choice on a given day in a given state i we can write, consistent with equation (2):

$$p_i^{SR}(t') = \tau_i^{SR}(t') + c_i^{SR}(t') = \tau_i^{SR}(t') + \alpha \cdot T(r_i(t'); K_i) + \begin{cases} \beta \cdot \left(t^{\#} - t'\right) & \text{if } t' \le t^{\#} \\ \gamma \cdot \left(t' - t^{\#}\right) & \text{if } t' > t^{\#} \end{cases} \tag{4}$$

Note that equation (4) only gives the cost and price components that are variable in the short-run. It is therefore not a standard short-run cost or price function, for which also the long-run cost components should be added, with $t^{\#}$ and arrival moments on all other days treated as given. Equation (4) is therefore the one that a traveller minimizes when choosing t'.

To obtain the function that travellers minimize when choosing the SRPAT $t^{\#}$, we add the various relevant long-run price components to the expected value of the price components identified in (4). Expected values will be denoted using the operator E[·], and the superscript *LR* is used to denote price components that include both the short-run components in (4) and the additional long-run components. Hence:

$$\begin{aligned} c^{LR}(t^{\#}) &= \mathrm{E}[c_i^{SR}(t_i'(t^{\#}), t^{\#})] + (1-\pi) \cdot a \cdot \alpha \cdot T(t_0'(t^{\#})) + \pi \cdot a \cdot \alpha \cdot T(t_1'(t^{\#})) \\ &\quad + \begin{cases} g \cdot \beta \cdot \left(t^{*} - t^{\#}\right) & \text{if } t^{\#} \le t^{*} \\ g \cdot \gamma \cdot \left(t^{\#} - t^{*}\right) & \text{if } t^{\#} > t^{*} \end{cases} \\ p^{LR}(t^{\#}) &= c^{LR}(t^{\#}) + \mathrm{E}[\tau_i^{SR}(t_i'(t^{\#}))] + \tau^{LR} \end{aligned} \tag{5}$$

For notational convenience, the arguments r and K are suppressed in T in (5) and only T's time-dependence is made explicit, while the notation $t_i'(t^{\#})$ reflects that the traveller takes into account, in the long-run problem, that departure times in the short-run problem will depend on the long-run choice of $t^{\#}$. Again, it should be emphasized that (4) and (5) should not be mistaken to represent conventional short-run and long-run cost functions. A short-run cost function for the departure-time choice on a particular day could be derived from (5) by treating $t^{\#}$ as well as the choices of t' on all other days, whether or not with the same realization of capacity, as given. The long-run cost function would follow from (5) after optimizing $t^{\#}$ as well as all t'. The reason for not writing out and using these functions explicitly is that these still ignore how interactions between drivers are essential for establishing equilibrium, making these functions of only very limited relevance. Phrased differently, for determining short-run and long-run equilibria, it is not just the individual's own choices of $t^{\#}$ and all t', minimizing her own generalized prices in (4) and (5), that should be identified, but also the mutual impacts of all drivers' decisions on each other's cost and price levels.

Finally, we simplify notation by setting $t^{*}=0$ without loss of generality, and will ignore any travel time costs related to T_f by setting it equal to zero. This means that (1) can no longer represent a conventional BPR function, but as T_f is just a fixed component for each trip, it plays no role of interest in the analysis and may as well be ignored.

Assuming for the time being that short-run equilibria are either entirely condensed or entirely dispersed – an assumption to be confirmed after the equilibrium distribution of $z(t^{\#})$ has been determined – three types of long-run equilibria and optima can in principle be distinguished. These are referred to as "Always Dispersed" (*AD*), for which dispersed short-run equilibria occur in both states 0 and 1; "Sometimes Dispersed" (*SD*), where there is a dispersed equilibrium in the high-capacity state 0 and a condensed one in state 1 (the opposite will not occur for obvious reasons); and "Never Dispersed" (*ND*), which means that a condensed short-run equilibrium applies in both states. The next section will discuss these three types of equilibria in detail.

3. No-toll equilibria, short-run optima, and long-run optima

3.1. Always Dispersed (AD) long-run no-toll equilibrium

In the AD equilibrium, $z(t^{\#})$ is so flat and dispersed that drivers will choose to arrive at their SRPAT in both short-run states. Quite intuitively, this type of equilibrium is more likely to occur, the lower g and therewith the long-run schedule delay cost. If g equals zero, there is no reason to have the SRPAT close to the LRPAT: by choosing one really far off, one could travel alone at SRPAT and avoid travel delays without incurring short-run schedule delay costs. The exact parameter conditions under which the *AD* equilibrium will occur, will be identified below. First we determine the equilibrium itself, for the no-toll situation.

Because every driver chooses the SRPAT as the arrival time in both states, the expected short-run cost consists of travel delay cost only:

$$\mathrm{E}\left[c^{SR}(t^{\#})\right] = (1-\pi)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi} + \pi\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_1}\right)^{\chi} \tag{6}$$

The long-run cost weigh these at their long-run value, and adds to this the long-run schedule delay cost:

$$c^{LR}(t^{\#}) = (1+a)\cdot\alpha\cdot z(t^{\#})^{\chi}\cdot\left((1-\pi)\cdot\left(\frac{1}{K_0}\right)^{\chi} + \pi\cdot\left(\frac{1}{K_1}\right)^{\chi}\right) + \begin{cases} -\beta\cdot g\cdot t^{\#} & \text{if } t^{\#}\le 0 \\ \gamma\cdot g\cdot t^{\#} & \text{if } t^{\#}>0 \end{cases} \tag{7}$$

(recall that $t^{*}=0$). In the long-run *AD* equilibrium, the distribution of $z(t^{\#})$ should be such that the time-derivative of (7), with respect to $t^{\#}$, is zero. This defines a partial differential equation for $z(t^{\#})$, the solution of which is given below, with the derivations relegated to Appendix A:

$$z(t^{\#}) = \begin{cases} \left(\dfrac{g\cdot\beta\cdot\left(t^{\#}-t_l\right)}{(1+a)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } t_l \le t^{\#} \le 0 \\ \left(\dfrac{g\cdot\gamma\cdot\left(t_{l'}-t^{\#}\right)}{(1+a)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } 0 < t^{\#} \le t_{l'} \end{cases} \tag{8}$$

where t_l is defined as the very first SRPAT chosen, and $t_{l'}$ as the very last one. These can be solved for as:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{(1+a)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}{g\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \tag{9}$$

Equation (9) has a structure similar to those for early and late intervals in Table 1 (after substitution of Ψ_i). Also note that $z(t^{\#})$ in (8) varies in such a way over time $t^{\#}$ that its time pattern is comparable to the reference arrival rates in Table 1. In particular, the time difference between the moment evaluated and the relevant endpoint of the total time interval is raised to the power of $1/\chi$. This makes the time patterns to be comparable, the possible differences only being the starting and ending moments and a multiplicative term that is constant over time. Because all distributions should have a total integral of N over the entire peak, a more dispersed, wider interval naturally implies a lower constant multiplicative term, and therewith a lower density at any given distance (in time) from the relevant end-point and a lower time-derivative of the density function. This explains why, with endogenous $z(\cdot)$, a short-run equilibrium will always be entirely condensed or entirely dispersed.

For the *AD* regime to form the relevant equilibrium, the distribution of $z^{\#}$ should be more dispersed than the distributions of both the reference arrival rates, implying that the following two inequalities should hold: $t_l \le t_q^1 \le t_q^0$ and $t_{l'} \ge t_{q'}^1 \ge t_{q'}^0$, where superscripts 0 and 1 refer to the reference equilibria for states 0 and 1. Likewise, $z(t^{\#})$ in (8) should be less steep than the reference arrival rates in Table 1, so that even in state 1, with flatter slopes of r_i, there is no incentive to reschedule the departure time away from $t^{\#}$: travel times should be too high to compensate for the increase in short-run schedule delay cost when moving towards the relevant endpoints of the peak (a move towards the centre is certainly unattractive, as both short-run schedule delay cost and travel delay costs would then be higher). The conditions on the endpoints of the distributions, and on the densities, imply the same conditions on parameters for *AD* to be the relevant equilibrium:

$$\frac{(1+a)\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}{g} \ge (1/K_1)^{\chi} \tag{10}$$

Quite intuitively, an "Always Dispersed" equilibrium is therefore more likely if a is relatively high compared to g, so that it becomes more attractive to avoid travel delays by accepting the inconvenience of the SRPAT deviating from the LRPAT. It is also more likely when π is relatively large and K_0 relatively small, so that a distribution of SRPAT's closer to r_0 brings relatively small gains in expected terms.

3.2. Sometimes Dispersed (SD) long-run no-toll equilibrium

When the inequality in (10) is violated, the density of $z(t^{\#})$ is higher, and its time window narrower, than that of the reference arrival rate $r_1(t')$. As long as it is sufficiently sparse to maintain the dispersed equilibrium in state 0 (see Section 3.3 below), there will be a Sometimes Dispersed (*SD*) long-run equilibrium. The expected short-run cost becomes:

$$\mathrm{E}\left[c^{SR}(t^{\#})\right] = (1-\pi)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi} + \pi\cdot\Psi_1\cdot\alpha + \begin{cases}\pi\cdot t^{\#}\cdot\beta & \text{if } t^{\#}\le 0\\ -\pi\cdot t^{\#}\cdot\gamma & \text{if } t^{\#}>0\end{cases} \tag{11}$$

where the last, negative scheduling term indicates that the short-run schedule delay costs in state 1 are lower than they would be in the model with identical preferred arrival times, for which the generalized cost is $\psi_i\cdot\alpha$. The difference is determined by the difference between $t^{\#}$ and t^{*}=0. Note that (11) assumes that all travellers face "condensed" conditions in state 1 and thus will not travel at $t^{\#}$ but instead closer to the relevant end-point of the short-run peak; this is consistent with the equilibrium pattern of $z(t^{\#})$ to be derived below. The corresponding long-run cost becomes:

$$\begin{aligned} c^{LR}(t^{\#}) &= (1-\pi)\cdot\alpha\cdot(1+a)\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi} \\ &\quad + \pi\cdot\Psi_1\cdot\alpha + \pi\cdot a\cdot\alpha\cdot\left(\frac{r_1(t_1'(t^{\#}))}{K_1}\right)^{\chi} + \begin{cases}(\pi-g)\cdot t^{\#}\cdot\beta & \text{if } t^{\#}\le 0\\ (g-\pi)\cdot t^{\#}\cdot\gamma & \text{if } t^{\#}>0\end{cases}\end{aligned} \tag{12}$$

Again, (12) implies a partial differential equation for $z(t^{\#})$, but the solution is not straightforward because of the second term on the second line. This term contains the arrival rate in the condensed state 1 at the arrival moment t' that corresponds in that state with $t^{\#}$: $r_1(t_1'(t^{\#}))$. The term captures the additional long-run valuation of travel delays in state 1. Because the moment of t' for any $t^{\#}$ can only be found after it has been determined also for all earlier $t^{\#}$, it seems a difficult problem to solve. But it turns out that a distribution of $z(t^{\#})$ that follows the by now familiar time pattern, with the difference between $t^{\#}$ and the relevant endpoint of the interval raised to the power of $1/\chi$, in fact gives the solution. In this solution, the ratio $r_1(t_1'(t^{\#}))/z(t^{\#})$ remains constant over time and that constant ratio will be denoted φ. Appendix A discusses the derivation, which results in:

$$z(t^{\#}) = \begin{cases}\left(\dfrac{(g-\pi)\cdot\beta\cdot(t^{\#}-t_l)}{(1-\pi)\cdot(1+a)\cdot\alpha\cdot(1/K_0)^{\chi}+\pi\cdot a\cdot\alpha\cdot(\varphi/K_1)^{\chi}}\right)^{\frac{1}{\chi}} & \text{if } t_l\le t^{\#}\le 0\\ \left(\dfrac{(g-\pi)\cdot\gamma\cdot(t_{l'}-t^{\#})}{(1-\pi)\cdot(1+a)\cdot\alpha\cdot(1/K_0)^{\chi}+\pi\cdot a\cdot\alpha\cdot(\varphi/K_1)^{\chi}}\right)^{\frac{1}{\chi}} & \text{if } 0<t^{\#}\le t_{l'}\end{cases} \tag{13}$$

where t_l and $t_{l'}$ now become:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{(1-\pi)\cdot(1+a)\cdot\alpha\cdot(1/K_0)^{\chi}+\pi\cdot a\cdot\alpha\cdot(\varphi/K_1)^{\chi}}{(g-\pi)\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \tag{14}$$

The ratio φ can be solved from:

$$-t_l(\varphi) = -\varphi\cdot t_{q,1} = \varphi\cdot\Psi_1\cdot\frac{\alpha}{\beta} \tag{15}$$

Equation (15) follows from the fact that if $r_1(t_1'(t^{\#}))/z(t^{\#}) = \varphi < 1$, the width of the distribution of $t^{\#}$ needs to be multiplied by a factor φ, too, to secure that the integrals of both distributions sum up to the same value of N. There appears to be no closed-form solution for φ, but for the numerical exercises to be presented below it could always be found without difficulty through numerical solution of (15). Regime *SD* will be relevant if the density of z falls between both reference arrivals rates, which means that besides a strict violation of (10), the following inequality needs to be satisfied:

$$\left(1/K_0\right)^{\chi} \le \frac{\left(1-\pi\right)\cdot\left(1+a\right)\cdot\left(1/K_0\right)^{\chi} + \pi\cdot a\cdot\left(\varphi/K_1\right)^{\chi}}{\left(g-\pi\right)} \tag{16}$$

For this inequality to be satisfied, it is sufficient if $(1+a)\cdot(1-\pi)/(g-\pi)\ge 1$, a condition easily fulfilled by the assumptions $a\ge 0$ and $0\le g\le 1$ (combined with $0\le\pi\le g\le 1$). This already suggests that the third possible regime, "Never Dispersed" (*ND*), to which we will turn next, will not occur in equilibrium under the assumed parameter restrictions.

3.3. Never Dispersed (ND) long-run no-toll equilibrium

The fact that the inequality in (16) is always satisfied, implies that the third possibility of a "Never Dispersed" (*ND*) long-run equilibrium will never be relevant. The intuition is quite simple. Assume it would exist. Then all early drivers would arrive before their SRPAT in both states (the reasoning for late drivers follows the same but 'mirrored' logic). Shifting $t^{\#}$ to an earlier moment reduces the expected short-run schedule delay cost in both states, therefore yielding a certain gain of β for a unit shift. Given that the equilibria are condensed, travel time costs will not change. And because the long-run schedule delay cost will increase by $g\cdot\beta$ only, with $g<1$, this move is beneficial as long as both states have a condensed equilibrium. One would thus expect that z will be at least as dense as the high-capacity arrival rate r_0, and an *ND* long-run equilibrium will never occur.

3.4. Long-run equilibria under first-best short-run congestion pricing

The same type of long-run equilibria can be identified for the case where, dependent on the state i, optimal congestion tolls are applied on the road. This means that regular time-dependent congestion tolls as defined in (3) are imposed, which for the BPR congestion function boils down to a consistent application of the following time-dependent tolls:

$$\tau(t') = \alpha\cdot\chi\cdot\left(\frac{r(t')}{K_i}\right)^{\chi} \tag{17}$$

The "Always Dispersed" (AD) long-run equilibrium under short-run optimal congestion pricing

To find the *AD* long-run equilibrium under short-run optimal congestion pricing, one should add (17) to the expected cost in equation (6) to obtain the expected short-run generalized price, which in turn implies the following long-run generalized price:

$$p^{LR}(t^{\#}) = \left(1+a+\chi\right)\cdot\alpha\cdot z(t^{\#})^{\chi}\cdot\left((1-\pi)\cdot\left(\frac{1}{K_0}\right)^{\chi}+\pi\cdot\left(\frac{1}{K_1}\right)^{\chi}\right)+\begin{cases}-\beta\cdot g\cdot t^{\#} & \text{if } t^{\#}\le 0\\ \gamma\cdot g\cdot t^{\#} & \text{if } t^{\#}>0\end{cases} \tag{18}$$

The structure of the implied partial differential equation that results from setting the time-derivative to zero is similar to that for the no-tolled equilibria in (7), and the solution is:

$$z(t^{\#}) = \begin{cases}\left(\dfrac{g\cdot\beta\cdot\left(t^{\#}-t_l\right)}{\left(1+a+\chi\right)\cdot\alpha\cdot\left(\left(1-\pi\right)\cdot\left(1/K_0\right)^{\chi}+\pi\cdot\left(1/K_1\right)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } t_l\le t^{\#}\le 0\\ \left(\dfrac{g\cdot\gamma\cdot\left(t_{l'}-t^{\#}\right)}{\left(1+a+\chi\right)\cdot\alpha\cdot\left(\left(1-\pi\right)\cdot\left(1/K_0\right)^{\chi}+\pi\cdot\left(1/K_1\right)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } 0<t^{\#}\le t_{l'}\end{cases} \tag{19}$$

The instants t_l and $t_{l'}$ can now be solved for as:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{\left(1+a+\chi\right)\cdot\alpha\cdot\left(\left(1-\pi\right)\cdot\left(1/K_0\right)^{\chi}+\pi\cdot\left(1/K_1\right)^{\chi}\right)}{g\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \tag{20}$$

The parametrical conditions under which this *AD* long-run equilibrium under first-best short-run congestion pricing is relevant, are also pretty similar to those in (10) for the untolled case:

$$\frac{\left(1+a+\chi\right)\cdot\left(\left(1-\pi\right)\cdot\left(1/K_0\right)^{\chi}+\pi\cdot\left(1/K_1\right)^{\chi}\right)}{g}\ge\left(1+\chi\right)\cdot\left(1/K_1\right)^{\chi} \tag{21}$$

and the interpretation is therefore also along the same lines.

The "Sometimes Dispersed" (SD) long-run equilibrium under short-run optimal congestion pricing

Similarly, we can characterize the *SD* long-run equilibrium under first-best short-run congestion pricing as an analytically close variant of the untolled case discussed in Section 3.2. For the long-run generalized price we find:

$$\begin{aligned}p^{LR}(t^{\#}) &= (1-\pi)\cdot\left(1+a+\chi\right)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi}\\ &\quad+\pi\cdot\Psi_1\cdot\alpha\cdot\left(1+\chi\right)^{\frac{1}{1+\chi}}+\pi\cdot a\cdot\alpha\cdot\left(\frac{r_1(t_1'(t^{\#}))}{K_1}\right)^{\chi}+\begin{cases}\left(\pi-g\right)\cdot t^{\#}\cdot\beta & \text{if } t^{\#}\le 0\\ \left(g-\pi\right)\cdot t^{\#}\cdot\gamma & \text{if } t^{\#}>0\end{cases}\end{aligned} \tag{22}$$

The solution for $z(t^{\#})$ becomes:

$$z(t^{\#}) = \begin{cases}\left(\dfrac{\left(g-\pi\right)\cdot\beta\cdot\left(t^{\#}-t_l\right)}{\left(1-\pi\right)\cdot\left(1+a+\chi\right)\cdot\alpha\cdot\left(1/K_0\right)^{\chi}+\pi\cdot a\cdot\alpha\cdot\left(\varphi/K_1\right)^{\chi}}\right)^{\frac{1}{\chi}} & \text{if } t_l\le t^{\#}\le 0\\ \left(\dfrac{\left(g-\pi\right)\cdot\gamma\cdot\left(t_{l'}-t^{\#}\right)}{\left(1-\pi\right)\cdot\left(1+a+\chi\right)\cdot\alpha\cdot\left(1/K_0\right)^{\chi}+\pi\cdot a\cdot\alpha\cdot\left(\varphi/K_1\right)^{\chi}}\right)^{\frac{1}{\chi}} & \text{if } 0<t^{\#}\le t_{l'}\end{cases} \tag{23}$$

where φ will have a different value than for the untolled case, and t_l and $t_{l'}$ become:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{(1-\pi)\cdot(1+a+\chi)\cdot\alpha\cdot(1/K_0)^{\chi}+\pi\cdot a\cdot\alpha\cdot(\varphi/K_1)^{\chi}}{(g-\pi)\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \quad (24)$$

The ratio φ again cannot be solved for in closed form, but follows from:

$$-t_l(\varphi) = -\varphi\cdot t_{q,1} = \varphi\cdot(1+\chi)^{\frac{1}{1+\chi}}\cdot\Psi_1\cdot\frac{\alpha}{\beta} \quad (25)$$

Regime *SD* will be relevant if the density of *z* falls between both reference arrivals rates, and its time window fits between the two reference windows, so that besides a strict violation of (21) the following inequality should hold:

$$(1+\chi)\cdot(1/K_0)^{\chi} \le \frac{(1-\pi)\cdot(1+a+\chi)\cdot(1/K_0)^{\chi}+\pi\cdot a\cdot(\varphi/K_1)^{\chi}}{(g-\pi)} \quad (26)$$

For this inequality to be satisfied, it is sufficient if $(1+a+\chi)\cdot(1-\pi)/(g-\pi)\ge(1+\chi)$, a condition that is again easily fulfilled by the assumptions $a\ge 0$ and $0\le g\le 1$ (combined with $0\le\pi\le g\le 1$).

The "Never Dispersed" (ND) long-run equilibrium under short-run optimal congestion pricing

Because (26) is always satisfied, the *ND* regime will also not occur in equilibrium under first-best short-run congestion pricing.

3.5. Long-run first-best optima

The central question of this paper is whether the long-run equilibrium that emerges under first-best short-run congestion pricing, as described in Section 3.4 above, constitutes a long-run first-best optimum. With the road tolls optimized to internalize the congestion externality by optimally affecting departure time decisions in the state-dependent short-run problems, there is another margin of behaviour, which remains only indirectly affected, and that is the choice of the SRPAT in the long-run problem. To answer the question of whether the endogenous equilibrium pattern of SRPAT's, so the distribution $z(t^{\#})$, that materializes under consistent application of first-best short-run congestion pricing is efficient, we equip the regulator with a third instrument. Besides the two state-dependent first-best short-run congestion tolls that we considered above, and that now will be denoted τ_i^{SR}, we assume that the regulator also has a perfectly flexible instrument available to affect the choice of the desired arrival time: a long-run tax $\tau^{LR}(t^{\#})$ on the choice of $t^{\#}$. Even though it may not be realistic to assume that such a perfect instrument could be applied, it offers an intuitive way of identifying the social desirability of affecting the choice of $t^{\#}$ under optimal short-run tolling. That is, any non-zero first-best optimal time pattern of $\tau^{LR}(t^{\#})$ identified when optimal short-run tolls τ_i^{SR} are in place reflects an apparent inefficiency in the decentralized choice of $t^{\#}$. It does so without introducing any second-best complications that more realistic instruments such as staggered work hour would typically introduce into the analysis. We will therefore

now derive the optimal time-pattern of $\tau^{LR}(t^{\#})$ for the two types of long-run equilibrium that may occur: "Always Dispersed" (*AD*) and "Sometimes Dispersed" (*SD*). Two convenient notational manipulations are introduced. Because in a dispersed equilibrium, travellers arrive at their desired arrival time, we may for such short-run equilibria substitute the desired arrival time $z^{\#}$ for the actual arrival time z'. For a condensed equilibrium, in the *SD* regime, we may use that when drivers maintain their order, there is a given correspondence between $z^{\#}$ and z', which we will denote as $z'(z^{\#})$. Both manipulations will be used below, with the advantage that we can optimize the long-run and short-run problem simultaneously through the determination of the optimal time pattern of SRPAT's, $z(t^{\#})$, alone.

The "Always Dispersed" (AD) long-run optimum

To identify the long-run optimum, we first write out the long-run cost in the *AD* regime, which is given by the expected value of the short-run cost plus the additional valuations that arise from the long-run cost components, where it is to be remembered that there are no short-run schedule delay costs because travellers arrive at their desired arrival time in either state:

$$c^{LR}(t^{\#}) = \alpha \cdot (1+a) \cdot \left((1-\pi) \cdot \left(\frac{z(t^{\#})}{K_0} \right)^{\chi} + \pi \cdot \left(\frac{z(t^{\#})}{K_1} \right)^{\chi} \right) + \begin{cases} -\beta \cdot g \cdot t^{\#} \\ \gamma \cdot g \cdot t^{\#} \end{cases} \tag{27}$$

The long-run marginal cost associated with a choice of $t^{\#}$ as the SRPAT therefore amounts to:

$$mc^{LR}(t^{\#}) = c^{LR}(t^{\#}) + (1+a) \cdot \alpha \cdot z(t^{\#}) \cdot \left((1-\pi) \cdot \frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})} + \pi \cdot \frac{\partial T_1(z(t^{\#}))}{\partial z(t^{\#})} \right) \tag{28}$$

Under short-run optimal congestion pricing, the expected value of the long-run generalized price becomes:

$$\begin{aligned} & p^{LR}(t^{\#}) = c^{LR}(t^{\#}) + (1-\pi) \cdot \tau_0^{SR}(t^{\#}) + \pi \cdot \tau_1^{SR}(t^{\#}) + \tau^{LR}(t^{\#}) \\ & \text{with:} \quad \tau_0^{SR}(t^{\#}) = \alpha \cdot z(t^{\#}) \cdot \frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})} \quad \text{and} \quad \tau_1^{SR}(t^{\#}) = \alpha \cdot z(t^{\#}) \cdot \frac{\partial T_1(z(t^{\#}))}{\partial z(t^{\#})} \end{aligned} \tag{29}$$

Comparing (28) and (29), it is then straightforward to derive the following long-run tax on the choice of $t^{\#}$ for the *AD* long-run optimum:

$$\tau^{LR}(t^{\#}) = a \cdot \alpha \cdot z(t^{\#}) \cdot \left((1-\pi) \cdot \frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})} + \pi \cdot \frac{\partial T_1(z(t^{\#}))}{\partial z(t^{\#})} \right) \tag{30}$$

Equation (30) shows that, also when optimal short-run congestion tolls are in place on the road in both states, there is still an uninternalized congestion externality involved in the choice of the SRPAT, and consequently the optimal long-run tax is unequal to zero. This answers the central question asked in this paper in the affirmative: it is not generally sufficient to apply only first-best optimal short-run when the objective is to achieve the first-best long-run outcome. The intuition, for this "Always Dispersed" long-run optimum, is as follows. The short-run tolls are set to optimize the use of the road conditional on the prevailing capacity. The corresponding toll rules are based on the short-run valuations, and therefore do not reflect the full long-run negative valuation by other drivers of an individual's choice to choose a certain SRPAT. This full valuation exceeds the short-run valuation, already internalized

through the short-run tolls, by a multiplicative factor *a*, and that is exactly the weighting of the value of time α applied in the long-run toll of (30).

Following the same steps as before, we can characterize the long-run equilibrium in closed-form as follows. First, substitution of (30) implies the following long-run generalized price:

$$p^{LR}(t^{\#}) = (1+a)\cdot(1+\chi)\cdot\alpha\cdot z(t^{\#})^{\chi}\cdot\left((1-\pi)\cdot\left(\frac{1}{K_0}\right)^{\chi}+\pi\cdot\left(\frac{1}{K_1}\right)^{\chi}\right)+\begin{cases}-\beta\cdot g\cdot t^{\#} & \text{if } t^{\#}\le 0\\ \gamma\cdot g\cdot t^{\#} & \text{if } t^{\#}>0\end{cases} \quad (31)$$

The solution for the implied partial differential equation is now:

$$z(t^{\#}) = \begin{cases}\left(\dfrac{g\cdot\beta\cdot(t^{\#}-t_l)}{(1+a)\cdot(1+\chi)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } t_l\le t^{\#}\le 0\\ \left(\dfrac{g\cdot\gamma\cdot(t_{l'}-t^{\#})}{(1+a)\cdot(1+\chi)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}\right)^{\frac{1}{\chi}} & \text{for } 0<t^{\#}\le t_{l'}\end{cases} \quad (32)$$

The instants t_l and $t_{l'}$ now become:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{(1+a)\cdot(1+\chi)\cdot\alpha\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}{g\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \quad (33)$$

The parametrical conditions under which this *AD* long-run equilibrium under first-best short-run congestion pricing is relevant are now given by:

$$\frac{(1+a)\cdot(1+\chi)\cdot\left((1-\pi)\cdot(1/K_0)^{\chi}+\pi\cdot(1/K_1)^{\chi}\right)}{g}\ge(1+\chi)\cdot(1/K_1)^{\chi} \quad (34)$$

Comparison with (21) shows that an *AD* constellation occurs more easily in the overall optimum than for optimal short-run tolling without the long-run tax. This is intuitive, as the long-run tax in (30) leads to a dispersion in SRPAT's.

The "Sometimes Dispersed" (SD) long-run optimum

Also the long-run optimum can have the "Sometimes Dispersed" configuration. To determine the optimal long-run tax τ^{LR}, we follow the same conceptual steps as above, but there is an important difference in the derivations: for the condensed state 1, the short-run optimum will be insensitive to a marginal change in the distribution of $z^{\#}$. Such a change will not lead to a different equilibrium dynamic pattern of travel delays, exactly because it is a condensed equilibrium, and the travel pattern is the reference pattern r_1^{FB} that would occur with identical SRPAT's t^*. In that state, a traveller's generalized cost and price is the same as what is in the basic model summarized in Table 1, minus the lower schedule delay cost due to the SRPAT $t^{\#}$ being closer to the arrival time t' than t^* is. Likewise, the short-run marginal cost in that state,

mc_1^{SR}, is independent of $z^{\#}$. This means that the generalized cost and price as a function of $t^{\#}$ now become:

$$c^{LR}(t^{\#}) = (1-\pi)\cdot(1+a)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi} + \pi\cdot\left(c_1^{SR} + a\cdot\alpha\cdot\left(\frac{r_1(t'(t^{\#}))}{K_1}\right)^{\chi}\right) + \begin{cases} -\beta\cdot(g-\pi)\cdot t^{\#} \\ \gamma\cdot(g-\pi)\cdot t^{\#} \end{cases} \tag{35}$$

$$\begin{aligned} p^{LR}(t^{\#}) = \tau^{LR} &+ (1-\pi)\cdot\left(\tau_0^{SR} + (1+a)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi}\right) \\ &+ \pi\cdot\left(p_1^{SR} + a\cdot\alpha\cdot\left(\frac{r_1(t'(t^{\#}))}{K_1}\right)^{\chi}\right) + \begin{cases} -\beta\cdot(g-\pi)\cdot t^{\#} \\ \gamma\cdot(g-\pi)\cdot t^{\#} \end{cases} \end{aligned} \tag{36}$$

Note that p_1^{SR} includes the short-run tolls τ_1^{SR} charged in that state. The difference between the generalized price in (36) and the long-run marginal cost can now be written as:

$$\begin{aligned} p^{LR}(t^{\#}) - mc^{LR}(t^{\#}) = \tau^{LR} &+ (1-\pi)\cdot\left(\tau_0^{SR} - (1+a)\cdot\alpha\cdot z(t^{\#})\cdot\frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})}\right) \\ &+ \pi\cdot\left(p_1^{SR} - mc_1^{SR} + a\cdot\alpha\cdot\left(\frac{r_1(t'(t^{\#}))}{K_1}\right)^{\chi}\right) \end{aligned} \tag{37}$$

The short-run tolls amount to:

$$\begin{aligned} \tau_0^{SR}(t^{\#}) &= z(t^{\#})\cdot\alpha\cdot\frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})} \\ \tau_1^{SR}(t') &= r(t')\cdot\alpha\cdot\frac{\partial T_1(r(t'))}{\partial r(t')} \end{aligned} \tag{38}$$

These tolls, combined with $p_1^{SR} = mc_1^{SR}$, imply that in order to equate (37) to zero, the following long-run tax τ^{LR} should be imposed:

$$\tau^{LR} = (1-\pi)\cdot(1+a)\cdot\alpha\cdot z(t^{\#})\cdot\frac{\partial T_0(z(t^{\#}))}{\partial z(t^{\#})} - \pi\cdot a\cdot\alpha\cdot T_1(t'(t^{\#})) \tag{39}$$

The first term corresponds to the dispersed short-run optimum in the high-capacity state, and has the same interpretation as the long-run toll for the "Always Dispersed" long-run optimum: it internalizes the additional congestion externality that occurs in the long run because the long-run value of time exceeds its short-run counterpart, which, in turn, is the one considered in the specification of short-run congestion tolls. The surprise is in the second, negative term. Whereas one might have expected, intuitively, that for a more heavily congested system – with a Sometimes Dispersed rather than an Always Dispersed short-run optimum – the need to also regulate the choice for SRPAT's would increase, the opposite is in fact true. That is, a positive term in the long-run toll for the low-capacity state in the *AD* optimum is replaced by a negative term for that same state in the *SD* equilibrium.

The intuition is that a marginal change in the pattern of short-run preferred arrival times does not change traffic conditions in the low-capacity state 1, because it is a condensed equilibrium. In other words, an individual's choice of $t^{\#}$ does not impose any travel delay

externality on other drivers in that state, unlike what is the case in the *AD* long-run optimum. At the same time, an individual's choice of $t^{\#}$ does affect her own expected travel times for the low capacity state 1, and therewith the long-run valuation of expected travel delays over state 1, $\pi \cdot a \cdot \alpha \cdot T_1(t'(t^{\#}))$. However, because the aggregate travel pattern and hence the total travel delay in state 1, and therewith also its long-run value, is insensitive to marginal changes in $z(t^{\#})$, this variation in generalized price over $t^{\#}$ does not reflect variation in marginal cost. Therefore, to induce the optimal density $z(t^{\#})$ from the perspective also of the dispersed high-capacity state 0, this term $\pi \cdot a \cdot \alpha \cdot T_1(t'(t^{\#}))$ is subtracted from the first term in the long-run tax τ^{LR} in (39). As a result, the long-run toll may be quite low, and may in fact even be negative.

Finally, we can again solve for the long-run equilibrium in closed-form. For the generalized prices, we substitute the tolls in (38) and (39), which imply that the regular short-run generalized price from Table 1 applies in state 1, corrected for the lower schedule delay cost in the short run optimum due to the dispersion in $t^{\#}$:

$$\begin{aligned} p^{LR}(t^{\#}) &= (1-\pi)\cdot(1+a)\cdot(1+\chi)\cdot\alpha\cdot\left(\frac{z(t^{\#})}{K_0}\right)^{\chi} \\ &\quad + \pi\cdot(1+\chi)^{\frac{1}{1+\chi}}\cdot\Psi_1\cdot\alpha + \begin{cases} -\beta\cdot(g-\pi)\cdot t^{\#} & \text{if } t^{\#}\le 0 \\ \gamma\cdot(g-\pi)\cdot t^{\#} & \text{if } t^{\#}>0 \end{cases} \end{aligned} \tag{40}$$

The solution for the implied partial differential equation is:

$$z(t^{\#}) = \begin{cases} \left(\dfrac{(g-\pi)\cdot\beta\cdot(t^{\#}-t_l)}{(1-\pi)\cdot(1+a)\cdot(1+\chi)\cdot\alpha\cdot(1/K_0)^{\chi}}\right)^{\frac{1}{\chi}} & \text{for } t_l \le t^{\#} \le 0 \\ \left(\dfrac{(g-\pi)\cdot\gamma\cdot(t_{l'}-t^{\#})}{(1-\pi)\cdot(1+a)\cdot(1+\chi)\cdot\alpha\cdot(1/K_0)^{\chi}}\right)^{\frac{1}{\chi}} & \text{for } 0 < t^{\#} \le t_{l'} \end{cases} \tag{41}$$

The following instants t_l and $t_{l'}$ can be found:

$$-t_l = \frac{\gamma}{\beta}\cdot t_{l'} = \left(N\cdot\frac{\gamma}{\gamma+\beta}\cdot\frac{1+\chi}{\chi}\cdot\left(\frac{(1-\pi)\cdot(1+a)\cdot(1+\chi)\cdot\alpha\cdot(1/K_0)^{\chi}}{(g-\pi)\cdot\beta}\right)^{\frac{1}{\chi}}\right)^{\frac{\chi}{1+\chi}} \tag{42}$$

The parametrical conditions that separates this *SD* long-run equilibrium under first-best short-run congestion pricing from an *ND* equilibrium are now given by:

$$\frac{(1-\pi)}{(g-\pi)}\cdot(1+a)\ge 1 \tag{43}$$

By the assumptions $a \ge 0$ and $g \le 1$, this is always satisfied, confirming that again an ND equilibrium will not occur.

4. A numerical example

4.1. Parametrization

This section presents some numerical results and sensitivity analysis that aim to provide insight into the determinants of the importance of applying the long-run tax instrument on top

of conventional short-run road tolls. The numerical model developed for this purpose is entirely consistent with the model employed in Section 3 above. The number of parameters to be set is not too high. The scheduling parameters are set at 10 for the value of time α; a value of schedule delay early β below α at 5; and the value of schedule delay late γ is highest of the three, at 20. The implied value of δ is 4. These relative magnitudes are reasonably in line with the original estimates of Small (1982). For the power of the BPR function the conventional value of χ = 4 is chosen for the base case. To generate sufficient stochasticity with a two-mass-points distribution, the capacity is assumed to be halved when there is an incident – think of a lane closure on a two-lane highway – from 10 000 to 5 000; and a rather high value of the probability is assumed, with π set at 0.25 in the base case. The total number of users is 10 000. Finally, the relative differences between short-run and long-run valuations are set at g = 0.5 and a =3 for the base case; where the former has a magnitude comparable to what was found in Peer *et al.* (2015), a is well below the factor 5 found in that study.

4.2. Illustrating equilibrium

Before moving to the sensitivity analyses, it is instructive to briefly consider equilibrium flows for a Sometimes Dispersed equilibrium, as displayed in Figure 1. The figure shows the two reference arrival rates. For the chosen parametrization, with π = 0.05, an *SD* equilibrium arises, with $z(\cdot)$ lying between $r_0(\cdot)$ and $r_1(\cdot)$. The actual arrival rate in state 0 is therefore given by $z(\cdot)$, not $r_0(\cdot)$. The shapes of the rates depicted in Figure 1 are typical for the equilibrium generated by the model, and follow from the functional forms identified in Table 1 and in Section 3, which raise the linear time difference with the long-run preferred arrival time to the power of $1/\chi$.

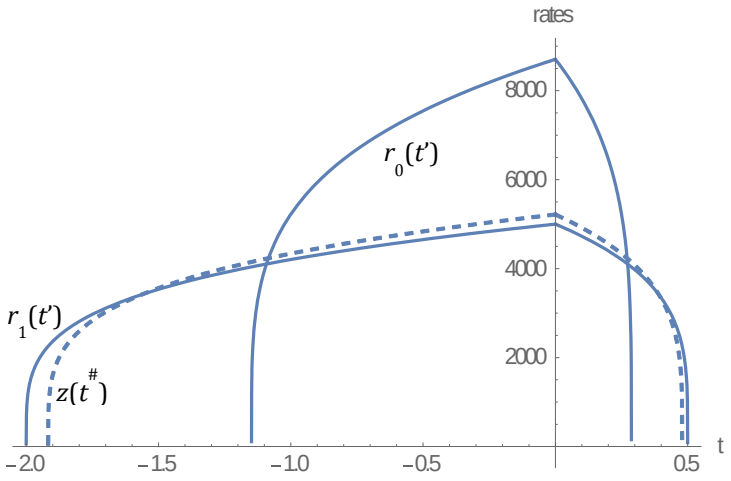


Figure 1. Illustrating equilibrium: $r_0(\cdot)$,$r_1(\cdot)$, and $z(\cdot)$ in a Sometimes Dispersed equilibrium (π=0.05)

4.3. Optimal pricing in the base-case model

Next, we turn to the impacts of optimal pricing. For the base-case parameterization of the model, with π set back to 0.25, Always Dispersed no-toll equilibria and optima apply. For our discussion, it is useful to distinguish between two types of optima. The first is the First-Best (*FB*) optimum, in which social welfare is maximized through application of the short-run

optimal congestion tolls and the long-run optimal tax. The second is the one in which only state-dependent short-run congestion tolls are used. It will be referred to as the Quasi First-Best (*QFB*) optimum, to reflect that the regulator applies the standard optimal short-run tolls ignoring how these should be adjusted in a second-best fashion to affect the choices of SRPAT's in absence of the long-run tax. And of course, besides these two optima, the No-Toll Equilibrium is the relevant reference. For the base-case parametrization, total social cost of travel including long-run schedule-delay cost amount to 59 443 in *NTE*; 48 759 in *QFB*; and 45 564 in *FB*. The so-called relative efficiency ω of *QFB*, defined as the fraction of *FB* cost savings it achieve, thus amounts to 0.77: the welfare gain from conventional congestion pricing is 77% of the welfare gain that can be achieved when the long-run scheduling decisions are also optimized using the long-run tax as the third instrument.

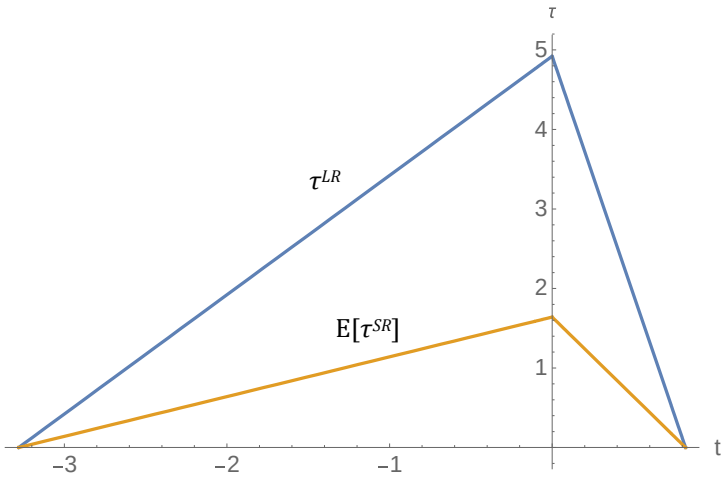


Figure 2. Optimal long-run tax and expected value of optimal short-run tolls in the FB optimum

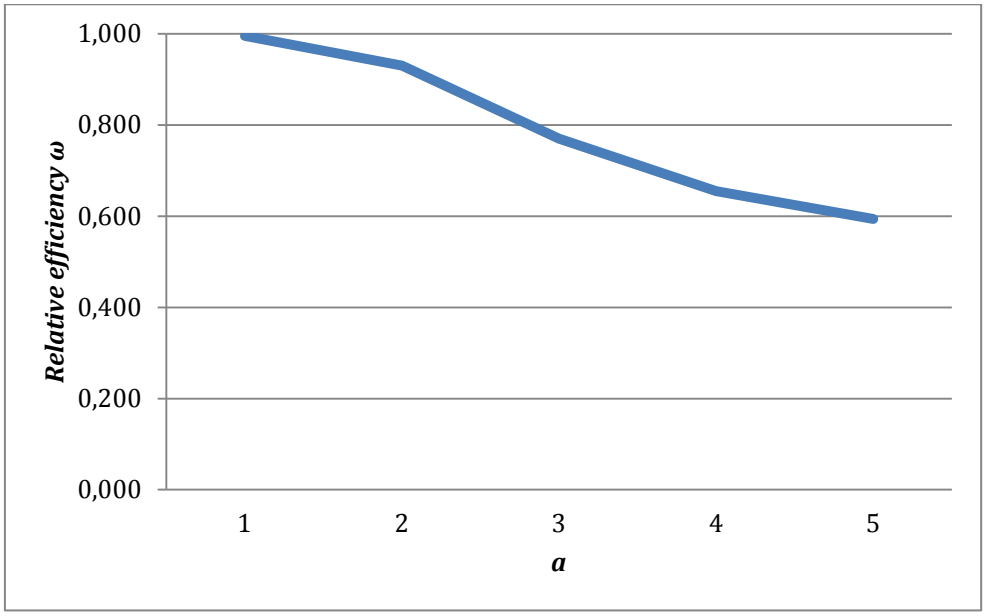
Figure 2 shows that in the *FB* optimum, the long-run tax exceeds the expected value of the short-run tolls, underlining the potential importance of this third instrument. Note that the time-patterns of tolls are linear, consistent with the results of the original Chu (1995) model summarized in Table 1. Despite the high level of τ^{LR} compared to τ^{SR}, dropping the instrument and readjusting the short-run tolls in response leads, as stated, to a welfare loss of 23%, which may be smaller than expected on the basis of Figure 2. The reason is that the short-run tolls will be higher than in Figure 2, with the expected value peaking around 4. This higher expected value works as a substitute for the long-run toll, limiting the welfare loss from losing that instrument.

4.4. Sensitivity analysis I: varying a

The first parameter we vary to gain deeper insight into the determinants of the relative efficiency ω of the *QFB* policy is the premium a that determines the difference between the long-run and short-run value of time. Figure 3 shows the results. As will be true for other sensitivity analysis below, the pattern of ω sometimes appears "kinked". The reason is that along the sensitivity analysis, either of the three relevant equilibria (*NTE*, *QFB* and *FB*) may

switch between Always Dispersed (*AD*) and Sometimes Dispersed (*SD*). For that reason, the figure carries a note explaining which types of equilibria apply in which parameter ranges.

Figure 3 shows that a higher value of a reduces the relative efficiency of *QFB*; quite intuitively so as the ignored long-run value of time becomes more important. For the point estimate of a from Peer *et al.* (2015), ω has dropped to 0.6, suggesting that ignoring the nature of scheduling decisions with routine formation may, indeed, cause substantial welfare losses.



Note: NTE is AD throughout; QFB is SD for a={1,2,3} and AD for a={4,5}; FB is SD for a=1 and AD for a={2,3,4,5}

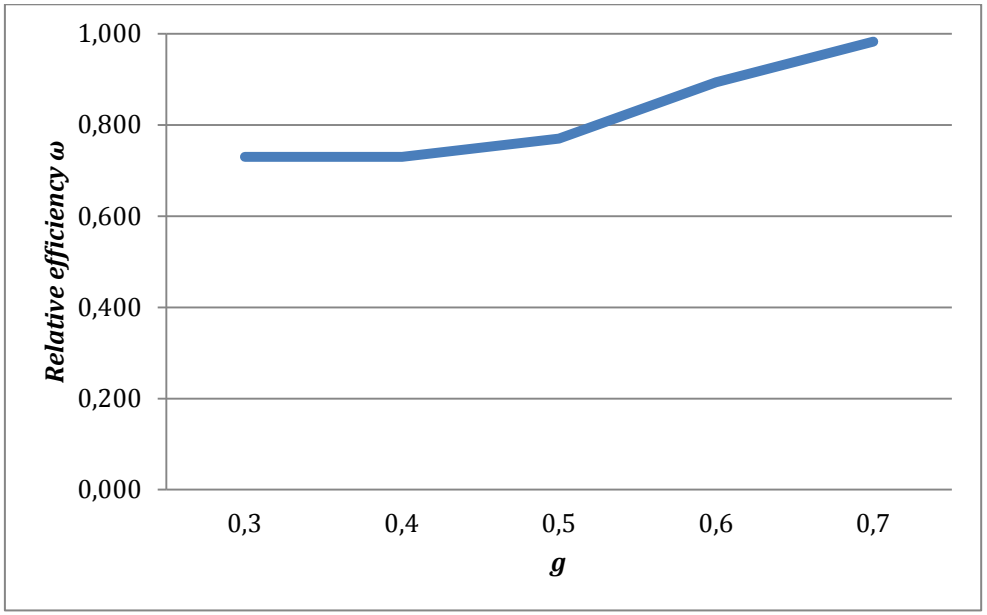
Figure 3. Varying a: relative efficiency ω

4.5. Sensitivity analysis II: varying g

Next, Figure 4 shows the relative efficiency ω of *QFB* for variation in the ratio of long-run versus short-run values of schedule delay, g. Again, we see a decline in ω when valuations differ stronger between the short run and the long run, now when moving to the left in the figure. The reason is that a low value of g means that it is relatively attractive to reduce congestion by having travellers change their routines, for which the long-run toll is the most effective instrument.

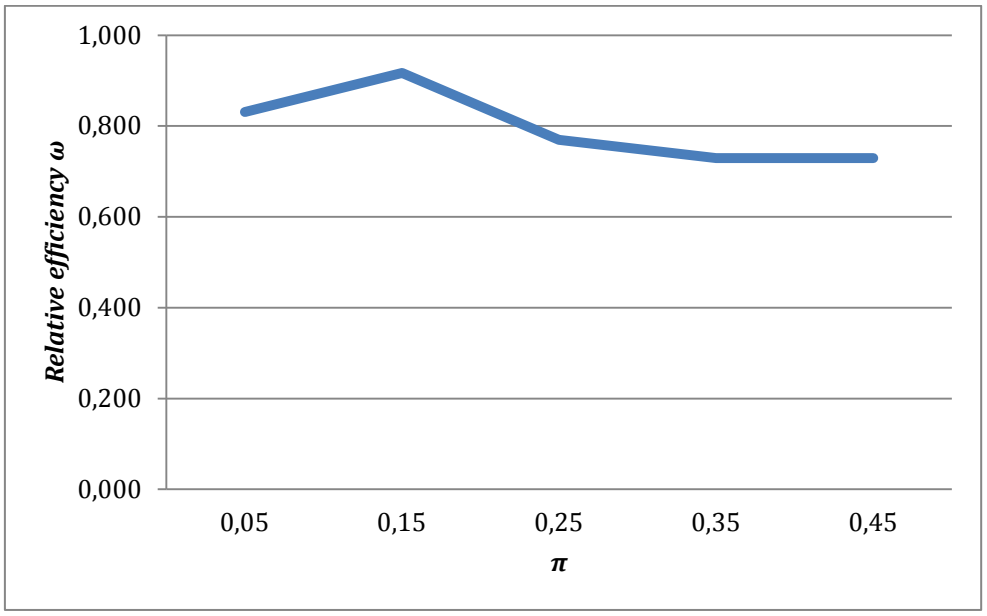
4.6. Sensitivity analysis III: varying π

A third parameter of interest is π, the probability of an incident and hence capturing the degree of stocahsticity in the system. Figure 5 gives the results. We now see a more volatile pattern, which has to do with how regimes change between *SD* and *AD* for the different equilibria. In particular, the switch for *FB* from *SD* to *AD* raises its cost discretely, which makes *QFB* (for which that switch occurs between 0.25 and 0.35) gain in terms of ω. Otherwise, we see ω fall with an increasing volatility, which is intuitive as the greater importance of state 1 means that routine scheduling becomes more distinct from departure time optimization in state 0, and hence more important for overall efficiency.



Note: NTE is AD throughout; QFB is AD for g={0.3,0.4} and SD for g={0.5,0.6,0.7}; FB is AD for g={0.3,0.4, 0.5,0.6} and SD for a=0.7

Figure 4. Varying g: relative efficiency ω



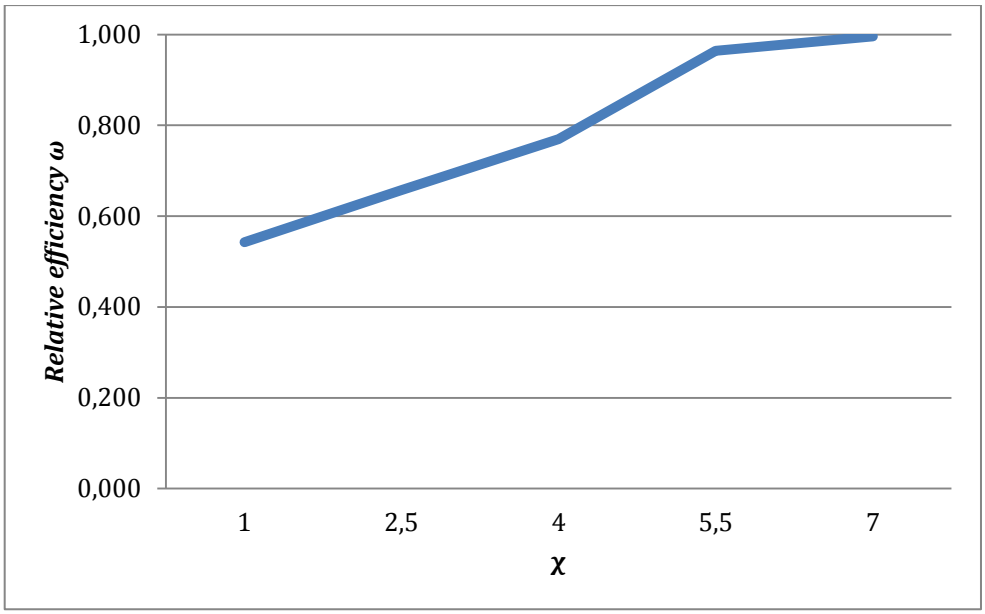
Note: NTE is SD for π=0.05 and AD for π={0.15,0.25,0.3, 0.45}; QFB is SD for π={0.05,0.15,0.25} and AD for π={0.35, 0.45}; FB is SD for π=0.05 and AD for π={0.15,0.25,0.3, 0.45}

Figure 5. Varying π: relative efficiency ω

4.7. Sensitivity analysis IV: varying χ

The final sensitivity analysis involves the curvature of the congestion function, as reflected by the power χ. Figure 6 displays the results, and shows how ω rises as the power χ increases. The intuition is now that the strong curvature of the congestion function means that only relatively small changes in flow are needed to solve most congestion. Once the short-run tolls

have achieved this, there is relatively little to gain from further changes in routines. The implication is that the explicit consideration of long-run scheduling decisions can be expected to bring greater benefits for congested networks, for which χ is typically lower as interactions at for instance crossings cause congestion to set in at relatively lower use levels, than for single facilities.



Note: NTE is AD throughout; QFB is AD for χ={1,2.5} and SD for χ={4,5.5,7}; FB is AD for χ={1,2.5,4,5.5} and SD for χ=7

Figure 6. Varying χ: relative efficiency ω

5. Conclusion

Recent empirical work has suggested that there is an important distinction between short-run and long-run scheduling behaviour of commuters. Peer *et al.* (2015) for example find that the average value of time when consumers form their routines in the long-run problem may exceed by a factor 6 the short-run value that governs departure time choice given these routines. For values of schedule delay, in contrast, the short-run value exceeds the long-run value, by a factor 2. And, when forming routines, consumers in fact choose a most preferred arrival time that may deviate from the value they would choose in absence of congestion because a change in routines may mean that shorter delays will be encountered.

This paper has investigated whether this distinction between short-run and long-run scheduling decisions affect optimal pricing of a congestible facility. The framework developed rests on a stochastic generalization of the Henderson-Chu dynamic model of flow congestion for describing short-run equilibria. This system was fully integrated with a dynamic model of endogenous routine formation.

The paper's main conclusion is that consistent application of short-run first-best optimal congestion pricing does not optimally decentralize the optimal formation of routines in the long-run problem. It is important to emphasize that the model does not assume

externalities in arrival timing that could arise because of, for example, the need to coordinate meetings (e.g. Fosgerau and Small, 2017). Such externalities can quite naturally be expected to call for corrective taxation on top of regular congestion taxes. But such market failure is not present in the model presented above, and the need to affect the choice of routines stems directly and exclusively from the structure of preferences.

A separate instrument is therefore needed to optimize routine formation. In the analysis, it was assumed that for this purpose, an optimal tax instrument is available to affect consumers' choices of the preferred arrival times for their short-run problems, their SRPAT's. In practice, it may not be easy to implement such a tax, since the choice of SRPAT's cannot be observed. The results indicate that in such cases, second-best instruments that would affect the choice of SRPAT's, such as subsidies on staggered work-hours arrangements, may be welfare enhancing also if short-run optimal road pricing is already in place.

Somewhat counterintuitively, it was found that the relevance of the long-run tax, as expressed by it level, is likely to be higher for less congested systems; in the terminology of this paper, when an Always Dispersed (*AD*) rather than a Sometimes Dispersed (*SD*) equilibrium applies. This means that travellers always arrive at their SRPAT. The intuition is that in *SD* equilibrium, marginal changes in the distribution of SRPAT's have now consequences for the dynamic pattern of arrivals; hence there is no benefit from inducing such marginal changes.

Finally, a numerical model provided insight into the determinants of the relative value added of the long-run tax instrument. This value added increases when discrepancies between short-run and long-run valuations increase, when the stochasticity increases, and when the congestion function is less strongly curved meaning that congestion sets in already far below capacity. The intuition behind these results appears quite general, and it may be expected that the results carry over to other model specifications.

Indeed, one may identify various avenues for further research. Within the context of the present model, it would be interesting to explore other congestion technologies, more sophisticated formulations of stochasticity, other types of scheduling models including the so-called *HW* model (Vickrey, 1973; Tseng and Verhoef, 2008), and externalities that may arise from arrival time decisions. Besides that, further empirical work on the nature of long-run and short-run consumer decision making seems highly relevant. This paper has demonstrated that a distinction between short-run and long-run decision making may justify the use of an additional instrument, focused on routine formation. Ignoring this may lead, according to the numerical examples given, lead to welfare losses of up to around 40% of the potential gains from optimal pricing. This seems to justify such further empirical work.

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.

Chu, X. (1995) “Endogenous trip scheduling: the Henderson approach reformulated and compared with the Vickrey approach” *Journal of Urban Economics* **37** (3) 324-343.

Fosgerau, M. and K.A. Small (2017) “Endogenous scheduling preferences and congestion” *International Economic Review* **58** (2) 585-615.

Henderson, J.V. (1974) “Road congestion: a reconsideration of pricing theory” *Journal of Urban Economics* **1** (3) 346-365.

Henderson, J.V. (1981) “The economics of staggered work hours” *Journal of Urban Economics* **9** (3) 349-364.

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.

Mun, S. (1994) "Traffic jams and the congestion toll" *Transportation Research* **28B** (5) 365-375.

Peer, S., and E.T. Verhoef (2013) “Equilibrium at a bottleneck when long-run and short-run scheduling preferences diverge” *Transportation Research Part B: Methodological* **57** 12-27.

Peer, S., E.T. Verhoef, J. Knockaert, P. Koster and Y. Tseng (2015) “Long-run vs. short-run perspectives on consumer scheduling: Evidence from a revealed-preference experiment among peak-hour road commuters” *International Economic Review* **56**(1) 303-323.

Small, K.A. (1982) “The scheduling of consumer activities: Work trips” *American Economic Review* **72**(3), 467–479.

Small, K.A. (2012) “Valuation of travel time” *Economics of Transportation* **1** (1-2) 2-14.

Tseng, Y. and E.T. Verhoef (2008) “Value of time by time of day: a stated-preference study” *Transportation Research* **42B** (7-8) 607-618.

Verhoef, E.T. (2000) “The implementation of marginal external cost pricing in road transport: long run *vs* short run and first-best *vs* second-best” *Papers in Regional Science* **79** 307-332.

Vickrey, W. (1973) “Pricing, metering, and efficiently using urban transportation facilities” *Highway Research Record* **476** 36-48.

Appendix A. Solving for the long-run equilibria

The distribution of SRPATS in the long-run equilibria in Section 3 are all found by solving the partial differential equation that is obtained when setting the time derivative of the long-run generalized price, p^{LR}, equal to zero. This generalized price always takes on a form that can compactly be written as:

$$p^{LR}(t^{\#}) = X \cdot z(t^{\#})^{\chi} + Y + \begin{cases} -d \cdot \beta \cdot t^{\#} \\ d \cdot \gamma \cdot t^{\#} \end{cases} \tag{A.1}$$

where the shorthand composite parameters X, Y and d vary between the long-run equilibria considered. Note that also for the Sometimes Dispersed equilibria with and without short-run tolling we obtain this form as soon as we assume upfront that in the solution to the original problem, a fixed ratio φ will exist between r_1 and z so that the former can be replaced using $r_1(t_1'(t^{\#})) = \varphi \cdot z(t^{\#})$ – an assumption that, of course, needs verification after the solution is obtained. It means that X in fact becomes a function of φ. Setting the time-derivative of (A.1) equal to zero yields the partial differential equation:

$$\dot{p}^{LR}(t^{\#}) = X \cdot \chi \cdot z(t^{\#})^{\chi-1} \cdot \dot{z}(t^{\#}) + \begin{cases} -d \cdot \beta \\ d \cdot \gamma \end{cases} = 0 \tag{A.2}$$

The following solution for $z(t^{\#})$ can be found:

$$z(t^{\#}) = \left(\frac{1}{X} \cdot \begin{cases} d \cdot \beta \cdot \left(t^{\#} - t_l \right) \\ d \cdot \gamma \cdot \left(t_{l'} - t^{\#} \right) \end{cases} \right)^{\frac{1}{\chi}} \tag{A.3}$$

The starting and ending moments of the distribution can be found by setting the integral of (A.3) between $t^{\#}=t_l$ and $t^{\#}=0$ equal to $\gamma/(\beta+\gamma)\cdot N$; and similarly, the integral of (A.3) between $t^{\#}=0$ and $t^{\#}=t_{l'}$ equal to $\beta/(\beta+\gamma)\cdot N$. This yields:

$$t_l = -\left(N \cdot \frac{\gamma}{\beta+\gamma} \cdot \frac{1+\chi}{\chi} \cdot \left(\frac{X}{d \cdot \beta} \right)^{\frac{1}{\chi}} \right)^{\frac{\chi}{1+\chi}}; \quad t_{l'} = \left(N \cdot \frac{\beta}{\beta+\gamma} \cdot \frac{1+\chi}{\chi} \cdot \left(\frac{X}{d \cdot \gamma} \right)^{\frac{1}{\chi}} \right)^{\frac{\chi}{1+\chi}} \tag{A.4}$$

It is easily checked that $-g\cdot\beta\cdot t_l = g\cdot\gamma\cdot t_{l'}$ applies in addition to $-\beta\cdot t_q = \gamma\cdot t_{q'}$, so that the assumed equilibrium shares of early and late SRPATS is indeed an equilibrium: the very first and very last driver have equal short-run schedule delay costs in both states, as well as equal long-run schedule delay costs. Because these individuals never face travel delays, they are equally well off, confirming equilibrium.

The critical conditions on parameters reported in Section 3 are subsequently found by comparing t_l and $t_{l'}$ in (A.4) with t_q and $t_{q'}$ for K_1 to determine the domains of Always Dispersed versus Sometimes Dispersed, and for K_0 to determine the domains of Sometimes Dispersed versus Never Dispersed.