

PRODUCT DIFFERENTIATION ON ROADS

Second-Best Congestion Pricing with Heterogeneity under Public and Private Ownership

Erik T. Verhoef^{1*} and Kenneth A. Small²

¹Department of Spatial Economics
Free University Amsterdam
De Boelelaan 1105
1081 HV Amsterdam
The Netherlands
Phone: +31-20-4446094
Fax: +31-20-4446004
E-mail: everhoef@econ.vu.nl

²Department of Economics
University of Irvine at California
Social Science Plaza
Irvine CA 92697-5100
USA
Phone: +1-949-824-5658
Fax: +1-949-824-2182
E-mail: ksmall@uci.edu

This version: 12/08/99

Key words: congestion, road pricing, networks, second-best

JEL codes: R41, R48, D62

Abstract

We explore the properties of various types of public and private pricing on a congested road network with heterogeneous users and allowing for elastic demand. Heterogeneity is represented by a continuum of values of time. The network consists of both serial and parallel links, which allows us to model second-best pricing restrictions on either complementary or substitute links, while still accounting for interaction between different groups on shared links (e.g. in city centres). We find that private (revenue-maximizing) pricing is much less efficient than public pricing, whether on the partial or the full network; but this difference is mitigated by the product differentiation made possible by heterogeneous users. Ignoring heterogeneity causes the welfare benefits of second-best pricing of one parallel link, a policy currently receiving favourable consideration, to be dramatically underestimated. Product differentiation produces some unexpected distributional effects, including the possibility that first-best pricing can result in one of the parallel routes being both more congested than without pricing.

*Erik Verhoef is affiliated as a research fellow to the Tinbergen Institute.

Erik Verhoef would like to thank UCI in general, and Ken Small, Linda Cohen and Iris Adam in particular, for their hospitality during his stay in the spring of 1999, when the first draft of this paper was written.

1 Introduction

Public-finance economists have long advocated Pigovian taxes and related ‘market-like’ policies to attain better pricing of goods supplied by the public sector. Most such policies are enacted on a piecemeal and limited basis, if at all. Cases in point are the marketable permits established by the US Clean Air Act of 1990, which apply only to sulphur oxides emitted by large sources, and several heavily restricted pollution trading schemes reviewed by Hahn (1989).

One of the best-studied applications of Pigovian taxes is road pricing. The economic fundamentals were well laid out by Pigou (1920), Knight (1924), Walters (1961), and Vickrey (1963, 1969). The concept is favoured by many transportation policy makers, but mainly in the form of experiments or demonstrations rather than full-scale applications (Small and Gómez-Ibáñez, 1998). Examples include toll rings in Norway, peak-period toll surcharges on certain French expressways, special tolled express lanes on two freeway segments in southern California, and a fully congestion-priced new expressway near Toronto.

This history suggests an increasing importance of partial congestion-pricing schemes, for which the applicable theory is that of the second best. Such schemes include privately or publicly operated toll roads parallel to unpriced highways, and a variant of high-occupancy vehicle lanes in which low-occupancy vehicles can use the premium lane by paying a toll.¹ Depending on the particular scheme, pricing may be prohibited on either substitute or complementary routes to the one that is priced, and may involve either social or private objectives. Thus the analysis requires a model with a variety of possible objectives and pricing constraints.

Furthermore, because much of the purpose is to test and shape public opinion, it is crucial to focus on distributional issues. Doing so turns out to have an additional advantage because one of the features of some demonstrations is to offer highly differentiated products, *e.g.* a free highly congested road or a very expensive free-flowing road. Not only does this turn out to substantially affect the choice of toll or even the preferred toll regime, but it is itself a potent political factor in the debate over such policies, exemplified by the term “Lexus Lane” applied by opponents of optional high-priced lanes with premium service.

In this paper, we directly and simultaneously address issues of second-best, public or private objectives, product differentiation, and distribution as they arise from limited road-pricing demonstrations. We use the results to consider how a project’s design affects its economic efficiency and political viability. Specifically, we design a model with the following features: (1) pricing of some links but not others; (2) heterogeneous population with respect to willingness to pay for travel-time savings; (3) product differentiation arising either from inherent traits (road length) or from endogenously-determined service quality (amount of congestion); and (4) a comparison between second-best and revenue-maximizing policies. The

¹ Both of the California projects are of this latter type, known as High Occupancy/Toll (HOT) Lanes. One (Interstate 15 near San Diego) is publicly operated while the other (State Route 91 in Orange County, near Los Angeles) is private.

model allows for elastic demand, allowing us to examine the two roles of second-best prices, as Ramsey-like prices targeted at less elastic market segments and as route-allocation devices aimed at minimizing user costs for a given demand. The numerical version of the model uses (for its base case) an empirically obtained distribution of values of time for morning peak road use, based on a recent questionnaire among morning peak road users in the Dutch Randstad area (Verhoef, Nijkamp and Rietveld, 1997). We analyse both substitute and complementary services to the one being priced by using a simple network that has both parallel and serial links and examining the results of various pricing restrictions on them. Such a set-up can represent, for example, parallel priced and unpriced arterials entering a city centre where users of both interact on congested streets. The existence of a complementary (serial) link turns out to add some interesting features to the policy evaluations.

A preview of especially interesting results: We find that ignoring heterogeneity in values of time may cause the welfare benefits of second-best policies to be dramatically underestimated, by a factor of nine in our base case. Private pricing is almost always worse than no pricing, except when a private route has significant free-flow speed advantages over the free parallel route. Dispersion makes first-best pricing quite dramatically anti-egalitarian, in that it may actually worsen the travel times faced by low-value-of-time users – a paradox explained by its effect of channelling these users onto just a portion of the total capacity but then applying a low price to them. Second-best pricing is much more egalitarian; however, welfare is greatly enhanced if instead of pricing just a small portion of the network, most capacity is priced with a small portion reserved as a free option. (Thus the special lanes should be free “Lemon Lanes” rather than high-priced “Lexus Lanes”). Finally, offering a differentiated product can produce the intriguing possibility that a second-best pricing policy may provide benefits to those who care least and to those who care most about service quality, while hurting those in the middle – probably not an ideal set-up for political success.

Such results pose challenges for the demonstration-project approach. There is a real danger that most of the hoped-for welfare benefits from pricing will be lost, or even turned into disbenefits; or that specific groups will incur perverse results such as both higher price and worse service. On the other hand, dispersion in preferences does offer the potential to reap substantial benefits through product differentiation, which lends itself to an experimental approach. Our model provides a flexible and realistic tool to study these advantages and disadvantages.

It is worth noting at this point that the correlation between value of time and income is actually far from perfect. One of the most striking findings from actual experience on SR-91 is that most people who use the express lanes do so intermittently, and the mix of incomes using the express and the free lanes on any given day overlaps considerably (Sullivan, 1998; Parkany, 1999). Thus, we caution against the tempting simplification (embodied in the “Lexus lane” epithet) that the value-of-time distribution represents the income distribution.

2 The analytical model

2.1 Prior literature

Most of the second-best literature addresses two parallel routes where one of the two routes is untolled. Lévy-Lambert (1968), Marchand (1968), and Verhoef, Nijkamp and Rietveld (1996) use the static model of Walters (1961) and Vickrey (1963), while Braid (1996) uses the dynamic bottleneck model of Vickrey (1969). The main conclusions are that the second-best toll trades off route split effects against overall demand effects; that this toll is usually considerably smaller than the first-best toll and may even be negative; and that second-best pricing often leads to much smaller welfare gains than first-best pricing. Liu and McDonald (1998) confirm these results for parameters designed to match one of the California pricing demonstration projects (SR-91 in Orange County). Yang and Huang (1999), finally, focus explicitly on HOT lanes, endogenizing the choice of being a car-pooler.

Revenue-maximizing congestion tolls for a single highway are derived by Edelson (1971). For two parallel roads, Verhoef *et al.* (1996) and Liu and McDonald (1998) find that a private operator will typically set a price much higher than the second-best toll and will achieve very much lower, usually negative, welfare gains. (Of course, this does not address the question of whether the new capacity could have been financed without the private enterprise.) A recent and probably the most complete analysis of private tolling on a parallel route network is given by De Palma and Lindsey (1999), who consider a variety of ownership regimes, including private and mixed duopolies.

As noted, however, these studies do not account for any product-differentiation advantages because they treat homogenous users. Very few studies of the two-route problem incorporate heterogeneity. Arnott, De Palma and Lindsey (1992) consider two user groups and two routes within the bottleneck model, but do not consider the case when one route cannot be priced. Small and Yan (1999) do consider such a case, also with just two discrete user groups. Less closely related is the analysis by Train, McFadden and Goett (1987) and by Train, Ben-Akiva and Atherton (1989) of electricity and telephone users, respectively, facing a voluntary choice among alternate rate schedules with different time-of-day characteristics.

Models that treat two discrete user groups, besides providing only a crude approximation to real heterogeneity, result in analytical difficulties because several types of equilibria can arise: pooled (where both groups use both routes), partially separated (where one group uses both routes but the other group uses only one route), and fully separated (where each group exclusively uses one route). In the present paper, we consider a continuum of groups. Only two types of equilibria can then occur: pooled (when tolls are absent or exactly equal on the two parallel routes), or fully separated (in all other cases). Moreover, using a continuum of values of time allows intermediate groups to be considered explicitly.

2.2 Basic set-up and equilibrium conditions

The network considered is shown in Figure 1. There is just one origin-destination pair, OD, connected by two routes: AC (consisting of links A and C) and BC (consisting of links B and C). The travel time on link L is a non-decreasing function of the level of use: $T_L(N_L)$ with

$T_L' \geq 0$ (primes are used to denote derivatives). Link L may have a toll, τ_L . Because there are three links but only two routes, there is one redundant toll: a constant can be subtracted from τ_A and τ_B , and added to τ_C , without affecting the price of either route. For convenience, we normalize τ_C to zero except when we wish to require the prices of the two routes to be equal, in which case we normalize $\tau_A = \tau_B = 0$ and allow τ_C to represent the single uniform price.

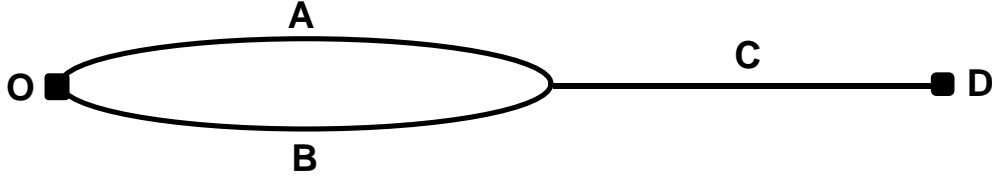


Figure 1. The network considered

We consider a continuum of values of time, α . For a traveller with value α , the travel cost on a link is $\alpha \cdot T_L$. The density function of users, by value of time, is denoted N_α ; that is, there are N_α users within an infinitesimally small range $[\alpha, \alpha + d\alpha]$. However, this density is endogenous due to the fact that for each value of time, there is a downward-sloping inverse demand function $D_\alpha(N_\alpha)$. Figure 2 shows an example of the resulting demand plane, depicting the one used in the numerical model in Sections 3 and 4.

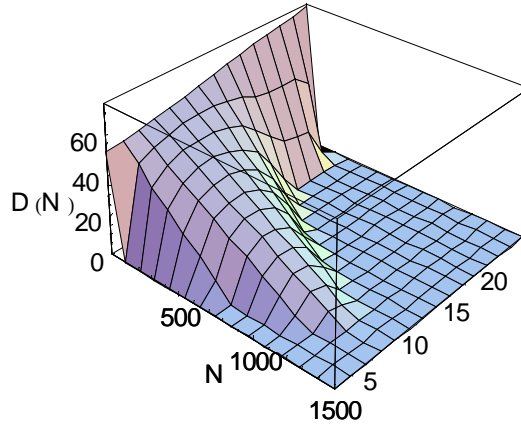


Figure 2. An inverse demand plane

Let $N_{\alpha L}$ and $N_{\alpha R}$ be the corresponding densities on link L and route R. The conditions of user equilibrium are the complementary slackness conditions of Wardrop (1952):

$$N_{\alpha R} \cdot (P_{\alpha R} - D_\alpha) = 0 \quad (1a)$$

$$N_{\alpha R} \geq 0 \quad (1b)$$

$$P_{\alpha R} - D_\alpha \geq 0 \quad (1c)$$

for every α and for both routes R, where $P_{\alpha R}$ is the ‘full price’ of using route R, defined as:

$$P_{\alpha R} \equiv \mathbf{a} \cdot (T_L + T_C) + \mathbf{t}_L + \mathbf{t}_C, \quad \{L, R\} = \{A, AC\}, \{B, BC\} \quad (1d)$$

These equations state that type- α users will use only the route(s) that have least cost to them, or will not travel at all if all routes have costs exceeding willingness to pay.

Formally, we must proceed differently in solving equations (1) depending on whether or not $\tau_A = \tau_B$. If $\tau_A = \tau_B$, positive use can occur on both roads only if travel times are equal, since otherwise all users would choose the road with the lower travel time. In that case, we need an additional condition to obtain a unique equilibrium. The one we choose is that $N_{\alpha A}/N_{\alpha B} = N_A/N_B$ for every α , which may be viewed as a mixed strategy Nash equilibrium.² This is a perfectly pooled equilibrium, which we can analyse by merging links A and B into a single link, D, whose travel time is simply a function $T_D(N)$ of total traffic N .³

When $\tau_A \neq \tau_B$, both routes can be used only when $T_A \neq T_B$, and there is a separated equilibrium. More precisely, for both routes to be used, $\text{sign}\{\tau_B - \tau_A\} = \text{sign}\{T_A - T_B\}$ is required. Since the difference in full prices on both routes can be written as $(\tau_A + \alpha \cdot T_A) - (\tau_B + \alpha \cdot T_B)$ for each α , the critical value of time α^* , for which drivers are indifferent between both routes, is:

$$\alpha^* = \frac{t_B - t_A}{T_A - T_B} \quad (2)$$

It is easily checked that, when $\tau_A < \tau_B$, link A is more attractive for all drivers with $\alpha < \alpha^*$, and link B for all drivers with $\alpha > \alpha^*$. Drivers with a relatively low value of time thus only use the link with the lower toll, and similarly for a relatively high value of time and the higher toll.

To complete the model, the following identities are added:

$$N_{aC} = N_{aA} + N_{aB} \quad (3)$$

$$N_L = \int_{\alpha_{\min}}^{\alpha_{\max}} N_{aL} \, d\alpha \quad (4)$$

where α_{\min} (α_{\max}) gives the minimum (maximum) value of time in the population. In the case where $\tau_A < \tau_B$, (4) implies:

$$N_A = \int_{\alpha_{\min}}^{\alpha^*} N_{aA} \, d\alpha \quad (4a)$$

² The reason that an extra condition is needed is that, although we know that all drivers with all relevant values of time *could* be present on both routes, they do not have to be. For example, with identical routes A and B, one of the possible equilibria satisfying (1) would be if all drivers with a value of time smaller than or equal to the median α were on the one route, and the rest on the other route. By contrast, in the symmetric Nash equilibrium every user plays the same strategy, which is the best strategy given that all other users play that same strategy. That strategy is to play a mixed strategy, choosing link A with a probability $p_A = N_A/(N_A + N_B)$. N_A and N_B are then determined from the condition that $T_A(N_A) = T_B(N_B)$. With a sufficiently large number of users, the equilibrium then indeed comes about, with all users playing the same strategy and having no strategy that would improve their pay-off. In particular, note that a driver applying a different value of p_A will have a marginally higher expected travel time if all other users stick to p_A .

³ This function is chosen to be consistent with an allocation $N = N_A + N_B$ such that $T_A(N_A) = T_B(N_B) = T_D(N)$ is satisfied. It has the property that:

$$\frac{1}{T_D'} = \frac{1}{T_A'} + \frac{1}{T_B'}$$

Note that, when one of the two links, say link A, is unused in the equilibrium because the free-flow travel time on A exceeds the equilibrium travel time on B, we would have $T_D(N) = T_B(N)$ and $T_D' = T_B'$.

$$N_B = \int_{a^*}^{a_{\max}} N_{aB} da \quad (4b)$$

2.3 Tolling regimes

Despite its simplicity, the network considered allows us to consider a large variety of toll regimes. Ignoring the possibility of private or mixed duopolies, as considered in De Palma and Lindsey (1999), we still have the seven possibilities of no tolls, private or public tolls on the entire network, on one of the parallel links (labelled B without loss of generality), or on the serial link. Table 1 summarizes these tolling regimes.

Abbreviation	Description	Tolls on:
NT	No Tolls	–
<i>Public tolling:</i>		
FB	<u>F</u> irst- <u>B</u> est tolls on the full network	A and B
SBPL	<u>S</u> econd- <u>B</u> est toll on one of the <u>P</u> arallel <u>L</u> inks	B
SBSL	<u>S</u> econd- <u>B</u> est toll on the <u>S</u> erial <u>L</u> ink	C
<i>Private tolling</i>		
PF	<u>P</u> rivate tolls on the <u>F</u> ull network	A and B
PPL	<u>P</u> rivate toll on one of the <u>P</u> arallel <u>L</u> inks	B
PSL	<u>P</u> rivate toll on the <u>S</u> erial <u>L</u> ink	C

Table 1. Tolling regimes

For the public operator, we assume that the objective is to maximize social welfare W . It is defined as total benefits based on the Marshallian consumers' surplus (in this case represented by the volume of the three-dimensional body below the demand plane over the relevant area) minus total cost. In this section, we label without loss of generality the link with the higher toll link B whenever $\tau_A \neq \tau_B$. When $\tau_A = \tau_B$, welfare can be written as:

$$W = \int_{a_{\min}}^{a_{\max}} \int_0^{N_a} D_a(n) dn da - \int_{a_{\min}}^{a^*} N_{aA} \cdot a \cdot T_A \left(\int_{a_{\min}}^{a^*} N_{aA} da \right) da - \int_{a^*}^{a_{\max}} N_{aB} \cdot a \cdot T_B \left(\int_{a^*}^{a_{\max}} N_{aB} da \right) da - \int_{a_{\min}}^{a_{\max}} N_a \cdot a \cdot T_C \left(\int_{a_{\min}}^{a_{\max}} N_a da \right) da \quad (5a)$$

where α^* is given by (2), and where we have used that $N_{\alpha B} = 0$ for $\alpha < \alpha^*$ and $N_{\alpha A} = 0$ for $\alpha > \alpha^*$.

When $\tau_A = \tau_B = 0$, so when only τ_C is used and the merged link D is considered, the objective is:

$$W = \int_{a_{\min}}^{a_{\max}} \int_0^{N_a} D_a(n) dn da - \int_{a_{\min}}^{a_{\max}} N_a \cdot a \cdot T_C \left(\int_{a_{\min}}^{a_{\max}} N_a da \right) da - \int_{a_{\min}}^{a_{\max}} N_a \cdot a \cdot T_D \left(\int_{a_{\min}}^{a_{\max}} N_a da \right) da \quad (5b)$$

With private tolling, the objective is assumed to be the maximization of total toll revenues, R . Using dummies δ_L to denote whether a toll is in operation on link L, this objective function can be written as:

$$R = \mathbf{d}_A \cdot \mathbf{t}_A \cdot \int_{\mathbf{a}_{\min}}^{\mathbf{a}^*} N_{aA} \, d\mathbf{a} + \mathbf{d}_B \cdot \mathbf{t}_B \cdot \int_{\mathbf{a}^*}^{\mathbf{a}_{\max}} N_{aB} \, d\mathbf{a} + \mathbf{d}_C \cdot \mathbf{t}_C \cdot \int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} N_a \, d\mathbf{a} \quad (6)$$

These objective functions can be turned into the relevant Lagrangians by adding terms representing the appropriate constraints, of which there is a continuum for all values of α . These constraints represent the requirement of free user choice as embodied in equation (1), and they are identical for the public and the private operator. When $\tau_A \neq \tau_B$, the constraints are represented by the following Lagrangian terms:

$$\begin{aligned} & + \int_{\mathbf{a}_{\min}}^{\mathbf{a}^*} \mathbf{I}_{aA} \cdot \left(\mathbf{a} \cdot T_A \left(\int_{\mathbf{a}_{\min}}^{\mathbf{a}^*} N_{aA} \, d\mathbf{a} \right) \right) d\mathbf{a} + \mathbf{a} \cdot T_C \left(\int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} N_a \, d\mathbf{a} \right) d\mathbf{a} + \mathbf{d}_A \cdot \mathbf{t}_A - D_a(N_a) \Big) d\mathbf{a} \\ & + \int_{\mathbf{a}^*}^{\mathbf{a}_{\max}} \mathbf{I}_{aB} \cdot \left(\mathbf{a} \cdot T_B \left(\int_{\mathbf{a}^*}^{\mathbf{a}_{\max}} N_{aB} \, d\mathbf{a} \right) \right) d\mathbf{a} + \mathbf{a} \cdot T_C \left(\int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} N_a \, d\mathbf{a} \right) d\mathbf{a} + \mathbf{d}_A \cdot \mathbf{t}_B - D_a(N_a) \Big) d\mathbf{a} \end{aligned} \quad (7a)$$

where $\lambda_{\alpha L}$ is the Lagrangian multiplier for the constraint implied by (1a) for those values of α having positive N_α . When $\tau_A = \tau_B = 0$, the Lagrangian term is:

$$+ \int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} \mathbf{I}_a \cdot \left(\mathbf{a} \cdot T_C \left(\int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} N_a \, d\mathbf{a} \right) \right) d\mathbf{a} + \mathbf{a} \cdot T_D \left(\int_{\mathbf{a}_{\min}}^{\mathbf{a}_{\max}} N_a \, d\mathbf{a} \right) d\mathbf{a} + \mathbf{t}_C - D_a(N_a) \Big) d\mathbf{a} \quad (7b)$$

For each of the schemes considered, the optimal tax rules can then be found by maximizing the appropriate Lagrangian, found by adding the relevant objective function (5a), (5b) or (6) to the relevant constraints (7a) or (7b). Unfortunately, we failed to find a closed form analytical solution for three of the more interesting cases, namely SBPL, PF and PPL, due to a discontinuity at α^* . For the cases where we did find closed form analytical solutions (FB, SBSL, PSL), the tax rules are rather straightforward generalizations of those applying in the much simpler case with only one single value of time, as given, for instance, in Verhoef *et al.* (1996). We therefore relegate the derivation and discussion of the first-order conditions, and of the tax-rules for the cases where they could be solved, to the Appendix. For the other cases, we devised a numerical algorithm which finds the toll that maximizes welfare (for SBPL) or revenue (for PF and PPL).

3 A numerical model: the base case

In this section we present a numerical model to assess and illustrate the economic properties of these regimes.

3.1 The cost side

The cost side of the model consists of link travel-time functions, describing travel times T_L as a function of usage N_L . The following general functional form is used:

$$T_L = T_{FL} + T_{FL} \cdot b \cdot \left(\frac{N_L}{K_L} \right)^k \quad (8)$$

where b and k are parameters, T_{FL} is the free-flow travel time on link L , and K_L is the ‘capacity’ of link L . Because there is in fact no maximum capacity for this type of travel time

function, ‘relative capacity’ would actually be a better term. With $b=0.15$ and $k=4$, as assumed throughout the simulations, this function is the well-known Bureau of Public Roads formula. For the base-case of the model, it is assumed that link B (the tolled link in case of parallel-route pricing) has 25%, and link A 75%, of their joint capacity (8 000 vehicles per hour). A and B have equal free-flow travel times (22.5 minutes). This could correspond to a four-lane highway with one lane subject to tolling. Furthermore, it is assumed that link C has the same capacity as A and B together, and that the free-flow travel time on C is 25% (*i.e.* 7.5 minutes) of the total free-flow travel time. The latter is set at 30 minutes, which is a reasonable average for commuting trips in The Randstad⁵ and many other urban areas, at least for commuters using main highways. Table 2 gives the base-case parameters for the three links:

	Link A	Link B	Link C
b	0.15	0.15	0.15
k	4	4	4
T_{FL}	0.375	0.375	0.125
K_L	6000	2000	8000

Table 2. The base-case parameters for the cost functions

3.2 The demand side

The base-case inverse demand plane is depicted in Figure 2 above. It is assumed that for every value of time, the demand function is linear over the relevant range (between the lowest and highest use levels considered), and can thus be written as:

$$D_a = m_a - d_a \cdot N_a \quad (9)$$

Functions m_α and d_α were calibrated so as to achieve three objectives: (1) a weighted demand elasticity (over all α) of -0.4 in the NT-equilibrium⁴; (2) a distribution of values of time in the NT-equilibrium similar to that found in an earlier stated preference study for the Dutch Randstad area (Verhoef *et al.*, 1997)⁵; and (3) NT travel times approximately double the free-flow travel time of 0.5 hours implied by Table 2. The following functions accomplished this:

$$m_a = 50 + a \quad (10a)$$

⁴ We calculated this elasticity assuming that the variable monetary cost of the trip is DfI 12 (6 litres of fuel times DfI 2) in the NT-equilibrium; *i.e.* both the demand plane and the cost level were shifted upwards by DfI 12 to calculate this generalized price elasticity. These variable monetary costs, however, are assumed to be constant over the various tolling regimes considered, and so are ignored in the simulations. Note that because the full price P_α increases with the value of time, so would the absolute value of demand elasticity *ceteris paribus* (*i.e.*, with equally steep demand functions).

⁵ The dashed line in the left panel in Figure 3 below shows this distribution (approximately). It was derived using 961 (93%) of the 1027 respondents for whom a value of time could be calculated, eliminating the 7% with the highest values of time so as to keep a compact distribution. A simple fourth-order polynomial was fitted on the histogram of values of time, split in 12 categories of size DfI 2 ($R^2=0.975$). Because of the selection, the average value of time used here is DfI 9.08, as opposed to DfI 10.92 for the full set of respondents. The distribution is clearly skewed to the left, which is what one would intuitively expect when values of time are non-negative.

$$d_a = \frac{-0.0434783}{-0.713714 + 0.705429 \cdot a - 0.0950357 \cdot a^2 + 0.00468093 \cdot a^3 - 0.000079 \cdot a^4} \quad (10b)$$

The values of time considered in the simulations range between a minimum of DFl 1.2 and a maximum of DFl 23.8 per hour.⁶

Although we do not claim to present a full-fledged empirical analysis here, we believe these parameters result in a base case that is representative of realistic morning peak traffic situations, and therefore allows a meaningful comparison of the comparative static properties of the equilibria arising under the various tolling regimes considered. In the next sub-section, such a comparison will be given.

3.3 General results: base case

Table 3 presents the equilibrium values of important variables for the various tolling regimes, using the base-case parameters discussed above.

The first-best (FB) policy produces substantial service differentiation, with travel time on link A 0.204 hours less than on link B. But this policy also produces some surprises. First, for these particular parameters it is the facility with the larger capacity (link A) that gets the premium service, in contrast to what one might expect from the analogy of first-class service on aeroplanes and trains. Second, although overall demand is reduced (by 14 percent) compared to the no-toll (NT) scenario, congestion on the lower-priced link is actually worse than with no tolls. This finding highlights the effect of differentiated pricing on optimal service levels, through the separation of users it induces. We note that when the portion of the trip on link C is taken into account, all users receive faster service in the first-best policy than in the no-toll policy – average time losses, compared to free flow travel times, reduce by 10 and 54 percent for users of the slower and faster parallel links, respectively. However, Section 4.2 presents an example where also the *total travel times* for low-value-of-time users actually *increase* in the first-best optimum.

A third surprise is how small the toll differentiation is: the tolls on links A and B differ from each other by only 15 percent. There are three reasons for this. First, although the average value of time of link-B users is smaller, there are more of them (per unit of capacity), and these two effects work in opposite directions on the externality cost of a trip. Second, link-B users interact with higher-value-of-time users on the shared link C, which further increases the marginal cost they impose. Finally, the travel-time functions are rather steep in the region of the FB solution. So even a modest difference in usage rates produces a substantial difference in service quality.

Given the limited degree of optimal toll differentiation, it is not too surprising that the uniform toll policy, SBSL, performs nearly as well in terms of efficiency. It achieves 92 percent of the maximum possible welfare gains, at a uniform toll quite close to the higher of the differentiated FB tolls. Although not shown in the table, one can readily see that most or all low-value-of-time users are worse off with a uniform toll policy than with FB because the

⁶ The exchange rate of the Dutch guilder is approximately DFl 2.2 ≈ 1 ≈ \$1.1.

former forces them to accept a higher service quality and higher price than they prefer. (We discuss the distributional effects at greater length in the next subsection.)

	NT	FB	SBPL	SBSL	PF	PPL	PSL	Free-flow
Rel. use A^a	1	0.812	1.046	0.854	0.498	1.117	0.527	
Rel. use B^a	1	1.003	0.831	0.854	0.616	0.533	0.527	
Rel. use C^a	1	0.860	0.992	0.854	0.527	0.971	0.527	
α^*	-	5.919	12.996	-	6.138	15.265	-	
Travel time A	0.729	0.529	0.798	0.563	0.397	0.926	0.402	0.375
Travel time B	0.729	0.733	0.544	0.563	0.426	0.404	0.402	0.375
Travel time C	0.243	0.189	0.239	0.188	0.134	0.230	0.134	0.125
Toll A	0	9.50	0	0	27.83	0	0	
Toll B	0	8.29	3.31	0	27.65	7.98	0	
Toll C	0	0	0	9.38	0	0	27.80	
Toll revenues	0	99 606	8703	101 484	185 603	13 468	185 487	
ω^b	0/0	1	0.229	0.920	-2.599	-0.272	-2.623	

^a Use relative to that in NT scenario. The latter is: 9501 on link A, 3167 on link B, 12669 on link C. As discussed in the text, the fact that these exceed link ‘capacity’ is entirely consistent with the power-law model of equation (8). The NT-use levels are probably best thought of as covering a peak period of about 1.5 hours.

^b Index of relative efficiency: increase in social welfare (compared to NT) as a share of the increase in social welfare (compared to NT) obtained in the first-best optimum

Table 3. Performance of the various toll regimes for the base-case parameters

By contrast, when only one of two parallel links can be priced (SBPL), with 25% of total capacity, less than one-fourth of the possible welfare gains are achieved. Consistent with the studies reviewed earlier, the second-best toll is much lower than first-best, only Df1 3.31. The reason is that now, welfare gains on link B have to be traded off against induced welfare losses on link A, as described in Verhoef *et al.* (1996). Nevertheless there is a surprise for second-best policy as well: as we shall see in Section 4.1, more than twice as great a welfare gain could be achieved with second-best parallel pricing by pricing the high-capacity section of the road instead of the low-capacity section. This result is consistent with the fact that with first-best pricing, it was the higher-capacity road that received the higher price.

We now turn to revenue-maximizing tolling by a private operator. Unrestricted tolling extracts a high social cost: welfare is substantially reduced, compared to no tolls. The loss is 27 percent of the maximum achievable gain when only link B can be priced (PPL), and a whopping 260 percent when both links can be priced (PF). The tolls are much higher than the corresponding second-best or first-best optimal tolls: more than twice as high for PPL as for SBPL, and around three times as high for PF as for FB. This is consistent with earlier results, although it is not necessarily the case that revenue-maximizing tolling would always lead to a decrease in social welfare (see Verhoef *et al.*, 1996, and De Palma and Lindsey, 1999). From the point of view of the private operator, being restricted to pricing only the smaller part of the roadway is very costly: it takes away 93 percent of the potential revenues.

There is a surprise in private tolling, as well: the toll differentiation in unrestricted private pricing is negligible. The reason is that the monopoly toll level has reduced total traffic

by so much (47 percent) that nearly all congestion is eliminated, rendering significant toll and service differentiation impossible. A corollary to this result is the negligible difference between policies PF and PSL (which differ only in allowing or not allowing the operator to charge differentially).

3.4 Welfare results: base case

The numerical simulations allow a detailed analysis of the welfare effects of the various tolling regimes, differentiated by value of time. In this sub-section, we present such results for two public tolling regimes: FB and SBPL. We ignore SBSL because of its close similarity to FB. Private tolling regimes are ignored mainly for reasons of space, further motivated by the fact that the results presented below will give quite some insight already into the corresponding private tolling regimes.

Figure 3 gives the absolute (left panel) and relative (right panel) levels of road use by value of time for the two policies. Since the usage under SBPL is close to the NT use levels – graphically indistinguishable in the left panel – the dashed line in the left panel also gives an accurate impression of the original distribution of values of time used. Relative use, in the right panel, is defined in the same way as in Table 3, *i.e.* relative to NT.

To understand the differences in use between the two tolling regimes, it is easier to start with Figure 4. This figure shows the changes in total (left panel) and average (right panel) consumers' surplus by value of time, compared to the surplus enjoyed in the NT-regime. For each value of time, the change in total consumers' surplus is given by the change in generalized costs (including the toll) for those users who remain on the road, plus the change in surplus for those who leave the road due to tolling (as described by the demand plane). Note that the welfare losses for the latter group are smaller than for the former. The average change is then defined as the change in the total surplus divided by the level of use in the NT-regime.

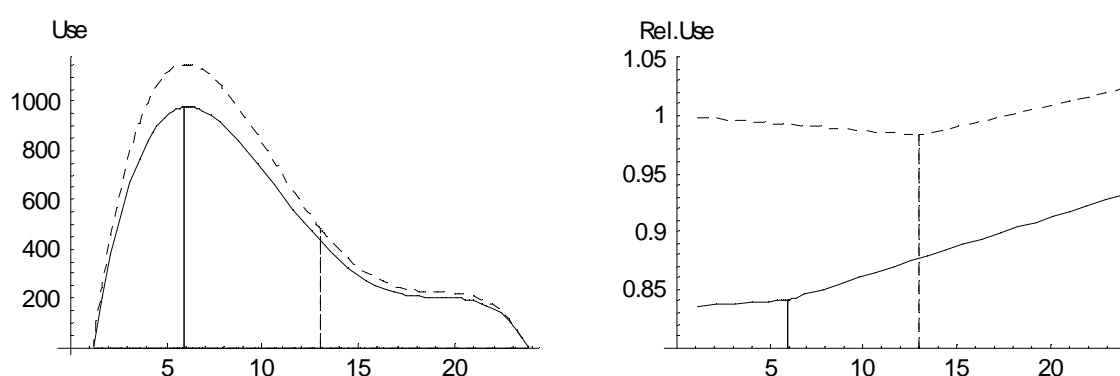


Figure 3. Absolute (left panel) and relative (right panel) use, under FB (solid) and SBPL (dashed) tolling; the vertical lines indicate \mathbf{a}^*

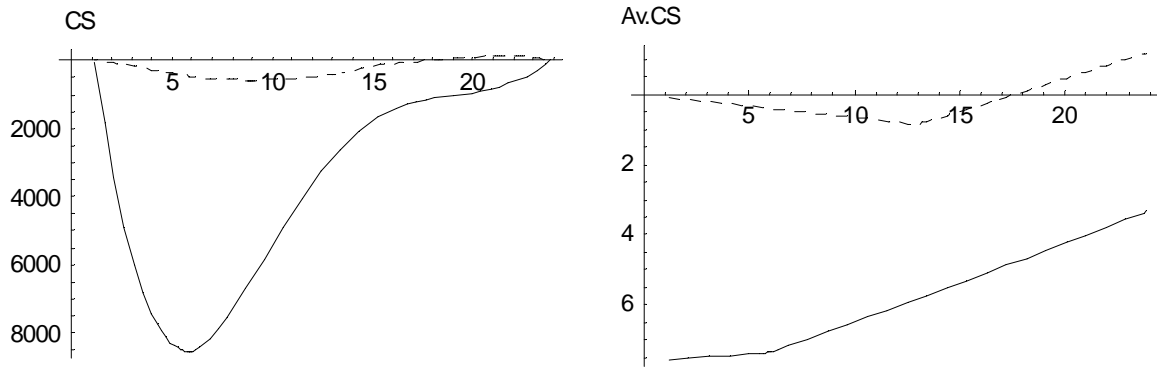


Figure 4. Absolute (left panel) and average (right panel) change in consumers' surplus, compared to NT, before tax recycling, under FB (solid) and SBPL (dashed) tolling

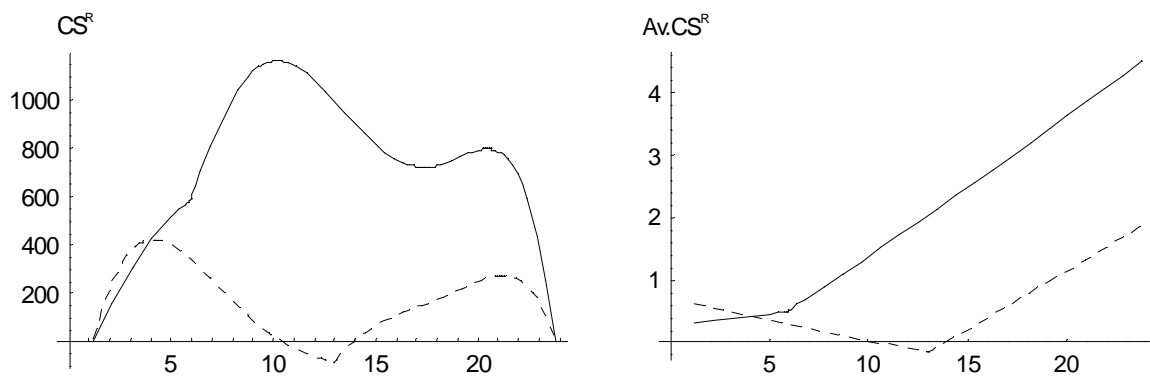


Figure 5. Absolute (left panel) and average (right panel) change in consumers' surplus, compared to NT, after non-differentiated tax recycling, under FB (solid) and SBPL (dashed) tolling

The changes in average consumers' surplus are closely related to the changes in generalized price. Under FB tolling, the loss in average surplus is strictly decreasing with the value of time. This result can be expected to be generally valid as long as the relative changes in use do not vary too greatly by value of time, so that the change in generalized costs dominates the distribution of average welfare effects. For each link individually, one would expect the average change in consumers' surplus to increase with the value of time; because the users with the critical value of time are indifferent between both routes, this result readily carries over to the full range of values of time. The kink at α^* can be explained by the fact that the ratio of toll paid and travel times gained is different between the two links. Figure 3 follows the same general pattern as Figure 4 simply because the change in consumers' surplus is closely related to the change in full price, which in turn determines the change in usage.

The pattern of changes in average consumers' surplus by values of time is notably different for the SBPL regime (dashed lines in Figures 3-5). In this case, users near the critical value of time suffer the largest average losses (Figure 4, right panel). This result is also likely to be valid quite generally; again as long as the change in generalized price is the dominating factor in the change of average consumers' surplus. The reason is that imposition of second-best tolling on link B substantially improves travel time for people taking that route while worsening it for those taking the other route. On the link without the toll, it is users near the

critical value of time who suffer most from this travel-time increase; while on the other link, it is users near the critical value who benefit least from the travel-time reduction. One could say that the policy caters to the more extreme users at both ends, leaving users in the middle disadvantaged. However, we note that none of the consumer-surplus changes are very large, the biggest loss amounting to just DfI 0.85 (US\$ 0.40) per trip. Again, the changes in relative use, depicted in Figure 3, are consistent with the changes in average consumers' surplus.

Note that the changes in consumers' surplus for SBPL are much smaller than under FB. Users with a relatively high value of time even benefit directly from SBPL. This helps explain why parallel-route pricing appears to be a socially more acceptable regime than first-best tolling. As expected, the relative attractiveness of both regimes is reversed as soon as redistribution of the toll revenues is considered. Figure 5 shows the change in total and average consumers' surplus by α after applying the simplest possible tax-recycling scheme: an equal redistribution to all initial road users. (This simply means an upward shift of the curves shown in the right panel of Figure 4.) Because revenue is much larger under first- than second-best pricing, the solid curve is shifted up by much more than the dashed curve, so that first-best pricing is now better than second-best pricing of link B for all but the very lowest-value-of-time users. In fact, with this toll redistribution, first-best pricing is welfare-enhancing for every user compared to no tolls. When these average surplus changes are multiplied by the level of usage shown in the left panel of Figure 3, the result is the curious double-peaked distribution of change in total consumer surplus shown on the left panel of Figure 5.

A simple variation on the redistribution scheme considered, favouring the lowest values of time relatively stronger, should make everybody significantly better off under FB than under SBPL. If the regulator fails to make this recycling known to and understood by the public at large beforehand, so that opinions are based on Figure 4 rather than Figure 5, FB will nevertheless evoke much stronger public opposition than SBPL will.

Figures 3-5 thus convey an interesting message on the distributional impacts of first-best tolling versus second-best tolling with an unpriced parallel substitute. The former would be regressive, unless a progressive tax-recycling scheme is used. Second-best tolling, however, would show a more complex distributional pattern. The greatest welfare losses now occur for those users with an intermediate (*i.e.* the critical) value of time, because even after (non-differentiated) tax recycling, these users remain worse off under SBPL.

It can be expected that the changes in average consumers' surplus by values of time under private tolling will show patterns comparable to those shown in the right panel of Figure 4, for the same reasons as outlined above. Of course, the absolute welfare losses will be larger; and since all the private tolling regimes generate net welfare losses, no redistribution (*e.g.* from auctioning franchise rights and using the proceeds to reduce taxes) could make everyone better off. In practice, private tolls are likely to be restricted by additional regulations, such as rate-of-return caps or direct price regulation. Our results provide support for some such restriction on private tolls, subject to the possibility that the revenues are needed to finance capacity as for example on California's SR91.

4 The impact of some key parameters on the performance of the tolling regimes

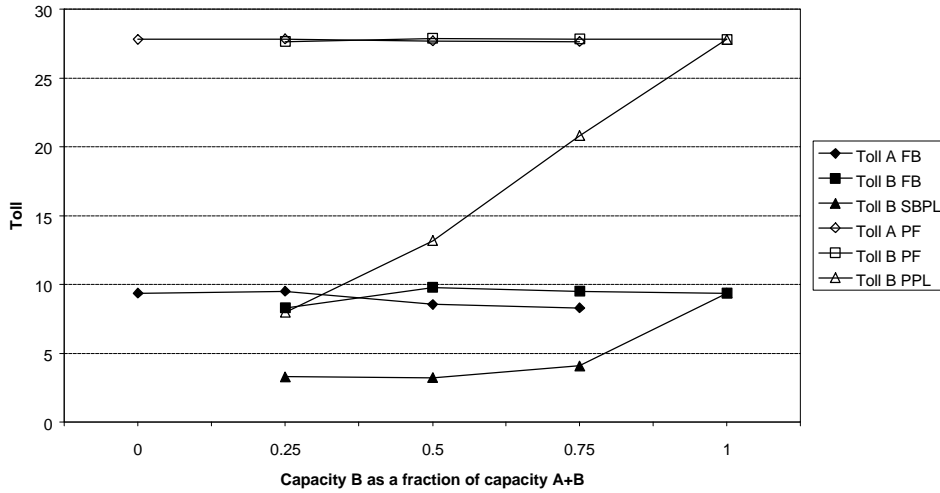
In this section, we present the results of a number of simulation exercises that assess the impact of key parameters upon the relative performance of the different tolling regimes. We start by varying parameters related to the cost side of the model, namely the capacities and lengths of the links. We next consider the impact of two characteristics of the demand side: the (weighted) demand elasticity, and the type of distribution of values of time.

4.1 Varying the relative capacities of the two parallel links

We first consider the impacts of increasing the fraction of the highway subject to tolling, keeping total joint capacity of A and B fixed. Figures 6a and 6b show the optimal tolls and the relative efficiency ω , respectively, if the capacity of B is increased, in 25% steps, from 0% to 100% of the joint capacity (recall that the base-case is at 25%). On the left-hand side of these figures, therefore, SBPL and PPL are identical to NT, because there is no parallel link to be tolled; whereas on the right-hand side, they are identical to SBSL and PSL, respectively, because there is only one route to be tolled. Furthermore, it should be no surprise that the absolute performance of SBSL and PSL remain completely unaffected by this variation. Finally, we observe that when link B has 0% or 100% of the total capacity, toll differentiation is not possible, rendering FB and PF identical to SBSL and PSL, respectively.

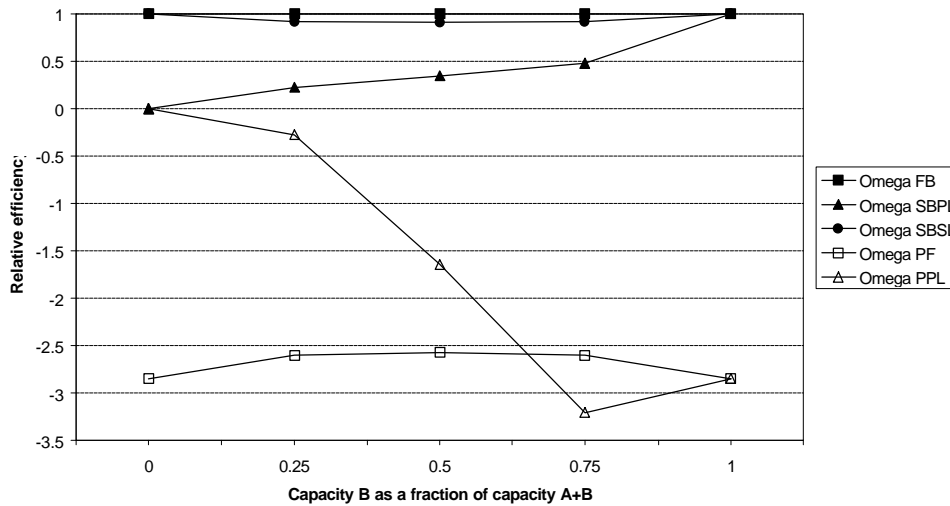
Unsurprisingly, the greatest impacts of capacity allocation occur for SBPL and PPL. The greater capacity of B implies that SBPL becomes relatively more efficient, because the importance of the unpriced substitute diminishes. Nevertheless, at 75% capacity, still only 48.5% of the possible welfare gains can be realized, because counter-productive spill-overs to the unpriced parallel route remain important. Still, these results suggest that from an efficiency viewpoint, and taking into account heterogeneity of users, one public ‘free-lane’ on a four-lane highway is preferable to one public ‘pay-lane’. In other words, it would be better to think of a priced system with a ‘life-line’ type of unpriced service available to those who most need it, rather than an unpriced system with special premium service for the elite.

The opposite holds for private tolling. The private operator, ignoring the efficiency aspects of spill-overs, increases the toll on the parallel route rapidly as its relative capacity increases, which dramatically increases the relative welfare losses from PPL. It is even more efficient to have the private operator controlling the entire road rather than just 75% of its capacity. This counterintuitive result, that in some cases it may be inefficient to restrict a monopolist’s market share when the alternative is not optimally priced, was found earlier in for instance Verhoef *et al.* (1996). The main intuition is that, with incomplete control, monopolist pricing creates a highly inefficient route split, which is avoided with full pricing. Finally, it can be observed that the relatively limited degree of price differentiation remains intact in Figures 6a and 6b, although the price differences reach their maxima at a fifty-fifty distribution of the relative capacities of A and B.



Note: For graphical clarity, tolls for SBSL and PSL, being close to those for FB and PF, are suppressed.

Figure 6a. Varying the relative capacities of the two parallel links: tolls



Note: For graphical clarity, relative efficiency for PSL, being close to that for PF, is suppressed.

Figure 6b. Varying the relative capacities of the two parallel links: relative efficiency

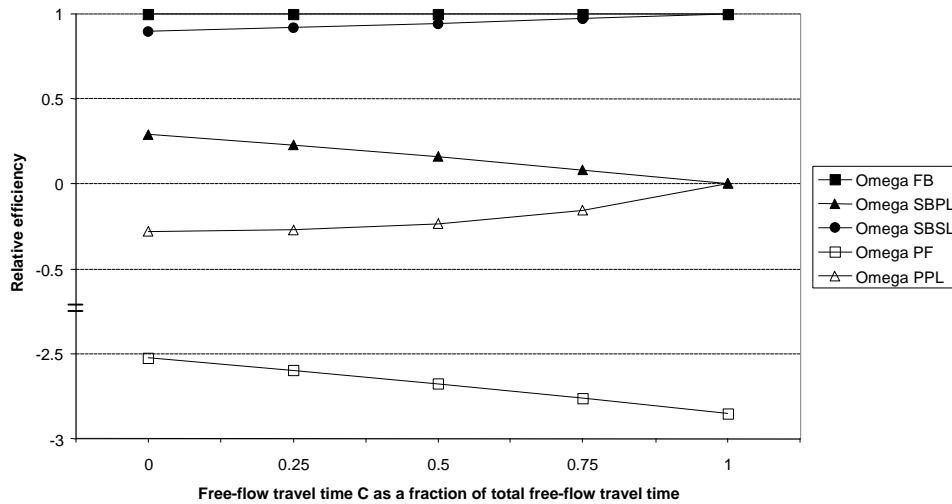
Together, these results constitute something of a reversal of the usual notions of the public sector. With product differentiation, it seems that if one insists on a system with both priced and unpriced alternatives, it is more efficient to allow a public operator to price most of the capacity, but a private operator to price only a small portion of it.⁷

4.2 Varying the relative length of the serial link

Most studies ignore the fact that users of two parallel routes will usually not be completed isolated but will often share some links, upstream or downstream of the split road section. A

⁷ The intersection of the lines representing τ_A and τ_B for FB in Figures 6a and 8a should not be misinterpreted as implying that at that specific combination of parameters, these two tolls should be equal. Rather, it is near the point where it makes no difference whether it is link A or link B that gets the higher toll, but still the tolls will be unequal. A similar argument holds for PF.

similar interaction is likely for other congestible facilities. Figure 7 shows how this feature affects the relative efficiency of the various tolling regimes considered. Along the horizontal axis, the relative length of the serial link C – represented by the free-flow travel time – is increased in 25% steps, keeping the total free-flow travel time constant. Note that FB and PF become identical to SBSL and PSL, respectively, when the relative length of C has become 1.



Note: For graphical clarity, relative efficiency for PSL, being close to that for PF, is suppressed.

Figure 7. Varying the relative length of the serial length: relative efficiency

As the relative length of the serial link increases, second-best toll differentiation becomes less viable. Both the public and the private toll on the parallel link (not shown diagrammatically) fall and approach zero as link C increases toward 1. As a result, the relative efficiency of these regimes approach zero as well. This is, from the efficiency perspective, bad news for the public toll and good news for the private toll.⁸ This finding suggests that the relative efficiency gains or losses from parallel route pricing are likely to be overstated in studies ignoring the existence of serial, common used links. For instance, ω_{SBPL} is equal to 0.29 when link C has zero length, but falls to 0.16 when C is equally long as A and B and to 0.08 when C is 3 times as long. A similar pattern would be found if instead of increasing the relative length of the serial link, its relative capacity were decreased.

The base-case result that FB tolling actually increases congestion (not shown in diagram) on link B, compared to no toll, remains true when link C has zero length. Therefore, product differentiation alone is strong enough to produce a result in which optimal pricing increases congestion for the entire trip by lower-value-of-time users, compared to no pricing. This can only mean that these users cause such large externalities when allowed to mix with other users (as in the NT case) that socially it becomes desirable to restrict the capacity available to them so much as to increase their travel times even while pricing them efficiently as a group. Of course, since FB pricing is a potential Pareto improvement, it remains true that these users

⁸ Of course, if link C were of zero length throughout, these tolls would also approach zero when the length of A and B were reduced to zero, but not as rapidly as in the present case; and the relative efficiencies of both SBPL and PPL would remain constant as long as the links would have a positive length.

could be made better off by some lump-sum redistribution of revenues. In practice, this result raises a strong political barrier to FB pricing – qualified, however, by a reminder that low-value-of-time users are not necessarily the same people from one day to the next.

4.3 Varying the relative length of the parallel links

It is of course possible that the two parallel links are not lanes of the same highway, but are separate roads instead. In that case, the parallel links need not be equally long. Figures 8a and 8b show how the tolls and the relative efficiency change if the free-flow travel times on links A and B are changed in opposite directions. The base case is now in the centre of the diagram. As the tolled link B becomes shorter when moving to the left, it becomes relatively more attractive; moreover, an equalization of marginal private costs implies that the marginal external costs are relatively higher on B. The tolls for link B therefore have the tendency to increase when moving to the left.

For SBPL, there will be a specific combination of parameters for which the second-best optimal toll is actually zero (this combination is not among the plotted points). This requires link B to be longer than link A. The two forces governing the second-best optimal level of the toll – reducing overall traffic, and diverting traffic from link A, where marginal external costs are higher, to link B – then exactly off-set each other. In this case, ω is zero, as the toll has a zero second-best optimal level. Beyond that point, a subsidy is welfare improving as we can see on the extreme right of Figure 8a (see also Verhoef *et al.*, 1996).

Toll differentiation naturally becomes more important when the two links are of different lengths: that is, when products vary in more dimensions than just amount of congestion.⁹ Consequently, ω_{SBSL} is largest when the free-flow travel times on links A and B are equal and decreases rapidly when moving to the edges of the diagram. The shorter link tends to get the higher price, and carries the higher value of time travellers, both for FB and for PF.¹⁰

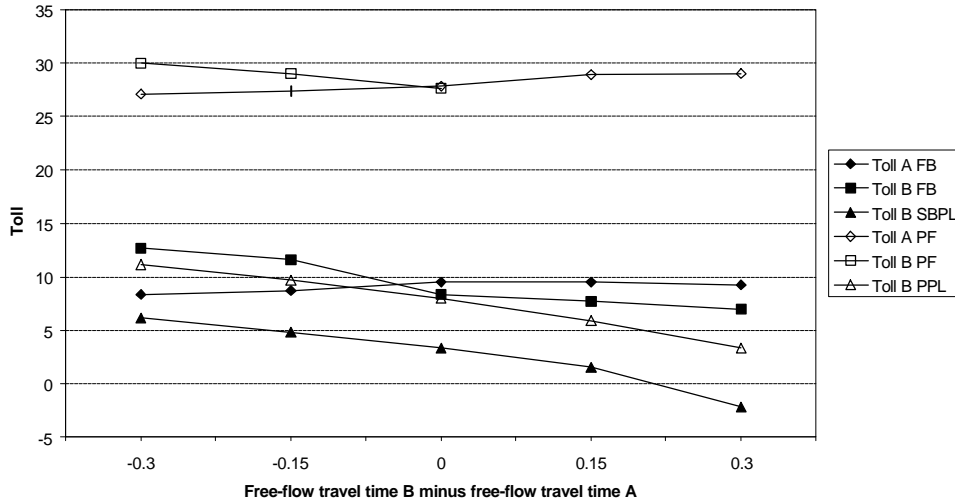
The relative efficiency of PPL declines somewhat more strongly than that of SBPL when moving to the right. In the range where a subsidy would be welfare enhancing when only link B can be tolled, ω for PPL remains low. It does not decrease any further though, since link B has become relatively so unattractive that the monopolist is quite ‘harmless’.

With the other private tolling policies (PF and PSL), the private operator actually has closed down link B at both observations to the right of the base-case by setting the tolls so that

⁹ This is illustrated by a curious result which appears when free-flow travel time is 0.3 hours less on A than on B. This case produces substantial price differentiation under FB pricing, as seen at the far right of Figure 8a. But the second-best serial pricing for this case (SBSL, not shown in the diagram) produces a toll that is lower than both FB tolls – in contrast to all other simulations, where the serial toll lies between the two FB tolls. The reason appears to be that SBSL pricing provides such an inferior option for high-value-of-time users, relative to FB, that it substantially reduces their proportion in the overall composition of traffic. This lowers the marginal cost imposed by any driver sufficiently to result in a second-best toll lower even than the lowest of the two first-best tolls.

¹⁰ It should be noted that the ω 's are in a sense ‘deflated’ when moving to either side of Figure 8b, since the welfare gain with FB increases, due to growing efficiency gains of toll differentiation. Therefore, the same *absolute* welfare change with any given policy would show as a smaller *relative* welfare change.

it is not used. The shorter and higher-capacity link A gets all the traffic, despite the toll. On the far left-hand side, in contrast, we witness an instance of private tolling leading to an efficiency gain (but only when restricted to link B).



Note: For graphical clarity, tolls for SBSL and PSL, being close to those for FB and PF, are suppressed.

Figure 8a. Varying the relative lengths of the parallel links: tolls

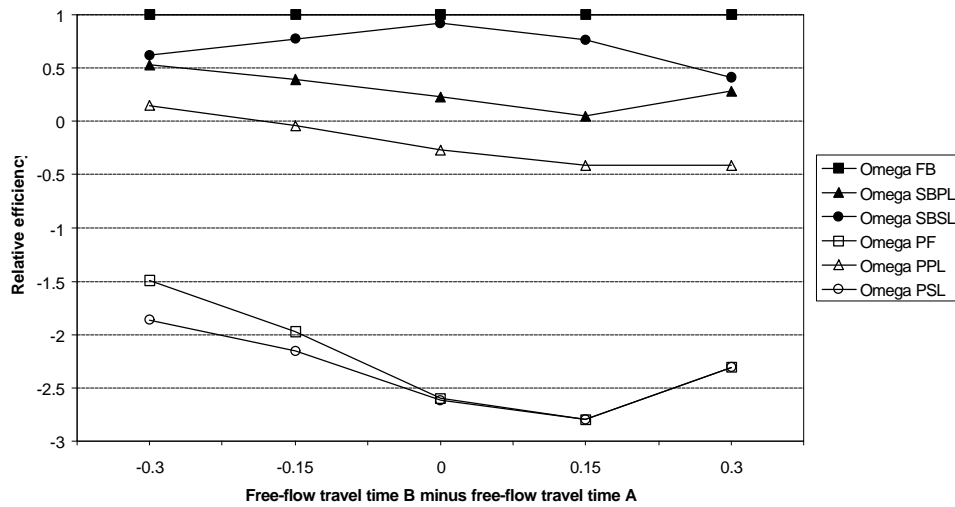


Figure 8b. Varying the relative lengths of the parallel links: relative efficiency

4.4 Varying the overall capacity of the network

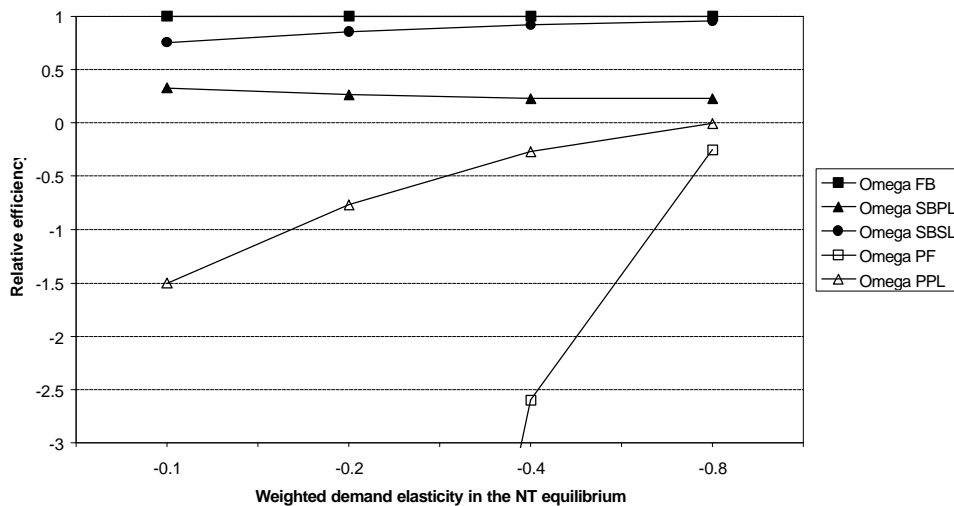
Next, we consider the effect of a simultaneous proportional increase of the three links' capacities. We examined the tolls for a total capacity of 6 000, the base-case of 8 000, then 10 000, and finally, after a jump, 100 000 vehicles per hour, all for the same demand plane. The results (not depicted graphically) show that the degree of toll differentiation (in FB and PF) increases with the equilibrium level of congestion. All public tolls approach zero as the capacity of the network approaches infinity, and congestion vanishes. Also the PPL toll approaches zero: with equally long links A and B, no one will use link B if it is tolled. With PF

and PSL, however, the private operator can still extract monopoly profits and relative efficiency ω for these two regimes approach minus infinity.¹¹

4.5 Varying the total (weighted) demand elasticity

In the next round of simulations, m_α and d_α (see equations (10ab)) were changed simultaneously so as to generate different weighted demand elasticities in the NT-equilibrium, keeping the total level of road use approximately fixed. The demand elasticities considered are those explained in the Footnote to Section 3.2, and values of approximately -0.1 , -0.2 , -0.4 (the base case), and -0.8 were produced. Figure 9 shows the effect on relative efficiency.

At a more inelastic demand, the welfare effects of monopolistic pricing become increasingly negative, as is well known from earlier studies (Verhoef *et al.*, 1996). Therefore, for PF and PSL, and to a lesser extent also for PPL, ω falls rapidly and at an increasing rate when moving leftwards. A new result is, however, that as demand becomes more inelastic, separation of traffic with different values of time becomes relatively more important for overall efficiency. Therefore, ω_{SBSL} decreases when moving to the left.



Note: For graphical clarity, relative efficiency for PSL, being close to that for PF, is suppressed.

Figure 9. Varying the weighted demand elasticity: relative efficiency

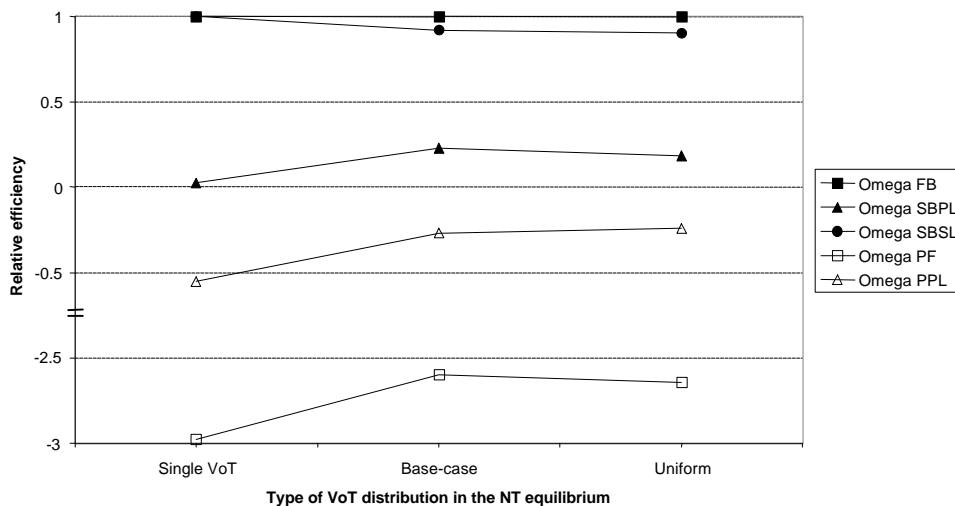
4.6 Varying the type of distribution of values of time

Finally, we consider the extent to which the results presented depend on the shape and degree of dispersion in the distribution of values of time. To that end, we reconsidered the base case assuming two other types of distribution. The exact distribution varies between equilibria, because drivers with a different value of time will generally respond differently to tolls (see *e.g.* Figure 3). We therefore considered the distribution in the NT-equilibrium. Two extreme alternatives were studied: a uniform distribution, implying an increase in the dispersion, and a

¹¹ We also used this variation to double-check the logic of our private tolls by confirming that, as expected when congestion is negligible, the monopolist operates at the point where the total demand elasticity (with respect to toll, not full price) is -1 .

single value of time, implying a complete elimination of dispersion. We considered ‘mean preserving’ changes, with the same average of DFI 9.08 per hour. For the uniform distribution, we accomplished this using an interval [1.2,16.96]. The total level of NT road use and the weighted demand elasticity in the NT-equilibrium remained the same.

Figure 10 shows the impacts on relative efficiency. Of course, the significance of toll differentiation disappears with a single value of time;¹² as a result, policies restricted to pricing just one parallel link perform considerably worse than in the base case. Thus ignoring heterogeneity may lead to serious underestimation of the efficiency of parallel link pricing. Of particular interest, ignoring heterogeneity would lead one to underestimate the relative efficiency of the SBPL policy by a factor of nine (0.025 compared to 0.229 in the base case). This establishes that product differentiation by congestion level is indeed critical to the evaluation of pricing policies that leave parallel roads unpriced.



Note: For graphical clarity, relative efficiency for PSL, being close to that for PF, is suppressed.

Figure 10. Varying the type of distribution of values of time: relative efficiency

At the other extreme, moving from the base-case to the uniform distribution produces slightly more toll differentiation in the FB case, and thus the second-best policies are slightly worse relatively. These latter differences are small, however, so we conclude that the results of this paper are not sensitive to the shape of the value-of-time distribution.

5 Conclusion

This paper has re-considered the standard parallel-route pricing problem in a significantly broader context. Heterogeneity of road users is accounted for by assuming a continuous distribution of values of time. In addition, a third, jointly shared serial link allows for interaction of users of the two routes and also permits analysis of pricing policies where

¹² This is true also of PSL, not shown in the figure, and of PF which, as noted earlier, produces very little toll differentiation even when there is dispersion in values of time.

differentiation is prohibited. A numerical model was calibrated using a standard type of congestion function, and an empirically obtained distribution of values of time.

A number of new results stand out. First, when heterogeneity of road users is considered, travel times in the first-best optimum might for some users actually be higher than in the no-toll equilibrium. This is caused by the trade-off between the value of travel time gains for users with a higher value of time against the losses for those with a lower value of time.

Second, the traditional focus of parallel-route pricing studies on the parallel roads only, with homogeneous users, leads to two opposing biases. Ignoring interaction of users on a third serial link causes benefits of second-best parallel-route pricing to be overestimated because users of the free lanes cause additional external congestion costs elsewhere (on the serial link, in our model). Ignoring heterogeneity, on the other hand, causes great underestimation of benefits, by a factor of nine in our base case, by ignoring the efficiency gains due to separation of traffic. Interestingly, it did not matter much to our results exactly what form of distribution was assumed to represent heterogeneity.

A third result considers the distribution of benefits and losses. First-best pricing, or pricing on the serial link only, combined with uniform tax redistribution, makes the users with the lowest values of time suffer the greatest average welfare losses, or enjoy the smallest gains. This pattern changes when parallel-route pricing is considered: then, the users with the critical value of time, being indifferent between the two routes, suffer most or gain least.

The degree of toll differentiation applied by either a public or a private operator is smaller than expected. It is especially small for the private operator, because at the revenue-maximizing toll nearly all congestion is eliminated. The importance of toll differentiation increases when demand becomes less elastic, and when the parallel links have different free-flow travel times.

For public parallel-route pricing, it turned out that ‘free-lanes’, where a small fraction of a highway remains untolled, is preferable to ‘pay-lanes’, where the opposite holds. The percentage welfare gain of the former is about double that of the latter. An opposite pattern was found for a private operator (although clearly the operator’s objective is better met by pricing more of the road). It may actually be efficiency enhancing to have the private operator tolling the entire highway rather than three out of four lanes, due to non-internalized spill-overs caused with incomplete control.

Finally, the results reconfirmed a more general insight that can be extracted also from other studies in second-best pricing. This is that the amount of information necessary for applying a policy instrument optimally seems to be increasing with the ‘imperfectness’ of this instrument. In a single value of time setting, the second-best tax rule for parallel route pricing implies that the regulator should not only observe the level of marginal external costs, but demand and cost elasticities as well. Our results show that especially for second-best pricing, it is in addition important to know the distribution of values of time. Second-best policies are not only less efficient *per se*, but may in addition yield even smaller net benefits because information needed to efficiently apply the instrument available is lacking.

References

- Arnott, R., A. de Palma and R. Lindsey (1992) "Route choice with heterogeneous drivers and group-specific congestion costs" *Regional Science and Urban Economics* **22** 71-102.
- Braid, R.M. (1996) "Peak-load pricing of a transportation route with an unpriced substitute" *Journal of Urban Economics* **40** (179-197).
- De Palma, A. and R. Lindsey (1999) "Private roads: competition under various ownership regimes" *Annals of Regional Science* forthcoming.
- Edelson, N.E. (1971) "Congestion tolls under monopoly" *American Economic Review* **61** (5) 872-882.
- Hahn, R. (1989) "Economic prescriptions for environmental problems: How the patient followed the doctor's orders" *Journal of Economic Perspectives* **3** (2) 95-114.
- Knight, F. (1924) "Some Fallacies in the Interpretation of Social Costs" *Quarterly Journal of Economics* **38** 582-606.
- Lévy-Lambert, H. (1968) "Tarification des services à qualité variable: application aux péages de circulation" *Econometrica* **36** (3-4) 564-574.
- Liu, L.N. and J.F. McDonald (1998) "Efficient congestion tolls in the presence of unpriced congestion: a peak and off-peak simulation model" *Journal of Urban Economics* **44** 352-366.
- Marchand, M. (1968) "A note on optimal tolls in an imperfect environment" *Econometrica* **36** (3-4) 575-581.
- Parkany, A.E. (1999) *Traveler responses to new choices: Toll vs. free alternatives in a congested corridor*, Ph.D. dissertation, Transportation Science, University of California at Irvine.
- Pigou, A.C. (1920) *Wealth and Welfare*. Macmillan, London.
- Small, K.A. and J.A. Gómez-Ibáñez (1998) "Road Pricing for Congestion Management: The Transition from Theory to Policy," in: *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility*, ed. by K.J. Button and E.T. Verhoef. Cheltenham, UK: Edward Elgar, pp. 213-246.
- Small, K.A. and J. Yan (1999) "The value of 'value pricing' of roads: Second-best pricing and product differentiation" working paper, University of California at Irvine.
- Sullivan, E. (1998) *Evaluating the impacts of the SR 91 variable-toll express lane facility: Final report*, report to California Department of Transportation. Dept. of Civil and Environmental Engineering, Cal Poly State University, San Luis Obispo, California.
- Train, K.E., D.L. McFadden and A.A. Goett (1987) "Consumer attitudes and voluntary rate schedules for public utilities" *Review of Economics and Statistics* **69** (3) 383-391.
- Train, K.E., M. Ben-Akiva and T. Atherton (1989) "Consumption patterns and self-selecting tariffs" *Review of Economics and Statistics* **71** (1) 62-73.
- Verhoef, E.T., P. Nijkamp and P. Rietveld (1996) "Second-best congestion pricing: the case of an untolled alternative" *Journal of Urban Economics* **40** (3) 279-302.
- Verhoef, E.T., P. Nijkamp and P. Rietveld (1997) "The social feasibility of road pricing: a case study for the Randstad area" *Journal of Transport Economics and Policy* **31** (3) 255-276.
- Vickrey, W.S. (1963) "Pricing in Urban and Suburban Transport" *American Economic Review, Papers and Proceedings* **53** (2) 452-465.
- Vickrey, W.S. (1969) "Congestion Theory and Transport Investment," *American Economic Review, Papers and Proceedings* **59** (2) 251-260.
- Walters, A.A. (1961) "The Theory and Measurement of Private and Social Cost of Highway Congestion" *Econometrica* **29** 676-699.
- Wardrop, J.G. (1952) "Some theoretical aspects of road traffic research" *Proceedings of the Institute of Civil Engineers*, **1(II)**, 325-378.
- Wardrop, J. (1952) "Some theoretical aspects of road traffic research" *Proceedings of the Institute of Civil Engineers* **1** (2) 325-378.
- Yang, H. and H.-J. Huang (1999) "Carpooling and congestion pricing in a multilane highway with high-occupancy-vehicles" *Transportation Research* **33A** 139-155.

Appendix: analytical derivation of optimal tolls in the various regimes

In this appendix, we consider the analytical derivation of the optimal tax rules for the various pricing regimes considered in the main text. These results provide insights into the solution and, in three cases (FB, SBSL and PSL), were used to calculate the numerical solutions.

A.1 FB and SBPL: public differentiated tolling

The Lagrangian Λ for schemes FB and SBPL results from adding the objective (5a) and the constraints (7a) (with $N_\alpha = N_{\alpha A} + N_{\alpha B}$). For FB we set the ‘toll-dummies’ $\delta_A = \delta_B = 1$ and $\delta_C = 0$, while for SBPL we set $\delta_B = 1$ and $\delta_A = \delta_C = 0$. The first-order conditions can be found by setting the partial derivatives of Λ with respect to each of the following variables equal to zero: $N_{\alpha L}$ (for all α present on L); $\lambda_{\alpha L}$ (for all α present on L); τ_A ; and τ_B . When taking these derivatives, equation (2) is substituted for α^* , which therefore depends on τ_A , τ_B , and all N_α (since every N_α appears in the argument of either T_A or T_B). We again assume without loss of generality that $\tau_B > \tau_A$ and we define dummy variable \mathbf{d}_{a^*} , which takes on the value of 1 only when $\alpha = \alpha^*$. The first-order conditions then imply (simplified by using the constraints):

$$\begin{aligned} \frac{\partial \Lambda}{\partial N_{aA}} = 0 \Rightarrow & \mathbf{d}_A \cdot \mathbf{t}_A - \int_{a_{\min}}^{a^*} N_{aA} \cdot a \cdot T'_A \, da - \int_{a_{\min}}^{a_{\max}} N_a \cdot a \cdot T'_C \, da \\ & + \int_{a_{\min}}^{a^*} \mathbf{I}_{aA} \cdot a \cdot (T'_A + T'_C) \, da + \int_{a^*}^{a_{\max}} \mathbf{I}_{aB} \cdot a \cdot T'_C \, da - \mathbf{I}_{aA} \cdot D'_a - \mathbf{d}_{a^*} \cdot \mathbf{I}_{a^*B} \cdot D'_{a^*} \\ & + \frac{T'_A \cdot (\mathbf{d}_B \cdot \mathbf{t}_B - \mathbf{d}_A \cdot \mathbf{t}_A)}{(T_A - T_B)^2} \cdot N_{a^*} \cdot X^* = 0 \quad \forall a \leq a^* \end{aligned} \quad (\text{A1a})$$

$$\begin{aligned} \frac{\partial \Lambda}{\partial N_{aB}} = 0 \Rightarrow & \mathbf{d}_B \cdot \mathbf{t}_B - \int_{a^*}^{a_{\max}} N_{aB} \cdot a \cdot T'_B \, da - \int_{a_{\min}}^{a_{\max}} N_a \cdot a \cdot T'_C \, da \\ & + \int_{a_{\min}}^{a^*} \mathbf{I}_{aA} \cdot a \cdot T'_C \, da + \int_{a^*}^{a_{\max}} \mathbf{I}_{aB} \cdot a \cdot (T'_B + T'_C) \, da - \mathbf{I}_{aB} \cdot D'_a - \mathbf{d}_{a^*} \cdot \mathbf{I}_{a^*A} \cdot D'_{a^*} \\ & - \frac{T'_B \cdot (\mathbf{d}_B \cdot \mathbf{t}_B - \mathbf{d}_A \cdot \mathbf{t}_A)}{(T_A - T_B)^2} \cdot N_{a^*} \cdot X^* = 0 \quad \forall a \geq a^* \end{aligned} \quad (\text{A1b})$$

$$\frac{\partial \Lambda}{\partial \mathbf{t}_A} = \int_{a_{\min}}^{a^*} \mathbf{I}_{aA} \, da + \frac{1}{(T_A - T_B)} \cdot N_{a^*} \cdot X^* = 0 \quad \text{iff} \quad \mathbf{d}_A = 1 \quad (\text{A2a})$$

$$\frac{\partial \Lambda}{\partial \mathbf{t}_B} = \int_{a^*}^{a_{\max}} \mathbf{I}_{aB} \, da - \frac{1}{(T_A - T_B)} \cdot N_{a^*} \cdot X^* = 0 \quad \text{iff} \quad \mathbf{d}_B = 1 \quad (\text{A2b})$$

with:

$$\begin{aligned}
X^* = & \alpha^* \cdot (T_A - T_B) + \int_{\alpha_{\min}}^{\alpha^*} N_{aA} \cdot \alpha \cdot T'_A \, d\alpha - \int_{\alpha^*}^{\alpha_{\max}} N_{aB} \cdot \alpha \cdot T'_B \, d\alpha \\
& - \int_{\alpha_{\min}}^{\alpha^*} I_{aA} \cdot \alpha \cdot T'_A \, d\alpha + \int_{\alpha^*}^{\alpha_{\max}} I_{aB} \cdot \alpha \cdot T'_B \, d\alpha
\end{aligned} \tag{A3}$$

The first two conditions (A1a) and (A1b) involve trading off the direct benefits of road use on the one route against the direct costs on that same route, as well as the indirect costs on the other. The direct costs are represented by the first two terms, which are familiar expressions reflecting the marginal external congestion costs imposed by a vehicle on all others using the same road. Note that the marginal benefits D_α and private travel costs $\alpha \cdot T$ do not appear in (A1a) and (A1b) because they were eliminated by substituting the constraints (7a) into the first-order conditions, causing the tolls τ to appear instead.

Next come four terms involving the Lagrangian multipliers λ_α , each of which gives the shadow price of a constraint which in simplified form is just $\alpha \cdot (T_L + T_C) + \tau_L = D_\alpha$ for which ever link L applies. If we think of D_α as containing an exogenous parameter shifting the inverse demand curve for α -type users downward, we see that λ_α represents the marginal impact on social welfare of such a demand shift. In the first-best optimum, FB, it will turn out that everyone is priced at marginal cost so a demand shift has no welfare impact at the margin and $\lambda_\alpha = 0$ for every α . In the second-best optimum SBPL, however, even users of the priced link are paying less than their marginal cost so there is positive social welfare from shifting their demand downward, hence $\lambda_\alpha > 0$ for all α . These three terms in equations (A1), then, show that in evaluating the marginal cost of a user with value of time α , one should also consider the indirect effects of this change upon road use by all other users, the latter being caused by the change in travel time (hence full prices) on the two alternative routes, plus an adjustment for the own elasticity of demand (relevant for both routes when α^* is considered). Note that these demand-related terms are the only ones that differ when comparing (A1a) or (A1b) for different values of α present on either link A or B. Therefore, the shadow prices $\lambda_{\alpha L}$ are inversely proportional to the steepness of the demand D_α : when α -users are less sensitive to price differentials, the shadow price $\lambda_{\alpha L}$ decreases in proportion.

The terms related to X^* , defined in (A3), reflect the welfare impact of induced marginal changes in α^* , again via induced changes in travel times. Equation (A3) shows that this impact includes the change in travel time for α^* -drivers transferred from link B to link A, the direct external congestion cost changes of such a transfer on both routes, and indirect welfare effects, like those just discussed.

Equations (A2a) and (A2b) show that when a toll can be charged on a given link, the shadow price for users of that link would average to zero except for the effect of induced shifts to and from the other link (by users with value of time α^*). When both links are tolled, adding (A2a) and (A2b) show that overall, the shadow prices average to zero. In fact, we already noted that they are identically zero in that case.

These equations exhibit a highly inconvenient discontinuity at α^* – which is why the dummy \mathbf{d}_{a^*} was needed. This discontinuity arises from the fact that a marginal increase of use by α^* -users on either route will affect marginal benefits on both routes. As a result, unless all λ 's are equal to zero, a closed-form analytical solution to (A1a)-(A2b) cannot be found.¹³ To see why, observe that we can solve all λ 's for $\mathbf{I}_{a^*A} + \mathbf{I}_{a^*B}$ from (A1a) and (A1b):

$$\mathbf{I}_a = (\mathbf{I}_{a^*A} + \mathbf{I}_{a^*B}) \cdot \frac{-D'_{a^*}}{-D'_a} \quad \forall a \neq a^* \quad (\text{A4})$$

Substituting (A4) into equations like (A1a) and (A1b) lead to problems of discontinuity at α^* . In the first-best case, because it can be shown that all λ 's are zero, the following intuitive tax-rules apply:

$$\mathbf{t}_A = \int_{a_{\min}}^{a^*} N_{aA} \cdot \mathbf{a} \cdot T'_A \, da + \int_{a_{\min}}^{a_{\max}} N_a \cdot \mathbf{a} \cdot T'_C \, da \quad (\text{A5a})$$

$$\mathbf{t}_B = \int_{a^*}^{a_{\max}} N_{aB} \cdot \mathbf{a} \cdot T'_B \, da + \int_{a_{\min}}^{a_{\max}} N_a \cdot \mathbf{a} \cdot T'_C \, da \quad (\text{A5b})$$

These tax rules simply state that each toll should be equal to the marginal external cost for that route. With optimal pricing on one route, the optimal price on the other can be determined independently, a normal consequence of the envelope theorem. We can also see that, with $\lambda_{\alpha}=0$, (A2) require $X^*=0$, which, from (A3), requires that for α^* -users the valued time difference between the two routes be exactly balanced by the difference in externality costs. With first-best tolls applying on both routes, this is indeed the case.

For SBPL, a closed-form analytical solution can be found only if it happens that $N_{a^*} = 0$, so that no one is indifferent and hence there are no direct spill-over effects between links A and B. We then end up with an independent first-best optimization problem for the priced link. (Similarly, for FB we would end up with two independent first-best optimization problems.) Such a case can only arise if the distribution of values of time is bimodal. It is for this reason that assuming two groups, each with a distinct value of time, permits an analytical solution as in Small and Yan (1999).

A.2 SBSL: Public undifferentiated tolling

The second-best public toll on the serial link can be found by solving the Lagrangian consisting of objective (5b) and constraints (7b). The optimal non-differentiating toll on link C can be shown to be equal to:

¹³ If one would ignore the terms with \mathbf{d}_{a^*} in (A1a) and (A1b), a closed-form solution can be found, but using the simulation model, it was found to produce second-best taxes considerably different from the optimal second-best taxes. Ignoring these terms is ignoring the important role that α^* drivers play, as the 'means of communication' between links A and B, in finding the optimal solution as represented in equation (A3). Comparable erroneous simplifications were tested and refuted for other cases where no closed-form solution can be found (PF and PPL).

$$\mathbf{t}_C = \int_{a_{\min}}^{a_{\max}} N_a \cdot \mathbf{a} \cdot T'_C \, da + \int_{a_{\min}}^{a_{\max}} N_a \cdot \mathbf{a} \cdot T'_D \, da \quad (\text{A6})$$

We expect this solution to provide typically lower welfare than that computed for the first-best problem, but in fact we need to check because the latter was derived on the assumption that the tolls were unequal. We accomplish this by showing that in SBSL, the same traffic flow can be accommodated at lower total cost by setting τ_A marginally lower and τ_B marginally higher than τ_C as defined by (A6). Doing so would lead to a separation of traffic at α^* , and would induce a marginal shift of users from link B to link A. For simplicity, suppose the two links are identical, so that $T_A=T_B$, $T'_A=T'_B$ and $N_A=N_B$ at the solution to SBSL. Denote the size of the shifted traffic as Δ^* . Because travel times are equal on both links, the change in total travel costs resulting from this marginal tax change can be written as:

$$\Delta^* \cdot \left(\int_{a_{\min}}^{a^*} \mathbf{a} \cdot N_a \cdot T'_A \, da - \int_{a^*}^{a_{\max}} \mathbf{a} \cdot N_a \cdot T'_B \, da \right) \quad (\text{A7})$$

The change in travel costs is thus equal to Δ^* times the difference in marginal external congestion costs. With $T'_A=T'_B$ and $N_A=N_B$ this change in cost is negative, because on route B, α will be higher. With different routes, the same type of proof can be given by setting the marginally higher toll on the link that carries more traffic in SBSL. It could be the case, however, that counter-examples can be constructed where differences in T'_A and T'_B happen to exactly off-set the differences in $\int \mathbf{a} \cdot N_a$ in the SBSL equilibrium.

A.3 PF and PPL: Private differentiated tolling

For PF and PPL, the Lagrangian consists of equations (6) plus (7a). Proceeding as in Section A.1, the first-order conditions imply:

$$\begin{aligned} \frac{\partial \Lambda}{\partial N_{aA}} = 0 \Rightarrow & \mathbf{d}_A \cdot \mathbf{t}_A + \int_{a_{\min}}^{a^*} \mathbf{l}_{aA} \cdot \mathbf{a} \cdot (T'_A + T'_C) \, da + \int_{a^*}^{a_{\max}} \mathbf{l}_{aB} \cdot \mathbf{a} \cdot T'_C \, da - \mathbf{l}_{aA} \cdot D'_a - \mathbf{d}_{a^*} \cdot \mathbf{l}_{a^*B} \cdot D'_{a^*} \\ & + \frac{T'_A \cdot (\mathbf{d}_B \cdot \mathbf{t}_B - \mathbf{d}_A \cdot \mathbf{t}_A)}{(T_A - T_B)^2} \cdot N_{a^*} \cdot \left((\mathbf{t}_B - \mathbf{t}_A) - \int_{a_{\min}}^{a^*} \mathbf{l}_{aA} \cdot \mathbf{a} \cdot T'_A \, da + \int_{a^*}^{a_{\max}} \mathbf{l}_{aB} \cdot \mathbf{a} \cdot T'_B \, da \right) = 0 \quad (\text{A8a}) \end{aligned}$$

$$\forall \mathbf{a} \leq \mathbf{a}^*$$

$$\begin{aligned} \frac{\partial \Lambda}{\partial N_{aB}} = 0 \Rightarrow & \mathbf{d}_B \cdot \mathbf{t}_B + \int_{a_{\min}}^{a^*} \mathbf{l}_{aA} \cdot \mathbf{a} \cdot T'_C \, da + \int_{a^*}^{a_{\max}} \mathbf{l}_{aB} \cdot \mathbf{a} \cdot (T'_B + T'_C) \, da - \mathbf{l}_{aB} \cdot D'_a - \mathbf{d}_{a^*} \cdot \mathbf{l}_{a^*A} \cdot D'_{a^*} \\ & - \frac{T'_B \cdot (\mathbf{d}_B \cdot \mathbf{t}_B - \mathbf{d}_A \cdot \mathbf{t}_A)}{(T_A - T_B)^2} \cdot N_{a^*} \cdot \left((\mathbf{t}_B - \mathbf{t}_A) - \int_{a_{\min}}^{a^*} \mathbf{l}_{aA} \cdot \mathbf{a} \cdot T'_A \, da + \int_{a^*}^{a_{\max}} \mathbf{l}_{aB} \cdot \mathbf{a} \cdot T'_B \, da \right) \quad (\text{A8b}) \\ = 0 \quad & \forall \mathbf{a} \geq \mathbf{a}^* \end{aligned}$$

$$\begin{aligned} \frac{\partial \Lambda}{\partial t_A} &= \int_{a_{\min}}^{a^*} N_{aA} da + \int_{a_{\min}}^{a^*} I_{aA} da \\ &+ \frac{1}{(T_A - T_B)} \cdot N_{a^*} \cdot \left((t_B - t_A) - \int_{a_{\min}}^{a^*} I_{aA} \cdot a \cdot T'_A da + \int_{a^*}^{a_{\max}} I_{aB} \cdot a \cdot T'_B da \right) = 0 \end{aligned} \quad (\text{A9a})$$

iff $d_A = 1$

$$\begin{aligned} \frac{\partial \Lambda}{\partial t_B} &= \int_{a^*}^{a_{\max}} N_{aB} da + \int_{a^*}^{a_{\max}} I_{aB} da \\ &- \frac{1}{(T_A - T_B)} \cdot N_{a^*} \cdot \left((t_B - t_A) - \int_{a_{\min}}^{a^*} I_{aA} \cdot a \cdot T'_A da + \int_{a^*}^{a_{\max}} I_{aB} \cdot a \cdot T'_B da \right) = 0 \end{aligned} \quad (\text{A9b})$$

iff $d_B = 1$

Again, the first-order conditions are hard to interpret, and we refer to Verhoef *et al.* (1996) for an interpretation of simpler versions. Roughly speaking, the first two conditions consider the direct and indirect effects of marginal changes of road use upon the objective of maximizing revenue, whereas the latter two help to define the Lagrangian multipliers in the (private) optimum considered. Neither PF nor PPL has a closed-form analytical solution.

A.6 PPS: Private undifferentiated tolling

The problem of a private toll on the serial link has a Lagrangian which combines equation (6) and (7b). The first-order conditions are:

$$\frac{\partial \Lambda}{\partial N_a} = t_C + \int_{a_{\min}}^{a_{\max}} I_a \cdot a \cdot (T'_C + T'_D) da - I_a \cdot D'_a = 0 \quad \forall a \quad (\text{A10a})$$

$$\frac{\partial \Lambda}{\partial t_C} = \int_{a_{\min}}^{a_{\max}} N_a da + \int_{a_{\min}}^{a_{\max}} I_a da = 0 \quad (\text{A10b})$$

Equation (A10a) can be solved for $\int I_a$ by using that $\lambda_{\alpha} \cdot D'_a$ is constant for all α . The following pricing rule can then be found:

$$t_C = \frac{\int_{a_{\min}}^{a_{\max}} N_a da}{\int_{a_{\min}}^{a_{\max}} \frac{1}{-D'_a} da} \cdot \left(1 + (T'_C + T'_D) \cdot \int_{a_{\min}}^{a_{\max}} \frac{a}{-D'_a} da \right) \quad (\text{A11})$$

This rule is a somewhat complicated variant of the standard revenue-maximizing toll on a congested road. It shares with earlier results (*e.g.* Edelson, 1971; Verhoef *et al.*, 1996) the feature that the toll is decreasing in the elasticity of demand (the monopolistic mark-up), and increases in the marginal external congestion costs.