Forthcoming, *Journal of Transport Economics and Policy*

PRODUCT DIFFERENTIATION ON ROADS:

Constrained Congestion Pricing with Heterogeneous Users

Erik T. Verhoef and Kenneth A. Small

This version: 7 July 2003

Address for correspondence: Erik T. Verhoef, Department of Spatial Economics, Free University Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands. Phone: +31-20-4446094. Fax: +31-20-4446004. E-mail: everhoef@econ.vu.nl.

Acknowledgement:
Erik Verhoef is at the Department of Economics, Free University Amsterdam, and at the Tinbergen Institute (Amsterdam/Rotterdam). Kenneth Small is at the Department of Economics, University of California at Irvine. This research was made possible by a fellowship to Erik Verhoef of the Royal Netherlands Academy of Arts and Sciences, and by a grant to Kenneth Small from the University of California Transportation Center. Small is also grateful to Resources for the Future, Washington, DC, for being host to a year's leave during which some of this work was completed. Erik Verhoef would like to thank UC Irvine for their hospitality when the first draft of this paper was written. We thank Amihai Glazer, C. Robin Lindsey, anonymous referees, and seminar participants for comments on earlier drafts.

Key words: congestion, road pricing, networks, second-best, value of time
JEL codes: R41, R48, D62

Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users

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 allows us to model certain features of real-world significance: pricing restrictions on either complementary or substitute links, as well as interactions between different user groups on shared links (e.g. in city centers). We find that revenue-maximizing pricing is much less efficient than welfare-maximizing pricing, whether restricted or unrestricted; but this difference is mitigated by the product differentiation made possible with heterogeneous users. Product differentiation also produces some unexpected distributional effects: those hurt most by pricing may be people with moderate rather than low values of time, and first-best pricing can cause congestion levels to *increase* for some users compared to no pricing. Ignoring heterogeneity causes the welfare benefits of a policy of current interest, namely second-best pricing of one of two parallel links, to be dramatically underestimated. Unlike first-best policies, second-best policies are in danger of losing much of their potential effectiveness if heterogeneity is ignored when setting toll levels.

1 Introduction

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 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 around city centres in Norway, the recent area based congestion charging scheme in central London, peak-period toll surcharges on certain French expressways, special tolled express lanes on two freeway segments in southern California, and a single congestion-priced expressway near Toronto.

This history suggests an increasing importance of partial rather than first-best congestion-pricing schemes. Such schemes include privately or publicly operated toll roads parallel to unpriced highways. Depending on the particular scheme, pricing may be prohibited on routes that are either substitutes for or complements of the one that is priced, and may involve either social or private objectives. Thus a comprehensive analysis requires a model permitting a variety of objectives and pricing constraints. Because much of the purpose of these schemes is to test and shape public opinion, distributional issues are often paramount. Focusing on these turns out to be quite interesting because some of these demonstrations offer highly differentiated products.

In this paper, we simultaneously address issues of second-best policy, public or private objectives, product differentiation, and distribution as they arise from constrained road pricing. We are interested in quantitative statements about the importance of various phenomena, such as user heterogeneity, and so rely heavily on a numerical version of our model; it uses (for its base case) an empirically obtained distribution of values of time for morning peak road users. We analyse both substitutes and complements to the link(s) being priced by using a simple network with both parallel and serial links. Such a set-up can represent, for example, parallel priced and unpriced arterials entering a city center where their users interact on congested streets.

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 drastically underestimated, by a factor of nine in our base case. Private (i.e. profit-maximizing) pricing is almost always worse than no pricing, except when a private route has significant free-flow speed advantages over the free parallel route. Heterogeneity makes first-best differentiated pricing strongly anti-egalitarian, so much so that it may actually worsen the travel times faced by low-value-of-time users even while requiring them to pay – 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 only a small portion reserved as a free option. 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 – hardly an ideal set-up for political success.

Such results pose challenges for the demonstration-project approach to pricing policy. 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 higher price and worse service at the same time. 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.

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 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) endogenize vehicle occupancy and allow for free carpool access to the tolled route.

Revenue-maximizing congestion tolls for a single highway are derived by Edelson (1971) and Mills (1981). When just one of two parallel roads can be priced, Verhoef *et al.* (1996) and Liu and McDonald (1998, 1999) find that the revenue-maximizing price is typically much higher than the second-best price and will achieve very much lower, usually negative, welfare gains. McDonald et al. (1999, pp. 122-124) derive the second-best toll on a link that has both an unpriced substitute and an unpriced complement; but they are unable to say whether the complementary link makes the toll higher or lower. De Palma and Lindsey (2000) consider a variety of ownership regimes, including private and mixed duopolies, both with and without constraints on pricing one of two parallel roads; they focus especially on the effects of time-varying demand patterns and corresponding time-varying tolls. Viton (1995) considers the prospects for a private operator to cover the cost of road construction, reaching optimistic conclusions due to the high toll that can be charged even when in close competition with a free public road.

Very few studies of the two-route problem incorporate heterogeneity in value of time, which turns out to have important implications within a second-best context. The few exceptions all lack some essential feature of our model. Arnott, De Palma and Lindsey (1992) consider two user groups and two routes within the bottleneck model; but they do not consider the case when only one route can be priced. Small and Yan (2001) do consider such a case, but also with just two discrete user groups. Mohring (1979) considers a continuous distribution of values of travel time; but in the context of competing bus and automobile modes; furthermore he does not analyze dispersion in value of time separately from mean value of time and therefore cannot investigate, as we can, the effects of dispersion separately from those of mean value of time. Less closely related are the analyses 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 due to several distinct types of pooled or separated equilibria. In the present paper, we consider instead a continuum of user types. Only two types of equilibria then occur: pooled (when tolls are absent or exactly equal on the two parallel routes), or fully separated (in all other cases).[1] Moreover, using a continuum of values of time allows intermediate groups to be considered explicitly.

---

[1] When tolls are zero or equal a partially separated equilibrium is also possible, but its characteristics are identical to the pooled equilibrium so we rule it out by assumption.

2.2 Basic set-up: network, demand, congestion, and equilibrium conditions

In order to focus on the role of heterogeneity and product differentiation, we specify preferences in considerable detail, and we use a network that is simple yet permits varying degrees of differentiation of trip conditions. We omit from our model a number of practical considerations which would affect policy conclusions for any specific facility. We do not include the costs of toll collection, and we consider only congestion among the many possible sources of difference between private and social cost—ignoring, for example, taxes, accident costs, air pollution, energy security, noise, and land-use impacts. We treat user preferences for travel as exogenous rather than derived, and capacities as given. Finally, we do not examine the political economy or industrial organization of public and private operation of highways; rather, we use "public" and "private" as shorthand for second-best optimization and revenue maximization, respectively. This means, of course that "public" operation wins any showdown by definition; but the interesting questions we explore are by how much, and depending on what factors?

The network 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 user evaluates a trip from O to D solely in terms of its "full price," which includes money cost and self-perceived time cost; in terms of this full price, the routes are assumed to be perfect substitutes. Congestion is represented by assuming that travel time on link L is a non-decreasing function of the number of users N_L who travel on that link: $T_L=T_L(N_L)$ with $T_L'\geq0$ (primes are used to denote derivatives).

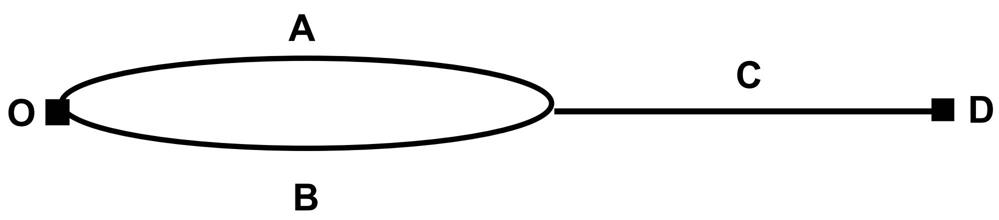


Figure 1. The network considered

Any link L may have a toll, τ_L. However, 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. The full price of a route from O to D consists of the sum over the links constituting that route.

The time-cost component of full price is fully determined by the travel time and a parameter α which we call "value of time." Thus for a traveller with value of time α, the travel cost on link L is $\alpha \cdot T_L$. User heterogeneity (other than that inherent in a downward-sloping demand curve) is represented by specifying a continuum of these values of time. We use N_α to denote the number of users travelling between O and D with value of time α, or more precisely, the density function of α across users; that is, there are $N_\alpha \cdot d\alpha$ users within an infinitesimally small range $[\alpha,\alpha+d\alpha]$ of user types. For each user type α, downward-sloping demand is represented by defining an inverse demand function $D_\alpha(N_\alpha)$, which can be viewed as stating the reservation "full price" of the marginal user of type α when there are N_α users of that type choosing to travel.

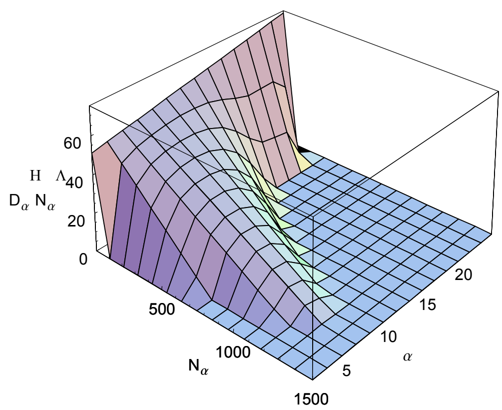


Figure 2. An inverse demand surface

Combining the inverse demand curves for the various values of α into a single diagram produces an inverse demand surface. Figure 2 shows one such surface, namely the one used in the numerical model of Sections 3 and 4. (We explain later how we derived Figure 2.) Intersecting this surface with a plane at constant α depicts a (linear) downward-sloping demand curve for that value of α. Intersecting it with a plane $D_\alpha=0$ depicts the density function of values of time in the population of people willing to travel when there is no time or money cost of doing so (that density function peaks at about α=6.4 DFl/hr).[2] Intersecting the surface with the plane $D_\alpha=0.972\alpha$ depicts the density function—peaking at 6.1 DFl/hr—of values of time of those willing to travel when there is no money cost but the time required is 0.972 hours (this happens to coincide with our base case without toll, and the curve was calibrated to reproduce an empirically derived value-of-time distribution for this particular case).

Variations across users in value of time may arise from many sources including income, gender, type of profession, and unobservable personal characteristics. We need not distinguish them here; in fact, we do not require even that the same individuals be ranked in the same order on different days, so long as the distribution is stable. In particular, we caution against the temptation to think of the value-of-time distribution as simply representing the income distribution; for example, observations on two southern California experiments suggest that the value of time that users exhibit in their choices is far from perfectly correlated with their income (Brownstone and Small, 2003).

We now consider user equilibrium in route choice. Each user is assumed to take prices and travel times on each link as given. Let $N_{\alpha L}$ and $N_{\alpha R}$ be the density functions of user types on link L and on route R. For each user type α and route R, these functions must satisfy the complementary slackness conditions of Wardrop (1952):

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

$$N_{\alpha R} \geq 0 \tag{1b}$$

$$P_{\alpha R} - D_\alpha \geq 0 \tag{1c}$$

where $P_{\alpha R}$ is the 'full price' of using route R, defined as:

$$P_{\alpha R} \equiv \alpha \cdot \left(T_L + T_C\right) + \tau_L + \tau_C, \quad \{L,R\} = \{A, AC\},\{B, BC\} \tag{1d}$$

These equations state that type-α users will use only the route(s) that have least full price to them, and that the reservation price of the marginal type-α user cannot exceed that full price.

Formally, we must proceed differently in solving equations (1) depending on whether or not $\tau_A=\tau_B$. When $\tau_A=\tau_B$, 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, which is entirely innocuous, is that $N_{\alpha A}/N_{\alpha B}=N_A/N_B$ for every α. This yields a perfectly pooled

[2] The exchange rate of the Dutch guilder in late 1999 was approximately DFl 2.2≈€1≈$1.

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 $\tau_A \neq \tau_B$, non-zero use of both routes can occur provided that $sign\{T_B - T_A\} = sign\{\tau_A - \tau_B\}$. This yields a separated equilibrium, in which users differ between the routes according to value of time. The difference in full price for user α can be written as $(\tau_A + \alpha \cdot T_A) - (\tau_B + \alpha \cdot T_B)$; therefore the critical value α^* for which users are indifferent between the routes is:

$$\alpha^* = \frac{\tau_B - \tau_A}{T_A - T_B} \tag{2}$$

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

To complete the model, the following identities are added:

$$N_{\alpha C} = N_{\alpha A} + N_{\alpha B} \tag{3}$$

$$N_L = \int_{\alpha_{\min}}^{\alpha_{\max}} N_{\alpha L}\, \mathrm{d}\alpha \tag{4}$$

where $\alpha_{\min}$ and $\alpha_{\max}$ are the minimum and maximum values of time in the population. In the case where $\tau_A < \tau_B$, (4) implies:

$$N_A = \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A}\, \mathrm{d}\alpha \tag{4a}$$

$$N_B = \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B}\, \mathrm{d}\alpha \tag{4b}$$

2.3 Tolling regimes

We now consider the problem of a private or public operator choosing a toll or set of tolls. It does so knowing that two simultaneous adjustments to the toll will take place: (a) individuals will choose routes, given tolls and travel times, according to equations (1)-(4); and (b) travel

[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 (a condition that easily yields a unique solution with the well-behaved congestions functions we use). It has the property that $(1/T'_D) = (1/T'_A) + (1/T'_B)$.

times on each link will adjust to the level of users, according to the equation $T_L = T_L(N_L)$ describing congestion.

We consider not individual tolls but rather toll regimes, i.e. rules for setting tolls. Our network allows us to analyze a wide variety of such regimes, of which we consider six, in addition to no tolls. These six are defined as the product of two possible objectives (public or private) and three possible choices of where tolls can be applied (entire network, parallel link B only, or serial link C only). These regimes are defined in Table 1, and may be described as follows.

Abbreviation	**Description**	**Tolls on:**
NT	No Tolls	–
Public tolling:		
FB	First-Best tolls on the full network	A and B
SBPL	Second-Best toll on one of the Parallel Links	B
SBSL	Second-Best toll on the Serial Link	C
Private tolling		
PF	Private tolls on the Full network	A and B
PPL	Private toll on one of the Parallel Links	B
PSL	Private toll on the Serial Link	C

Table 1. Tolling regimes

With public tolling, the objective is to maximize net social welfare. Net social welfare is defined as the volume below the inverse demand surface of Figure 2 less total costs.[4] In the unconstrained first-best (FB) regime where the entire network priced, welfare is maximized by setting prices on the two parallel links (recalling that a toll on the serial link is then redundant). In the two second-best (SB) regimes, only a single price can be set, either on one of the two parallel links (SBPL) or on the serial link (SBSL).

The cost part of the objective takes a slightly different mathematical form depending whether the resulting equilibrium is separated or pooled, as described earlier. When $\tau_A < \tau_B$, there is a separated equilibrium defined by the critical value of time α^* given by (2), and the objective can be written as:

[4] Equivalently, net social welfare is equal to Marshallian consumer surplus plus revenues. It would be possible, in the current framework, to define a social welfare function reflecting distributional concerns; but we think it is more useful to use one that identifies the distributional effects of a policy but does not in itself have a redistributional objective—*i.e.* it would not call for individual-specific tolls if there were no congestion.

$$W = \int_{\alpha_{\min}}^{\alpha_{\max}} \int_{0}^{N_\alpha} D_\alpha(n)\,\mathrm{d}n\,\mathrm{d}\alpha - \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A} \cdot \alpha \cdot T_A\left(\int_{\alpha_{\min}}^{\alpha^*} N_{aA}\,\mathrm{d}a\right)\mathrm{d}\alpha - \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B} \cdot \alpha \cdot T_B\left(\int_{\alpha^*}^{\alpha_{\max}} N_{aB}\,\mathrm{d}a\right)\mathrm{d}\alpha$$
$$- \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T_C\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a\,\mathrm{d}a\right)\mathrm{d}\alpha \tag{5a}$$

In (5a) we have used the equilibrium results, described earlier, that $N_{\alpha B}=0$ for $\alpha<\alpha^*$, $N_{\alpha A}=0$ for $\alpha>\alpha^*$, and $N_{\alpha C}=N_{\alpha A}+N_{\alpha B}=N_\alpha$. When $\tau_A=\tau_B=0$, there is a pooled equilibrium in which links A and B can be treated as a merged link D with their combined capacity; in that case the middle two terms on the right-hand side of (5a) are replaced by:

$$- \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T_D\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a\,\mathrm{d}a\right)\mathrm{d}\alpha \tag{5b}$$

With private tolling, the objective is to maximize total toll revenues, R. Again, this can be done in three ways: private tolling on the full network (PF), on one parallel link only (PPL), or on the serial link only (PSL). Using dummy variables δ_L to denote whether or not a toll is in operation on link L, this objective function can be written as:

$$R = \delta_A\tau_A \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A}\,\mathrm{d}\alpha + \delta_B\tau_B \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B}\,\mathrm{d}\alpha + \delta_C\tau_C \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha\,\mathrm{d}\alpha \tag{6}$$

This equation holds also when $\tau_A=\tau_B=0$, but in that case the first two terms on the right-hand side are zero so there is no need to define α^*.

As is common in normative pricing models, it is simpler to maximize the objective by choosing the numbers of travellers on each route rather than by choosing the price directly. In that way each of the four constrained pricing regimes and also the no toll (NT) regime can be represented by a continuum of constraints, one for each value of α, with each constraint representing the requirement of user equilibrium as embodied in equation (1a). When $\tau_A<\tau_B$ (recalling that we can then normalize $\tau_C=0$), the constraints are then represented by adding the following Lagrangian terms to the objective function:

$$+ \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A} \cdot \left[\alpha \cdot T_A\left(\int_{\alpha_{\min}}^{\alpha^*} N_{aA}\,\mathrm{d}a\right) + \alpha \cdot T_C\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a\,\mathrm{d}a\right) + \delta_A \cdot \tau_A - D_\alpha(N_\alpha)\right]\mathrm{d}\alpha$$
$$+ \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B} \cdot \left[\alpha \cdot T_B\left(\int_{\alpha^*}^{\alpha_{\max}} N_{aB}\,\mathrm{d}a\right) + \alpha \cdot T_C\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a\,\mathrm{d}a\right) + \delta_A \cdot \tau_B - D_\alpha(N_\alpha)\right]\mathrm{d}\alpha \tag{7a}$$

where $\lambda_{\alpha L}$ is the Lagrangian multiplier for the constraint (1a) for those values of α having positive $N_{\alpha L}$. The round brackets in (7a) represent the functional relationship defining congestion $T_L(N_L)$ on link L. When $\tau_A=\tau_B=0$, (7a) is replaced by:

$$+\int_{\alpha_{\min}}^{\alpha_{\max}} \lambda_\alpha \cdot \left[\alpha \cdot T_C\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a \,\mathrm{d}a \right) + \alpha \cdot T_D\left(\int_{\alpha_{\min}}^{\alpha_{\max}} N_a \,\mathrm{d}a \right) + \tau_C - D_\alpha(N_\alpha) \right] \mathrm{d}\alpha \qquad (7b)$$

For those tolling regimes (FB, SBSL, PSL) where the resulting toll formula is in closed form, the tax rules are rather straightforward generalizations of those applying with only a single value of time, as given in Verhoef *et al.* (1996). We therefore relegate the derivation and discussion of the first-order conditions and toll formulas to a separate appendix available from the authors upon request. For the other three regimes (SBPL, PF, PPL), the discontinuity at α^* prevented us from finding a closed form analytical solution for the optimal toll, so instead we devised a numerical algorithm to maximize the objective function.

3 A numerical model: the base case

In this section we present a numerical model to assess and illustrate the economic properties of these tolling 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 functional form used is

$$T_L = T_{FL} \cdot \left[1 + b \cdot \left(\frac{N_L}{K_L} \right)^k \right] \qquad (8)$$

where b and k are parameters, T_{FL} is the free-flow travel time on link L, and K_L is conventionally called the 'capacity' of link L. (Because there is no maximum flow for this type of congestion function, 'relative capacity' would actually be a better term.) This functional form has been used extensively for analysis of congestion and seems to fit actual data fairly well (Small, 1992, pp. 70-72). We choose b=0.15 and k=4 throughout our simulations, making it the well-known formula of US Bureau of Public Roads (1964). For capacities K_L, we assume in our base case that link A has three-fourths, and link B one-fourth, of their joint capacity, which for convenience we set it at 8,000 vehicles per hour. We assign this same joint capacity to link C. We also assign free-flow travel times of 22.5 minutes to links A and B, and 7.5 minutes to link C. Hence the setup could represent a four-lane highway with tolling possible along three-fourths of its distance.[5] Table 2 summarizes these base-case parameters.

	Link A	Link B	Link C
b	0.15	0.15	0.15
k	4	4	4
$\mathbf{T_{FL}}$ (hr)	0.375	0.375	0.125
$\mathbf{K_L}$ (veh/hr)	6000	2000	8000

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

3.2 The demand side

The base-case inverse demand surface, depicted in Figure 2, is determined as follows. For every value of time, the demand function is taken to be linear over the relevant range (between the lowest and highest use levels considered):

$$D_\alpha = m_\alpha - d_\alpha \cdot N_\alpha \tag{9}$$

Functions m_α and d_α are calibrated to achieve three objectives: (i) a weighted demand elasticity (over all α) of –0.4 in the NT equilibrium;[6] (ii) travel time in the base-case no-toll regime equal to twice the free-flow travel time; and (iii) 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).[7] The following functions achieve these objectives:

$$m_\alpha = 50 + \alpha \tag{10a}$$

$$d_\alpha = \frac{0.0434783}{-0.713714 + 0.705429 \cdot \alpha - 0.0950357 \cdot \alpha^2 + 0.00468093 \cdot \alpha^3 - 0.000079 \cdot \alpha^4} \tag{10b}$$

(These same functions m_α and d_α are retained for the sensitivity analyses as well, except for cases that explicitly vary the demand characteristics, even though the functions no longer

---

[5] A common approximation for freeway capacities is 2000 vehicles per hour per lane. For more detailed discussions of capacity, see Small (1992, pp. 61-68) or Transportation Research Board (1998).

[6] See, for example, Verhoef et al. (1996) for evidence on this elasticity, which is with respect to full price. In calculating it from the demand surface, we include in the full price a variable monetary cost set to Dfl 12 per trip (6 litres of fuel at price Dfl 2/liter). This variable monetary cost, however, is assumed constant over the various tolling regimes considered, and so is ignored in the simulations.

[7] This distribution was derived using 961 (93%) of the 1027 respondents for whom a value of time could be calculated: the 7% with the highest values of time were discarded 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 Dfl 2 (R^2=0.975). Because of the selection, the average value of time used here is Dfl 9.08, as opposed to Dfl 10.92

yield precisely the results described in (i)-(iii) above.) The values of time α considered in the simulations range between a minimum of DFl 1.2 and a maximum of DFl 23.8 per hour, with a weighted average value of DFl 9.08 in the base case described below.

3.3 General results: base case

Table 3 presents results for the various tolling regimes using these base-case parameters. Welfare results are summarized by an index ω showing a given policy's welfare gain (compared to no tolls) as a fraction of the maximum possible such welfare gain.

	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* (DFl/hr)	-	5.919	12.996	-	6.138	15.265	-	
Travel time A (hr)	0.729	0.529	0.798	0.563	0.397	0.926	0.402	0.375
Travel time B (hr)	0.729	0.733	0.544	0.563	0.426	0.404	0.402	0.375
Travel time C (hr)	0.243	0.189	0.239	0.188	0.134	0.230	0.134	0.125
Toll A (DFl)	0	9.50	0	0	27.83	0	0	
Toll B (DFl)	0	8.29	3.31	0	27.65	7.98	0	
Toll C (DFl)	0	0	0	9.38	0	0	27.80	
Toll revenues (DFl)	0	99606	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. The latter increase is DFl 16743, or DFl 1.32 per user in the NT equilibrium.

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

The first-best (FB) policy produces substantial service differentiation, with travel on link A 28 percent faster than on link B. But this policy also produces some surprises. First, welfare is maximized when the facility with the larger capacity (link A) gets the premium service, in contrast to what one might expect from the analogy of first-class service on airplanes and trains. Second, although overall demand is reduced (by 14 percent) compared

for the full set of respondents. As we note later, this distribution is approximately that shown as the dashed line in Figure 4.

to the no-toll (NT) regime, congestion on the lower-priced link is actually worse than with no tolls. In the base case, this paradox disappears when the portion of the trip on link C is taken into account – all users then receive faster service in the first-best policy than in the no-toll policy. However, Section 4.2 presents an example where even the *total* travel time for the lower-priced link actually *increases* with optimal tolling. Apparently product differentiation is a strong motivation here, calling for a rather low optimal service quality for the segment of the population with lower values of time.

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 two 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.

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 uniform policy 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.)

By contrast, when only one of the parallel links can be priced (SBPL), namely the one 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. The reason is that now, welfare gains on link B from raising its price have to be traded off against welfare losses of spill-over traffic onto link A. 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 related to the fact that with first-best pricing, it was the higher-capacity road that received the higher price.

We now turn to tolling by a private operator. Unrestricted revenue-maximizing tolling extracts a high social cost: welfare is substantially lower than with no tolls at all, especially when both links can be priced (PF). This is because the tolls are set 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.[8] Of course, revenues are correspondingly greater in just those cases where the price is set much higher than is optimal, and this must be taken into account in policy design if some of the capacity is to be financed by toll revenues.

There is a surprise in private tolling, as well: the toll differentiation in unrestricted private pricing (policy PF) 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, making significant service differentiation impossible.

3.4 Distributional results: base case

The numerical simulations allow us to calculate the distribution of welfare effects of the various tolling regimes across people with different values of time. In this sub-section, we present such results for just two public tolling regimes: FB and SBPL.

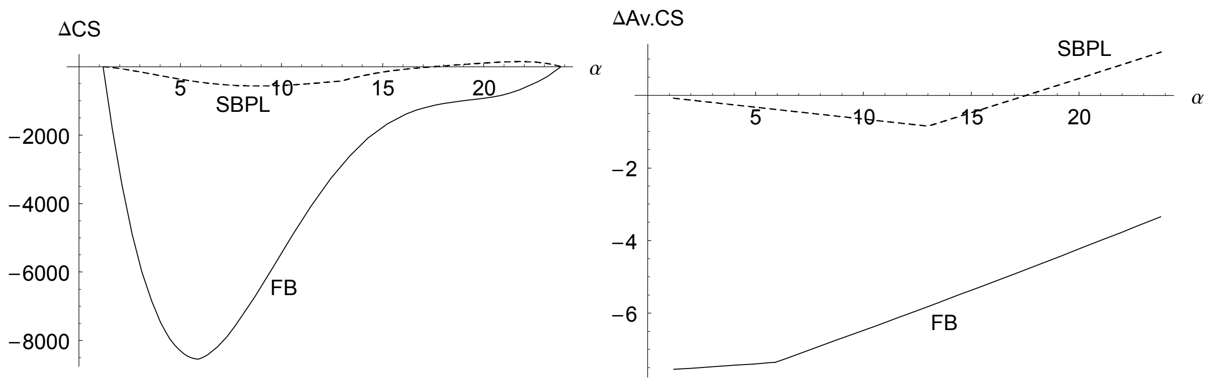


Figure 3. Total (left panel) and average (right panel) change in consumers' surplus, compared to NT, before tax recycling

Figure 3 shows the changes in total and average consumer surplus by value of time, compared to the NT regime.[9] For each value of time, the total change in consumer surplus is given by the change in full price for those users who remain on the road, plus the change in surplus for those who leave the road due to tolling. The average change is defined as the total change divided by the level of use in the NT regime. The figure shows that under first-best

---

[8] See Verhoef et al. (1996), De Palma and Lindsey (2000), and Small and Yan (2001).

[9] Units are total consumer surplus in DFl per unit interval of value of time (the latter in DFl/hr).

tolling (solid line), the average loss in surplus is smaller for people with higher value of time. This result arises, of course, because the price increase is offset by a travel-time decrease, which is valued more by such people. The kink seen in the right panel, which occurs at α^*, is due to the fact that the ratio of toll paid to travel time gained differs across the two parallel links.

Figure 4 shows the levels of road use by value of time for the same two policies.[10] Since the usage under SBPL is very close to that under NT, the dashed line in the left panel also gives a good 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. The right panel in Figure 4 follows the same general pattern as that in Figure 3, simply because the change in consumer surplus is closely related to the change in full price, which in turn determines the change in usage.

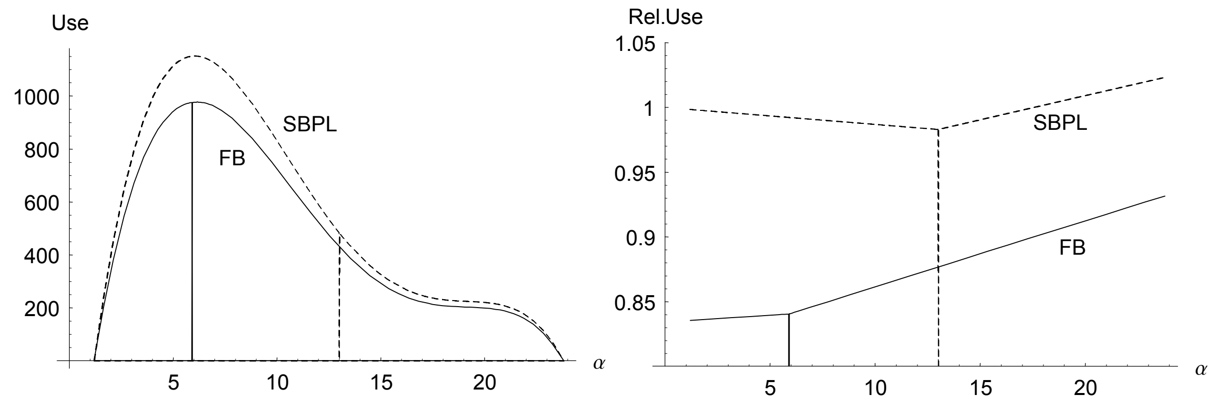


Figure 4. Total (left panel) and relative (right panel) use; the vertical lines indicate α^*

For the SBPL regime (dashed lines in Figures 3 and 4), it is then users near the critical value of time who suffer the largest average losses (Figure 3, right panel). The reason is that imposing second-best tolling on link B improves travel time for people taking that route, while worsening it for those taking the other route. Users near the critical value of time benefit least among those choosing the priced link from its travel-time reduction, and suffer most among those choosing the unpriced link from its travel-time increase. One could say that the policy caters to the more extreme users, leaving those in the middle disadvantaged. However, none of the consumer-surplus changes are very large, the biggest loss amounting to just DFl 0.85 (US$ 0.40) per trip. These changes are much smaller than under FB, and users with the highest values of time even benefit directly from SBPL, i.e. they are better off regardless of use of toll revenues. This helps explain why parallel-route pricing appears to be more politically acceptable than first-best tolling.

As expected, the relative attractiveness of the FB and SB regimes may be reversed by redistributing the toll revenues. Figure 5 shows the changes in total and average consumer surplus after applying the simplest possible tax-recycling scheme: an equal redistribution to all initial road users. This simply means an upward shift of each of the curves shown in the right panel of Figure 3. 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 for all but the very lowest-value-of-time users. Furthermore, first-best pricing is now welfare-enhancing for every user compared to no tolls.[11] 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.

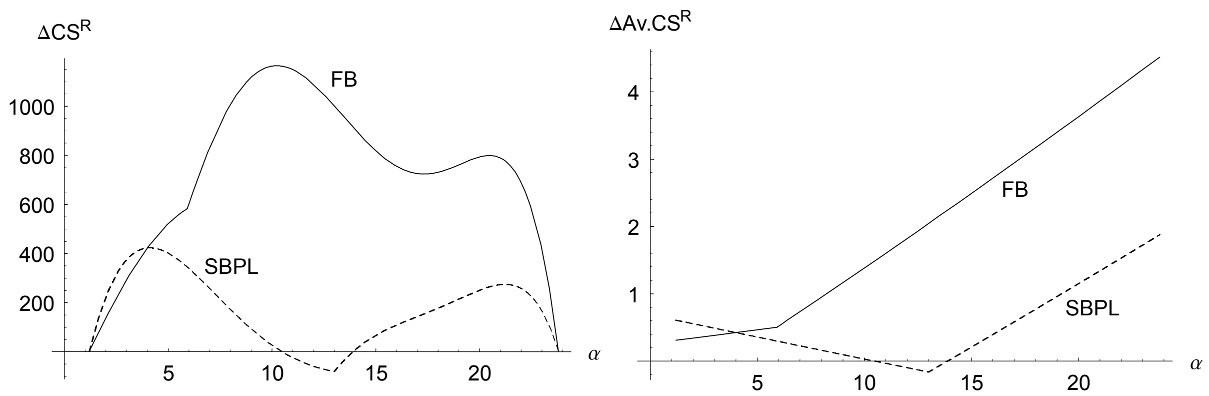


Figure 5. Total (left panel) and average (right panel) change in consumers' surplus, compared to NT, after non-differentiated tax recycling

Under private tolling, it can be expected that the distribution of changes in average consumer surplus will show patterns comparable to those shown in the right panel of Figure 3, 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 could make everyone better off. In practice, private tolls are likely to be restricted by additional

---

[10] Units are numbers of users per unit interval of value of time (the latter in DFl/hr).

[11] A similar result in a mode choice context was observed by Small (1983).

regulations, such as rate-of-return caps or direct price regulation; our results provide support for some such restriction.

4 Sensitivity analysis

In this section, we 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, while holding the demand surface invariant. 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.

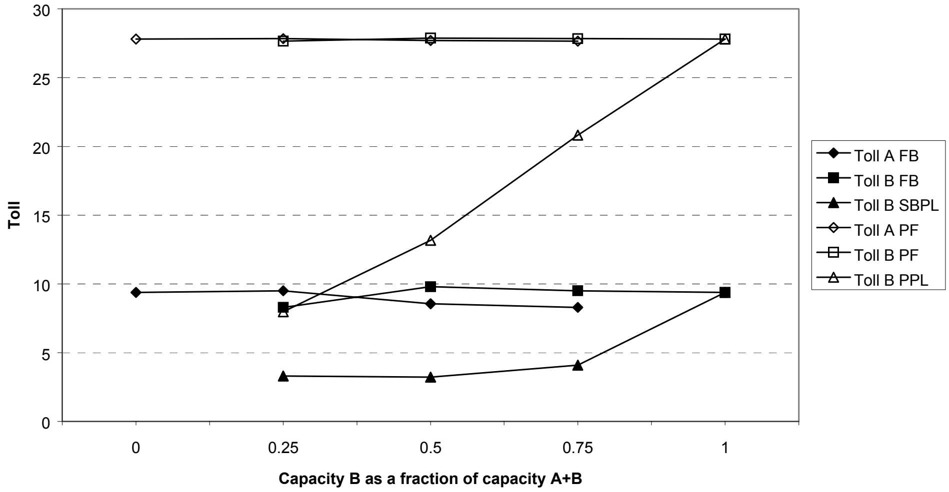
4.1 Varying the relative capacities of the two parallel links

We first consider the impacts of increasing the fraction of the highway subject to tolling, keeping total joint capacity of links 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%).[12] Unsurprisingly, the greatest impacts of capacity allocation occur for those policies constraining a parallel link, namely SBPL and PPL. For public tolling, greater capacity of B makes the second-best policy (SBPL) relatively more efficient, because the importance of the unpriced substitute is diminished; at 75% capacity, nearly half the possible welfare gains are realized. 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. This substantially increases the relative welfare losses from PPL, at least up to 75% of capacity. Oddly, once at least 75 % of capacity is allocated to a private operator it is better that all capacity be so allocated; this counterintuitive result, also found by Verhoef *et al.*

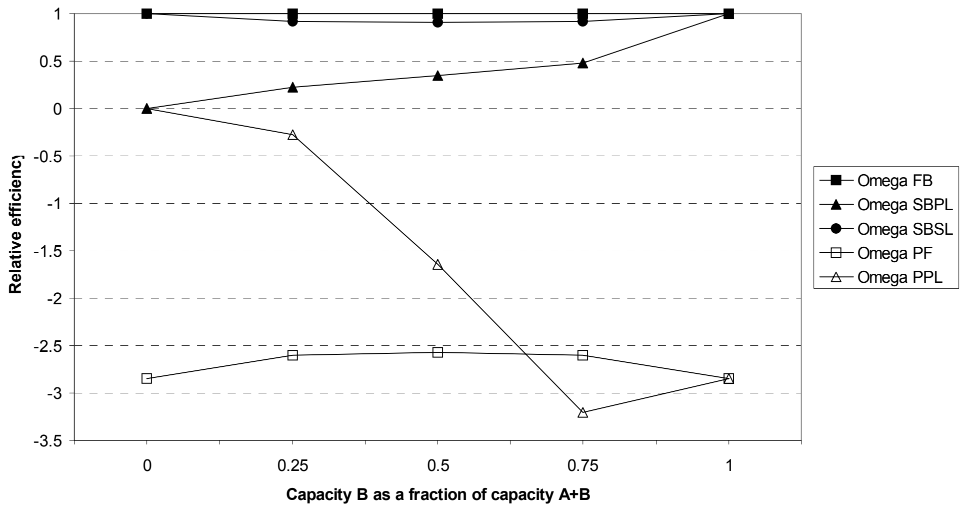
[12] On the left-hand side of these figures, therefore, SBPL and PPL are identical to NT, because no capacity is tolled; whereas on the right-hand side, they are identical to SBSL and PSL, respectively, because all the capacity is tolled. At both extremes, toll differentiation is impossible, so FB is identical to SBSL and PF to PSL.

(1996), occurs because full control of the network avoids inefficient route splits. Finally, the finding of relatively limited price differentiation under FB pricing remains intact.[13]



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

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 surpressed.

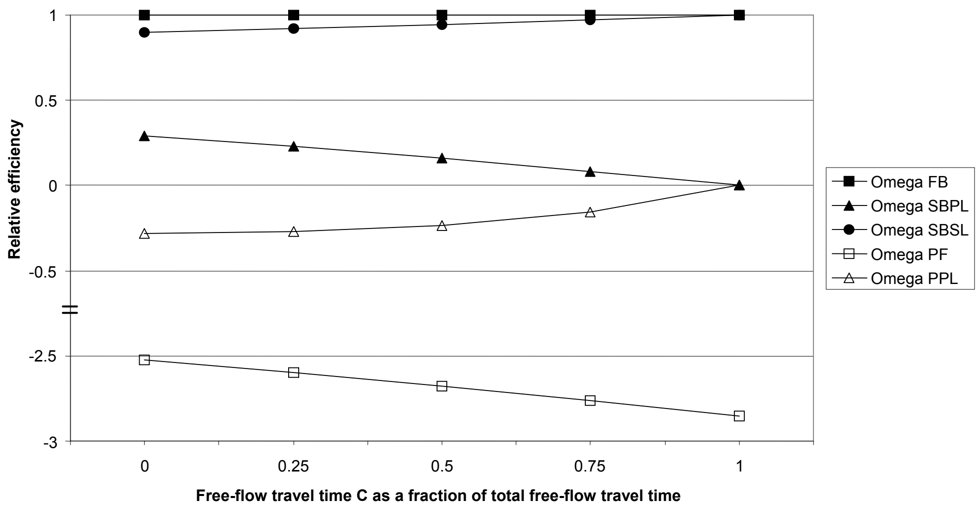
Figure 6b. Varying the relative capacities of the two parallel links: relative efficiency

[13] The FB scheme will have differentiated tolls throughout. The intersection of the lines representing τ_A and τ_B in Figure 6a results from graphical interpolation only, and is near the point where it becomes more efficient to charge a higher toll on link B than on link A, instead of the other way around. A similar argument holds for PF, and also for FB in Figure 8a below.

Together, these results contradict the idea that efficiency always increases monotonically with the degree of privatization. If one insists on a system with both unpriced and priced alternatives, it is more efficient to allow a public operator to price most of the capacity, but a private operator only a small portion of it instead.

4.2 Varying the relative length of the serial link

Most studies ignore the likelihood that users of two parallel routes will not be completely isolated, but rather will share some links upstream or downstream of the split road section. 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 its free-flow travel time – is increased in 25% steps, keeping the total free-flow travel time constant. Note that when the relative length of C has become 1, the parallel links effectively disappear so FB becomes identical to SBSL and PF to PSL.



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

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

As the relative length of the serial link increases, second-best toll differentiation becomes less viable, so both the public and private tolls on the parallel link fall (even per kilometer) and approach zero. As a result, the relative efficiencies of these regimes approach zero as well. From a societal point of view, this is bad news in the case of the public toll and good news in the case of the private toll. 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. Similarly ω_{PPL} changes from –0.28 to –0.16 over the same interval. 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 can cause optimal pricing to increase the travel times of lower-value-of-time users, compared to no pricing. Of course, since FB pricing leads to a potential Pareto improvement, it remains true that these users could be made better off by some lump-sum redistribution of revenues. In practice, this result raises a strong political barrier to optimal pricing – qualified, however, by a reminder that low-value-of-time users are not necessarily the same people from one day to the next.

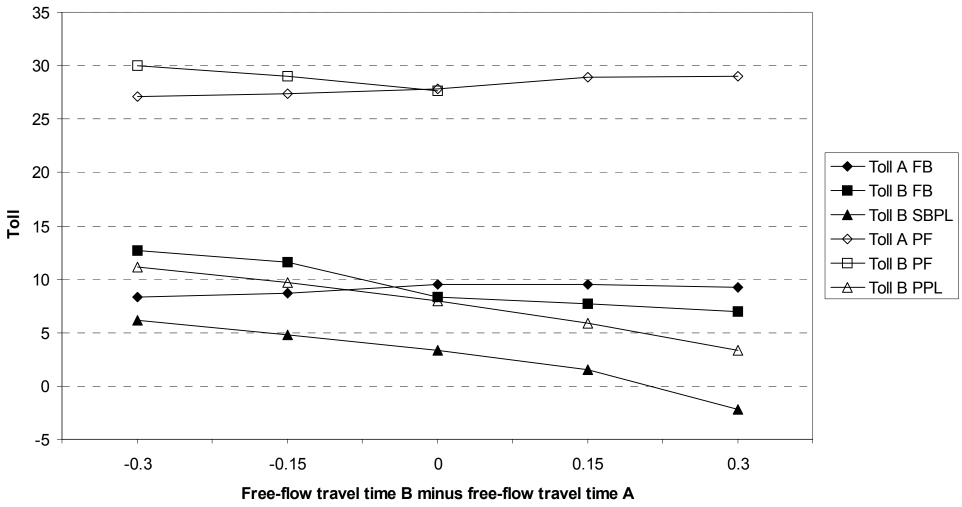
4.3 Varying the relative length of the parallel links

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 have equal free-flow times. An example is a toll road that parallels an arterial with at-grade intersections.

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 smaller-capacity link (B) becomes shorter when moving to the left, it requires a relatively higher marginal external cost or a higher toll in order to equalize marginal private costs on the two links. The tolls for link B therefore have the tendency to increase when moving to the left, and to decrease – even becoming negative – when moving to the right.[14]

Toll differentiation naturally becomes more important when the two links are of different lengths: that is, when products vary in more dimensions that just amount of

[14] 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_{SBPL} is zero. Beyond that point, as Figure 8a shows, a subsidy is welfare improving.



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

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

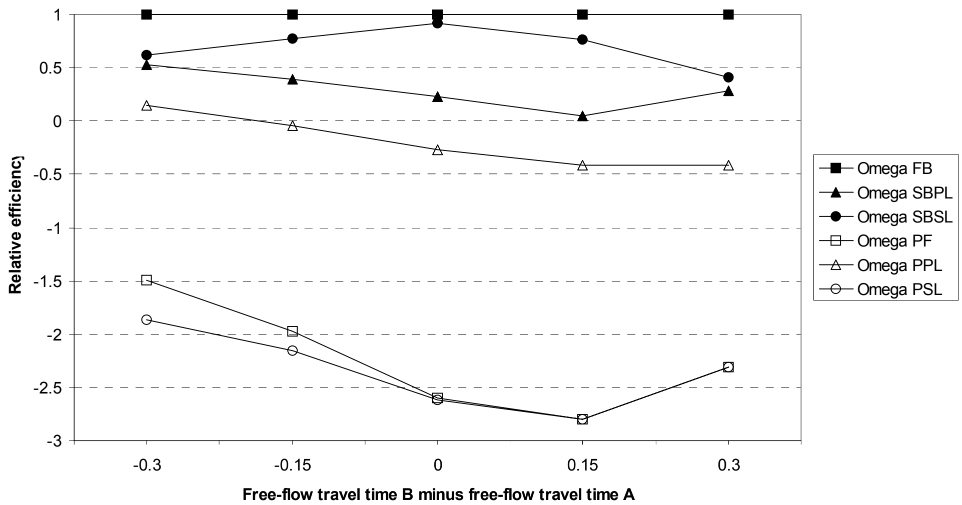


Figure 8b. Varying the relative lengths of the parallel links: relative efficiency

congestion.[15] Consequently, the potential welfare gain from fully optimal pricing (FB) increases as free-flow travel times become more unequal. Furthermore, when link B is shorter

---

[15] 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 either FB toll – 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.

(left side of Figure 8), there is less disadvantage to being unable to price link A, so the relative welfare gain from SBPL also rises – to just over 50% at a 0.3 hours free-flow travel time difference. A similar result is also found by Verhoef *et al.* (1996, Figures 2 and 5) and Liu and McDonald (1999, Table 1 and p. 187).

Another consequence is that equal prices on the parallel links become increasingly unsatisfactory as the links become more unequal. As a result, ω_{SBSL} decreases rapidly when moving to the edges of the diagram. The shorter link tends to get the higher price for both FB and PF.[16]

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'. On the far left-hand side, in contrast, we witness an instance of private tolling on link B leading to an efficiency gain. 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 link B is not used.

4.4 Varying the overall capacity of the network

Next, we consider the effect of a simultaneous proportional increase of the three links' capacities. Since the demand function is unchanged, this process effectively varies the amount of congestion. We examined the tolls for four values of total capacity: 6000, 8000 (the base case), 10,000, and 100,000 vehicles per hour, all for the same demand surface. 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, as well as the PPL toll, approach zero as the capacity of the network approaches infinity and congestion vanishes. With PF and PSL, however, the private operator can still extract monopoly profits by tolling, leading to tolls and welfare losses which do not approach zero.[17]

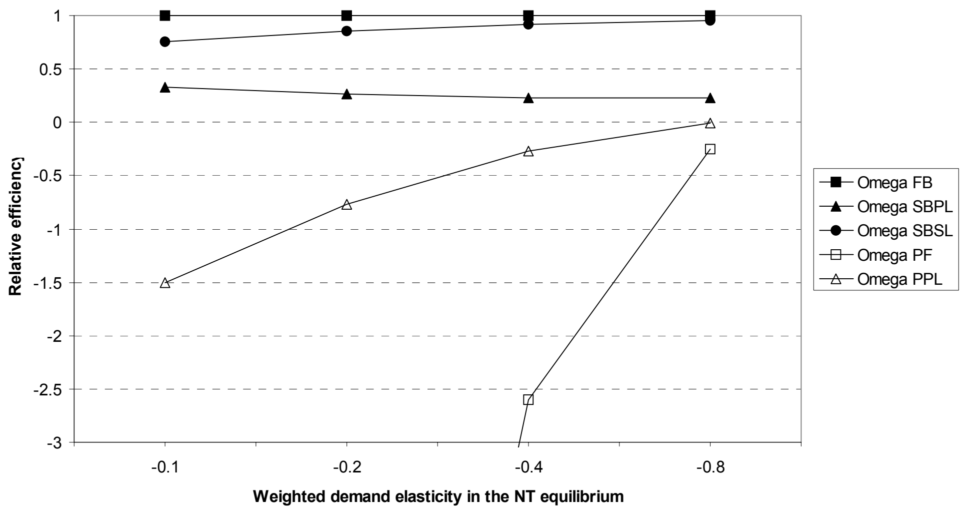
---

[16] Note 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.

[17] 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.

4.5 Varying the total (weighted) demand elasticity

In the next round of simulations, m_α and d_α in equations (10) 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 calculation of demand elasticity is explained in the first footnote to Section 3.2.) 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.



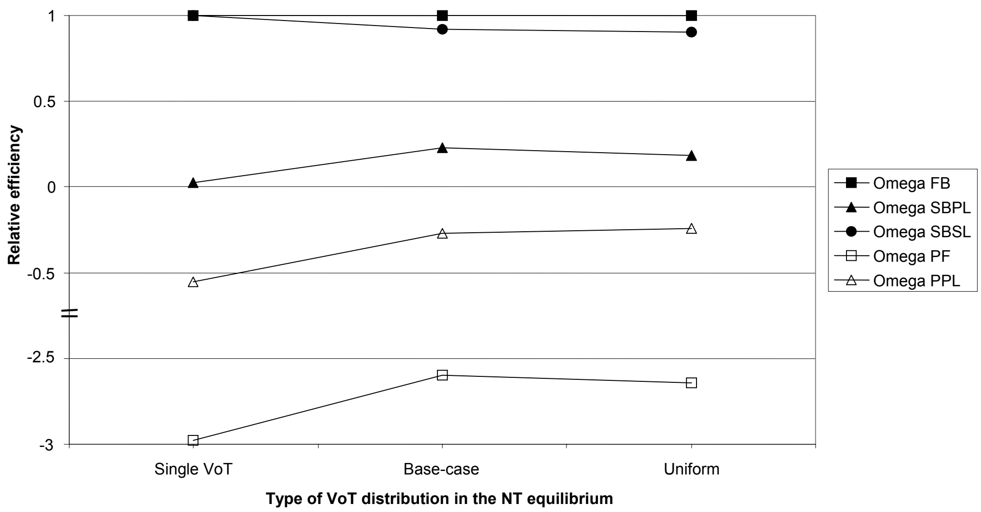
Note: For graphical clarity, relative efficiency for PSL, being close to that for PF, is surpressed.

Figure 9. Varying the weighted demand elasticity: 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 for PPL, ω falls rapidly and at an increasing rate when moving leftwards. A new result, however, is also seen: as demand becomes more inelastic, separation of traffic with different values of time becomes relatively more important for overall efficiency. Therefore, ω_{SBPL} increases and ω_{SBSL} decreases when moving to the left.

4.6 Varying the type of distribution of values of time

Finally, we consider the extent to which the results presented depend on the distribution of values of time. To that end, we redo the base case with two alternative types of distribution: a uniform distribution (which has greater variance of values of time than the base case distribution) and a degenerate distribution with a single value of time. We calibrate on the distribution in the NT equilibrium, since the exact distribution varies between equilibria (see Figure 4). We keep the same weighted average value of time of DFl 9.08 per hour, again in the NT equilibrium; for the uniform distribution, we accomplish this using an interval [1.20,16.96]. The height and price-slope of the demand surface are calibrated to keep total road use and weighted demand elasticity in the NT equilibrium the same as in the base case.



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

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

Figure 10 shows the impacts on relative efficiency. Of course, the significance of toll differentiation disappears with a single value of time;[18] 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, as suggested also by Small and Yan (2001). 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. 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 exact shape of the value-of-time distribution.

What if an erroneous assumption of a single value of time is carried through to the toll-setting stage? The second-best toll for parallel-route pricing (SBPL, single value of time) is only DFl 0.88, about 27% of the true second-best toll of DFl 3.31 (shown in Table 3). The

[18] 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.

actual use of this smaller toll when true heterogeneity exists, as in our base case, would lead to a relative welfare gain of ω=0.103. This is 45% of the welfare gain from the correctly calculated toll, which is ω=0.229 (again as shown in Table 3). Therefore, a regulator knowing the average value of time but ignoring its dispersion when setting the toll could lose about half of the already limited efficiency gains possible from parallel route pricing.

For first-best pricing, in contrast, the predicted optimal toll when ignoring heterogeneity is DFl 9.19, not very different from the truly optimal differentiated tolls of DFl 8.29 and 9.50. The relative welfare gain, applying the former toll, is ω=0.9199; that is, the inefficiency from ignoring heterogeneity is only eight percent. Furthermore, the best one can do with a single toll is ω=0.9203 (the value for SBPL from Table 3). Therefore so long as both parallel links are being priced, the inefficiency from ignoring heterogeneity is almost entirely from adopting uniform pricing, which may actually be optimal once collection costs are accounted for; the further inefficiency from calculating the wrong uniform toll is negligible.

This reconfirms an insight from earlier studies: second-best taxes are not only by definition less efficient than first-best taxes, but in addition are harder to implement optimally because they require more information. First-best tax rules require knowing only the level of marginal external costs in the final equilibrium. The second-best tax rule for parallel-route pricing, as derived for example by Verhoef *et al.* (1996), requires that the regulator also know the demand and cost elasticities. Our results show that in addition it is important to know the distribution of values of time. When such information is lacking or ignored, the resulting inefficiency from non-optimal toll levels is much greater than for first-best taxes.

5 Conclusion

This paper has reconsidered the road-pricing problem in a significantly broader context. We treat partial network pricing in a flexible way by considering two parallel routes followed by a shared link. We account for heterogeneity of users by assuming a continuous distribution of values of time. These innovations capture aspects of real applications of pricing, and they turn out to have significant effects.

Several new results stand out. First, when heterogeneity of road users is considered, travel times in the first-best optimum might actually be higher on one of the routes than in the no-toll equilibrium. This is caused by the use of differentiated tolls to provide a higher-quality service on link A by crowding link B even more.

Second, the most common approach to analyzing the benefits of parallel-route pricing creates two opposing biases. On the one hand, using two parallel routes but ignoring the interaction of users on other parts of the network (link C in our model) causes benefits of second-best pricing to be overstated, because users of the free lanes cause additional external congestion costs elsewhere. On the other hand, ignoring user heterogeneity causes benefits of second-best policies to be understated, by a factor of nine in our base case, because significant efficiency gains due to separation of traffic are omitted. Interestingly, it does not matter much to our results exactly what form the heterogeneity takes.

A third result concerns the distribution of benefits and losses. Under first-best pricing, users with the lowest values of time suffer the greatest average welfare losses or enjoy the smallest average gains. Many discussions of the politics of road pricing have focused on this point. However, the pattern changes when close substitutes of the priced good remain free: then, the users with intermediate values of time suffer most or gain least. It is as though we were to offer airline travellers only propeller planes or supersonic jets; this would cater to the extremes, but a lot of people would want something in between. To the extent that democratic processes cater to median preferences, this may help explain why pricing policies for congestible public facilities have made less political headway than other market-oriented reforms.

Fourth, the degree of toll differentiation that maximizes either welfare or revenue in an unconstrainted setting is smaller than expected. The importance of toll differentiation increases when demand becomes less elastic, and when the parallel links have different free-flow travel times.

Finally, the results confirm a more general insight from studies in second-best pricing: the amount of information required to apply a policy instrument to best advantage increases with the 'imperfectness' of this instrument. For the case considered here, this information includes the distribution of values of time and the demand elasticities of users having different values of time. Thus, second-best policies require considerable sophistication in order to achieve their theoretical benefits.

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).

Brownstone, David, and Kenneth A. Small (2003) "Valuing Time and Reliability: Assessing the Evidence from Road Pricing Demonstrations," working paper, Dept. of Economics, Univ. of California at Irvine (May).

De Palma, A. and R. Lindsey (2000) "Private roads: competition under various ownership regimes" *Annals of Regional Science* **34** (1) 13-35.

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, N.L. 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.

Liu, N.L. and J.F. McDonald (1999) "Economic efficiency of second-best congestion pricing schemes in urban highway systems" *Transportation Research* **33B** 157-188.

Marchand, M. (1968) "A note on optimal tolls in an imperfect environment" *Econometrica* **36** (3-4) 575-581.

McDonald, John F., Edmond L. d'Ouville, and Louie Nan Liu (1999) *Economics of Urban Highway Congestion and Pricing*. Kluwer, Boston.

Mills, D.E. (1981) "Ownership arrangements and congestion-prone facilities" *American Economic Review, Papers and Proceedings* **71** (3) 493-502.

Mohring, Herbert (1979) "The benefits of reserved bus lanes, mass transit subsidies, and marginal cost pricing in alleviating traffic congestion" *Current Issues in Urban Economics*, ed. by Peter Mieszkowski and Mahlon Straszheim. Johns Hopkins, Baltimore, pp. 165-195.

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

Small, Kenneth A. (1983) "The incidence of congestion tolls on urban highways," *Journal of Urban Economics* **13** (1) 90-111.

Small, Kenneth A. (1992) *Urban Transportation Economics*. Harwood Acaemic Publishers, Chur, Switzerland.

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 (2001) "The value of 'value pricing' of roads: second-best pricing and product differentiation" *Journal of Urban Economics* **49** (2) 310-336.

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.

Transportation Research Board (1998) *Highway Capacity Manual: Special Report 209* (3rd edition, 1997 update). National Research Council, Washington, D.C.

US Bureau of Public Roads (1964) *Traffic Assignment Manual*, US Bureau of Public Roads, Washington, D.C.

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.

Viton, P.A. (1995) "Private roads" *Journal of Urban Economics* **37** (3) 260-289.

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** (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.

Not for publication

Appendix to: "Product Differentiation on Roads: Constrained Congestion Pricing with Heterogeneous Users"

Erik T. Verhoef and Kenneth A. Small

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 δ_{α^*}, which takes on the value of 1 only when $\alpha = \alpha^*$. The first-order conditions then imply (simplified by using the constraints):

$$\frac{\partial \Lambda}{\partial N_{\alpha A}} = 0 \Rightarrow \delta_A \cdot \tau_A - \int_{\alpha_{\min}}^{\alpha^*} N_{aA} \cdot a \cdot T_A' \,\mathrm{d}a - \int_{\alpha_{\min}}^{\alpha_{\max}} N_a \cdot a \cdot T_C' \,\mathrm{d}a$$
$$+ \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot \left(T_A' + T_C'\right) \mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot T_C' \,\mathrm{d}a - \lambda_{\alpha A} \cdot D_\alpha' - \delta_{\alpha^*} \cdot \lambda_{\alpha^* B} \cdot D_{\alpha^*}' \tag{A1a}$$
$$+ \frac{T_A' \cdot \left(\delta_B \cdot \tau_B - \delta_A \cdot \tau_A\right)}{\left(T_A - T_B\right)^2} \cdot N_{\alpha^*} \cdot X^* = 0 \quad \forall\, \alpha \le \alpha^*$$

$$\frac{\partial \Lambda}{\partial N_{\alpha B}} = 0 \Rightarrow \delta_B \cdot \tau_B - \int_{\alpha^*}^{\alpha_{\max}} N_{aB} \cdot a \cdot T_B' \,\mathrm{d}a - \int_{\alpha_{\min}}^{\alpha_{\max}} N_a \cdot a \cdot T_C' \,\mathrm{d}a$$
$$+ \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot T_C' \,\mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot \left(T_B' + T_C'\right) \mathrm{d}a - \lambda_{\alpha B} \cdot D_\alpha' - \delta_{\alpha^*} \cdot \lambda_{\alpha^* A} \cdot D_{\alpha^*}' \tag{A1b}$$
$$- \frac{T_B' \cdot \left(\delta_B \cdot \tau_B - \delta_A \cdot \tau_A\right)}{\left(T_A - T_B\right)^2} \cdot N_{\alpha^*} \cdot X^* = 0 \quad \forall\, \alpha \ge \alpha^*$$

$$\frac{\partial \Lambda}{\partial \tau_A} = \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A} \,\mathrm{d}\alpha + \frac{1}{(T_A - T_B)} \cdot N_{\alpha^*} \cdot X^* = 0 \quad \textit{iff} \quad \delta_A = 1 \tag{A2a}$$

$$\frac{\partial \Lambda}{\partial \tau_B} = \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B} \,\mathrm{d}\alpha - \frac{1}{(T_A - T_B)} \cdot N_{\alpha^*} \cdot X^* = 0 \quad \textit{iff} \quad \delta_B = 1 \tag{A2b}$$

with:

$$\begin{aligned} X^* = {} & \alpha^* \cdot (T_A - T_B) + \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A} \cdot \alpha \cdot T_A' \,\mathrm{d}\alpha - \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B} \cdot \alpha \cdot T_B' \,\mathrm{d}\alpha \\ & - \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A} \cdot \alpha \cdot T_A' \,\mathrm{d}\alpha + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B} \cdot \alpha \cdot T_B' \,\mathrm{d}\alpha \end{aligned} \tag{A3}$$

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 δ_{α^*} 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.[19] To see why, observe that we can solve all λ's for $\lambda_{\alpha^* A} + \lambda_{\alpha^* B}$ from (A1a) and (A1b):

$$\lambda_a = \left(\lambda_{\alpha^* A} + \lambda_{\alpha^* B}\right) \cdot \frac{-D'_{\alpha^*}}{-D'_a} \quad \forall\, a \neq \alpha^* \tag{A4}$$

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:

$$\tau_A = \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A} \cdot \alpha \cdot T'_A \, \mathrm{d}a + \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T'_C \, \mathrm{d}\alpha \tag{A5a}$$

$$\tau_B = \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B} \cdot \alpha \cdot T'_B \, \mathrm{d}\alpha + \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T'_C \, \mathrm{d}\alpha \tag{A5b}$$

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

---

[19] *If* one would ignore the terms with δ_{α^*} 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. Comparable erroneous simplifications were tested and refuted for other cases where no closed-form solution can be found (PF and PPL).

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 $N_{\alpha^*} = 0$, 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 (2001).

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:

$$\tau_C = \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T_C' \, \mathrm{d}\alpha + \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \cdot \alpha \cdot T_D' \, \mathrm{d}\alpha \tag{A6}$$

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:

$$\Delta^* \cdot \left(\int_{\alpha_{\min}}^{\alpha^*} \alpha \cdot N_\alpha \cdot T_A' \, \mathrm{d}\alpha - \int_{\alpha^*}^{\alpha_{\max}} \alpha \cdot N_\alpha \cdot T_B' \, \mathrm{d}\alpha \right) \tag{A7}$$

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 α will be higher on route B. 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 $\int \alpha \cdot N_\alpha$ 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:

$$\frac{\partial \Lambda}{\partial N_{\alpha A}} = 0 \Rightarrow \delta_A \cdot \tau_A + \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot (T_A' + T_C')\,\mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot T_C'\,\mathrm{d}a - \lambda_{\alpha A} \cdot D_\alpha' - \delta_{\alpha^*} \cdot \lambda_{\alpha^* B} \cdot D_{\alpha^*}'$$
$$+ \frac{T_A' \cdot (\delta_B \cdot \tau_B - \delta_A \cdot \tau_A)}{(T_A - T_B)^2} \cdot N_{\alpha^*} \cdot \left((\tau_B - \tau_A) - \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot T_A'\,\mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot T_B'\,\mathrm{d}a \right) = 0 \quad \text{(A8a)}$$
$$\forall\, \alpha \le \alpha^*$$

$$\frac{\partial \Lambda}{\partial N_{\alpha B}} = 0 \Rightarrow \delta_B \cdot \tau_B + \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot T_C'\,\mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot (T_B' + T_C')\,\mathrm{d}a - \lambda_{\alpha B} \cdot D_\alpha' - \delta_{\alpha^*} \cdot \lambda_{\alpha^* A} \cdot D_{\alpha^*}'$$
$$- \frac{T_B' \cdot (\delta_B \cdot \tau_B - \delta_A \cdot \tau_A)}{(T_A - T_B)^2} \cdot N_{\alpha^*} \cdot \left((\tau_B - \tau_A) - \int_{\alpha_{\min}}^{\alpha^*} \lambda_{aA} \cdot a \cdot T_A'\,\mathrm{d}a + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{aB} \cdot a \cdot T_B'\,\mathrm{d}a \right)$$
$$= 0 \quad \forall\, \alpha \ge \alpha^* \quad \text{(A8b)}$$

$$\frac{\partial \Lambda}{\partial \tau_A} = \int_{\alpha_{\min}}^{\alpha^*} N_{\alpha A}\,\mathrm{d}\alpha + \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A}\,\mathrm{d}\alpha$$
$$+ \frac{1}{(T_A - T_B)} \cdot N_{\alpha^*} \cdot \left((\tau_B - \tau_A) - \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A} \cdot \alpha \cdot T_A'\,\mathrm{d}\alpha + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B} \cdot \alpha \cdot T_B'\,\mathrm{d}\alpha \right) = 0 \quad \text{(A9a)}$$
$$\textit{iff} \quad \delta_A = 1$$

$$\frac{\partial \Lambda}{\partial \tau_B} = \int_{\alpha^*}^{\alpha_{\max}} N_{\alpha B}\,\mathrm{d}\alpha + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B}\,\mathrm{d}\alpha$$
$$- \frac{1}{(T_A - T_B)} \cdot N_{\alpha^*} \cdot \left((\tau_B - \tau_A) - \int_{\alpha_{\min}}^{\alpha^*} \lambda_{\alpha A} \cdot \alpha \cdot T_A'\,\mathrm{d}\alpha + \int_{\alpha^*}^{\alpha_{\max}} \lambda_{\alpha B} \cdot \alpha \cdot T_B'\,\mathrm{d}\alpha \right) = 0 \quad \text{(A9b)}$$
$$\textit{iff} \quad \delta_B = 1$$

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:

$$\frac{\partial \Lambda}{\partial N_\alpha} = \tau_C + \int_{\alpha_{\min}}^{\alpha_{\max}} \lambda_a \cdot a \cdot \left(T_C' + T_D'\right) \mathrm{d}a - \lambda_\alpha \cdot D_\alpha' = 0 \quad \forall \alpha \tag{A10a}$$

$$\frac{\partial \Lambda}{\partial \tau_C} = \int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \,\mathrm{d}\alpha + \int_{\alpha_{\min}}^{\alpha_{\max}} \lambda_\alpha \,\mathrm{d}\alpha = 0 \tag{A10b}$$

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

$$\tau_C = \frac{\int_{\alpha_{\min}}^{\alpha_{\max}} N_\alpha \,\mathrm{d}\alpha}{\int_{\alpha_{\min}}^{\alpha_{\max}} \frac{1}{-D_\alpha'} \,\mathrm{d}\alpha} \cdot \left(1 + \left(T_C' + T_D'\right) \cdot \int_{\alpha_{\min}}^{\alpha_{\max}} \frac{\alpha}{-D_\alpha'} \,\mathrm{d}\alpha\right) \tag{A11}$$

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 decreases with the elasticity of demand (the monopolistic mark-up), and increases with the marginal external congestion costs.