THE GENERALIZED SECOND-BEST NETWORK CONGESTION PRICING PROBLEM

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

This version: 30/06/00

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

Abstract

This paper considers the second-best problem where not all links of a congested transportation network can be tolled. The paper builds on earlier work in which the second-best tax rule for this problem was derived for general static networks, so that the solution presented is valid for any graph of the network, and for any set of tolling points available on that network. The solution is now applied in an illustrative simulation model, in which various second-best problems can be studied that might arise with the implementation of different archetype pricing schemes. Apart from the benchmark of first-best pricing, these include for instance a toll-cordon, parking policies in the city centre, and pay-lanes and 'free-lanes' on major roads feeding into the city. An exploratory analysis is given of a possible method for selecting the optimal location of toll points in case not all links can be tolled.

*The author is affiliated to the Tinbergen Institute, Keizersgracht 482, 1017 EG Amsterdam. The research of Erik Verhoef has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

1. Introduction

Second-best issues in transport regulation have received ample attention in the recent literature. This is often motivated by the observation that the first-best policy for a congested road network – tolls equal to marginal external costs on each individual link – is a rather theoretical construct. Various considerations often lead transport regulators to consider second-best solutions only, in which not every single link of a transport network can be tolled. Such considerations may include the costs for additional tolling points, such as the equipment required in case of electronic toll collection, as well as the possible desire to start with demonstration projects before implementing road pricing on a system-wide scale. Examples of the resulting second-best congestion pricing schemes include 'pay-lanes', such as used at various sites in the US, and 'toll cordons' around city centres, which have for instance been in operation in various Scandinavian cities (see Small and Gomez-Ibañez, 1998, for a recent review of applications of and experiments with road pricing). With a sufficiently broad definition of the concept of a transport network, however, also parking charges can be regarded as an example of the same type of problem: public parking space could be considered as a tolled link in a network, parallel to an untolled link representing free private parking. And, when adding 'virtual links' to a network, which involve no actual travelling but only a possible toll, even policies like area permits appear to belong to the same class of second-best congestion pricing problems.

A classic example of this type of second-best problems concerns the two-route problem, where an untolled alternative road is available parallel to a toll road. This problem has for instance been studied by Lévy-Lambert (1968), Marchand (1968), and more recently Braid (1996), Verhoef, Nijkamp and Rietveld (1996). De Palma and Lindsey (1999), and Verhoef and Small (1999), considered the same network for investigating various related second-best problems. Glazer and Niskanen (1992) study second-best optimal parking fees for a city centre where through-traffic as well as road users with access to private parking places cannot be charged. And, as a somewhat different type of second-best problem, Braid (1989) and Arnott, De Palma and Lindsey (1990) consider 'flat' single valued tolls for a dynamic bottleneck. A recurring result is that second-best tax-rules – set so as to maximize social welfare given the persistence of the second-best distortion – are generally different from the simple Pigouvian rule (Verhoef, Nijkamp and Rietveld, 1995).

Verhoef (2000) offered a general solution for the second-best problem where not all links of a congested transportation network can be tolled. The solution is 'general' in the sense that it is valid for any possible graph of the network, for any possible sub-set of links that can be tolled, and with elastic origin-destination (OD) demands. Yan and Lam (1996), without deriving or using the analytic solution to the problem, discussed algorithms to find the solution to a similar problem in the context of inelastic demands, but allowing for stationary state queues.

In this paper, the solution proposed by Verhoef (2000) is tested and explored further using a medium-sized network. The network used is primarily designed to capture the most important types of network spill-overs that may be relevant in the design of second-best congestion pricing schemes, while limiting the number of links so that the possible danger of the complexity of the network clouding essential insights is minimized. However, at the same time, to secure some immediate policy relevance, a network configuration is used that can describe some archetype second-best policies in a reasonably accurate way. In particular, attention will be paid to area licences, parking charges, pay-lanes and ‘free-lanes’ (being the mirror image of pay-lanes, so that instead of only one lane on a highway being tolled, only one lane remains untolled), toll-rings, and of course first-best pricing. The network could be considered as an admittedly very abstract representation of the current morning peak situation in and around a city like Amsterdam, a city for which various types of congestion pricing schemes have recently been proposed, some of which still being under serious consideration.

Apart from illustrating the general methodology of determining optimal second-best tolls for a given set of toll-points, the paper will also consider the important question of which links to select in case only a limited number of toll points can be selected. A simple procedure for selecting the optimal toll points is proposed, which works surprisingly well in the simple network used. The paper will give ample attention to practical aspects, such as the efficiency and speed of algorithms and indicators, that will be important when applying the second-best tolls and link-selection procedures in larger transport network models.

The plan of the paper is as follows. Section 2 briefly reviews the procedure for determining second-best optimal tolls in a generalized network, as proposed by Verhoef (2000). Section 3 presents the simulation model, and discusses the qualitative properties of various second-best optima resulting from the archetype pricing schemes mentioned above. Section 4 proceeds by considering the problem of selecting optimal toll points. Section 5 concludes and gives some directions for further research.

2. A general characterization of the problem[1]

The analysis in this section pertains to a general transportation network $\mathcal{G}$ with continuous numbers of users. This network consists of a set of nodes and a set of directed links (arcs). Any pair of distinct nodes can be an origin-destination (OD-)pair, and the demand for trips between such an OD-pair is not restricted to be perfectly inelastic. Apart from having a possibly different willingness to pay for making a trip, and possibly different nodes of origin and destination, all (potential) users of the network are assumed to be identical. The following notation will be used (where primes denote derivatives):

$\mathcal{N}$ the set of nodes in the network

$\mathcal{I}$ the set of OD-pairs, denoted i=1,…,I

N_i the continuous number of users (or OD-flow) for OD-pair i, with $N_i \geq 0$

$D_i(N_i)$ the inverse demand function for trips for OD-pair i, with $D_i' \leq 0$

$\mathcal{J}$ the set of directed links in the network, denoted j=1,…,J

N_j the continuous number of users (or link-flow) on link j, with $N_j \geq 0$

[1] This section draws heavily from Verhoef (2000).

$c_j(N_j)$ the average cost function for the use of link j, with $c_j'\geq 0$

$\mathcal{P}$ the set of non-cyclical paths in the network, denoted p=1,…,P

N_p the continuous number of users (or path-flow) for path p, with $N_p\geq 0$

$\mathcal{P}_i$ the set of non-cyclical paths for OD-pair i, denoted p_i=1,…,P_i

δ_{jp} a dummy that takes on the value of 1 if link j belong to path p, and a value of 0 otherwise

δ_j a dummy that takes on the value of 1 if a toll can be charged on link j, and a value of 0 otherwise

f_j the level of the toll on link j if δ_j=1

δ_{ip} a dummy that takes on the value of 1 if p$\in\mathcal{P}_i$ and $\sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j(N_j)+\delta_j\cdot f_j\right)-D_i(N_i)\leq 0,$ and a value of 0 otherwise

Most of these variables are self-explanatory; the last dummy δ_{ip} can be interpreted as a dummy identifying (when equal to 1) the 'relevant paths' in a network: those paths for which the equilibrium cost level is equal to the minimum possible equilibrium costs for OD-pair i (see below for further explanation). It is assumed that that all relevant functions $D_i(N_i)$ and $c_j(N_j)$ are continuous and smooth. The cost functions represent generalized user costs including monetized time costs, and are upward sloping in case of congestion. In the analysis below, congestion is assumed to be link-specific. In case of a dynamic generalization of the present model, for instance based on Vickrey's (1969) model of bottleneck congestion, account should indeed be taken of the possibility that in case of an arrival rate of users at the tail of a link exceeding its capacity, queuing will occur, and will directly affect the cost levels at upstream links. For a static model, however, which by definition cannot give a meaningful representation of cases where arrival rates exceed capacities anyway (Verhoef, 1999), the assumption that congestion is link-specific may often be acceptable.

Since we are dealing with a static network, the use of a link is defined as:

$$N_j=\sum_{p=1}^{P}\delta_{jp}\cdot N_p \tag{1}$$

An important equilibrium concept is Wardrop's (1952) first principle, stating that for every OD-pair i the costs for used paths must be the same and that there are no unused paths with strictly lower costs. For the general case where the demand functions $D_i(N_i)$ are not necessarily perfectly inelastic, this can be represented according to the following complementary slackness equilibrium conditions (see, for instance, Smith, 1979):

$$N_p\geq 0;\quad \sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j+\delta_j\cdot f_j\right)-D_i\geq 0\quad\text{and}\quad N_p\cdot\left(\sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j+\delta_j\cdot f_j\right)-D_i\right)=0$$
$$\forall\ p\in\mathcal{P}_i \tag{2}$$

(the arguments in the cost and demand functions are dropped whenever this does not lead to confusion). Compared with the case of inelastic demands, equation (2) therefore adds the economic equilibrium principle that marginal benefits should be equal to marginal private costs to the standard Wardrop (1952) condition. The fact that Wardrop's principle allows a formulation of network problems in terms of variational inequalities (Kinderlehrer and Stampacchia, 1980) has been recognized by for instance Dafermos (1980) and Nagurney (1993). Inspection of (2) reveals that the dummy variable δ_{ip} discussed earlier takes on the value of 1 only if path p from the set $\mathcal{P}_i$ is among those that may be used in the equilibrium by travellers between OD-pair i. Such paths with $\delta_{ip}=1$ will be called 'relevant paths' in the sequel. However, for some of the relevant paths, N_p actually still may be equal to zero in the equilibrium, as will become clear when the uniqueness of the various variables in an equilibrium is considered below. First, however, a final identity can be given, equating the usage for a OD-pair to the sum of usage on all relevant paths connecting that OD-pair:

$$N_i = \sum_{p=1}^{P} \delta_{ip} \cdot N_p \tag{3}$$

Under rather general conditions, a transportation network as described above can be expected to have a unique equilibrium in OD-flows (the vector $\mathbf{N_i}$) and link-flows (the vector $\mathbf{N_j}$) for a given set of tolls f_j, in particular if $D_i'(N_i)<0$ and $c_j'(N_j)>0$ for all relevant i and j over the relevant ranges (see, for instance, De Palma and Nesterov, 1998). It will be assumed throughout this paper that such a unique solution exists. However, this does not imply that the solution will be necessarily unique also in path-flows (the vector $\mathbf{N_p}$), nor in (first-best or second-best) optimal toll levels (the vector $\mathbf{f_j}$) (Dafermos, 1973). Path flows are for instance not unique when users from different OD-pairs share a part of the network where they can choose between two parallel links with equal travel costs. Evidently, interchanging two users – one from each parallel link – will then leave the equilibrium in terms of link-flows and OD-flows intact, but may alter the equilibrium in terms of path-flows. Next, tolls may not be unique when, for instance, on an intersection of tolled links, no ('active') origin or destination node is located. A constant can then be added to the tolls on the links feeding into the intersection, and subtracted from the tolls originating from that intersection, without changing the equilibrium (see also Verhoef, 2000).

We now turn to the problem of finding the second-best optimal congestion tolls in the case that tolls can be charged only on a given subset of links. As a matter of fact, the first-best problem where tolls can be charged on all links is, of course, a special case of this general second-best problem. It is assumed that, given the second-best constraint, the regulator sets tolls so as to maximize social welfare, defined as total benefits minus total costs. Benefits are determined according to the Marshallian measure. The regulator therefore has to solve the problem that can be represented by the following Lagrangian:

$$\begin{aligned}\Lambda = &\sum_{i=1}^{I} \int_0^{\sum_{p=1}^{P}\delta_{ip}\cdot N_p} D_i(x_i)dx_i - \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P} \delta_{jp}\cdot\delta_{ip}\cdot N_p\cdot c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{jq}\cdot\delta_{kq}\cdot N_q\right)\\ &+\sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{ip}\cdot\lambda_p\cdot\left[\sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{jq}\cdot\delta_{kq}\cdot N_q\right)+\delta_j\cdot f_j\right)-D_i\left(\sum_{q=1}^{P}\delta_{iq}\cdot N_q\right)\right]\end{aligned} \tag{4}$$

The first set of terms represent total benefits, summed over all OD-pairs; note that the total OD-flow is determined according to (3). The second set of terms represent total costs, summed over all links in the network; note that the total link-flow is determined according to (1). The third set of terms represent the constraints caused by the equilibrium conditions that for each relevant path, the marginal benefits will be equal to the average costs plus the fees incurred on the links making up that path. Note that these constraints are consistent with (2), and that λ_p denotes the Lagrangian multiplier associated with the constraint for path p. The inclusion of the dummies δ_{ip}, or δ_{iq} when the index q is used to denote paths for notational reasons, secures that in the determination of the necessary first-order conditions for a local optimum only the relevant paths are considered (note that, also for notational reasons, the index k, when used, denotes OD-pairs). The following first-order conditions can be derived (where arguments in demand and cost functions are dropped for notational convenience):

$$\begin{aligned}\frac{\partial \Lambda}{\partial N_p} &= \sum_{i=1}^{I} \delta_{ip} \cdot D_i - \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \sum_{k=1}^{I} \sum_{q=1}^{P} \delta_{jq} \cdot \delta_{kq} \cdot N_q \cdot c'_j \right) \\ &+ \sum_{k=1}^{I} \sum_{q=1}^{P} \delta_{kq} \cdot \lambda_q \cdot \left(\sum_{j=1}^{J} \delta_{jp} \cdot \delta_{jq} \cdot c'_j \right) - \sum_{i=1}^{I} \sum_{q=1}^{P} \delta_{ip} \cdot \delta_{iq} \cdot \lambda_p \cdot D'_i = 0 \quad \forall\ p \text{ with } \delta_{ip} = 1\end{aligned} \tag{5}$$

$$\frac{\partial \Lambda}{\partial f_j} = \sum_{i=1}^{I} \sum_{p=1}^{P} \delta_{ip} \cdot \delta_{jp} \cdot \lambda_p = 0 \quad \forall\ j \text{ with } \delta_j = 1 \tag{6}$$

$$\frac{\partial \Lambda}{\partial \lambda_p} = \sum_{j=1}^{J} \delta_{jp} \cdot \left(c_j + \delta_j \cdot f_j \right) - \sum_{i=1}^{I} \delta_{ip} \cdot D_i = 0 \quad \forall\ p \text{ with } \delta_{ip} = 1 \tag{7}$$

Notwithstanding the fact that the second-best equilibrium may not be unique in path-flows as pointed out above, equations (5) show that the first-order conditions with respect to path-flows are used to solve the problem. Path-flows give the necessary connection between the benefit side (in terms of OD-flows) and the cost side (in terms of link-flows) in the model. It may in particular be noted that the value of the derivative in (5) is independent of the specific distribution of users from a given OD-pair over the various possible paths, as long of course as the equilibrium conditions shown in equation (2) hold, since the relevant terms only depend on either OD-flows or link-flows, which will all remain the same for any of the possible equilibria in terms of path-flows.

Verhoef (2000) considered the analytical solution for the system of equations (5)-(7), which, as can be expected, turns out to involve tedious expressions. The Lagrangian multipliers λ_p play an important role in the solution. These multipliers can be interpreted as the ‘shadow price of non-optimal pricing’ in the second-best optimum – which in fact follows directly from the specification of the Lagrangian (4). Under first-best pricing, these multipliers would each be equal to zero. Under second-best pricing, the tolls that can be controlled are set in such a way that the sum of the multipliers that can be directly affected is zero. For further details and interpretation, see Verhoef (2000).

For the application of (5)-(7) in larger networks, no such analytical expressions for the multipliers λ_p and tolls f_j have to be used. Instead, one can rely on the numerical solution of the system of equations that follows from substitution of (7) into (5) for each relevant path:

$$\begin{aligned}&\sum_{j=1}^{J}\delta_{jp}\cdot\left(\delta_j\cdot f_j-\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{jq}\cdot\delta_{kq}\cdot N_q\cdot c'_j\right)+\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{kq}\cdot\lambda_q\cdot\left(\sum_{j=1}^{J}\delta_{jp}\cdot\delta_{jq}\cdot c'_j\right)\\&-\sum_{i=1}^{I}\sum_{q=1}^{P}\delta_{ip}\cdot\delta_{iq}\cdot\lambda_p\cdot D'_i=0\quad\forall\ p\text{ with }\delta_{ip}=1\end{aligned}\tag{8}$$

$$\frac{\partial\Lambda}{\partial f_j}=\sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{ip}\cdot\delta_{jp}\cdot\lambda_p=0\quad\forall\ j\text{ with }\delta_j=1\tag{9}$$

(where (9) is identical to (6)). For a given network equilibrium in terms of use levels N, equations (8) and (9) define a system in a number of linear equations equal to the number of relevant paths plus the number of relevant tolls, in the same amount of unknowns (the λ_p's and f_j's). A general algorithm for finding a second-best equilibrium would be:

1. compute, for a given network equilibrium with given use levels, consistent with given second-best tolls (possibly 0 in the first iteration), the solution to the system (8)-(9);
2. implement the implied tolls f_j in the network to find a new network equilibrium.

Steps 1 and 2 can then be repeated until convergence.

3. A numerical simulation model

3.1. Description of the network and the non-intervention equilibrium

In this section, the solution to the generalized second-best congestion pricing problem presented above is tested and explored further using a medium-sized network. The network is primarily designed to capture the most important types of network spill-overs that may be relevant in the design of second-best congestion pricing schemes, while limiting the number of links so that the possible danger of the complexity of the network clouding essential insights is minimized. However, at the same time, a network configuration is used that can describe some archetype second-best policies in a reasonably accurate way. The resulting network consists of 10 links, 3 of which are virtual, and is depicted in Figure 1.

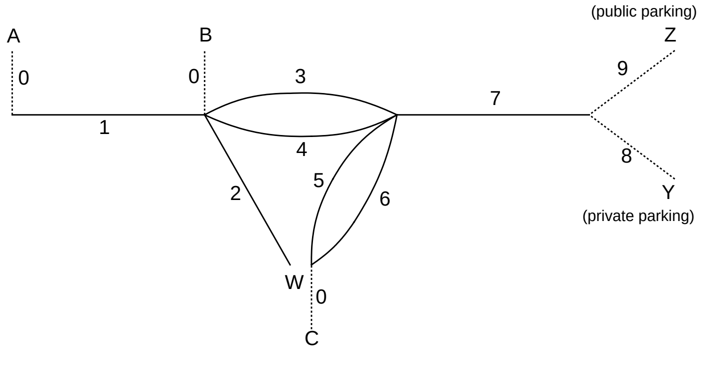

Figure 1. The network used for the simulations

The network has 3 origin-nodes: A, B and C, and two ‘real’ destinations: W and Y&Z. The latter denotes the bigger city, which is split into two possible destination-nodes: Y and Z. This distinction is made to enable consideration of public parking charges, which will only affect that part of the traffic using public parking space (node Z), as opposed to private parking space (Y). We thus have 8 OD-pairs: AW, AY, AZ, BW, BY, BZ, CY and CZ.

Links 1-7 are real links; the dotted links 0, 8 and 9 are ‘virtual links’, on which no real traffic costs are incurred and only possibly a toll is charged. Links 8 and 9 denote the use of a private versus public parking space, respectively. This is assumed to involve a negligible amount of travelling; hence the choice for representing this part of the trip by a virtual link. Note that the representation conveniently assumes that people cannot switch from public to private parking, or *vice versa*; a formulation with a virtual link connecting Y and Z could be used to endogenize the choice of public versus private parking. Link 0, attached to every possible origin node, represents the possibility of using area licences: a fixed toll for travelling, independent of the route and length of the trip followed. Finally, two pairs of parallel links are included in the network: 3&4, and 5&6. Such pairs could either represent the existence of minor roads parallel to highways, or – as will be the case in the simulations below – could be used to investigate pay-lanes or free-lanes. For every trip terminating in either Y or Z, there are therefore two possible routes, and the total number of paths in the network is equal to 14. Table 1 shows the incidence between OD-flows, path-flows and link-flows in the simulation model.

OD-pairs	AW	AY		AZ		BW	BY		BZ		CY		CZ	
	(40, .035)	(55, 0.045)		(55, 0.045)		(20, .013)	(35, 0.02)		(35, 0.02)		(35, 0.02)		(35, 0.02)	
Links:	AW	AY3	AY4	AZ3	AZ4	BW	BY3	BY4	BZ3	BZ4	CY5	CY6	CZ5	CZ6
Paths:														
0 (-,-)	*	*	*	*	*	*	*	*	*	*	*	*	*	*
1 (2.5, 0.001)	*	*	*	*	*									
2 (2.5, 0.001)	*					*								
3 (2.5, 0.002)		*		*			*		*					
4 (2.5, 0.0006667)			*		*			*		*				
5 (2.5, 0.002)											*		*	
6 (2.5, 0.0006667)												*		*
7 (2.5, 0.003)		*	*	*	*		*	*	*	*	*	*	*	*
8 (-,-)		*	*				*	*			*	*		
9 (-,-)				*	*				*	*			*	*

Note: Parameters of (link) cost functions and (OD) demand functions are given in brackets (intercept, slope)

Table 1. Incidence of OD-flows, path-flows and link-flows in the simulation model

The simulation network thus seems to capture the most important types of network complexities that would be relevant for various forms of second-best pricing. These include in particular: the existence of parallel connections (3&4 and 5&6); the use of serial links; the fact that drivers using a certain link may have trips with different lengths and routes, normally involving different total marginal external costs (*e.g.*, the users on 7); the fact that not all users of a network need to terminate their trips in the same city (W *versus* Y&Z) (which is in fact a variant of the previous complexity); the fact that some pricing policies may not affect all users with the same route in terms of real links (users on link 8 *versus* 9); and the existence of multiple paths for OD-pairs (all OD-pairs except AW and BW). At the same time, the network can describe some archetype second-best policies:

1. area licences, which can be represented with a toll on link 0;
2. parking charges, which can be represented with a toll on link 9;
3. pay-lanes, which can be represented with a toll on links 3 and/or 5, provided links 3 and 4 together (and 5 and 6 together) are modelled to represent a highway, with 3 (5) representing one lane;
4. 'free-lanes' (being the mirror image of pay-lanes, so that instead of only one lane on a highway being tolled, only one lane remains untolled), which can be represented with a toll on links 4 and/or 6 under the same modelling assumptions as under 3;
5. a toll-ring around the big city, implying an equal access fee for all users to destinations Y or Z, which can be represented with a toll on link 7.

For the simulation model, it is assumed that all demand and cost functions are linear. Under the assumed parameters (see Table 1), an equilibrium results which is characterized by the following OD-flows: N_{AW}=865, N_{AY}=901, N_{AZ}=901, N_{BW}=1188, N_{BY}=1285, N_{BZ}=1285, N_{CY}=1328, and N_{CZ}=1328. For the parallel routes, capacity ratios of 1:3 are assumed, yielding a 1:3 equilibrium route split. One out of four lanes would thus be the pay-lane or free-lane, when relevant. From the two origin nodes that have flows to both destination cities, more than two thirds of the traffic goes to the bigger city Y&Z. For convenience of checking results, the traffic going to the bigger city is assumed to be equally divided among public and private parking. Note that every real link has the same free-flow costs of 2.5[2], which with a value of time of 10 would mean 15 minutes of travelling. Travel costs in the non-intervention equilibrium are around one-and-a-half to two times as high, with values of c_1=5.17, c_2=4.55, c_3=c_4=4.69, c_5=c_6=3.83 and c_7=4.71. Demand elasticities in the non-intervention equilibrium are in the order of -0.3 to -0.35.

3.2. The welfare effects of some archetype second-best policies

Solving the set of equations defined by (8)-(9), and using the algorithm described just below (9), the optimal second-best tolls can be derived for any possible combination of links that can be tolled (the speed of convergence will be discussed in Section 3.3 below). Table 2 shows for 9 archetype tolling policies the main results: use levels relative to the use in the non-intervention equilibrium at the link- and OD-level, toll levels, and an efficiency index ω.

Before discussing the qualitative properties of these equilibria, it can be noted that the simulation model gives an opportunity to test the validity of the optimal tax rules that can be

[2] The model is calibrated to produce monetary values in Dutch Guilders (DFl). The exchange rate of the Dutch guilder in mid 2000 was approximately DFl 2.2≈€1≈$0.98.

derived from the system of equations (8)-(9) (see Verhoef, 2000), in a network that seems to capture the most important types of network complexities. A simple test was performed, involving small variations of each second-best toll level generated (keeping other tolls at the second-best optimal level, when relevant). The resulting welfare level was in all cases found to be below the level obtained in the relevant second-best optimum. This validates the optimality of the second-best taxes.

As an illustration, Figure 2 shows the welfare gain relative to the gain obtained in the second-best optimum with two free-lanes, for tolls varying from 0% to 200% of the second-best optimal levels. The centre of the diagram, with both tolls set optimally, is indeed the second-best welfare optimum. The figure further demonstrates that the objective function is strictly concave with respect to both tolls, and, as a result, relatively flat near the second-best optimum. This suggests that small errors in toll prediction are relatively unimportant; that is, a 1% further deviation has a greater negative impact on efficiency, the further the toll is from its second-best optimal value. Similar results were found for all other schemes considered.[3]

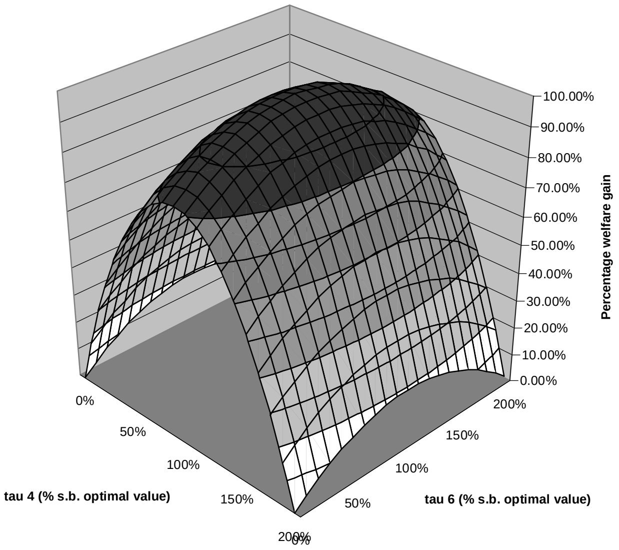

Figure 2. Relative welfare with varying tolls for the 'Two free-lanes' scheme

Despite the artificial character of the simulation model, some of the qualitative properties of the equilibria are worthy of some further elaboration, in particular because the specific second-best distortions arising with each of the archetype policies are typical for these policies, and are likely to occur also in more realistic networks. A meaningful assessment can be given by comparing the results of a second-best scheme to two bench-marks: the non-intervention equilibrium and the first-best optimum (which involves reductions in usage with 10-13% for all links and all OD-pairs; see Table 2). The relative performance of the various second-best

[3] The nearly perfect symmetry displayed in the figure is caused by the linearity of demand and cost functions, and is unlikely to carry over to more general formulations.

schemes is represented with an efficiency index ω, which gives the welfare gain relative to the gain that is achieved with first-best pricing.

	First-best	Two pay-	Two free-	Two	Highway	Highway	Toll-ring	Area	Parking
		lanes	lanes	highways	34	56		licences	charges
Tolled links	all	3,5	4,6	3,4,5,6	3,4	5,6	7	0	9
n_{AW}	0.881	1.000	1.001	1.005	1.005	1.000	1.005	0.903	1.002
n_{AY}	0.871	0.999	0.991	0.907	0.905	1.002	0.920	0.931	1.008
n_{AZ}	0.871	0.999	0.991	0.907	0.905	1.002	0.920	0.931	0.913
n_{BW}	0.896	1.000	1.000	1.000	1.000	1.000	1.000	0.797	1.000
n_{BY}	0.874	0.998	0.985	0.846	0.844	1.003	0.868	0.882	1.009
n_{BZ}	0.874	0.998	0.985	0.846	0.844	1.003	0.868	0.882	0.859
n_{CY}	0.899	0.999	0.993	0.899	1.006	0.894	0.869	0.884	1.008
n_{CZ}	0.899	0.999	0.993	0.899	1.006	0.894	0.869	0.884	0.863
n_0	0.884	0.999	0.992	0.909	0.939	0.970	0.909	0.883	0.955
n_1	0.874	0.999	0.994	0.939	0.938	1.001	0.947	0.922	0.974
n_2	0.890	1.000	1.000	1.002	1.002	1.000	1.002	0.842	1.001
n_3	0.873	0.927	1.185	0.871	0.869	1.002	0.889	0.902	0.945
n_4	0.873	1.022	0.922	0.871	0.869	1.002	0.889	0.902	0.945
n_5	0.899	0.943	1.151	0.899	1.006	0.894	0.869	0.884	0.935
n_6	0.899	1.018	0.941	0.899	1.006	0.894	0.869	0.884	0.935
n_7	0.883	0.999	0.990	0.882	0.921	0.962	0.882	0.895	0.941
n_8	0.883	0.999	0.990	0.882	0.921	0.962	0.882	0.895	1.008
n_9	0.883	0.999	0.990	0.882	0.921	0.962	0.882	0.895	0.874
f_0	redundant							3.459	
f_1	2.331								
f_2	1.827								
f_3	1.908	0.209		4.477	4.462				
f_4	1.908		0.574	4.477	4.462				
f_5	1.194	0.099		3.054		3.025			
f_6	1.194		0.280	3.054		3.025			
f_7	1.861						3.893		
f_8	redundant								
f_9	redundant								3.861
ω	1	0.009	0.072	0.806	0.607	0.195	0.780	0.882	0.387
ω^{elas2}	1	0.017	0.127	0.802	0.596	0.194	0.778	0.881	0.379

Notes: n is defined as use-level relative to use in the non-intervention equilibrium

ω is the 'index of relative welfare improvement': the welfare gain relative to the gain that is achieved with first-best pricing

ω^{elas2} is ω for the parametrization yielding the same non-intervention use levels at double demand elasticities for each OD-pair

Table 2. Characteristics of second-best optima for some archetype policies

By far the least efficient pricing variant considered turns out to be the imposition of two pay-lanes on the highways leading to the big city. Consistent with earlier findings (*e.g.* Verhoef and Small, 1999), relatively low second-best optimal tolls are found, yielding in the present case an efficiency improvement of less than one percent of the theoretically possible gains. Verhoef and Small (1999) have demonstrated that this is probably an underestimate, due to the neglect of differences in values of time in the present model, but even when the said underestimation would amount to a factor 9, as in the numerical model used by Verhoef and Small (1999), pay-lanes would still only yield a meagre less than 9% efficiency gain. The mirror-image of pay-lanes, a system of 'free-lanes' where only one lane of the highway remains untolled, performs markedly better with an ω of 0.07, but is still by far only the one-but-least-efficient tolling scheme. The reason that these policies are so inefficient is that they cause serious spill-overs to unpriced parallel routes: the policy variant where tolls apply on the full capacity of the highway(s) perform much better ('Two highways', 'Highway 34' and 'Highway 56'). Under the assumed conditions, the benefits of simultaneous pricing of parallel links of a highway are thus highly 'super-additive' (the benefits of joint implementation exceed the sum of the benefits of implementation in isolation), because the said distortionary spill-overs are avoided with pricing of the full capacity.

The fact that the efficiency gains for 'Highway 34' and 'Highway 56' are nearly additive is indicative for the limited degree of interaction between users of these highways: only on link 7. This interaction is reflected in the second-best toll levels, which are slightly lower when only the highway alone is tolled (4.46 and 3.02 for Highway 34 and 56, respectively), compared to the tolls applying when both are subject to tolling (4.48 and 3.05). This toll reduction reflects that a reduction of usage of one highway will, through reduced congestion on link 7, induce some extra traffic and hence extra congestion on the other highway.[4]

A specific variant of highway tolling is the toll-ring, which can be seen as highway pricing where the tolls for users of both highways are restricted to be equal. As a result, the toll differentiation of 4.48 on Highway 34 versus 3.05 on Highway 56 in 'Two Highways' is no longer possible, and an intermediate second-best optimal toll of 3.89 results. Under the assumed parameters, the resulting welfare loss is only limited: ω only reduces from 0.802 for 'Two highways' to 0.778 for 'Toll-ring'.

Closely related to 'Toll-ring', in fact, is 'Parking charges': a toll on link 9, carrying exactly half of the original users of link 7 in the no-toll equilibrium. Leading to the same type of second-best distortions as just described for single highways – tolling of users of the virtual link 9 induces extra use by users of the virtual link 8 through reduced congestion elsewhere in the network – this policy leads to a second-best toll and a welfare gain less (albeit slightly) than half that of 'Toll-ring': f=3.86 and ω=0.379.

Finally, the policy of 'Area licences', implying a non-differentiated toll of 3.46 for all users, turns out to be the most efficient second-best variant for the assumed network and parameters. Apparently, the inability to differentiate tolls among users is under the

[4] Note that when two parallel links are priced simultaneously, identical optimal tolls are invariably found. This is not the result of an exogenous constraint on these tolls, but results from the fact that with linear cost functions with equal intercepts, and with a single value of time, equalization of marginal costs (a property of optimality) means that also average costs are equalized. This, in turn, implies equal tolls (which are equal to the difference between marginal and average cost). When heterogeneity with respect to value of time were introduced, however, toll differentiation would become beneficial for efficiency (Verhoef and Small, 1999).

circumstances considered less detrimental to overall efficiency than network spill-overs that arise with the other types of second-best pricing considered, despite the considerable variation in total tolls that can be found for different paths in the first-best optimum (varying from 1.83 for users for OD-pair BW to 6.10 for users for OD-pairs AY and AZ).

The relative performance of the various schemes is of course crucially dependent on the assumed network configuration (*e.g.* in terms of availability of parallel routes, implied interactions, and relative lengths of links and paths) and parameters (*e.g.* implied capacities, relative use levels, cost- and demand elasticities). Under the assumed conditions, interactions between users from different OD-pairs do not seem too important a source of second-best distortions, which is exemplified by the performance of ‘Highway 34’ or ‘Highway 56’ relative to ‘Two highways’, and ‘Parking charges’ relative to ‘Toll ring’. On the other hand, distortions resulting from non-optimally parallel links seem a rather important source of second-best distortions; witness the relative performance of ‘Two pay-lanes’ and ‘Two free-lanes’ compared to that of ‘Two highways’. Both features could be the result of a relatively inelastic demand. In the first case, reduction of usage for one OD-pair would then invoke not too much use from other OD-pairs. In the second case, pricing on one parallel link will most importantly lead to extra use of the other link, rather than to a reduction in overall use (see also, in Table 2, the rows reflecting relative use levels).

It is therefore worthwhile to see to what extent the results change when the non-intervention demand elasticities are doubled, by simultaneously changing the slopes and intercepts of the demand functions, keeping all other parameters and equilibrium use levels constant. The first-best equilibrium in this case involves reductions in use between 16 and 21% of the no-toll use levels, for each OD-pair and for each link. The bottom row in Table 2 shows the ω’s under those revised demand elasticities. It turns out that the effect is relatively strongest for the parallel link charges, with an increase of the relevant ω’s to a level of around a factor two of the original level – implying, however, still rather low levels of relative efficiency gains. For the other schemes, the effects are minimal.

The results thus seem reasonably robust for changes in demand elasticities, and seem to be driven primarily by the assumed network configuration, the base-case use levels and the base-case congestion levels. Insofar as these would be considered representative for an existing network, the ω’s presented may provide a reasonably accurate impression of the relative performance of the second-best tolling schemes considered. In particular, it can be noted that the two sets of demand elasticities considered define a range (from –0.3 to –0.6) that is generally considered representative for morning peak road usage.

It might be hypothesized that the assumption of linear cost functions would systematically discriminate against the parallel link tolling schemes considered. Linear congestion functions may underestimate the welfare gains from the first marginal reductions in road usage, if the cost functions are actually steeper near the non-intervention use levels than is assumed with linear cost functions. However, at the same time, this would imply that the additional costs from adverse route switching due to parallel link pricing would also be underestimated using linear cost functions. Therefore, it is doubtful whether this assumption would really strongly discriminate against these schemes, in particular when bearing in mind the limited efficiency gains from such policies that are predicted in other models, using other types of cost functions (*e.g.* Verhoef and Small, 1999). As indicated earlier, a stronger source of underestimation of the benefits of parallel route pricing probably results from the assumed homogeneity of users with respect to the value of time.

In conclusion, this section has shown that the second-best tax rules implied by the set of equations (8)-(9) indeed lead to a second-best optimum. The qualitative properties of the second-best equilibria described here capture the most important sources of distortions associated with each of the archetype policies considered. The quantitative properties, summarized in the efficiency index ω, are of course valid only for the assumed network and parameters.

3.3. Convergence

From a modelling perspective, an important question concerns the speed of convergence of the algorithm described just below (9). Whereas the small network considered here easily allows 1000 iterations or more within the time span of one minute on a modern PC, things become different when the second-best taxes are to be calculated for a large empirical network model. It will then often be important to restrict the number of iterations needed to find that network equilibrium for which the second-best tolls (and Lagrangian multipliers) satisfy equations (8) and (9).

Iteration	*Pay-lane 3*	*Pay-lane 3*#	*Highway 34*	*Parking charges (link 9)*	*Area licence (link 0)*	*First-best (sum of tolls)*	*Serial links 79*	
							f$_7$	*f*$_9$
0	0	0	0	0	0	0	0	0
1	0.4148	0.2074*	4.9590	4.0860	3.8840	13.8576	4.3798	0.0000*
2	0.0077	0.2093	4.4071	3.8474*	3.4071	12.0033	3.8317	..
3	0.4072	..	4.4686*	3.8614	3.4656*	12.2540*	3.9003*	..
4	0.0151	..	4.4617	3.8605	3.4585	12.2200	3.8917	..
5	0.3999	..	4.4625	3.8606	3.4593	12.2246	3.8927	..
6	0.0223	..	4.4624	..	3.4592	12.2240	3.8926	..
7	0.3929	..	..	..	..	..	..	..
8	0.0292	..	..	..	..	..	..	..
9	0.3862	..	..	..	..	..	..	..
10	0.0358	..	..	..	..	..	..	..

Note: The underlined value marks the iteration for which the toll has reached its equilibrium value in whole cents. A column ends when the toll does no longer change at 4 digit precision.

*Iteration for which the toll is within ± 1% of equilibrium value.

#Alternative in which after each iteration, instead of using the newly predicted toll level, the average of this newly predicted toll level and the previous toll level is used to calculate the next network equilibrium.

Table 3. Convergence of the general algorithm

Table 3 shows the sequences of tolls in subsequent iterations that were found for a number of tolling schemes, partly overlapping with those in Table 2, starting with initial tolls equal to zero. The starred values in the columns denote the iteration for which the toll (and the sum of tolls in case of first-best pricing) has approached the second-best equilibrium value within an accuracy limit of ± 1%. For all but one policies (Pay-lane 3, see below), this is the case after three iterations or less. Usually, one iteration later the toll has reached its equilibrium value in whole cents – a very practical measure of accuracy, which however of course strongly depends on local circumstances, among which the currency used. Interestingly, even for a policy with two serial links (7 and 9), the algorithm converges very rapidly, even though such a policy may seem prone to inefficiently slow convergence due to the possible danger of alternating higher and lower toll levels for the two serial links.

Table 3 thus shows that the use of the system of equations (8)-(9) and the simple algorithm described just below (9) provide a rather efficient way of finding second-best optimal tolls in most cases, that probably cannot easily be improved upon – unless, of course, the mathematical software used could cope with the original set of (possibly partly non-linear) equations (5)-(7), implying that a second-best equilibrium could be found in one calculation. The intuition behind this efficiency is that in each iteration, all relevant information affecting the second-best level of a toll, insofar as available in the computed equilibrium, is used as efficiently as possible to predict new toll levels. This, for instance, secures that f_9 is kept at a level equal to zero in the serial links example. As demonstrated by ‘Highway 34’, ‘First-best’ and ‘Serial Links 79’, this picture does not change when multiple instead of single tolls are used. Important qualifications, however, are that convergence is likely to require more iterations when the network becomes larger (although the impact of a given toll decreases with the distance from the tolled link) and, in particular, when demand and cost functions are non-linear. In particular, in each iteration, the predicted tolls are based on the local slopes of demand and cost functions (compare (8)-(9)). The error in the prediction may be expected to increase when these slopes are not constant.

As stated, an exception to the rather efficient performance is given by the pay-lane policy, for which the ± 1% accuracy limit would only be reached after no less than some 250 iterations, during which the predicted toll consistently alternately overshoots and undershoots the second-best equilibrium level. Apart from being a particularly inefficient policy, pay-lanes thus also make the general algorithm particularly inefficient (comparable patterns were found for the other pay-lane and for free-lanes). The inefficiency of the algorithm can be repaired by using a minor variation on the general algorithm, in which after each iteration instead of using the newly predicted toll level, the average of this newly predicted toll level and the previous toll level is used to calculate the next network equilibrium. As shown in the second column, the toll is then predicted within a ± 1% accuracy range already after the first iteration. This procedure – or yet another variation thereof – is likely to be useful for larger networks, too. Fortunately, links exhibiting this type of behaviour can be identified after 2 iterations already, because of the low value of the second predicted toll relative to the first prediction.

4. The optimal selection of toll-points

In the practical design of second-best tolling schemes, a question probably equally important to that of finding the second-best optimal tolls for a given set of toll-points, involves the question of which links to toll in the first place. Especially in larger networks, this may not be easy to decide on the basis of logical reasoning, if anything due to the large number of possibilities,[5] and to the complicated interactions that may exist between tolls on different links. It is therefore interesting to consider the question of whether on the basis of the information available in the existing equilibrium, it is possible to predict which link(s) would be the 'best' one(s) to select for the implementation of tolling points. 'Best', in this analysis, refers to the highest positive impact on social surplus; obviously, political, social or practical considerations may sometimes lead to the selection of different toll-points.

It should be emphasized that this analysis is relevant only for situations where it is impracticable to follow the obviously most reliable route of selecting optimal toll points, namely the calculation of the second-best optima for each possible combination of t tolls in a J-link network, and comparing the welfare levels to select the optimal combination.

This section deals with this problem of selecting the optimal t toll-points in a J-link network. The question of optimizing t itself is left aside; this problem could be formalized in a straightforward manner by considering the marginal cost for implementing an additional toll-point. Section 4.1 starts with the selection of a single first toll-point, and Section 4.2 proceeds with the selection of multiple toll-points.

4.1. The optimal first toll-point

Verhoef (2000) discussed possible indicators for the selection of a first toll-point in a network, which can be calculated on the basis of 'out-of-equilibrium' values of λ_p and f_j, as they can be approximated in the initial no-toll equilibrium. One of the hypotheses was that in practice, the product of two indicators, to be discussed below, may in fact perform best. In this section, this particular hypothesis will be considered further, and will be tested in the numerical network presented in the previous section. As will become clear below, the great advantage in a computational sense of the proposed procedure is that, instead of having to calculate J second-best network equilibria, only 1+J systems of linear equations will have to be solved.

The proposed indicator I_x predicts the welfare gain from implementing a second-best toll on link x, starting from the no-toll equilibrium, as half the product of two terms. The first is F_x, representing the level of the second-best toll f_x as it is predicted in the no-toll equilibrium. The second is L_x, representing the (marginal) impact on social surplus of a marginal increase in the toll on that link, evaluated in the no-toll equilibrium. Hence, $I_x=½\cdot F_x\cdot L_x$. The intuition behind the indicator I_x is simple. As L_x, to be defined precisely in equation (9′) below, represents for each possible toll level the (marginal) gain in social surplus due to marginal increases in the toll, the total gain in social surplus ΔW_x from using the toll optimally at a level f_x can be represented as:[6]

[5] With t tolls on an J-link network, the number of combinations is $\begin{pmatrix} J \\ t \end{pmatrix} = \frac{J!}{t!\cdot(J-t)!}$. For 3 tolls on a 100-link network, this already implies 161700 combinations, and for 50 tolls no less than $1.0\cdot 10^{29}$ combinations.

[6] For each possible toll level on link x, the term L_x consists of the sum of Lagrangian multipliers. Such multipliers in general represent the impact on the objective of a marginal loosening of the constraint, with

$$\Delta W_x = \int_0^{f_x} L_x(z)\,dz \tag{10}$$

The first-order condition in (9) implies that, evaluated in the second-best optimum, L_x according to (9′) will be equal to zero; *i.e.* L_x=0 in the upper limit of integration in (10). A simple linear approximation of ΔW_x as defined in (10) would therefore be ½·L_x·F_x. If the relative error of this prediction would be equal for all possible tolls, the indicator I_x would correlate perfectly with the welfare gains that can be realized using a toll on link x.

L_x can be calculated, for a given network equilibrium, as a variant on the first-order condition (9). Specifically, for a given initial network equilibrium, first define L_p (for each relevant path) as the value found after simultaneous solution of the set of equations:

$$\begin{aligned}&\sum_{j=1}^{J}\delta_{jp}\cdot\left(\delta_j\cdot f_j-\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{jq}\cdot\delta_{kq}\cdot N_q\cdot c_j'\right)+\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{kq}\cdot L_q\cdot\left(\sum_{j=1}^{J}\delta_{jp}\cdot\delta_{jq}\cdot c_j'\right)\\&-\sum_{i=1}^{I}\sum_{q=1}^{P}\delta_{ip}\cdot\delta_{iq}\cdot L_p\cdot D_i'=0\quad\forall\ p\ \text{with}\ \delta_{ip}=1\end{aligned}\tag{8′}$$

(where, starting with the no-toll equilibrium, δ_j=f_j=0 for all j). Note that the number of equations in the set (8′) is equal to the number of unkowns, L_p (note in particular that all use levels N_p are treated as given). Both are equal to the number of relevant paths in the initial equilibrium considered. L_p can thus be interpreted as the 'out-of-equilibrium value' of the Lagrangian multiplier λ_p used in (4). With L_p thus calculated for every relevant path, L_x for link x can then be found as the following variant on (9):

$$L_x=\sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{ip}\cdot\delta_{xp}\cdot L_p \tag{9′}$$

Next, F_x represents the predicted second-best optimal toll level on link x, if that link were the only link to be tolled in the entire network. Such a prediction can be found from the solution to the set of equations given by (8)-(9) with (9) included only for the link x considered, starting with zero tolls. F_x is therefore the same as the prediction of f_x in iteration 1 in Table 3.

Note that the possible indicator L_x alone, when calculated for the non-intervention equilibrium, would fail to take account of the specific distortions resulting from second-best tolls on a specific link, as the condition defining the optimal use of a specific second-best toll, (9), has not been used. Similarly, the possible indicator F_x alone only reflects the predicted toll level, without taking into account the predicted welfare gain that may be realized with it. Finally, as an aside, note that L_x can only be derived from (8′), and not from the λ_p's found in the first-iteration solution of (8)-(9), as in the latter case equation (9) would imply a zero value of L_x.

optimal adjustments in all other choice variables (in this case: the levels of use N). A marginal loosening of the constraint can in this case be interpreted as a marginal increases in the relevant toll; compare the formulation of the Lagrangian in (4).

	Link 0	*Link 1*	*Link 2*	*Link 3*	*Link 4*	*Link 5*	*Link 6*	*Link 7*	*Link 8*	*Link 9*
Rank ΔW_x	1	3	6	9	7	10	8	2	4	4
	(ω=0.88)	(ω=0.43)	(ω=0.13)	(ω=0.01)	(ω=0.06)	(ω=0.00)	(ω=0.01)	(ω=0.78)	(ω=0.39)	(ω=0.39)
Rank I_x	1	3	6	9	7	10	8	2	4	4
$\Delta W_x/I_x$ [a]	0.89	0.92	0.91	0.52	0.54	0.54	0.54	0.89	0.95	0.95
$\Delta W_x/I^*_x$ [b,c]	1.00	1.00	1.00	1.02	1.00	1.07	1.01	1.00	1.00	1.00
$\Delta W_x/I^{II}_x$ [d,e]	0.95	0.96	0.96	1.01	0.94	1.06	0.97	0.95	0.97	0.97

[a] Correlation coefficient ΔW_x and I_x: 0.9987

[b] I^*_x is I_x calculated using the true second-best toll (hence, $I^*_x=L_x\cdot f_x$)

[c] Correlation coefficient ΔW_x and I^*_x: 1.0000

[d] I^{II}_x is I_x calculated using the average of the predictions for f_x in the first and second iteration (hence, $I^{II}_x=L_x\cdot½\cdot(F_x^1+F_x^2)$)

[e] Correlation coefficient ΔW_x and I^{II}_x: 0.9999

Table 4. Performance of the toll-selection indicators I_x and I^{II}_x

Table 4 shows the performance of the toll-selection indicator I_x for each of the links in the network presented in Section 3. The ω's shown in the first row indicate that the efficiency gains vary considerably between the 10 possibilities, with ω varying from 0.002 for link 5 to 0.882 for link 0, so that the indicator can be tested for a wide variety of types of links and efficiency levels. I_x perfectly predicts the ranking of the 10 links. However, the relative size of the welfare gains is predicted less accurately, with a correlation coefficient of 0.9987.

Problems are caused, again, by the two sets of parallel links 3&4, and 5&6. The third row of Table 4 shows that, whereas links 0, 1, 2, 7, 8 and 9 exhibit a rather constant ratio between ΔW_x and I_x (between 0.89 and 0.95), the parallel links 3-6 have a distinctly lower (but again constant) ratio of 0.52-0.54. This is consistent with the findings in Table 3, which shows that F_x – the prediction of f_x in the first iteration – is much further above the true second-best optimal value of the toll for the parallel link 3 than for other links (similar patterns were found for links 4-6, not shown in Table 3). As a result, I_x would be overestimated, and $\Delta W_x/I_x$ underestimated, for links 3-6 (relative to the estimates for other links) in Table 4. This explanation is supported by the values of $\Delta W_x/I^*_x$ shown in the fourth row of Table 4, where I^*_x is I_x calculated using the true second-best toll (hence, $I^*_x=L_x\cdot f_x$), instead of using its first-iteration projected value F_x. I^*_x shows a much smaller variation than I_x, and is close to 1 for all links. As a result, the correlation coefficient between ΔW_x and I^*_x is 1.0000 (and the ranking is predicted perfectly). As an aside, note that the values of $\Delta W_x/I^*_x$ near 1 support the intuition given earlier for the indicator I_x.

On the basis of the small network used here, I_x therefore appears a relatively accurate indicator for the selection of a first toll point in a network. It predicts the ranking of links perfectly, and the relative welfare gains reasonably well. There is however a consistent overestimation of the predicted welfare gains for links having parallel connections. The indicator nevertheless takes on relatively low values for such links, so that the chances of accidentally selecting such a link are not too large. Nevertheless, this implies that in larger networks, the ranking of such links is likely to be consistently upwardly biased (that is, the predicted ranking will be closer to 1 than the actual ranking).

It should be re-emphasized that I_x can be calculated after the solution of only 1+J systems of linear equations: system (8′) once, and system (8)-(9) for each j, and requires the calculation of no network equilibria other than the initial no-toll equilibrium. The use of this indicator will therefore be attractive only if this procedure requires considerably less time than finding the optimal toll for each of the J links, which involves the calculation of a multiple of J of network equilibria, and an equal amount of solutions of systems of linear equations (8)-(9) to calculate new toll levels for each iteration – the multiplicative factor in 'the multiple of J' depending on the speed of convergence of the general algorithm for the particular network.

Especially for larger networks, a relevant question therefore is whether a pragmatic compromise between these two strategies can be found, which further improves the performance of the indicator I_x as shown in Table 4, without requiring the calculation of a multiple of J network equilibria. As the fourth row in Table 4 suggests, it would in particular be beneficial to have better estimates of f_x than the first-iteration predictor F_x.

One possibility would be to perform, for each link, the general algorithm for finding the second-best optimal toll described under equation (9) for 'one-and-a-half iteration', and to use in the calculation of I_x, instead of F_x, the average of F_x^1 (F_x as found in the first iteration, starting from the initial equilibrium), and the value F_x^2 found by solution of (8)-(9) starting from the equilibrium with a toll $f_x=F_x^1$. This implies a revised indicator I^{II}_x based on two toll estimates: $I^{II}_x=L_x\cdot½\cdot(F_x^1+F_x^2)$. The bottom row in Table 4 shows the results of using this revised indicator I^{II}_x. It performs markedly better than I_x, bringing the ratio $\Delta W_x/I^{II}_x$ for the pairs of parallel links 3&4 and 5&6 much more in line with those for the other links. As a result, the correlation coefficient between ΔW_x and I^{II}_x has gone up to 0.9999 (and the ranking is again predicted perfectly). A side-advantage of using the revised indicator I^{II}_x would be that it immediately allows the identification of links suffering from slowly converging second-best tolls, which will have a small ratio F_x^2/F_x^1. This is useful information, as it both may help designing an efficient algorithm for finding second-best optima (as explained in Section 3), and as it is useful information for procedures for selecting sets of links to be tolled, as will become clear in Section 4.2 below.

Finally, it should be emphasized that the indicators I_x and I^{II}_x both use linear predictions from the no-toll equilibrium. The linearity of cost and demand functions used in the simulation model may therefore lead to an overestimation of the accuracy of the indicator.

4.2. The selection of multiple toll-points

When, instead of a single toll, the optimal locations for a set of t>1 tolls have to be determined, the size of the problem increases rapidly, due to the sheer number of combinations that can be chosen from. The question is whether accurate procedures can be developed to identify the t links in a J-link network for which the implementation of t second-best tolls would lead to the highest possible efficiency gain, without having to go through the possibly enormous task of calculating J!/(t!·(J–t)!) second-best optima. On the basis of the discussion in the previous sub-section, three possible strategies can be identified, which will be discussed in order of increasing computational burden.

	link 0	*link 1*	*link 2*	*link 3*	*link 4*	*link 5*	*link 6*	*link 7*	*link 8*	*link 9*
link 0	0.88									
link 1	0.95	0.43								
link 2	0.91	0.46	0.13							
link 3	0.88	0.44	0.14	0.01						
link 4	0.89	0.46	0.19	0.61	0.06					
link 5	0.88	0.44	0.14	0.01	0.06	0.00				
link 6	0.88	0.45	0.15	0.02	0.07	0.19	0.01			
link 7	0.91	0.93	0.91	0.78	0.78	0.78	0.78	0.78		
link 8	0.89	0.67	0.52	0.39	0.41	0.39	0.39	0.78	0.39	
link 9	0.89	0.67	0.52	0.39	0.41	0.39	0.39	0.78	0.78	0.39

Note: The terms in diagonal cells denote relative efficiency with single-link tolling on the associated link

Table 5. Performance of the possible combinations of two toll-points (ω)

The analysis and discussion will be restricted to the situation where only two toll-points (denoted x and y) can be implemented on the network used in this paper. This allows consideration of the most important issues, while keeping the presentation manageable in the sense that only ½·10·9=45 possible combinations have to be considered. Moreover, with two toll-points, already 95% of the possible efficiency gains are achieved in the network used. Table 5 shows the relative performance for each of these combinations, expressed in the efficiency indicator ω, with the ω's in the 10 diagonal cells representing the performance of the relevant toll used in isolation, as a single toll. As the addition of an extra toll point can never lead to a reduction in social welfare – the extra toll can always be kept at a zero level, which for instance is optimal when adding a toll-point on link 8 or 9 to a second-best toll on link 7 – each diagonal cell has the minimum score for the associated row and column.

The five most favourable combinations are, in order of decreasing efficiency: (0,1), (1,7), and, *ex equo*, (0,2), (0,7) and (2,7). Inspection of the network reveals that the latter three possibilities should indeed be equivalent, each allowing differentiated tolls for the two main destinations W and Y&Z. The four least efficient combinations, not surprisingly, involve the four possible combinations of parallel links 3-6 for which the parallel-link problem is not avoided: (3,5), (3,6), (4,5) and (4,6). Whereas the most efficient combination already achieves 95% of the maximum possible welfare gains, the relative efficiency for these latter options does not exceed 8%. With the combinations considered, there thus seems to be sufficient differentiation in efficiency to test the performance of the three strategies for selecting the optimal combination.

4.2.1. Strategy 1: selecting the t toll-points with the highest score I_x for implementation in isolation

The simplest possible strategy for selecting the t>1 best performing toll-points would be to choose those t links that have the highest predicted scores I_x (or I^{II}_x, when available) for the implementation as a single toll-point. The simplicity of this procedure stems from the fact that interactions between tolls in a network are fully ignored. These interactions may in reality lead to ‘sub-additivity’ or ‘super-additivity’ of the benefits of implementing an extra toll-point. Sub-additivity (the benefits of joint implementation are smaller than the sum of the benefits of implementation in isolation) is likely to occur with tolls on serial links; compare for instance links 7 and 8 in Table 5. Super-additivity (the benefits of joint implementation exceed the sum of the benefits of implementation in isolation) may for instance result with tolls implemented on parallel links; compare for instance links 3 and 4 in Table 5.

This means that this strategy could only be reliable when the interaction between tolls is small. This could for instance be the case when the number of toll-points t is small relative to the number of links J, when the links with the highest scores I_x or I^{II}_x are sufficiently far apart to prevent strong interactions, and when these links do not suffer from reduced efficiency due to the existence of parallel routes so that biased scores I_x are unlikely to occur. Based on the indicator I_x (as opposed to I^{II}_x), the indicator $I_{xy}^{s1}=I_x+I_y$ (where s1 denotes ‘strategy 1’) would then predict the welfare gains of using toll-points x and y together.

The network considered in this paper is too small, and hence the interactions are too strong, to make this strategy appear very reliable. Using the indicator I_{xy}^{s1}, the following combinations are predicted as the most efficient ones (in order of decreasing predicted efficiency I_{xy}^{s1}): (0,7), (0,1), (0,8) and (0,9), and (1,7). The truly third option is ranked first, and the truly first option is ranked second. The correlation coefficient between predicted and true efficiency for all possible combinations of two toll-points is 0.9152, which is probably not high enough to make this indicator seem sufficiently reliable. The obvious advantage of this procedure, however, is that no additional calculations are needed, compared to those necessary for the selection of an optimal first toll-point. There are thus two reasons why this strategy may become relatively more attractive for larger networks: the implied savings in the amount of additional calculations (compared to strategies 2 and 3 below) become more significant, and the degree of interactions between tolls probably becomes less important, as a network becomes bigger.

4.2.2. Strategy 2: selecting toll-points one-by-one, taking previously selected tolls as given

A second possible strategy involves a step-by-step approach, in which the optimal next toll-point is selected *given* the selection of the previous toll-point(s), and *given* the second-best optimal toll level(s) applying in the second-best equilibrium with these previous toll(s) set optimally. After the determination of the second-best network equilibrium given the previous toll(s), this requires the same procedure and hence the same amount of calculations (minus the number of links already tolled) as the procedure for selecting the first toll-point, discussed in Section 4.1. Interactions between the existing toll(s) and the new toll are taken into account in a one-directional way: the (predicted) tolls for possible next toll-points are optimally adjusted to the existence and level(s) of the previous toll(s), which themselves are treated as given.

As a consequence, the score for a given combination of t tolls will generally depend on the order in which the tolls are assumed to be implemented; in particular, on the question of which one was the last toll-point added. For the assessment of the performance of this indicator in the simulation model, for each combination (x,y) the indicator I_{xy}^{s2} is determined for the sequence where first a toll is implemented on the link l for which I_l is higher. This mimics the order in which tolls would be selected in practice, using this indicator. Labelling the link with the first chosen toll x and the other link y, the indicator I_{xy}^{s2} is then defined as $I_{xy}^{s2}=\Delta W_x+I_{yx}$, where I_{yx} denotes the predicted welfare gain from implementing a second-best optimal toll on link y, keeping the toll f_x fixed at its previous second-best optimal level. The indicator thus defined correctly predicts (0,1) as the optimal combination of 2 toll-points for the network considered here, followed by (1,7) and (2,7). The combinations (0,2) and (0,7) are ranked fourth and fifth, whereas they in reality should share the third rank with (2.7). The correlation coefficient between predicted and true efficiency is 0.9798, which is considerably higher than for strategy 1.

This second strategy thus adds tolls one by one, by predicting the best next toll-point given the true second-best optimum when using the toll-points already selected. A possible drawback of this method would be that toll-points that seems relatively efficient in the beginning of this procedure may in fact become less attractive when the total number of toll-points t increases. Checks could be built in the procedure to account for this possibility. For instance, when the target number of tolls t is reached, a possible test would involve the removal of previously selected toll-points to see whether these links are still predicted as the most efficient option. Another drawback is that for the selection of t tolls, t-1 second-best network equilibria have to be calculated. A third drawback, already mentioned, involves the fact that interactions between tolls are only taken into account in a one-directional manner when predicting the next toll-point. This issue is dealt with more explicitly in the third possible strategy for selecting multiple toll-points.

4.2.3. Strategy 3: selecting the set of toll-points with the highest predicted score when implemented simultaneously

The third strategy is designed to fully account for interactions between tolls. This strategy would calculate for each possible combination of t possible toll-points in a J-link network the predicted efficiency gain from simultaneous implementation. This can be done by including (9) for each of the t links in the specific combination considered when solving (8)-(9) to find F_x and F_y. This results in predictions for tolls for the links considered that fully take account of interactions between the tolls. The predicted tolls can then be multiplied with L_x and L_y as resulting from the solution of (8′)-(9′). The resulting indicator $I_{xy}^{s3}=F_x \cdot L_x+F_y \cdot L_y$ is a straightforward generalization of I_x introduced in Section 4.1, and will therefore henceforth simply be denoted I_{xy}.

Among the three strategies considered, this one appears to perform best, with the first five combinations ranked perfectly, and a correlation coefficient of predicted and true welfare gains of 0.9987 – coincidentally the same value that was found for I_x applied to single links. Table 6 shows the performance of this indicator in terms of $\Delta W_{xy}/I_{xy}$. While for most combinations, this ratio is rather constant in the range 0.88-0.94, deviations are found again for combinations involving parallel links. Fortunately, however, by far not all combinations involving at least one of the links 3-6 have a $\Delta W_{xy}/I_{xy}$ ratio outside the range 0.88-0.94

mentioned. In contrast, four out of the in total only five ‘strongly deviating ratios’, marked in bold in Table 6, involve paired combinations of one link of Highway 34 and one of Highway 56. Despite the upward biased predictions of efficiency gains of using these combinations, these five problematic combinations are ranked 37, 42, 43, 44 and 45 (out of 45) by the indicator I_{xy}, which makes the erroneous selection of such combinations highly unlikely. Moreover, when links suffering from close parallel substitutes are already identified earlier, as suggested in Section 4.1, an extra safety check could easily be built into the procedure to identify less reliable predictions.

	link 0	*link 1*	*link 2*	*link 3*	*link 4*	*link 5*	*link 6*	*link 7*	*link 8*
link 1	0.88								
link 2	0.89	0.91							
link 3	0.89	0.91	0.87						
link 4	0.89	0.88	**0.75**	0.90					
link 5	0.89	0.91	0.90	**0.52**	**0.54**				
link 6	0.89	0.90	0.85	**0.53**	**0.54**	0.93			
link 7	0.89	0.88	0.89	0.89	0.89	0.89	0.89		
link 8	0.89	0.91	0.93	0.94	0.90	0.94	0.94	0.89	
link 9	0.89	0.91	0.93	0.94	0.90	0.94	0.94	0.89	0.89

Notes: Correlation coefficient ΔW_{xy} and I_{xy}: 0.9987
Bold figures represent ratios outside the range 0.88-0.94

Table 6. Performance of the toll-selection indicator I_{xy}: $\Delta W_{xy}/I_{xy}$

With these considerations kept in mind, the indicator I_{xy} appears a sufficiently reliable indicator for the selection of multiple toll-points – at least based on the results obtained with the present network. Among the three possible strategies considered, this third strategy seems the one preferable on theoretical grounds – in particular the fact that interactions between tolls are taking into account – as well as the one performing best in terms of the correlation between predicted and true efficiency gains of the various combinations of toll-points considered. The main disadvantage is the large number of calculations that will have to be performed: $1+J!/(t!\cdot(J-t)!)$ solutions to systems of linear equations; the 1 representing (8′)-(9′), and the $J!/(t!\cdot(J-t)!)$ representing (8)-(9) (with (9) included for each of the t links belonging to the specific combination considered). However, an advantage compared to strategy 2 is that no network equilibria have to be calculated to determine the set of suggested toll-points.

4.3. *The selection of toll-points: concluding comments*

The indicator I_x proposed in Section 4.1 for the selection of a first single toll point appears a reasonably accurate measure, with a correlation coefficient between true and predicted welfare gains of 0.9987. Whether it would be worthwhile to use the computationally more demanding indicator I^*_x instead is a question that is difficult to answer in general, as it will depend on the network used. A pragmatic compromise would be to calculate I_x for all links, and I^*_x for a certain subset of most preferable links identified on the basis of I_x. However, the calculation of

I^*_x for all links has the advantage that links suffering from slow convergence and overestimated welfare gains from tolling due to the presence of parallel links can immediately be identified.

For the determination of multiple toll-points, comparable pragmatic solutions may be used in practice. Three strategies were identified, where the computational burden seems to be increasing with the quality of the prediction (although for some models, strategy 2 may actually turn out to be more demanding than strategy 3). An obvious possibility is to limit the set of possible combinations of t toll-points in a J-link network on the basis of strategy 1, and to use strategy 3 for this reduced set of combinations. Alternatively, one could start with the set of t links suggested by strategy 1, calculate the second-best equilibrium for this combination, and monitor the extent to which implied second-best taxes deviate from the predictions consistent with strategy 1. Second-best taxes markedly lower than these predictions would reflect a likely overestimation of the predicted welfare gains with strategy 1. One could then either look for sets of toll-points skipping these links, or could identify the relevant unpriced parallel links causing the deviation, and apply extra tolls on these links – possibly in exchange for those toll-points that had the lowest scores in strategy 1 if the number of t toll-points is a hard constraint.

The conclusion is therefore that, provided used with care, the toll-selection procedures discussed may be helpful in identifying those links for which the implementation of toll-points may lead to relatively large efficiency gains. The computational advantage realized by avoiding the calculation of J!/(t!·(J–t)!) second-best network equilibria comes at the price of a below unity correlation coefficient between predicted and true welfare gains. However, the results are encouraging enough to justify further testing or even application of these indicators – and the pragmatic compromises mentioned just above – in large network models.

5. Conclusion

This paper considered the generalized second-best network congestion pricing problem, in which not all links of a network can be tolled, so that the standard first-best solution of tolls equal to marginal external costs for all links is not a relevant policy option. A simulation model was used, designed to capture the most important types of possible network complications while allowing for a meaningful consideration of some archetype second-best policies that are often used or proposed for real road transport networks. Using this model, the general solution proposed by Verhoef (2000) was validated, in the sense that this solution was indeed found to produce second-best optimal tolls for the second-best policies considered. The simulation model confirmed earlier findings that parallel route pricing schemes – in particular 'pay-lanes' – constitute a relatively inefficient type of second-best congestion pricing. 'Free-lanes', although still not very efficient, at least lead to higher welfare gains than 'pay-lanes', and might therefore often offer an alternative to pay-lanes that is preferable on efficiency grounds.

The paper considered a number of aspects of the general problem and its solution that would be relevant when studying this type of second-best pricing in larger networks. First of all, the proposed algorithm for finding a second-best optimum appears relatively efficient, in the sense that for most tolls after two or three iterations, the optimal second-best toll is approached with an accuracy of more than 99%. An exception to this rule is given by links suffering from the availability of parallel substitutes. However, a simple variation on the general algorithm was found to behave as efficiently as the general algorithm does for links not having such substitutes.

In the practical design of second-best tolling schemes, a question probably equally important to that of finding the second-best optimal tolls for a given set of toll-points, involves the question of which links to toll in the first place. Especially in larger networks, this may not be easy to decide on the basis of logical reasoning, if anything due to the large number of possibilities, and to the complicated interactions that may exist between tolls on different links. The paper therefore considered the question of whether on the basis of the information available in the existing equilibrium, it is possible to predict which link(s) would be the most efficient one(s) to select for the implementation of tolling points. The basic procedure suggested for one link and the theoretically most correct variation thereof for multiple links performed rather well, with correlation coefficients of predicted and true welfare gains exceeding 0.99. For the case of single links, a more accurate indicator was suggested, and for the case of multiple links, computationally less demanding indicators were put forward. However, on the basis of the current network, it is hard to make a definite assessment of the relative performance the different indicators for other types of networks.

This brings us to the directions for further research. Two possibilities seem particularly worth further explorations. The first one involves the use and further testing of the proposed methodology and indicators in larger networks, possibly involving non-linear demand and cost-functions. It would be interesting to see to what extent the generally favourable results reported here are due to these two features of the network used. One might suspect that in particular the assumed linearity of the cost and demand functions may lead to an overestimation of the efficiency of the algorithm and procedures considered. A second topic for further study would involve the introduction of theoretical refinements in the general problem set-up and network model used in this paper. These refinements could in particular involve the introduction of dynamics and the consideration of heterogeneous traffic. The results presented in this paper seem encouraging enough to justify such further research.

References

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

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

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

Dafermos, S. (1973) "Toll patterns for multiclass-user transportation networks" *Transportation Science* **7** 211-223.

Dafermos, S. (1980) "Traffic equilibrium and variational inequalities" *Transportation Science* **14** 42-54.

De Palma, A. and R. Lindsey (1999) "Private roads: competition under various ownership regimes" *Annals of Regional Science* forthcoming.

De Palma, A. and Y. Nesterov (1998) "Optimization formulations and static equilibrium in congested transportation networks" Paper presented to the 8th WCTR-conference, 12–17 july 1998, Antwerp, Belgium.

Glazer, A. and E. Niskanen (1992) "Parking fees and congestion" *Regional Science and Urban Economics* **22** 123-132.

Kinderlehrer, D. and G. Stampacchia (1980) *An Introduction to Variational Inequalities and Their Applications* Academic Press, New York.

Lévy-Lambert, H. (1968) "Tarification des services à qualité variable: application aux péages de circulation" *Econometrica* **36** (3-4) 564-574.

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

Nagurney, A. (1993) *Network Economics: A Variational Inequality Approach* Kluwer Academic Publishers, Dordrecht.

Small, K.A. and J.A. Gomez-Ibañez (1998) "Road pricing for congestion management: the transition from theory to policy". In: K.J. Button and E.T. Verhoef (1998) *Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility* Edward Elgar, Cheltenham (forthcoming).

Smith, M.J. (1979) "The marginal cost pricing of a transportation network" *Transportation Research* **13B** 237-242.

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

Verhoef, E.T. (2000) "Second-best congestion pricing in general static transportation networks with elastic demands" Unpublished paper, Free University Amsterdam.

Verhoef, E.T., P. Nijkamp and P. Rietveld (1995) "Second-best regulation of road transport externalities" *Journal of Transport Economics and Policy* **29** (2) 147-167.

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. and K.A. Small (1999) "Product differentiation on roads: second-best congestion pricing with heterogeneity under public and private ownership" Discussion paper TI 99-066/3, Tinbergen Institute, Amsterdam-Rotterdam.

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

Wardrop, J. (1952) "Some theoretical aspects of road traffic research" *Proceedings of the Institute of Civil Engineers* **1** (2) 325-378.

Yan, H. and W.H.K. Lam (1996) "Optimal road tolls under conditions of queueing and congestion" *Transportation Research* **30A** (5) 319-332.