SECOND-BEST CONGESTION PRICING IN GENERAL STATIC TRANSPORTATION NETWORKS WITH ELASTIC DEMANDS

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/vakgroep/re/members/everhoef/et.html

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

Abstract

This paper studies the second-best problem where not all links of a congested transportation network can be tolled. The second-best tax rule for this problem is 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. It is demonstrated that the solution obtained indeed generalizes a number of known second-best tax rules presented earlier in the literature, for specific cases of the general problem discussed in the present paper. Finally, it is demonstrated that, for instance by using the concept of 'virtual links', the analysis can be applied rather easily also to a broader class of second-best problems in static networks.

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

1. Introduction

Pigouvian marginal external cost pricing (Pigou, 1920) is widely accepted among transport economists as the first-best bench-mark solution in the regulation of road transport externalities. It is, however, almost equally commonly recognized that the necessary assumptions for the practical applicability of this standard Pigouvian tax rule will seldom, if ever, be met in reality. These assumptions include, for instance, that optimal charging mechanisms be available, allowing the regulator to set perfectly differentiated taxes for all road users and on all links of the network; that first-best conditions prevail throughout the economic environment to which the transport system under consideration belongs; and the assumption of perfect information for all users of the system as well as for the regulator.

Indeed, such assumptions are quite unrealistic, and second-best issues in transport regulation have accordingly received ample attention in the literature. For instance, Wilson (1983), and d'Ouville and McDonald (1990) study optimal road capacity with sub-optimal congestion pricing. Braid (1989) and Arnott, De Palma and Lindsey (1990) consider uniform or step-wise pricing of a bottleneck. Arnott (1979) and Sullivan (1983) look at congestion policies through urban land-use strategies. A classic problem in the second-best regulation in road transport 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 also by Braid (1996) and Verhoef, Nijkamp and Rietveld (1996). 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. Verhoef, Emmerink, Nijkamp and Rietveld (1996) consider second-best congestion tolls under conditions of stochastic congestion and imperfect information. A recurring results in the studies mentioned here, as well as in other studies, 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).

The present paper aims to offer a general solution for the second-best problem where not all links of a congested transportation network can be tolled. A specific application of this problem is the two-route problem just mentioned. Numerous other applications can however be thought of. This type of problem will become increasingly relevant from a practical viewpoint when the foreseen introduction of electronic road charging for a growing number of urban areas becomes reality (Small and Gomez-Ibañez, 1998). For example the determination of optimal cordon charges, for a toll-ring around a city centre, is a special case of the general problem studied in this paper. Often, this type of second-best problems may be 'self-imposed' by the regulator, in particular when it is considered inefficient to collect charges on all links of a network, rather than on a subset of links only. This could be the case if relatively high costs are associated with installing the additional tolling equipment, while only relatively low social benefits would be expected to arise from having the additional tolls available. Especially with electronic tolling, such a cost structure may often be the rule rather than the exception.

The analysis to follow considers congestion as the only relevant externality, and is cast in terms of a road network. The purpose is to derive the second-best optimal tax rules that would apply for any set of toll-points on any congested transportation network. The analysis pertains to static networks only, and assumes deterministic equilibria with perfect information. Generalizations to dynamic transportation networks, and to networks with imperfect information and stochasticity, are left as important topics for future research.

The paper is organized as follows. The next section introduces the notation, and discusses some important features related to the uniqueness of equilibrium values of some key variables in general transportation networks. Section 3 presents the second-best optimization problem and its solution, and in addition considers the related important question of the optimal location of additional toll-points. Section 4 shows that the general solution obtained indeed is a generalization of earlier results in the literature, and presents some further possible applications of the general model. Section 5 concludes.

2. A general characterization of the problem

The analysis in this paper pertains to a general transportation network 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):

N the set of nodes in the network

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 $D_i' \leq 0$

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 $N_j \geq 0$

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

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 $N_p \geq 0$

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 P_i and

$$\sum_{j=1}^{J} d_{jp} \cdot \left(c_j(N_j) + d_j \cdot f_j \right) - D_i(N_i) \leq 0,$$ and a value of 0 otherwise

The relevance and interpretation of the last of these variables will become clear in the discussion of the equilibrium conditions for the network below. 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. The more general case, where the travel time on a link may also depend on the usage of other links, is presented in the Appendix. It turns out to be a straightforward generalization of the analysis to be presented below. 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 preceding links. For a static model, however, which can by definition not give a meaningful representation of cases where arrival rates exceed capacities anyway (Verhoef, 1998), the assumption that congestion is link-specific may often be acceptable, unless intersections are considered to be an important source of congestion (see the Appendix).

Because every path p connects one unique OD-pair, defined by the nodes at the tail of the first arc and the head of the last arc, we have:

$$P = \sum_{i=1}^{I} P_i \qquad (1)$$

Since we are dealing with a static network, we also have:

$$N_j = \sum_{p=1}^{P} d_{jp} \cdot N_p \qquad (2)$$

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):

$$N_j \geq 0; \quad \sum_{j=1}^{J} d_{jp} \cdot \left(c_j + d_j \cdot f_j\right) - D_i \geq 0 \quad \text{and} \quad N_j \cdot \left(\sum_{j=1}^{J} d_{jp} \cdot \left(c_j + d_j \cdot f_j\right) - D_i \right) = 0$$

$$p \;\; \mathsf{P}_i \qquad (3)$$

(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 (3) therefore adds the economic equilibrium principle that marginal benefits should be equal to marginal private costs to the standard Wardrop 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 (3) reveals that the dummy variable δ_{ip} defined earlier takes on the value of 1 only if path p from the set P_i is among those that may be used in the equilibrium by travellers between OD-pair i. Such paths with δ_{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:

$$N_i = \sum_{p=1}^{P} d_{ip} \cdot N_p \tag{4}$$

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 $\mathbf{f_j}$). This can be illustrated by considering the simple network with four links (I-IV), connecting two OD-pairs A-C and B-C, shown in Figure 1. Note that both OD-pairs have two paths: in the last part of the trip, either link III or IV can be chosen by both types of drivers (‘A-drivers’ and ‘B-drivers’). It is assumed that the regulator can set tolls on each of the four links. Consider the first-best optimum, where the tolls f_j are each set equal to the marginal external congestion costs (MEC_j) on these links, and suppose that the two OD-flows and the four link-flows are all positive.

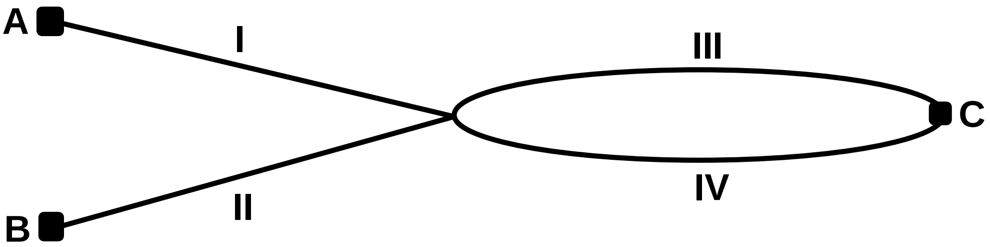

Figure 1. Non-uniqueness of path-flows and link-tolls in a simple network

It is then straightforward that this equilibrium is not unique in path-flows: after interchanging an A-driver on link III with a B-driver on link IV, the same equilibrium in OD-flows and link-flows (and hence also in terms of total as well as marginal benefits and costs) results, although the equilibrium has changed in terms of path-flows. Also, the equilibrium is not unique in link-tolls. In particular, the same optimum can be realized by any set of tolls according to:

$$\begin{aligned} f_I &= MEC_I + x \\ f_{II} &= MEC_{II} + x \\ f_{III} &= MEC_{III} - x \\ f_{IV} &= MEC_{IV} - x \end{aligned} \tag{5}$$

where x can have any value. Of course, for other network configurations, the equilibrium can be unique also in terms of path-flows and link-tolls. The latter can be verified in the example above by adding a third group of users, also having C as their destination, but having the intersection of the four links as their node of origin. Therefore, the main point here is only that for general networks, one cannot be sure that a unique solution in terms of path-flows and (first-best or second-best) link-tolls exists. This, in turn, will be reflected in the general solution to be derived below.

3. Solving the second-best optimization problem

The stage is now set to derive 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, actually only 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. Using the notation and assumptions presented in the previous section, the regulator therefore has to solve the problem that can be represented by the following Lagrangian:

$$\begin{aligned}\Lambda = \sum_{i=1}^{I}\int_{0}^{\sum_{p=1}^{P} d_{ip}\cdot N_p} D_i(x_i)dx_i - \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P} d_{jp}\cdot d_{ip}\cdot N_p\cdot c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} d_{jq}\cdot d_{kq}\cdot N_q\right) \\ +\sum_{i=1}^{I}\sum_{p=1}^{P} d_{ip}\cdot l_p\cdot\left[\sum_{j=1}^{J} d_{jp}\cdot\left(c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P} d_{jq}\cdot d_{kq}\cdot N_q\right)+d_j\cdot f_j\right)-D_i\left(\sum_{q=1}^{P} d_{iq}\cdot N_q\right)\right]\end{aligned} \tag{6}$$

The first set of terms represent total benefits, summed over all OD-pairs; note that the total OD-flow is determined according to (4). 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 (2). 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 (3), and that λ_p denotes the Lagrangian multiplier associated with the constraint for path p. These multipliers will be discussed in further detail below. Finally, it ought to be noted that 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 – which either are used or could be used in the second-best equilibrium – are considered. Note that, also for notational reasons, the index k, when used, denotes OD-pairs. The following necessary first-order conditions can now be derived (where arguments in demand and cost functions are again dropped for notational convenience):

$$\begin{aligned}\frac{¶\Lambda}{¶N_p} = \sum_{i=1}^{I} d_{ip}\cdot D_i - \sum_{j=1}^{J} d_{jp}\cdot\left(c_j + \sum_{k=1}^{I}\sum_{q=1}^{P} d_{jq}\cdot d_{kq}\cdot N_q\cdot c'_j\right) \\ +\sum_{k=1}^{I}\sum_{q=1}^{P} d_{kq}\cdot l_q\cdot\left(\sum_{j=1}^{J} d_{jp}\cdot d_{jq}\cdot c'_j\right)-\sum_{i=1}^{I} d_{ip}\cdot l_p\cdot D'_i = 0 \qquad \forall\ p\,\text{with}\, d_{ip} = 1\end{aligned} \tag{7}$$

$$\frac{\P\Lambda}{\P f_j} = \sum_{i=1}^{I}\sum_{p=1}^{P} d_{ip}\cdot d_{jp}\cdot l_p = 0 \qquad \forall\ j\,\text{with}\,d_j = 1 \tag{8}$$

$$\frac{\P\Lambda}{\P l_p} = \sum_{j=1}^{J} d_{jp}\cdot\left(c_j + d_j\cdot f_j\right) - \sum_{i=1}^{I} d_{ip}\cdot D_i = 0 \qquad \forall\ p\,\text{with}\,d_{ip} = 1 \tag{9}$$

Note that, notwithstanding the fact that the second-best equilibrium may not be unique in path-flows as pointed out in the previous section, equations (7) indicate that the first-order conditions with respect to path-flows should be 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 (7) 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 (3) 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.

Substitution of (9) into (7) for each p for which δ_{ip}=1 subsequently yields the following expression for the Lagrangian multipliers λ_p:

$$l_p = \frac{\sum_{j=1}^{J} d_{jp}\cdot\left(\sum_{q=1}^{P} d_{jq}\cdot N_q\cdot c'_j\right) - \sum_{q=1,q\neq p}^{P} l_q\cdot\left(\sum_{j=1}^{J} d_{jp}\cdot d_{jq}\cdot c'_j\right) - \sum_{j=1}^{J} d_{jp}\cdot d_j\cdot f_j}{\sum_{j=1}^{J} d_{jp}\cdot c'_j - \sum_{i=1}^{I} d_{ip}\cdot D'_i} \tag{10}$$

$$\forall\ p\ \text{with}\ d_{ip} = 1 \quad \text{and} \quad \forall\ q\ \text{with}\ d_{iq} = 1$$

These Lagrangian multipliers, when being unequal to zero, cause the second-best solution to be inferior to the first-best case where tolls can be set on all links. The fact that these multipliers would be zero in the first-best case can most easily be verified by rewriting (6) as if path-tolls f_p could be charged for all paths. This would yield, in place of (8), λ_p=0 for all relevant paths, and path tolls equal to the sum of marginal external congestion costs on all links used in that path (given by the first of the three terms in the numerator of (10)). This, in turn, can be realized with link tolls each equal to the marginal external congestion costs for that link.

The Lagrangian multipliers λ_p can thus be interpreted as the 'shadow price of non-optimal pricing' in the second-best optimum – which, of course, already followed from the specification of the Lagrangian (6) from the outset. Although for a general network, no analytical solution exists with each relevant λ_p explicitly solved for the other relevant λ_p's (or λ_q's as they are labelled in (10)), it can be noted that for equations (10) will make up a system of X equations, generally linearly independent, in X unkowns (the λ_p's), where X denotes the number of relevant paths in the second-best optimum. Hence, for a given problem, these multipliers can be solved for, independent of the value of the other multipliers. The reason that no general analytical solution can be given is, of course, that the expression for this solution will depend on the specific network, the tolling points, and the relevant paths.

A further inspection of equations (10) allows the identification of the terms affecting the size of the 'shadow price of non-optimal pricing' for a relevant path in the second-best optimum. Focusing first on the first and third term in the numerator, this shadow price appears to be increasing in the extent to which the marginal external congestion costs caused (the first term) exceeds the sum of total tolls paid during the trip (the third term). This is confirm intuition. The middle term shows that, because we are in a second-best optimum, also indirect effects count. In particular, the shadow price is decreasing in the extent to which the presence of users from the path considered prevents users from other non-optimally priced relevant paths to use the network. The associated term is in the first place increasing in the shadow prices for these other paths. This reflects that the 'reward' (that is, the reduction in the own shadow price) becomes larger, the larger the shadow prices for the other affected groups are. The term is also increasing in the slope of the average cost functions on the relevant links in the second-best optimum, which reflects that this effect becomes more important as the cost levels on these links are more strongly dependent on link usage. Note that λ_p may have either sign – which was in fact already implied by the first-order condition (8)[1]. Finally, the denominator of (10) shows that the 'shadow price of non-optimal pricing' for a specific relevant path is decreasing in the sensitivity of the path flow to distorted prices in the second-best optimum. If either the demand for the OD-pair or the 'supply' (represented by the link-cost functions) is fully inelastic, the multiplier vanishes. As (8) shows, an important feature of the second-best optimum is that the sum of the relevant λ_p's be minimized, which rather intuitively reflects the goal of minimizing the overall distortions due to imperfect pricing.

Finally, substitution of (10) into (8) gives the following expression for the second-best optimal congestion fees:

$$f_j = \frac{\sum_{p=1}^{P} d_{jp} \cdot \dfrac{\sum_{m=1}^{J} d_{mp} \cdot \left(\sum_{q=1}^{P} d_{mq} \cdot N_q \cdot c'_m \right) - \sum_{q=1,q\neq p}^{P} l_q \cdot \left(\sum_{m=1}^{J} d_{mp} \cdot d_{mq} \cdot c'_m \right) - \sum_{m=1,m\neq j}^{J} d_{mp} \cdot d_m \cdot f_m}{\sum_{m=1}^{J} d_{mp} \cdot c'_m - \sum_{i=1}^{I} d_{ip} \cdot D'_i}}{\sum_{p=1}^{P} \dfrac{d_{jp}}{\sum_{m=1}^{J} d_{mp} \cdot c'_m - \sum_{i=1}^{I} d_{ip} \cdot D'_i}} \qquad (11)$$

$$\forall\ j \text{ with } d_j = 1 \quad \text{and} \quad \forall\ p \text{ with } d_{ip} = 1 \quad \text{and} \quad \forall\ q \text{ with } d_{iq} = 1$$

where, for notational reasons, the index m, when used, denotes links. After the discussion of (10), the interpretation of (11) is actually more easy to give than may seem at first sight. First, the first term $\Sigma_p\delta_{jp}$ in the numerator of (11) shows that only the relevant paths using the link j should be considered directly in the determination of the second-best toll f_j – although, via the terms λ_q in the numerator's numerator, other relevant paths may of course indirectly affect the level of f_j. As a matter of fact, this term in the numerator's numerator again gives the difference between an OD-flow's 'generalized marginal external costs' (corrected for the indirect effect

[1] This is actually also the reason for using the dummy variables δ_{ip} and writing the problem as a Lagrangian instead of using a Kuhn-Tucker formulation, in which case the resulting multipliers would be restricted to be positively signed.

on the usage by other relevant paths) and the total tolls (but now net of the specific toll f_j itself), closely resembling the term already encountered in the numerator of (10). Finally, the further structure of (11) shows that the second-best optimal toll on a link should be a weighted average of the sum of the generalized marginal external costs, minus the tolls paid on other links, for the relevant OD-flows (possibly passing that link). The weights are increasing with the sensitivity of the OD-flow to prices in the second-best optimum. This effect reflects what was already observed in the discussion of equation (10).

The structure of (11) shows that the second-best tolls may need not be independent in the second-best optimum, as was illustrated already for the simple network considered in the previous section. An important question, however, is whether the first-order conditions (7)-(9) and the implied second-best values of the relevant λ_p's and f_j's imply a unique local second-best optimum in terms of OD-flows and link-flows (assuming that the second-order conditions are fulfilled). It is questionable whether a general answer to this question can be given, as it may depend on the exact shape of the network, the selected tolling points, and the shape of the demand and cost functions. One has to be modest, therefore, and it should be emphasized that the tax rules implied by (11) give necessary conditions for a local and a global second-best optimum in a general transportation network only, rather than sufficient conditions. Under quite general circumstances, however, one would expect only few (if more than one) second-best equilibria supported by taxes as given in (11) to exist, and considering only those equilibria where such taxes apply will generally greatly reduce the task of finding the second-best optimum for a given problem.[2]

Finally, an important question that is closely related to the above analysis concerns the optimal location of additional toll-points. This question will be relevant not only when an existing tolling system can be extended to cover a larger part of the network, but of course also when an entirely new tolling system can be installed. Clearly, the most preferable way of selecting the optimal location for a single next toll-point would be to calculate the level of welfare under second-best tolling according to (11), for having a toll added on each possible, as yet untolled link. The optimal next toll-point is the one yielding the highest welfare improvement (which could of course be compared with the costs of adding the toll-point to guarantee an efficient investment). For larger networks, however, this procedure may require many calculations, in particular if the problem is somewhat more general, and the optimal location for a (possibly also optimized) number of additional toll-points should be determined.

In such cases, the calculation of 'shadow tolls' ϕ_j for those links j on which no tolls can be set in the second-best optimum, according to the expression given for f_j in equation (11), may provide a rather efficient way of selecting the link for which it could be most efficient to

[2] A general numerical procedure for finding such a second-best optimum in a given transport network model could be based on the following sequence: (step 1) start with zero tolls and calculate the equilibrium; (step 2) calculate the out-of-equilibrium values of the λ_p's for all relevant paths for this equilibrium according to (10); (step 3) calculate the implied out-of-equilibrium tolls for the relevant toll points according to (11); (step 4) apply these tolls by adding them to the (perceived) link costs and calculating the new equilibrium; (step 5) check for convergence and go back to (step 2) if the system has not yet converged.

add the next toll. In particular, the link for which $|\phi_j|$ is maximized can be identified as the one for which a marginal change in the zero toll level gives the highest net social benefits. This may often be the link for which also in the new second-best equilibrium, with a toll added on that link, the total social welfare improvement is the largest.[3] Often, this link-selection procedure may offer a quick and reasonably accurate manner to select the optimal location for a next toll-point, or – by applying it sequentially – a number of toll-points. Nevertheless, as with the uniqueness of local optima, also for this matter one cannot be sure whether the suggested link-selection procedure will always be optimal.

4. Comparing the general solution with earlier results

It is instructive to validate the tax-rule (11) by comparing it to results reported earlier in the literature for second-best problems that are special cases of the general problem considered here. Three such cases will be considered: first-best tolling, the standard two-route problem, and the parking problem studied by Glazer and Niskanen (1992). The section concludes with some thoughts on further possible applications of the general model presented above.

4.1. First-best tolling

The most straightforward special case of the general second-best problem discussed is in fact the first-best problem where tolls can be set on all links. For that case, one would expect equation (11) to be consistent with the simple Pigouvian rule equating the tax to the marginal external congestion costs for each link, as given in (12):

$$f_j = \sum_{p=1}^{P} d_{jp} \cdot N_p \cdot c'_j \tag{12}$$

To show that this indeed is the case, first observe that – as was argued before – al λ_p's are equal to zero in the first-best case. Hence, the second term in the numerator's numerator in (11) drops out in the first-best case. Equation (11) then reduces to an expression stating that the toll on a link should be equal to the weighted average (over all relevant paths using j) of the difference between the total marginal external congestion costs for that path (over the entire trip, so over all links including link j itself) minus the tolls paid on all links other than j itself. Evidently, the simple Pigouvian tax in (12) is consistent with this rule (as pointed out in Section 2, for many networks the first-best solution needs not be unique in link-tolls).

4.2. The standard two-route problem

The standard two-route problem concerns a network as illustrated in Figure 2, where an untolled route (U) exists parallel to a tolled route (T), and both routes connect the same single origin-destination pair (AB). As mentioned, this problem has been studied by Lévy-Lambert (1968), Marchand (1968), Braid (1996), and Verhoef, Nijkamp and Rietveld (1996). These studies have shown that the optimal second-best one-route toll for route T can be written as:

[3] Note in particular that adding a toll on a link for which $\phi_j=0$ will yield no welfare improvement at all, since the optimal toll for this link will then also be $f_j=0$, and the same second-best equilibrium will necessarily result.

$$f_T = N_T \cdot c'_T - N_U \cdot c'_U \cdot \frac{-D'_{AB}}{c'_U - D'_{AB}} \tag{13}$$

Equation (13) shows that this toll should be equal to the marginal external congestion costs on the tolled route minus a term consisting of a fraction (between 0 and 1) of the marginal external congestion costs on the untolled route. Note that (13) may imply a zero or even a negative second-best optimal toll. For a further interpretation of (13), see for instance Verhoef, Nijkamp and Rietveld (1996); the important question now is whether the general tax rule in (11) is consistent with (13).

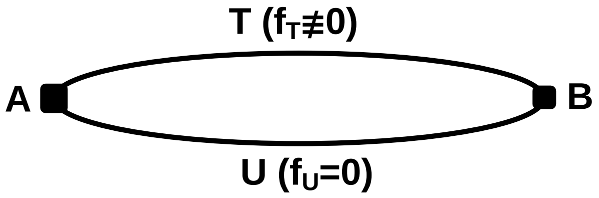

Figure 2. The standard two-route problem

To demonstrate that this is the case, it is sufficient to observe that for this problem, the two paths coincide with the two links, so that the necessary first-order conditions (7)-(9) become[4]:

$$\frac{\P\Lambda}{\P N_T} = D_{AB} - c_T - N_T \cdot c'_T + l_T \cdot c'_T - \left(l_T + l_U\right) \cdot D'_{AB} = 0 \tag{14a}$$

$$\frac{\P\Lambda}{\P N_U} = D_{AB} - c_U - N_U \cdot c'_U + l_U \cdot c'_U - \left(l_T + l_U\right) \cdot D'_{AB} = 0 \tag{14b}$$

$$\frac{\P\Lambda}{\P f_T} = l_T = 0 \tag{15}$$

$$\frac{\P\Lambda}{\P l_T} = c_T + f_T - D_{AB} = 0 \tag{16a}$$

$$\frac{\P\Lambda}{\P l_U} = c_U - D_{AB} = 0 \tag{16b}$$

Fully consistent with (10), using (15), (16b) and (14b) λ_U can then be solved as:

$$l_U = \frac{N_U \cdot c'_U}{c'_U - D'_{AB}} \tag{17}$$

and substitution of (15), (16a) and (17) into (14a) gives the desired result given in (13).

[4] These first-order conditions only differ from those in Verhoef, Nijkamp and Rietveld (1996) in the sense that the sign of the constraints and hence of the Lagrangian multipliers are now opposite to the signs in the original formulation. This, of course, does not affect the result.

4.3. Parking policies

A final example concerns the problem of optimal parking fees for congestion management in the case where a subset of road users do not have to pay this fee, for example because they have access to private parking places. This problem was studied by Glazer and Niskanen (1992). By adding two ‘virtual links’ to the original one-link network, the problem can be represented as a network problem that allows the solution of the optimal second-best parking fee using the methodology presented in the previous sections.

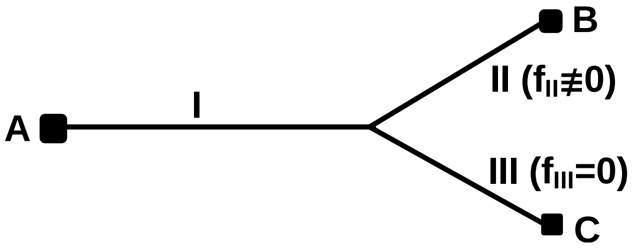

Figure 3. The Glazer and Niskanen (1992) parking problem in a network representation

Figure 3 shows the resulting three-link network, connecting two OD-pairs AB and AC, where B denotes priced parking space and C free parking space. Congestion only occurs on the shared link I. The virtual links II and III are costless, but a fee can of course be charged on link II. We then have the ingredients to solve the parking problem given above. Hence, the resulting second-best parking fee again should be consistent with (11). To demonstrate that this is the case, it is sufficient to observe that for this two-path problem, the two paths coincide with the two OD-pairs, so that the necessary first-order conditions (7)-(9) now become (note that $c_{II}=c'_{II}=c_{III}=c'_{III}=0$):

$$\frac{\P\Lambda}{\P N_{AB}} = D_{AB} - c_I - N_I \cdot c'_I + \left(l_{AB} + l_{AC}\right) \cdot c'_I - l_{AB} \cdot D'_{AB} = 0 \qquad (18a)$$

$$\frac{\P\Lambda}{\P N_{AC}} = D_{AC} - c_I - N_I \cdot c'_I + \left(l_{AB} + l_{AC}\right) \cdot c'_I - l_{AC} \cdot D'_{AC} = 0 \qquad (18b)$$

$$\frac{\P\Lambda}{\P f_{II}} = l_{AB} = 0 \qquad (19)$$

$$\frac{\P\Lambda}{\P l_{AB}} = c_I + f_{II} - D_{AB} = 0 \qquad (20a)$$

$$\frac{\P\Lambda}{\P l_{AC}} = c_I - D_{AC} = 0 \qquad (20b)$$

where $N_I=N_{AB}+N_{AC}$. Fully consistent with (10), using (19), (20b) and (18b) λ_{AC} can then be solved as:

$$l_{AC} = \frac{N_I \cdot c'_I}{c'_I - D'_{AC}} \qquad (21)$$

and substitution of (19), (20a) and (21) into (18a) gives:

$$f_{II} = N_I \cdot c_I' \cdot \frac{-D_{AC}'}{c_I' - D_{AC}'} \qquad (22)$$

showing that the optimal second-best congestion toll is now a fraction of the marginal external congestion costs on the congested link, where the fraction depends on the demand elasticity for the untolled users of this link (see Glazer and Niskanen, 1992, and Verhoef, Nijkamp and Rietveld, 1995, for further discussions). Also in this case therefore, the second-best optimal congestion toll turns out to be a specific case of the general solution given in (11).

4.4. Some further possible applications of the general model

The use of the concept of virtual links in the final case above in fact demonstrates how easily the general model in equations (6)-(11) can be adapted to allow consideration also of different types of second-best policies in a network environment, other than the pure problem caused by the physical joint existence of tolled and untolled links in the network.

Consider, for instance, the use of peak-hour permits. With such a policy, road users would have to purchase a permit before they are allowed to use the road. However, once they do have such a permit, there would be no further restriction on the use of the network. To determine the second-best optimal price for such permits for a given road network, the regulator therefore has to solve the second-best problem that is caused by the fact that the same single 'toll' (that is: the price of the permit) applies for drivers using different paths, and hence generally causing different levels of marginal external costs. The only adaptation that needs to be made to solve this particular problem using the general network model presented in equations (6)-(11), is to add one single virtual link, with zero costs, on which the regulator can set a toll. This virtual link should be added to all paths, and the optimal toll can then be derived directly according to (11).[5]

A second application that can be mentioned is the use of distance-based tolls. If the regulator can set only a toll level per vehicle-kilometre travelled, while marginal external congestion costs per vehicle-kilometre vary over the network, another second-best problem results. Also this problem can be solved using the proposed model. To do so, rewrite the original tolls f_j as $L_j \cdot f$, where L_j denotes the length of link j. Observe that the original first-order condition (8) is then replaced by:

$$\frac{\P\Lambda}{\P f} = \sum_{i=1}^{I}\sum_{p=1}^{P} d_{ip} \cdot l_p \cdot \sum_{j=1}^{J} d_{jp} \cdot d_j \cdot L_j = 0 \qquad (23)$$

(Note that the formulation allows cases where the per vehicle-kilometre toll is not charged on all links. This could be relevant when the scheme applies to a certain area only, and some users originate from outside this area). In the second-best solution, now the *weighted* sum of the λ_p's

[5] If the regulator wants to use a system of tradeable peak-hour permits, according to the same principles but distributed initially for free, in fact exactly the same problem as described in the main text has to be solved. The second-best optimal number of permits to be issued will then be equal to the number of trips made in the second-best optimum considered in the main text; and the equilibrium price of the permits will be equal to the second-best toll. This holds true, of course, only under the assumption of zero transaction costs.

should be equal to zero, where the weights increase in the number of tolled kilometres for a path. The problem can then be solved analogous to the discussion in Section 3.

Clearly, the two second-best problems just mentioned do not have simple analytical solutions. However, the reason for mentioning these problems here was merely to illustrate that the network model presented in Sections 2 and 3 can easily be extended to deal also with different classes of second-best problems in static transportation networks, other than the pure problem caused by the physical joint existence of tolled and untolled links in one network.

5. Conclusion

This paper presented a general solution for the problem of second-best congestion tolling in static transportation networks where not all links can be tolled. With the existing plans for introducing electronic road pricing in many urban areas throughout the world, this type of problem is likely to become very important in the near future, in particular because it will often be considered inefficient (or unmanageable) to install the necessary equipment in all existing links. For small networks, the second-best tax rule may still yield analytically tractable congestion tolls, as was shown in the previous section where three special cases of the general problem were discussed. Due to the occurrence of all sorts of cross-effects between tolled and untolled links, however, the tax rules will become solvable via numerical modelling only as the networks considered become more realistic and, as a consequence, larger. Nevertheless, the analysis presented has provided the necessary conditions for second-best optimality in such large networks, that can directly be applied regardless the size and the shape of the network. Moreover, it was demonstrated that, for instance by using the concept of 'virtual links', the analysis can even be applied rather easily to different classes of second-best problems in static networks as well.

In solving the general problem, an important set of variables used were the Lagrangian multipliers representing the 'shadow price of non-optimal pricing' for tolled and untolled links. The latter only play an indirect role in the second-best tax rule, reflecting indirect spill-over effects and interdependencies in networks that ought to be considered in second-best regulation. The former (the multipliers for the tolled links) are used directly in the optimization, in the sense that the absolute value of their sum is minimized. It was shown that in the first-best solution, these multipliers will all, individually, be equal to zero.

It was argued that the application of the second-best tax rule also for untolled links may often provide an intuitive guideline for selecting the particular link for which it is economically most beneficial to add the subsequent toll-point. In particular in large, realistic networks, where it is computationally too demanding to explicitly calculate the impact of adding a toll on each of the as yet untolled links, such procedures may be helpful.

A number of future research topics can be mentioned. The first of these would be the derivation of second-best tolls in general dynamic networks and under conditions of uncertainty. Another important topic is to test the applicability of the tax rule derived in existing static network models. A third topic, related to this, would be to investigate the possibility of deriving general conditions under which networks will have a unique local second-best local optimum.

References

Arnott, R.J. (1979) "Unpriced transport congestion" *Journal of Economic Theory* **21** 294-316.

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

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.

d'Ouville, E.L. and J.F. McDonald (1990) "Optimal road capacity with a suboptimal congestion toll" *Journal of Urban Economics* **28** 34-49.

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

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.

Sullivan, A.M. (1983) "Second-best policies for congestion externalities" *Journal of Urban Economics* **14** 105-123.

Verhoef, E.T. (1998) "Time, speeds, flows and densities in static models of road traffic congestion and congestion pricing" *Regional Science and Urban Economics* forthcoming.

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

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.

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.

Wilson, J.D. (1983) "Optimal road capacity in the presence of unpriced congestion" *Journal of Urban Economics* **13** 337-357.

Appendix: Relaxing the assumption of link-specific congestion

The assumption of strictly link-specific congestion, that was made for the general model presented in the main text, may become problematic if *intersections* are to be modelled more realistically. The formulation used in the main text could only be applied directly to intersections if all users of an intersection to the same extent suffer from, and contribute to congestion on that intersection. In that case, the intersection could of course be treated as another link, just like the other links. In more realistic formulations, however, the representation with only link-specific congestion can in fact be considered as too restrictive. This can be illustrated by considering the intersection depicted in Figure A.1, where two two-way streets intersect. The dashed links, showing the 12 possible ways of using the intersection, cross each other in various cases, and hence direct congestion cost interdependencies between links are very likely to exist.[6] Moreover, the size of the (marginal) cost interdependency certainly needs not be constant over all pairs of links on the intersection. For instance, two different 'turns-to-the-right' will hardly hinder each other, while 'turns-to-the-left' and 'straight-ons' will generally be more conflicting (note that it is assumed that drivers use the right side of the road).

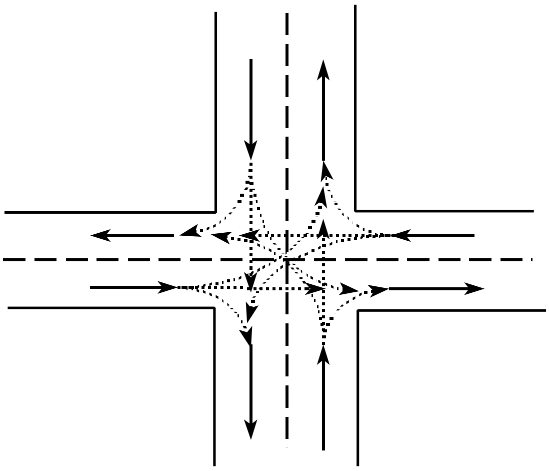

Figure A.1 Direct cost interdependencies between links on a simple intersection

Fortunately, it is rather straightforward to incorporate the implied direct cost interdependencies in the main model presented in equations (6)–(11). Doing so of course further complicates the analysis and the various expressions, but the main conclusions remain similar. We therefore present the equivalent expressions (A6)–(A11) below, for the more general case where the average user costs on link x may possibly depend on the level of usage N_j on all other links j, without further detailed comments.

First, the Lagrangian (6) can now be written as:

[6] It is very important to distinguish between these *direct cost interdependencies* between links, and the *indirect cost interdependencies* between links, that were mentioned also in the main text. Direct cost interdependencies result from the technical, direct interactions between users from different links; indirect cost interdependencies are caused by the equilibrating behaviour of users as described in Wardrop's first principle, leading to equalized equilibrium cost levels for all used routes between given OD-pairs.

$$\begin{aligned}\Lambda = \sum_{i=1}^{I}\int_{0}^{\sum_{p=1}^{P}\delta_{ip}\cdot N_p} D_i(x_i)dx_i - \sum_{j=1}^{J}\sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{jp}\cdot\delta_{ip}\cdot N_p\cdot c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{xq}\cdot\delta_{kq}\cdot N_q \;\forall\, x=1,\dots,J\right)\\ +\sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{ip}\cdot\lambda_p\cdot\left[\sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j\left(\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{xq}\cdot\delta_{kq}\cdot N_q \;\forall\, x=1,\dots,J\right)+\delta_j\cdot f_j\right)-D_i\left(\sum_{q=1}^{P}\delta_{iq}\cdot N_q\right)\right]\end{aligned} \tag{A6}$$

Note that the only difference between (6) and (A6) is that in (A6), c_j possibly depends on the link-flow on all links in the network.

The following necessary first-order conditions can now be derived:

$$\begin{aligned}\frac{\partial\Lambda}{\partial N_p} = \sum_{i=1}^{I}\delta_{ip}\cdot D_i - \sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j + \sum_{x=1}^{J}\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{xq}\cdot\delta_{kq}\cdot N_q\cdot\frac{\partial c_x}{\partial N_j}\right)\\ +\sum_{k=1}^{I}\sum_{q=1}^{P}\delta_{kq}\cdot\lambda_q\cdot\left(\sum_{j=1}^{J}\delta_{jp}\cdot\sum_{x=1}^{J}\delta_{xq}\cdot\frac{\partial c_x}{\partial N_j}\right)-\sum_{i=1}^{I}\delta_{ip}\cdot\lambda_p\cdot D_i' = 0 \qquad \forall\, p \text{ with } \delta_{ip}=1\end{aligned} \tag{A7}$$

$$\frac{\partial\Lambda}{\partial f_j} = \sum_{i=1}^{I}\sum_{p=1}^{P}\delta_{ip}\cdot\delta_{jp}\cdot\lambda_p = 0 \qquad \forall\, j \text{ with } \delta_j = 1 \tag{A8}$$

$$\frac{\partial\Lambda}{\partial \lambda_p} = \sum_{j=1}^{J}\delta_{jp}\cdot\left(c_j+\delta_j\cdot f_j\right)-\sum_{i=1}^{I}\delta_{ip}\cdot D_i = 0 \qquad \forall\, p \text{ with } \delta_{ip} = 1 \tag{A9}$$

Note that only (A7) has changed, reflecting that the marginal external costs of a trip may now concern more links, namely also those links x for which the cross effect $\partial c_x/\partial N_j$ is positive for any link j which is part of the path p. The second line of (A7) shows that, for the same reason, a larger number of other path-flows may now be affected by marginal changes in N_p.

Substitution of (A9) into (A7) for each p for which $\delta_{ip}=1$ subsequently yields the following expression for the Lagrangian multipliers λ_p:

$$\lambda_p = \frac{\sum_{j=1}^{J}\delta_{jp}\cdot\left(\sum_{x=1}^{J}\sum_{q=1}^{P}\delta_{xq}\cdot N_q\cdot\frac{\partial c_x}{\partial N_j}\right)-\sum_{q=1,q\neq p}^{P}\lambda_q\cdot\left(\sum_{j=1}^{J}\delta_{jp}\cdot\sum_{x=1}^{J}\delta_{xq}\cdot\frac{\partial c_x}{\partial N_j}\right)-\sum_{j=1}^{J}\delta_{jp}\cdot\delta_j\cdot f_j}{\sum_{j=1}^{J}\delta_{jp}\cdot\sum_{x=1}^{J}\delta_{xp}\cdot\frac{\partial c_x}{\partial N_j}-\sum_{i=1}^{I}\delta_{ip}\cdot D_i'} \tag{A10}$$

$$\forall\, p \text{ with } \delta_{ip}=1 \quad\text{and}\quad \forall\, q \text{ with } \delta_{iq}=1$$

As in the model in the main text, also here the system of equations (A10) should in principle have a unique solution for each λ_p, because it again makes up a system of X equations, generally linearly independent, in X unkowns (the λ_p's), where X denotes the number of relevant paths in the second-best optimum.

Finally, substitution of (A10) into (A8) gives the following expression for the second-best optimal congestion fees:

$$f_j = \frac{\sum_{p=1}^{P} d_{jp} \cdot \dfrac{\sum_{m=1}^{J} d_{mp} \cdot \left(\sum_{x=1}^{J} \sum_{q=1}^{P} d_{xq} \cdot N_q \cdot \dfrac{\P c_x}{\P N_m} \right) - \sum_{q=1,q\neq p}^{P} l_q \cdot \left(\sum_{m=1}^{J} d_{mp} \cdot \sum_{x=1}^{J} d_{xq} \cdot \dfrac{\P c_x}{\P N_m} \right) - \sum_{m=1,m\neq j}^{J} d_{mp} \cdot d_m \cdot f_m}{\sum_{m=1}^{J} d_{mp} \cdot \sum_{x=1}^{J} d_{xp} \cdot \dfrac{\P c_x}{\P N_m} - \sum_{i=1}^{I} d_{ip} \cdot D_i'}}{\sum_{p=1}^{P} \dfrac{d_{jp}}{\sum_{m=1}^{J} d_{mp} \cdot \sum_{x=1}^{J} d_{xp} \cdot \dfrac{\P c_x}{\P N_m} - \sum_{i=1}^{I} d_{ip} \cdot D_i'}} \tag{A11}$$

$$\forall\ j \text{ with } d_j = 1 \quad \text{and} \quad \forall\ p \text{ with } d_{ip} = 1 \quad \text{and} \quad \forall\ q \text{ with } d_{iq} = 1$$

which is again quite similar to (11).

Indeed, in general, the extension presented in this appendix is, from the analytical viewpoint, only a minor one. The interpretation of the model's solution as given in (A10) and (A11) is largely analogous to the interpretation of the model with only link-specific congestion, with the main differences being the generally increased number of path-flows that are affected by each individual path-flow, and the fact that the congestion effects $\partial c_x/\partial N_j$ need of course not be equal for each j (for a given x), whereas in the simpler model only terms c'_j played a role.

However, notwithstanding the modest extension from an analytical viewpoint, the generalized results given in this appendix makes the model presented in the main text of course applicable to an even wider range of network problems, which justifies the discussion just given.