TECHNICAL WORKING PAPER SERIES

WHEN ARE ANONYMOUS CONGESTION CHARGES CONSISTENT WITH MARGINAL COST PRICING?

Richard Arnott
Marvin Kraus

Technical Working Paper No. 154

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
April 1994

This paper is part of NBER's research program in Public Economics. Any opinions expressed are those of the authors and not those of the National Bureau of Economic Research.

NBER Technical Working Paper #154
April 1994

WHEN ARE ANONYMOUS CONGESTION CHARGES CONSISTENT WITH MARGINAL COST PRICING?

ABSTRACT

There are constraints on pricing congestible facilities. First, if heterogeneous users are observationally indistinguishable, then congestion charges must be anonymous. Second, the time variation of congestion charges may be constrained. Do these constraints undermine the feasibility of marginal cost pricing, and hence the applicability of the first-best theory of congestible facilities? We show that if heterogeneous users behave identically when using the congestible facility and if the time variation of congestion charges is unconstrained, then marginal cost pricing is feasible with anonymous congestion charges. If, however, the time variation of congestion charges is constrained, optimal pricing with anonymous congestion charges entails Ramsey pricing.

Richard Arnott
Department of Economics
Boston College
Chestnut Hill, MA 02167
and NBER

Marvin Kraus
Department of Economics
Boston College
Chestnut Hill, MA 02167

When Are Anonymous Congestion Charges Consistent with Marginal Cost Pricing?

With any congestible facility, whether it be a road, a swimming pool, or a telephone network, there are certain constraints on congestion pricing. For one thing, congestion charges can be differentiated across heterogeneous users only on the basis of *observable* differences. For another, there may be constraints on the time variation of congestion charges. Do these constraints undermine the feasibility of marginal (social) cost pricing? Put alternatively, what is the scope of application of the first-best theory of congestible facilities (which includes for example the well-known result that, for a congestible facility which exhibits constant long run average cost, the revenue from the optimal congestion charge exactly covers capacity costs when the facility's capacity is optimal)?

We came to this issue via consideration of urban auto congestion pricing, which is being actively discussed in Europe (see May (1993) for a review) and will presumably be seriously considered in North America before too long, particularly if the European policy experiments are successful. We posed the following problem: Suppose that rush-hour commuters behave the same way in traffic, but differ from one another in *observationally indistinguishable* ways such as work starting time or the shadow value of time. Then tolling (congestion pricing) must be anonymous. *Does the anonymity of tolling in this situation preclude the possibility of marginal cost pricing?* If it does, then the design of the optimal toll is an exercise in the theory of the second best, and the optimal toll will be some variant of Ramsey pricing. A toll is said to have the *form* $\Omega(\cdot)$ if it is a member of the family of time-dependent toll functions $\tau(t) = \Omega(t) + k$, where k, a constant, is the toll *level*. *Does the answer depend on the form of the toll?* For example, does constraining the toll to be uniform over the rush hour affect the answer?

Prior work throws some light on the issue. Arnott, de Palma, and Lindsey (1993) demonstrated, *in the context of Vickrey's (1969) bottleneck model of morning rush-hour auto congestion, that with identical commuters, marginal cost pricing can be achieved independently of the form of the toll.* The argument, extended to allow for other congestion technologies, is as follows: Consider two points, A and B, where consumers live and work, respectively. A and B are connected by a single congestible road, and driving is the only way to commute. Consumers have a common work starting time t^*. A consumer who arrives at work before or after t^* incurs a *schedule delay cost.* User cost equals schedule delay cost plus travel time cost. And trip price equals user cost plus the toll. Each commuter decides when to depart from home so as to minimize trip price. The crucial point is that *in equilibrium the trip price must be constant over the departure interval and at least as high outside this interval.* Otherwise, some commuters would change their departure times. Given any number of daily trips, solving for equilibrium entails solving for the time pattern of departures, which determines a time pattern of travel time cost and of schedule delay cost, such that the equilibrium trip-price condition is satisfied. In the bottleneck model, equilibrium is unique, so that total user costs can be determined as a function of the number of daily trips and the form of the toll, i.e. $TC = TC(N, \Omega)$, where N is the number of daily trips. Marginal social cost is then determined as $MSC(N, \Omega) = \partial TC(N, \Omega)/\partial N$. Thus, marginal social cost, too, is constant over the equilibrium departure interval. Now consider some arbitrary change in the toll level, k. For any given value of N, this has no effect on the equilibrium departure pattern and thus on marginal social cost. However, the equilibrium trip price goes through the same change as k. Thus, regardless of the form of the toll, k can be set so that price equals marginal social cost. Furthermore, with this toll level, since each commuter pays the marginal social cost of a trip, the toll equals the congestion externality throughout the departure interval.

Arnott and Kraus (1993) extended the bottleneck analysis of Arnott, de Palma, and Lindsey (1993) to treat heterogeneous commuters who behave the same way in traffic. Their argument with two commuter types (indexed by $i = 1, 2$) goes as follows. Given any anonymous toll function $\tau(t) = \Omega(t) + k$, total user costs in equilibrium can be determined as a function of the number of daily trips of each type, from which the corresponding marginal social cost functions, $MSC_i(N_1, N_2, \Omega)$, can be determined. Also, the trip price for each type, p_i, must be the same throughout its (equilibrium) departure interval and as least as high outside its interval. Now suppose that the toll level is adjusted such that the price of a type 1 trip equals its marginal social cost, implying that the toll equals the congestion externality of a type 1 trip throughout the departure interval for 1's. The issue is whether this results in the price of a type 2 trip equaling its marginal social cost. Sufficient conditions for this are that in equilibrium the two types have a common departure time, $\tilde{t}$, and that at $\tilde{t}$ both types' congestion externality be the same. Where $c_i(t)$ is user cost and $v_i(t)$ the congestion externality, this result follows from the following chain of equalities:

$$\begin{aligned}
p_2 &= p_2(\tilde{t}) && \text{(since there are type 2 departures at } \tilde{t}\text{)}\\
&= c_2(\tilde{t}) + \tau(\tilde{t}) && \\
&= c_2(\tilde{t}) + v_1(\tilde{t}) && \text{(since } \tau(t) = v_1(t) \text{ at any departure time used by type 1's)}\\
&= c_2(\tilde{t}) + v_2(\tilde{t}) && \text{(since } v_2(\tilde{t}) = v_1(\tilde{t})\text{)}\\
&= MSC_2(\tilde{t}) && \\
&= MSC_2. && \text{(since there are type 2 departures at } \tilde{t}\text{)}
\end{aligned}$$

In that paper it is shown that these conditions are satisfied with the optimal time-varying toll, which has both the optimal form and the optimal level, but not generally otherwise. The reason is that with the optimal time-varying toll, the congestion externality is *anonymous* -- $v_1(t) = v_2(t)$ for all t -- but with other toll forms the congestion externality is generally nonanonymous. The results can be extended to an arbitrary number of commuter types. Thus, *in the bottleneck model with heterogeneous commuters who behave the same way in traffic, marginal cost pricing is possible with the optimal anonymous time-varying toll, but not generally with anonymous tolls of other forms.*

The possibility of marginal cost pricing in these circumstances therefore turns on whether the congestion externality is anonymous or nonanonymous. How is it possible that two drivers who depart at the same time and who behave the same way in traffic can generate different congestion externalities? Perhaps the simplest way to illustrate this is with a bottleneck model with two commuter types and no toll.

All commuters have the same work starting time, t^*, and cannot arrive late. The user cost functions are

$$c_i(t) = \alpha_i(\text{travel time }(t)) + \beta(\text{time early }(t)),$$

where t is departure time, α is the shadow cost of travel time, and β the shadow cost of time early. Thus, the two types differ only in terms of α. Assume that $\alpha_1 > \alpha_2 > \beta$ -- type 1 commuters dislike being stuck in traffic more than type 2's, and both types' shadow cost of travel time exceeds the shadow cost of time early. The congestion technology is simple. There is a single bottleneck with capacity s -- only s commuters can pass through the bottleneck per unit time. If the arrival rate at the bottleneck exceeds s, a queue develops. Travel time equals queuing time. Since there is no toll, trip price equals user cost. Thus, in equilibrium all commuters of a given type will incur the same user cost. Since commuters who depart later arrive less early, they must incur greater travel costs and hence face a longer queue than those who depart earlier. All type 1 commuters, who dislike queuing more, depart earlier in the morning rush hour when the queue is shorter. Figure 1 characterizes equilibrium in terms of cumulative departures and arrivals over the rush hour. The initial equilibrium cumulative departure schedule is given by *abcd*. Cumulative arrivals are given by *ad*. The departure interval for type 1 commuters is $[t_o, \tilde{t}]$; the departure interval for 2's is $[\tilde{t}, t_f]$. The important feature to note is that to satisfy the equal-trip-price conditions, the queue (given by the vertical distance between the cumulative departure and cumulative arrival schedules) grows more rapidly over the type 2 departure interval than over the type 1 departure interval. Equilibrium requires that queuing *cost* grow at a certain rate. Since queuing cost equals the shadow cost of queuing time times queuing time, queuing time -- and therefore queue length -- must grow faster when the shadow cost of queuing time is smaller.

Now consider taking away one type 1 commuter at $\tilde{t}$ and adding a type 2 commuter at $\tilde{t}$. This causes the equilibrium cumulative departure schedule to change to *ab'c'd*, implying that all other type 2 commuters face a longer queue. Otherwise, the equilibrium is unchanged. This implies that a type 2 commuter imposes a larger congestion externality at $\tilde{t}$ than a type 1 commuter. Thus, two drivers who depart at the same time and who behave the same way in traffic can generate different

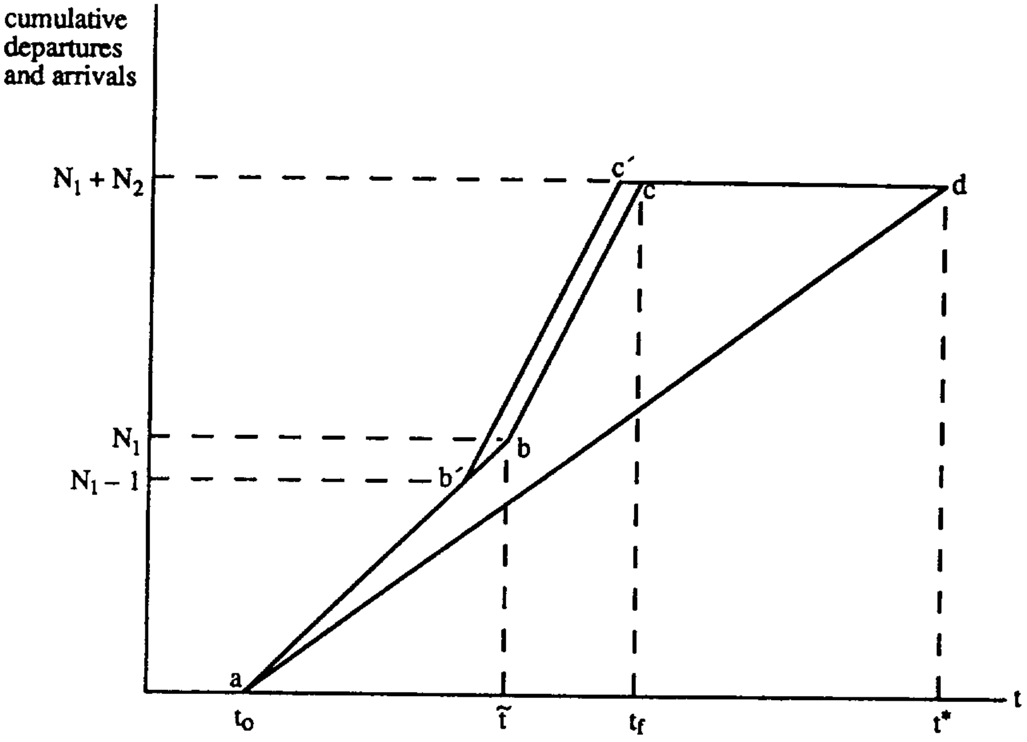

Figure 1. The No-Toll Equilibrium in the Bottleneck Model

Notes: *ad* has a slope of s. *ab* and *bc* have slopes of $\alpha_1 s/(\alpha_1 - \beta)$ and $\alpha_2 s/(\alpha_2 - \beta)$, respectively.

congestion externalities because they cause the queue to evolve in different ways in order to satisfy the equal-trip-price conditions.

With the optimal time-varying toll, there is no queue. Consequently, the cause of the difference in the two types' congestion externality disappears, so that the congestion externality is anonymous.

The obvious question is: To what extent do the results from the traffic bottleneck model generalize to other congestion technologies and to other congestible facilities? We examine this issue using a rather general model of congestion. In the course of our analysis, we generalize existing theorems concerning the extent to which the revenue from the optimal congestion charge covers capacity costs when the facility's capacity is optimal.

We show that the results do indeed generalize. With users who behave identically when using a congestible facility but differ in other, unobservable ways, the time-varying congestion externality is anonymous *with the efficient time pattern of utilization.* Setting an anonymous congestion charge equal at all points in time to the corresponding congestion externality decentralizes the efficient time pattern of utilization. Under these circumstances, anonymous congestion charges *are* consistent with marginal cost pricing. When, however, the form of the toll is constrained, the time pattern of utilization is not efficient, the congestion externality is generally nonanonymous, and anonymous congestion charges are inconsistent with marginal cost pricing.[1]

1. The Model

A congestible facility can be used by G types of consumers. The number of daily utilizations by consumers of type i (i = 1, ..., G) that begin by time of day t is denoted by $N_i(t)$. Let t_0 and t_f denote the earliest and latest times of day, respectively, that a utilization by any consumer begins. Then, for all i = 1, ..., G, the number of daily utilizations by consumers of type i is given by $N_i(t_f)$.

For a utilization beginning at time t, a consumer incurs a cost (excluding the congestion charge) given by $c_i(t, n(t), N(t), x(t), s)$, where i is the consumer's type; $N(t) = (N_1(t), ..., N_G(t))$;

$n(t) = (n_1(t), ..., n_G(t)) = (\dot{N}_1(t), ..., \dot{N}_G(t))$;[2] $x(t) = (x_1(t), ..., x_J(t))$, a vector of state variables at time t; and s is the facility's capacity. The state equations for $x_1, ..., x_J$ are given by

$$\dot{x}_j(t) = g_j(t, n(t), N(t), x(t), s) \qquad j = 1, ..., J. \tag{1}$$

$x(t_o)$ is given, while $x(t_f)$ is free. Capacity costs are given by K(s).

Observe that different consumer types may differ in terms of their cost functions or how their usage directly affects other consumers' costs or the state of the congestible facility. Later, we shall particularize the treatment of congestion to consider groups that are observationally indistinguishable and whose usage affects other users' costs and the state of the facility in the same way.

The specification of the congestion technology is as general as possible, consistent with congestion being described by a set of state variables. The interpretation of the state variables depends on context; possible state variables include air quality, temperature, blocking rate, load, ...

A type i utilization beginning at time t is subject to a congestion charge $\tau_i(t)$, which for the moment can differ across consumer types. The price of the utilization is given by

$$p_i(t) = c_i(t, n(t), N(t), x(t), s) + \tau_i(t). \tag{2}$$

Consumers choose when to begin utilizations so as to minimize price. Let t' be a time chosen by a type i consumer to begin a utilization. Then $p_i(t') = p_i^*$, where p_i^* is the minimum of $p_i(t)$ over all t.

The number of daily utilizations by type i consumers, $N_i(t_f)$, is given by the demand relationship

$$N_i(t_f) = D_i(p_i^*). \tag{3}$$

This assumes that the benefit from a utilization is independent of the time of usage. Put alternatively, utilizations at different points in time are treated as perfect substitutes in demand. Preferences regarding time of usage are captured in the cost function. Denoting the inverse of D_i by P_i,

$$p_i^* = P_i(N_i(t_f)). \tag{4}$$

Let T_i denote the set of times at which type i consumers begin utilizations. The economy's equilibrium conditions are that, for all i = 1, ..., G,

$$p_i(t) = P_i(N_i(t_f)) \qquad t \,\varepsilon\, T_i \tag{5a}$$

$$p_i(t) \geq P_i(N_i(t_f)) \qquad t \not\varepsilon\, T_i. \tag{5b}$$

Special cases. The treatment of congestion has as special cases both bottleneck and static congestion. To see this, consider two points, A and B, where consumers live and work, respectively. A and B are connected by a single road, and driving is the only way to commute. For consumers of type i, work starts at t_i^*. Let d(t) denote travel time for a consumer who leaves home at time t. If $t + d(t) \neq t_i^*$, the individual's arrival is early or late. Besides travel time costs, individuals incur schedule delay costs for arriving at work early or late.

With the simplest form of bottleneck congestion, the road has a single bottleneck having a well-defined capacity -- the maximum rate at which cars can pass through the bottleneck per hour. If arrivals at the bottleneck ever exceed this rate, then a queue forms. An individual's queuing time is his travel time from A to B.

Under these assumptions, d(t) = h(t)/s, where h(t) is the length of the queue at time t, and s is the bottleneck's capacity. As for queue length,

$$\dot{h}(t) = n_1(t) + \ldots + n_G(t) - s \qquad \text{for} \quad h(t) > 0 \tag{6a}$$

$$= \max\{n_1(t) + \ldots + n_G(t) - s, 0\} \qquad \text{for} \quad h(t) = 0, \tag{6b}$$

with $h(t_o) = 0$; $n_1(t), \ldots, n_G(t)$ and t_o are defined as above.

Now consider the general treatment of congestion with x(t) a scalar. We want to make x(t) correspond to d(t).[3] It is easily checked that bottleneck congestion corresponds to the special case of the general treatment of congestion in which $x(t_o) = 0$,

$$\dot{x}(t) = (n_1(t) + \ldots + n_G(t) - s)/s \qquad \text{for} \quad x(t) > 0 \tag{7a}$$

$$= \max\{n_1(t) + \ldots + n_G(t) - s, 0\}/s \qquad \text{for} \quad x(t) = 0, \tag{7b}$$

and $c_i(\cdot)$ is of the form $c_i(t, x(t))$.

With static congestion and the road having a uniform capacity of s,[4] d(t) is given by

$$d(t) = f(n_1(t) + \ldots + n_G(t), s), \tag{8}$$

where $f(\cdot)$ is homogeneous of degree zero. Thus, static congestion corresponds to the special case of the general treatment of congestion in which $c_i(\cdot)$ is of the form $c_i(t, n_1(t) + \ldots + n_G(t), s)$ and is homogeneous of degree zero in $n_1(t) + \ldots + n_G(t)$ and s. Introducing x(t) is unnecessary.

2. Analysis

2.1 *The Social Optimum*

Let the benefits to type i consumers be given by

$$\int_0^{N_i(t_f)} P_i(q)dq. \tag{9}$$

Society's problem can then be stated as

$$\max_{s, t_o, t_f, n(\cdot)} \quad \sum_{i=1}^{G} \int_0^{N_i(t_f)} P_i(q)dq - K(s) - \int_{t_o}^{t_f} \sum_{i=1}^{G} n_i(t)c_i(t, n(t), N(t), x(t), s)dt \tag{10}$$

s.t.

$$\dot{N}_i(t) = n_i(t) \qquad i = 1, \ldots, G \tag{11}$$

$$\dot{x}_j(t) = g_j(t, n(t), N(t), x(t), s) \qquad j = 1, \ldots, J, \tag{12}$$

with initial conditions

$$N_i(t_o) = 0 \qquad i = 1, \ldots, G \tag{13}$$

$$x_j(t_o) = x_j^o \qquad j = 1, \ldots, J, \tag{14}$$

for given $x_1^o, \ldots, x_J^o$, and inequality constraints

$$n_i(t) \geq 0 \qquad i = 1, \ldots, G. \tag{15}$$

(10)-(15) define an optimal control problem with state variables $N_1, ..., N_G$ and $x_1, ..., x_J$ and control variables $n_1, ..., n_G$. Assume that $c_1(\cdot), ..., c_G(\cdot), g_1(\cdot), ..., g_J(\cdot)$ and $K(\cdot)$ have continuous first derivatives, and define the Hamiltonian

$$H(t, n(t), N(t), x(t), \mu(t), \lambda(t), s) = -\sum_{i=1}^{G} n_i(t)c_i(t, n(t), N(t), x(t), s) + \sum_{i=1}^{G} \mu_i(t)n_i(t) + \sum_{j=1}^{J} \lambda_j(t)g_j(t, n(t), N(t), x(t), s), \qquad (16)$$

where $\mu_i(t)$ and $\lambda_j(t)$ are the costate variables corresponding to (11) and (12) respectively. The first-order conditions are

$$\frac{\partial H}{\partial n_{i'}} = -c_{i'} - \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} + \mu_{i'} + \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} \leq 0 \qquad i' = 1, ..., G \qquad (17)$$

$$n_{i'}\frac{\partial H}{\partial n_{i'}} = n_{i'}(-c_{i'} - \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} + \mu_{i'} + \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}}) = 0 \qquad i' = 1, ..., G \qquad (18)$$

$$\dot{\mu}_{i'} = -\frac{\partial H}{\partial N_{i'}} = \sum_i n_i \frac{\partial c_i}{\partial N_{i'}} - \sum_j \lambda_j \frac{\partial g_j}{\partial N_{i'}} \qquad i' = 1, ..., G \qquad (19)$$

$$\dot{\lambda}_{j'} = -\frac{\partial H}{\partial x_{j'}} = \sum_i n_i \frac{\partial c_i}{\partial x_{j'}} - \sum_j \lambda_j \frac{\partial g_j}{\partial x_{j'}} \qquad j' = 1, ..., J \qquad (20)$$

$$P_i(N_i(t_f)) = \mu_i(t_f) \qquad i = 1, ..., G \qquad (21)$$

$$\lambda_j(t_f) = 0 \qquad j = 1, ..., J \qquad (22)$$

$$H(t_f, n(t_f), N(t_f), x(t_f), \mu(t_f), \lambda(t_f), s) = 0 \qquad (23)$$

$$H(t_o, n(t_o), N(t_o), x(t_o), \mu(t_o), \lambda(t_o), s) = 0 \qquad (24)$$

$$-K'(s) - \int_{t_o}^{t_f} (\sum_i n_i \frac{\partial c_i}{\partial s} - \sum_j \lambda_j \frac{\partial g_j}{\partial s})dt = 0 \qquad (25)$$

as well as (11)-(15).

2.2 *Decentralization of the Optimum*

There are two distinct ways to identify the decentralizing congestion charges, depending on how congestion externalities are conceptualized. The first approach is based on the left-hand side of (17). At any time t, this expression equals the marginal net social benefit of a type i′ utilization beginning at t, conditional on the values of state variables at t, but allowing the controls to optimally readjust at times greater than t (see, e.g., Dorfman (1969)). Because of the conditioning on state variables at t, we term this the *ceteris paribus* marginal net social benefit. In the absence of a congestion charge, a type i′ utilization beginning at t has a marginal net *private* benefit given by $P_{i'}(N_{i'}(t_f)) - c_{i'}(t)$. Subtracting this from the *ceteris paribus* marginal net social benefit and reversing the sign gives the *ceteris paribus* congestion externality. Denoting the *ceteris paribus* congestion externality by $v_{i'}(t)$,

$$v_{i'}(t) = \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} - \mu_{i'} - \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} + P_{i'}(N_{i'}(t_f)). \tag{26}$$

We now show that this is the optimal congestion charge.

Proposition 1. An optimum to (10)-(15), conditional on s, can be decentralized with the *ceteris paribus* congestion charge $\tau_{i'}(t) = v_{i'}(t)$, where $v_{i'}(t)$ is evaluated at the conditional optimum.

Proof. See the Appendix.

Remark 1. In the special case of static congestion, (26) reduces to the classic congestion externality

$$v_{i'}(t) = \sum_i n_i \frac{\partial c_i}{\partial n_{i'}}. \tag{27}$$

To see this, recall from Section 1 that when congestion is static the only state variables of the model are $N_1, ..., N_G$ and that these do not appear as arguments of the cost functions $c_i(\cdot)$. Then the third term on the right-hand side of (26) can be dropped, and (19) reduces to $\dot{\mu}_{i'} = 0$, implying that $\mu_{i'}$ in (26) is equal to $\mu_{i'}(t_f)$. The result then follows from (21).

The second approach is based on the equivalence between the optimal control problem defined by (10)-(15) and a two-stage optimization problem in which $N_1(t_f)$, ..., $N_G(t_f)$ are given at the first stage and optimized at the second stage. The problem at the first stage is to minimize the last term in (10), subject to (11)-(15) and given values for $N_1(t_f)$, ..., $N_G(t_f)$ and s. The minimum value function for this problem is the facility's short run cost function, which we denote by $C(N_1(t_f), ..., N_G(t_f), s)$. At the second stage, values for $N_1(t_f)$, ..., $N_G(t_f)$ and s are determined to maximize net social benefits, written as

$$\sum_{i=1}^{G} \int_0^{N_i(t_f)} P_i(q)dq - K(s) - C(N_1(t_f), ..., N_G(t_f), s). \tag{28}$$

For later use, we note that this requires

$$P_i(N_i(t_f)) - \partial C/\partial N_i(t_f) = 0 \qquad i = 1, ..., G. \tag{29}$$

Starting from a solution to the stage one problem, what is the marginal cost of a type i′ utilization beginning at time t when the utilization times of other consumers are adjusted, if necessary, to minimize costs? If $t \varepsilon T_{i'}$, this *mutatis mutandis* marginal cost must equal $\partial C/\partial N_{i'}(t_f)$, while if $t \not\varepsilon T_{i'}$, it cannot be less than this value. The *mutatis mutandis* congestion externality is simply the difference between the *mutatis mutandis* marginal cost and the marginal user's own cost, $c_{i'}(t)$. Denoting the *mutatis mutandis* congestion externality by $\varphi_{i'}(t)$,

$$\varphi_{i'}(t) = \partial C/\partial N_{i'}(t_f) - c_{i'}(t) \qquad \text{for} \quad t \varepsilon T_{i'} \tag{30a}$$

$$\geq \partial C/\partial N_{i'}(t_f) - c_{i'}(t) \qquad \text{for} \quad t \not\varepsilon T_{i'} \tag{30b}$$

The relationship between $\varphi_{i'}(t)$ and $v_{i'}(t)$ is given by:

Lemma 1. At an optimum to (10)-(15), conditional on s, the *ceteris paribus* congestion externality $v_{i'}(t)$ is equal to the *mutatis mutandis* congestion externality $\varphi_{i'}(t)$ for all $t \varepsilon T_{i'}$.

Proof. See the Appendix.

It can then easily be shown that

Proposition 2. An optimum to (10)-(15), conditional on s, can be decentralized with the *mutatis mutandis* congestion charge $\tau_{i'}(t) = \varphi_{i'}(t)$, where $\varphi_{i'}(t)$ is evaluated at the conditional optimum.

Proof. See the Appendix.

We next introduce the assumption

(A.1) For all i = 1, ..., G, $c_i(\cdot)$ is of the form $c_i(t, n_1(t) + ... + n_G(t), N_1(t) + ... + N_G(t), x(t), s)$, while for all j = 1, ..., J, $g_j(\cdot)$ takes the form $g_j(t, n_1(t) + ... + n_G(t), N_1(t) + ... + N_G(t), x(t), s)$.

(A.1) corresponds to our assumption in the earlier discussion that commuters behave the same way in traffic. It is a standard assumption in models in which distinct consumer types are introduced to represent differences in time valuation or, in commuting models, work start time.

Our key result is:

Lemma 2. Suppose that (A.1) holds. Then at an an optimum to (10)-(15), conditional on s, the *ceteris paribus* congestion externality $v_{i'}(t)$ is the same for all consumer types i′.

Proof. From (26) and (21) we have that at any time t′,

$$v_{i'}(t') = \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} - \mu_{i'} - \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} + \mu_{i'}(t_f) \tag{31}$$

$$= \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} - \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} + \int_{t'}^{t_f} \dot{\mu}_{i'} dt. \tag{32}$$

From (32) and (19),

$$v_{i'}(t') = \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} - \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} + \int_{t'}^{t_f} (\sum_i n_i \frac{\partial c_i}{\partial N_{i'}} - \sum_j \lambda_j \frac{\partial g_j}{\partial N_{i'}}) dt. \tag{33}$$

But under (A.1) each term on the right-hand-side of (33) is independent of i′. ▮

The intuition for Lemma 2 goes as follows. First, $v_{i'}(t)$ consists partly of an externality on consumers who begin utilizations at time t, and partly of an externality on consumers who begin utilizations later than t. The first of these is given by the first term on the right-hand-side of (26). Under (A.1), $n_1(t), ..., n_G(t)$ enter the user cost functions only through their sum, which makes this term anonymous. The externality on consumers who begin utilizations later than t comes about through the effect that $n_{i'}(t)$ has on $x_1, ..., x_J$ and $N_{i'}$ at a time shortly after t. The part of this that operates through $x_1, ..., x_J$ is given by $-\sum_j \lambda_j \partial g_j / \partial n_{i'}$, since λ_j is the marginal social value of x_j (Dorfman (1969)). (A.1) also assumes that $n_1(t), ..., n_G(t)$ enter the functions g_j only through their sum, so that this too is anonymous. Finally, the marginal social value of $N_{i'}$ is $\mu_{i'}$. The effect that operates through $N_{i'}$ is therefore $-(\mu_{i'} - P_{i'}(N_{i'}(t_f))$. The additivity conditions in (A.1) involving $N_1(t), ..., N_G(t)$ are sufficient to make this anonymous.

Together with Proposition 1, Lemma 2 implies the main result of the paper:

Proposition 3. Suppose that (A.1) holds. Then an optimum to (10)-(15), conditional on s, can be decentralized with anonymous congestion charges.

Proof. Proposition 3 is a direct implication of Proposition 1 and Lemma 2.

When there are constraints on the time variation of the congestion charge, it is not in general possible to achieve marginal cost pricing for all types. Determination of the optimal congestion charge (subject to the constraints on its time variation) is then an exercise in the theory of the second best. The optimal congestion charge will entail Ramsey pricing, under which the absolute value of the deviation from marginal cost will "on average" be higher for types with less elastic demand.

2.3 *Self-Financing Results*

In this section, we develop two self-financing results analogous to those developed by Mohring and Harwitz (1962) and Strotz (1965) for the traditional model of a congestible facility. Both depend on the assumption

(A.2) $c_1(\cdot), \ldots, c_G(\cdot)$ and $g_1(\cdot), \ldots, g_J(\cdot)$ are homogeneous of degree zero in n(t), N(t) and s.

For the case in which a nonanonymous congestion charge is possible, we have:

Proposition 4. Suppose that (A.2) holds and that s and $\tau_1(\cdot), \ldots, \tau_G(\cdot)$ are unconstrained optimal. Then the ratio of receipts from congestion charges to capacity costs is given by the elasticity of capacity costs with respect to capacity.

Proof. See the Appendix.

The intuition for Proposition 4 is as follows. The facility is an input in a joint production process in which utilizations are viewed as distinct outputs if they either are made by different consumer types or begin at different times. The output levels of the process are given by $n_1(t), \ldots, n_G(t)$ for all $t \varepsilon [t_o, t_f]$. Suppose that a scale factor θ which differs infinitesimally from one is applied to all output levels and the facility's capacity. This scales $N(\cdot)$ by θ, so, by (A.2), $x_1, \ldots, x_J$ and $c_1, \ldots, c_G$ are unchanged at all times. Total user costs are therefore scaled by θ.

Suppose, for example, that the elasticity of capacity costs with respect to capacity is equal to one. Then capacity costs are also scaled by θ, and production involves constant long run ray average costs.[5]

From (5a) and (29), all outputs whose production levels are positive are priced at long run marginal cost.[6] With constant long run ray average costs, this implies a total value of output equal to the total cost of output. The total value of output is the sum of user costs and congestion charges, while the total cost of output is the sum of user costs and capacity costs. Congestion charges and capacity costs are therefore equal.

Now suppose that the congestion charge is constrained to be anonymous. If (A.1) holds, then by Proposition 3 this is a nonbinding constraint. It follows immediately from Proposition 4 that

Proposition 5. Suppose that (A.1) and (A.2) hold and that s and $\tau_1(\cdot)$, ..., $\tau_G(\cdot)$ are optimal subject to the constraint $\tau_1(\cdot) = ... = \tau_G(\cdot)$. Then the ratio of receipts from congestion charges to capacity costs is given by the elasticity of capacity costs with respect to capacity.

Remark 2. In the basic bottleneck model, the social optimum involves a zero queue length and a full bottleneck at all times from t_o to t_f. Thus the state function associated with driving time has only one-sided derivatives with respect to n_1, ..., n_G (see (7b)). As a result, the derivative $\partial H/\partial n_i'$ (see (17)) that defines the *ceteris paribus* marginal net social benefit does not exist, and the *ceteris paribus* congestion externality is undefined. Thus, the line of argument developed in this paper does not apply to the bottleneck model.

The *mutatis mutandis* congestion externality remains well-defined, however. Arnott and Kraus (1993) use this definition of the congestion externality to prove that Propositions 2-5 hold for the bottleneck model. (A.2) holds with travel time rather than queue length as a state variable.

Remark 3. Propositions 4 and 5 generalize in a straightforward way to the case in which s is a vector of capacity variables. In the generalized propositions, the facility's cost-recovery ratio is given by the *elasticity of scale* of K(s) with respect to s.

3. Conclusion

This paper addressed the following issue: Suppose that the users of a congestible facility differ from one another in unobservable ways, so that congestion charges must be anonymous. Under what circumstances is marginal (social) cost pricing possible? To focus on the central issues, we shall assume in the following discussion that users behave the same way when using the facility (If this condition does not hold -- if, for example, there are good drivers and bad drivers who cannot be distinguished -- it should be obvious that marginal cost pricing is not possible). What distinguishes different user types are differences in time valuation and in the most preferred time of use.

The single-period version of the traditional model of a congestible facility, which assumes static congestion and does not explicitly account for schedule delay costs, would give the following answer: Since each type's user cost depends only on the total number of utilizations and not on their composition by type, each utilization imposes the same congestion externality. Consequently, application of an anonymous Pigouvian congestion charge equal to the common congestion externality achieves marginal cost pricing. To allow for time-varying congestion, divide the congestion period up into intervals. For each user type, each interval has its own demand function which depends on that type's utilization prices in all of the intervals (reflecting cross-price effects). Since each utilization is assumed to contribute to congestion in only one interval, the above argument for the feasibility of marginal cost pricing applies for each interval separately.

This treatment of time-varying congestion is unsatisfactory in ignoring that congestion in a time interval spills over into the subsequent time interval through stock or state variables, the most obvious example being a queue. In the present paper, we examined the feasibility of marginal cost pricing for a congestible facility when users differ in unobservable ways, employing a model which not only accounts for such stock congestion effects but also explicitly models users' schedule delay costs. Our main result was that, with an anonymous congestion charge, marginal cost pricing is feasible when the time variation of the congestion charge is unconstrained but not generally otherwise.

What is the significance of this result? On the positive side, it indicates that unobservable differences between individuals do not undermine the feasibility of marginal cost pricing. Thus, first-best rules for optimal capacity and first-best results on self-financing apply. On the negative side, our main result implies that when the time variation of congestion charges is constrained, the unobservability of user characteristics in general renders marginal cost pricing infeasible. Not only does this generate inefficiency, but also calculation of the optimal congestion charge over time (subject to the constraints on its time variation -- an exercise in Ramsey pricing) as well as of optimal capacity becomes considerably more difficult and informationally demanding. Thus, our paper makes a theoretical case for flexible, time-varying congestion charges.

Appendix

Proof of Proposition 1. Suppose that $\tau_{i'}(t) = v_{i'}(t)$ and that (10)-(15) is at an optimum, conditional on s. The proof amounts to showing that (5a)-(5b) hold.

Given any consumer type i′, first consider $t \in T_{i'}$. This is equivalent to $n_{i'}(t) > 0$, so from (18) we have that at t,

$$-c_{i'} - \sum_i n_i \frac{\partial c_i}{\partial n_{i'}} + \mu_{i'} + \sum_j \lambda_j \frac{\partial g_j}{\partial n_{i'}} = 0. \tag{34}$$

From (26), (34) and $\tau_{i'}(t) = v_{i'}(t)$,

$$c_{i'} + \tau_{i'}(t) = P_{i'}(N_{i'}(t_f)). \tag{35}$$

Finally, from (2),

$$p_{i'}(t) = P_{i'}(N_{i'}(t_f)). \tag{36}$$

If $t \notin T_{i'}$, then (17) holds at t. Starting from this relationship and carrying out the same steps as above gives

$$p_{i'}(t) \geq P_{i'}(N_{i'}(t_f)). \quad ■$$

Proof of Lemma 1. For $t \in T_{i'}$, (34) holds, from which (26) can be rewritten

$$v_{i'}(t) = P_{i'}(N_{i'}(t_f)) - c_{i'}(t). \tag{37}$$

Using (29), this becomes

$$v_{i'}(t) = \partial C/\partial N_{i'}(t_f) - c_{i'}(t), \tag{38}$$

which is the same expression as in (30a). ■

Proof of Proposition 2. Suppose that $\tau_{i'}(t) = \varphi_{i'}(t)$ and that (10)-(15) is at an optimum, conditional on s. As with Proposition 1, the proof amounts to showing that (5a)-(5b) hold.

Given any consumer type i′, first consider $t \in T_{i'}$. By Lemma 1, $\varphi_{i'}(t) = v_{i'}(t)$, which together with Proposition 1, establishes (5a).

If $t \notin T_{i'}$, then from (30b), $\varphi_{i'}(t) \geq \partial C/\partial N_{i'}(t_f) - c_{i'}(t)$. Using (29) and $\tau_{i'}(t) = \varphi_{i'}(t)$, this implies

$$\tau_{i'}(t) \geq P_{i'}(N_{i'}(t_f)) - c_{i'}(t). \tag{39}$$

(5b) then follows from (39) and (2). ▮

Proof of Proposition 4. Multiplying (25) by s,

$$-sK'(s) - \int_{t_o}^{t_f} \Big(\sum_i n_i s\frac{\partial c_i}{\partial s} - \sum_j \lambda_j s\frac{\partial g_j}{\partial s}\Big)dt = 0. \tag{40}$$

Expanding (18) and summing over i′,

$$-\sum_{i'} n_{i'}c_{i'} - \sum_{i'}\sum_{i} n_{i'}n_i\frac{\partial c_i}{\partial n_{i'}} + \sum_{i'} n_{i'}\mu_{i'} + \sum_{i'}\sum_{j} n_{i'}\lambda_j\frac{\partial g_j}{\partial n_{i'}} = 0. \tag{41}$$

Reversing the order of summation in (41),

$$-\sum_{i'} n_{i'}c_{i'} - \sum_{i} n_i \sum_{i'} n_{i'}\frac{\partial c_i}{\partial n_{i'}} + \sum_{i'} n_{i'}\mu_{i'} + \sum_{j} \lambda_j \sum_{i'} n_{i'}\frac{\partial g_j}{\partial n_{i'}} = 0. \tag{42}$$

Multiplying (19) by $N_{i'}$, summing over i′, and reversing the order of summation,

$$\sum_{i'} N_{i'}\dot{\mu}_{i'} - \sum_{i} n_i \sum_{i'} N_{i'}\frac{\partial c_i}{\partial N_{i'}} + \sum_{j} \lambda_j \sum_{i'} N_{i'}\frac{\partial g_j}{\partial N_{i'}} = 0. \tag{43}$$

Adding (42) and (43),

$$-\sum_{i'} n_{i'}c_{i'} + \sum_{i'} n_{i'}\mu_{i'} + \sum_{i'} N_{i'}\dot{\mu}_{i'} - \sum_{i} n_i \sum_{i'} \Big(n_{i'}\frac{\partial c_i}{\partial n_{i'}} + N_{i'}\frac{\partial c_i}{\partial N_{i'}}\Big)$$
$$+ \sum_{j} \lambda_j \sum_{i'} \Big(n_{i'}\frac{\partial g_j}{\partial n_{i'}} + N_{i'}\frac{\partial g_j}{\partial N_{i'}}\Big) = 0. \tag{44}$$

Integrating (44) from t_o to t_f and adding to (40),

$$-\int_{t_o}^{t_f} \sum_{i'} n_{i'}c_{i'}\,dt + \int_{t_o}^{t_f} \sum_{i'} n_{i'}\mu_{i'}\,dt + \int_{t_o}^{t_f} \sum_{i'} N_{i'}\dot{\mu}_{i'}\,dt - \int_{t_o}^{t_f} \sum_{i} n_i \Big[\sum_{i'} \Big(n_{i'}\frac{\partial c_i}{\partial n_{i'}} + N_{i'}\frac{\partial c_i}{\partial N_{i'}}\Big) + s\frac{\partial c_i}{\partial s}\Big]dt$$
$$+ \int_{t_o}^{t_f} \sum_{j} \lambda_j \Big[\sum_{i'} \Big(n_{i'}\frac{\partial g_j}{\partial n_{i'}} + N_{i'}\frac{\partial g_j}{\partial N_{i'}}\Big) + s\frac{\partial g_j}{\partial s}\Big]dt = sK'(s). \tag{45}$$

By (A.2), both of the expressions that appear in square brackets in (45) are equal to zero, and (45) reduces to

$$-\int_{t_o}^{t_f} \sum_{i'} n_{i'} c_{i'} \, dt + \int_{t_o}^{t_f} \sum_{i'} n_{i'} \mu_{i'} \, dt + \int_{t_o}^{t_f} \sum_{i'} N_{i'} \dot{\mu}_{i'} \, dt = sK'(s). \qquad (46)$$

The third term can be rewritten,

$$\int_{t_o}^{t_f} \sum_{i'} N_{i'} \dot{\mu}_{i'} \, dt = \sum_{i'} \int_{t_o}^{t_f} N_{i'} \dot{\mu}_{i'} \, dt$$

$$= \sum_{i'} \left(N_{i'} \mu_{i'} \Big|_{t_o}^{t_f} - \int_{t_o}^{t_f} n_{i'} \mu_{i'} \, dt \right) \qquad \text{(using integration by parts)}$$

$$= \sum_{i'} P_{i'}(N_{i'}(t_f)) N_{i'}(t_f) - \int_{t_o}^{t_f} \sum_{i'} n_{i'} \mu_{i'} \, dt \qquad \text{(using (13) and (21)),}$$

from which (46) becomes

$$\sum_{i'} P_{i'}(N_{i'}(t_f)) N_{i'}(t_f) - \int_{t_o}^{t_f} \sum_{i'} n_{i'} c_{i'} \, dt = sK'(s). \qquad (47)$$

The left-hand side of (47) is the aggregate revenue from congestion charges; the proof is completed by dividing (47) by K(s). ▮

Remark. Allowing for kinks in the functions $N_i(\cdot)$ requires only minor changes in the preceding proof. We assume that there can be a finite number of times $t_1, \ldots, t_L$ for which a kink exists along the optimal time path of one or more of the N_i's. Then in the preceding proof, each integral over $[t_o, t_f]$ must be written instead as a sum of integrals over the subintervals $[t_o, t_1]$, $[t_1, t_2], \ldots, [t_L, t_f]$. As a result, the integration by parts step following (46) generates an expression which includes

$$\sum_{i'} \left(N_{i'} \mu_{i'} \Big|_{t_o}^{t_1^-} + N_{i'} \mu_{i'} \Big|_{t_1^+}^{t_2^-} + \ldots + N_{i'} \mu_{i'} \Big|_{t_{L-1}^+}^{t_L^-} + N_{i'} \mu_{i'} \Big|_{t_L^+}^{t_f} \right), \qquad (48)$$

where a minus indicates that the evaluation is to be carried out as a limiting operation from the left, and a plus from the right. Optimality now requires that for all $i' = 1, ..., G$, $\mu_{i'}$ is continuous at each of the times $t_1, ..., t_L$, from which (48) reduces to

$$\sum_{i'} N_{i'}\mu_{i'}\Big|_{t_o}^{t_f}, \tag{49}$$

which is the same term as before.

References

Arnott, R., de Palma, A. and Lindsey, R. (1993). "A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand," *American Economic Review*, 83, 161-179.

Arnott, R. and Kraus, M. (1993). "Financing Capacity in the Bottleneck Model," mimeo.

Dorfman, R. (1969). "An Economic Interpretation of Optimal Control Theory," *American Economic Review*, 59, 817-831.

Henderson, J.V. (1981). "The Economics of Staggered Work Hours," *Journal of Urban Economics*, 9, 349-364.

May, A.D. (1993). "Transportation Research Board Study of Urban Transportation Congestion Pricing: Potential of Next-Generation Technology," mimeo.

Mohring, H. (1970). "The Peak Load Problem with Increasing Returns and Pricing Constraints," *American Economic Review*, 60, 693-705.

Mohring, H. and Harwitz, M. (1962). *Highway Benefits: An Analytical Framework*, Northwestern University Press, Evanston.

Strotz, R.H. (1965). "Urban Transportation Parables," in J. Margolis, ed., *The Public Economy of Urban Communities*, Resources for the Future, Washington.

Vickrey, W.S. (1969). "Congestion Theory and Transport Investment," *American Economic Review Proceedings*, 59, 251-260.

Footnotes

[1] To those familiar with the traditional model of a congestible facility (e.g. Mohring (1970)), the results from the bottleneck model must seem peculiar. One reason is that the traditional model implicitly *assumes* that the congestion externality is the same for users who behave the same way when using the congestible facility. The traditional model starts with a user cost function for each group: $c_i^t = c_i(n_1^t, ..., n_G^t, s)$, where n_i^t is the number of utilizations by type i users during time interval t. It then *assumes* that if different types behave the same way when using the facility, $c_i^t = c_i(n_1^t + ... + n_G^t, s)$. Then total user costs during the interval are $\sum_i n_i^t c_i(n^t, s)$, where $n^t = \sum_i n_i^t$, so that $v_i^t = \sum_{i'} n_{i'}^t \partial c_{i'}(n^t, s)/\partial n^t$ -- the congestion externality is anonymous. This result stems from the assumption that user cost in a time interval depends only on the number of utilizations in that interval. Hence, it ignores the dynamic nature of congestion -- that history matters, that past usage affects current congestion, as manifested by a queue for example. Another difference between the traditional model and the bottleneck model is the treatment of demand. In the traditional model, utilizations in different time intervals are treated as different and imperfectly substitutable commodities, while in the bottleneck and related models (*dynamic, structural models of a congestible facility*) utilizations beginning at different times are treated as perfect substitutes with time-of-use preferences being captured via user costs. According to the traditional model, with users who behave the same way when using the facility, marginal cost pricing can be achieved with an optimal time-varying toll but not generally otherwise, whether there is one or many user types. We believe, as is argued in Arnott, de Palma, and Lindsey (1993), that the dynamic structural models of congestible facilities are superior, since they model explicitly the congestion technology and time-of-use decisions, which enforces conceptual consistency.

[2] If type i consumers begin utilizations only over some subinterval of $[t_0, t_f]$, then $N_i(\cdot)$ may have a kink at either the beginning or the end of the subinterval. In a Remark following the proof of Proposition 4 in the Appendix, we show that allowing for kinks in the functions $N_i(\cdot)$ does not affect our results.

3. $x(t)$ can also be made to correspond to $h(t)$; our reason for choosing $d(t)$ will become clear below.

4. In footnote 1, we discussed the traditional model of a congestible facility. That model incorporates static congestion, but also assumes, in contrast to our specification, that utilizations at different points in time are imperfect substitutes.

   Henderson (1981) presents a model which combines static congestion with the assumption that utilizations taken at different points in time are perfect substitutes in demand.

5. Having capacity change at the same proportionate rate as output levels in the argument is justified by the Envelope Theorem.

6. When evaluated at the optimal capacity, $\partial C/\partial N_i(t_f)$ is long run marginal cost.