

**A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand**



Richard Arnott; Andre de Palma; Robin Lindsey

*The American Economic Review*, Vol. 83, No. 1 (Mar., 1993), 161-179.

Stable URL:

<http://links.jstor.org/sici?sici=0002-8282%28199303%2983%3A1%3C161%3AASMOPC%3E2.0.CO%3B2-Q>

*The American Economic Review* is currently published by American Economic Association.

---

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/aea.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact [jstor-info@umich.edu](mailto:jstor-info@umich.edu).

# A Structural Model of Peak-Period Congestion: A Traffic Bottleneck with Elastic Demand

By RICHARD ARNOTT, ANDRÉ DE PALMA, AND ROBIN LINDSEY\*

*This paper considers the modeling of road congestion subject to peak-load demand. The standard model contains ambiguities and is poorly specified. These problems can be eliminated by working with a structural model that explicitly treats the congestion technology and drivers' behavioral decisions. The paper provides a detailed analysis of a particular structural model—William Vickrey's model of bottleneck congestion in the morning rush-hour auto commute, extended to treat elastic (i.e., price-sensitive) demand—and examines some economic implications of the structural approach. (JEL R41)*

This paper considers the modeling of road congestion subject to peak-load demand. It argues that a properly specified model should be *structural*, that is, be derived explicitly treating the congestion technology and consumers' behavioral decisions.

The standard model of a facility subject to peak-period congestion is specified as follows (e.g., Dennis J. Aigner and Joseph G. Hirschberg, 1985; Ronald R. Braeutigam, 1989).<sup>1</sup> The period of use is divided into

finite time intervals. A separate analysis is conducted in price–quantity space for each of these intervals. The quantity is the number of facility users. Price is the sum of the user's toll or fee and his private cost (net of the toll, but incorporating congestion). Marginal social cost is the increase in social cost in that interval from an additional user in that interval. Both private and marginal social cost are viewed as depending on the number of users and capacity and as being determined by the technology. The position of the demand curve depends on both the time interval and the equilibrium prices for the other time intervals (because of intertemporal substitution). Equilibrium is at the point of intersection of the demand and supply (relating price to the number of facility users) curves, and the social optimum at the point of intersection of the demand (marginal-social-benefit) and marginal-social-cost curves.

This specification contains a number of ambiguities. To illustrate, consider a stylized model of traffic in the morning rush hour in which congestion occurs at only a single bottleneck, the entry point to the central business district. This bottleneck has fixed capacity, and if the number of drivers arriving at the bottleneck exceeds this capacity a queue forms. Even for this simple

\*Arnott: Department of Economics, Boston College, Chestnut Hill, MA, 02167-3806; de Palma: Department of Civil Engineering, Northwestern University, Evanston, IL, 60208-3109; Lindsey: University of Alberta, Edmonton, Alberta, Canada, T6G 2H4. Financial support from NSERC, NATO, and NSF SES-8912335 is gratefully acknowledged. We also thank seminar participants at Universitat Autònoma de Barcelona, Calgary, Harvard, Laval, Princeton, Saskatchewan, and Trent universities, the London School of Economics, and the Séminaire René Roy, Paris, as well as Marvin Kraus, for helpful comments. This paper was redrafted while the authors were visiting GREQE, Marseilles; we thank the group there for its hospitality.

<sup>1</sup>Early works on peak-load pricing include Jules Dupuit (1849), Arthur C. Pigou (1920), and Frank Knight (1924); all of these authors recognized the application of the principle of marginal-cost pricing to traffic congestion. Major contributions to the modern theory of peak-load pricing include Marcel Boiteux (1949), Peter O. Steiner (1957), Oliver E. Williamson (1966), and Michael A. Crew and Paul R. Kleindorfer (1986). Important papers applying the theory to urban transportation include Alan A. Walters (1961), Herbert D. Mohring and Mitchell Harwitz (1962), Robert H.

Strotz (1965), Marvin C. Kraus et al. (1976), and Theodore E. Keeler and Kenneth A. Small (1977).

example, there are several respects in which it is not clear how to apply the standard model. First, what is the appropriate measure of the number of users in a time interval? Is it the number of drivers joining the queue or the number passing through the bottleneck? This ambiguity stems from a failure to model the congestion technology explicitly. Second, the private cost in an interval depends not only on the number of users in that interval, but also on queue length, which reflects congestion in previous intervals. This illustrates a point that transport engineers have stressed: that congestion is inherently a dynamic phenomenon. Third, and relatedly, the addition of a driver in an interval increases the queue length and hence social cost not just in that interval, but in other intervals as well. This list of ambiguities suggests that the standard model of peak-load congestion in which the period of use is divided into intervals is poorly specified.

In recent years a new structural model of peak-period congestion that does not contain these ambiguities has been extensively developed. The main innovation is to treat explicitly the user's behavioral decisions. The literature has focused on the time-of-use decision, whereby the user trades off the cost of using the facility at an inconvenient time against the congestion cost of using the facility when it is crowded. The structural approach also provides an explicit treatment of the congestion technology.

The seminal paper is one by William Vickrey (1969), and the example we gave above is based on his model. He examined equilibrium with a fixed number of identical drivers on a point-input, point-output road in the morning rush hour, along which there is a single bottleneck of fixed capacity. If the arrival rate at the bottleneck exceeds capacity, a queue develops. All drivers wish to arrive at work at the same time. (Actually, Vickrey [1969] considered a situation with a distribution of work start times, but the bulk of the literature assumes a common start time.) This is physically impossible. As a result, each driver faces a trade-off in deciding when to leave home. If she leaves early, she faces no queue but arrives at work inconveniently early; if she leaves so

as to arrive on time, she faces a long queue; and if she leaves late, she faces no queue, but arrives inconveniently late. Equilibrium obtains when the queue length over time is such that no driver can reduce her trip price by changing her departure time. Vickrey not only solved for the no-toll equilibrium, but also determined the social optimum and the time-varying toll which decentralizes it and made some astute observations concerning optimal capacity and equilibrium with two parallel roads.

The Vickrey (1969) model has been elaborated in a large number of papers in transportation science (e.g., Chris Hendrickson and George Kocur, 1981), but with the exception of papers by the present authors (e.g., de Palma and Arnott, 1986; Arnott et al., 1989, 1990; Arnott and Kraus, 1990) and by Ralph M. Braid (1989) and Yuval Cohen (1987), it has been ignored by economists.<sup>2</sup> This is unfortunate since the rich economic implications of the model have not received the attention they merit. In this paper, we extend the model, focusing on its economics. Our paper makes four contributions to the literature on peak-period congestion.

First, the paper illustrates the power of the structural approach in analyzing alternative pricing regimes and provides the first treatment of elastic (i.e., price-sensitive) trip demand in the context of the Vickrey (1969) bottleneck model (but see footnote 2). Equilibrium is characterized for four pricing regimes (in order of increasing sensitivity: no toll, the optimal uniform toll, the optimal step ["coarse"] toll, and the optimal time-varying ["fine"] toll) with capacity exogenous and then with capacity chosen optimally. The four regimes are then compared.

<sup>2</sup>Arnott et al. (1990) formalizes Vickrey's (1969) model, retaining his assumption of inelastic trip demand, and extends it to treat a coarse toll and to solve for optimal capacity. Cohen (1987) and Arnott et al. (1989) extend the Vickrey model to allow for driver heterogeneity. Braid (1989), who wrote his paper independently of ours, compares the social optimum and competitive equilibrium when trip demand is elastic and capacity fixed, with a general trip-cost function. Finally, Arnott and Kraus (1990) examines Ramsey pricing, employing the bottleneck model.

Among other results, we show that, with demand elasticity less than unity, optimal capacity is larger the less sensitive is the (optimal) toll.

Second, the paper provides insight into the appropriate use and interpretation of the standard model of a congestible facility with peak-period congestion. In the Vickrey (1969) model the departure rate from home over the rush-hour period adjusts such that trip price is constant over the period. Equilibrium is therefore determined simultaneously over the entire period, and it makes no sense to compute equilibrium over only part of the period. Thus, the standard procedure of dividing up the period of use into intervals and solving separately for equilibrium in each interval is conceptually unsound. We show, however, that the standard model can be interpreted as providing a reduced-form representation of the structural Vickrey model (since price and the number of users are endogenous, "semi-reduced" form is perhaps more appropriate terminology), as long as the entire rush-hour period is treated as a single interval.

Third, the results of the paper reinforce those in Arnott et al. (1990) in suggesting that the gains from the efficient pricing of road congestion may be considerably greater than those estimated in previous studies (reviewed in Arnott et al. [1990]). Pricing influences an individual's choice of home location, lot size, trip frequency, mode and route choice, and departure time. Also, by influencing land rents, the form and level of urban transport pricing affects urban spatial structure. The literature in urban economics has treated the rush hour as a period of fixed length with uniform congestion. In doing so, it has been able to incorporate all the margins of adjustment noted above except individuals' departure-time decisions. That this paper and Arnott et al. (1990) obtain efficiency gains from road tolling that are substantially greater than those previously estimated suggests that a considerable fraction of the gains result from the change induced in the time pattern of road usage over the peak period. This in turn suggests that more serious consideration should be given to pricing schemes in urban transport that smooth the peak; it also suggests that

the benefits from employing technologically sophisticated pricing schemes for urban roads, of the sort advocated for many years by Vickrey (1963, 1971) and discussed in Sanford F. Borins (1986) and Kraus (1989), may have been substantially underestimated.

Fourth, the paper discusses the degree to which roads should be self-financing. Mohring and Harwitz (1962) and Strotz (1965) (MHS hereafter) proved that if the total cost associated with a road (private costs plus capacity cost) doubles when both capacity and usage double, then revenue from the optimal toll equals the cost of constructing and maintaining optimal capacity, so that the road should be self-financing. MHS also examined the cases in which total cost is homogeneous of degree greater and less than one in capacity and usage.<sup>3</sup> We show that the MHS results extend to the basic bottleneck model independently of the tolling regime, as long as the toll is set optimally conditional on the regime. This result has important implications. For example, if toll collection on roads is prohibitively expensive and a gasoline tax is the only policy instrument available to the government which affects the price of a trip, then with optimal capacity (conditional on a gasoline tax) the tax rate should be raised to the point at which revenues from the tax satisfy the MHS conditions.

We shall develop our analysis using the Vickrey (1969) model with identical individuals. Subsequently, we shall discuss briefly how our results extend to heterogeneous individuals. Finally, we shall comment on how the structural approach to modeling urban road congestion can be adapted to develop structural models of other congestible facilities.

Section I examines the economics of the basic Vickrey (1969) model of road congestion. Equilibria with elastic trip demand and fixed capacity for the four pricing regimes are solved for and compared in Section II.

<sup>3</sup>Mohring and Harwitz (1962) and Strotz (1965) developed their models specifically in the context of transport congestion. Williamson (1966) derived the same result in a more specific model, but with a broader context.

The analysis is extended to treat optimal capacity in Section III. Section IV investigates the self-financing results in this context. A numerical example is presented in Section V. Generalizations are discussed in Section VI, and concluding comments are given in Section VII.

### I. Review of the Basic Model: Fixed Number of Drivers

#### A. The Characteristics of Demand and the Congestion Technology

Every morning  $N$  identical individuals travel from home to work along the same road, each in her own car. Travel is uncongested except at a single bottleneck through which at most  $s$  cars can pass per unit time; if the arrival rate at the bottleneck exceeds  $s$ , a queue develops. Thus, the capacity constraint is a flow constraint, while the queue discipline is first-come, first-served (FIFO).

Travel time from home to work is

$$(1) \quad T(t) = T^f + T^v(t)$$

where  $T^f$  is fixed travel time,  $T^v$  is variable travel time, and  $t$  is departure time from home. Without affecting results of interest, we set  $T^f = 0$ ; thus an individual arrives at the bottleneck as soon as she leaves home and arrives at work immediately upon leaving the bottleneck. Let  $D(t)$  be the queue length (i.e., number of cars). Then, an individual's queuing time equals queue length at the time she joins the queue divided by bottleneck capacity:

$$(2) \quad T^v(t) = \frac{D(t)}{s}.$$

Let  $\hat{t}$  denote the most recent time at which there was no queue and let  $r(t)$  be the departure rate function (from home). Then,

$$(3) \quad D(t) = \int_{\hat{t}}^t r(u) du - s(t - \hat{t}).$$

The price of a trip is the sum of private cost and the toll:

$$(4) \quad p(t) = C(t) + \mathcal{T}(t).$$

Private cost, in turn, is taken to be linear in travel time and schedule delay (time early or time late):

$$(5) \quad C(t) = \alpha T^v(t) + \beta(\text{time early}) \\ + \gamma(\text{time late})$$

where  $\alpha$  is the shadow cost of travel time,  $\beta$  is the unit cost of arriving early at work, and  $\gamma$  is the unit cost of arriving late. In accordance with empirical results (Small, 1982), we assume that  $\gamma > \alpha > \beta$ . Following convention, we call the cost of arriving at work early or late "schedule delay cost" and call  $T^v(t)$  "travel time cost." The desired arrival time is  $t^*$ . Thus, "time early" is  $\max[0, t^* - t - T^v(t)]$  and "time late" is  $\max[0, t + T^v(t) - t^*]$ .

Each individual decides when to leave home. In doing so, she trades off travel time, schedule delay, and the toll. Equilibrium obtains when no individual can decrease her trip price by altering her departure time, taking all other drivers' departure times as fixed. Thus, the equilibrium is a pure-strategy Nash equilibrium with departure times as the strategy variables. This seems to be the natural equilibrium concept to employ in this context.<sup>4</sup>

#### B. The No-Toll Equilibrium

Let  $t_q$  be the beginning of the rush hour, let  $t_q'$  be the end, and let  $\tilde{t}$  be the departure time for on-time arrival [ $\tilde{t} = t^* - T^v(\tilde{t})$ ]. For early arrival [ $t \in [t_q, \tilde{t}]$ ], the equal-trip-price condition is

$$(6) \quad p(t) = \bar{p} = \alpha T^v(t) + \beta[t^* - t - T^v(t)].$$

<sup>4</sup>In this context, pure strategies appear to be more realistic than mixed strategies, since most individuals prefer a routine—to leave home at the same time every day. The mixed-strategy equilibrium is the same if aggregate stochasticity is ignored. Moshe Ben Akiva et al. (1986) demonstrate convergence to the no-toll Nash equilibrium via a specific adjustment process in which, out of equilibrium, individuals adjust their departure times according to a particular probabilistic decision rule.

Differentiation of (6) yields

$$(7) \quad \frac{dT^v(t)}{dt} = \frac{\beta}{\alpha - \beta}.$$

From (3),

$$(8) \quad \frac{dD(t)}{dt} = r(t) - s.$$

Combining (2), (7), and (8) gives

$$(9) \quad r(t) = \frac{\alpha s}{\alpha - \beta} \quad \text{for } t \in [t_q, \tilde{t}].$$

By a similar argument, it can be shown that

$$(10) \quad r(t) = \frac{\alpha s}{\alpha + \gamma} \quad \text{for } t \in [\tilde{t}, t_{q'}].$$

Thus, the queue length evolves over the rush hour such that the equal-trip-price condition is satisfied.

The individuals who depart at the beginning and end of the rush hour incur only schedule delay cost, which must be equal in equilibrium. Since arrivals are continuous over the rush hour, the length of the rush hour is  $N/s$ . These results together imply

$$(11a) \quad t_q = t^* - \left(\frac{\gamma}{\beta + \gamma}\right) \frac{N}{s}$$

$$(11b) \quad t_{q'} = t^* + \left(\frac{\beta}{\beta + \gamma}\right) \frac{N}{s}.$$

Furthermore, it is easily shown that

$$(12) \quad \tilde{t} = t^* - \left(\frac{\beta}{\alpha}\right) \left(\frac{\gamma}{\beta + \gamma}\right) \frac{N}{s}.$$

The solution is depicted in Figure 1. The vertical distance between the cumulative departures schedule and the cumulative arrivals schedule is queue length, and the horizontal distance is travel time [ $D(t')$  and  $T^v(t')$ , respectively, in the figure]. The queue builds up linearly from  $t_q$  to  $\tilde{t}$ , and then dissipates linearly until it disappears at  $t_{q'}$ .

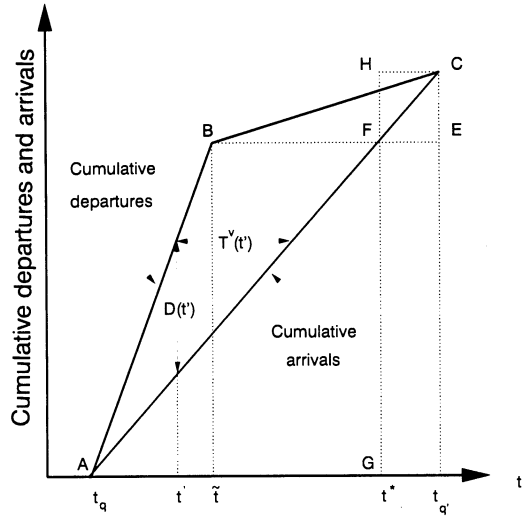


FIGURE 1. THE NO-TOLL EQUILIBRIUM

Let  $TTC$  be total travel time cost, let  $SDC$  be total schedule delay cost, and let  $TC$  be total travel cost. In Figure 1,  $TTC$  is  $\alpha$  times area  $ABCA$ , and  $SDC$  is  $\beta$  times area  $AFGA$  plus  $\gamma$  times area  $CFHC$ . It is straightforward to show that

$$(13) \quad TTC^e(N, s) = SDC^e(N, s) = \frac{\delta}{2} \left(\frac{N^2}{s}\right)$$

$$TC^e = \delta \left(\frac{N^2}{s}\right)$$

where

$$(14) \quad \delta \equiv \frac{\beta\gamma}{\beta + \gamma}$$

and superscript  $e$  denotes the no-toll equilibrium.<sup>5</sup>

Since there is no toll, price equals private cost, which is the same for everyone and equals average total travel cost,  $ATC \equiv$

<sup>5</sup>Note that all these aggregates are independent of  $\alpha$ . The start and end of the rush hour are independent of  $\alpha$ , and therefore, so is total schedule delay cost. Since  $TTC^e = SDC^e$ ,  $TTC^e$  and  $TC^e$  too are independent of  $\alpha$ . A doubling of  $\alpha$  halves queue length, preserving travel time cost.

TC/ $N$ ; that is,

$$(15) \quad p^e = C^e = ATC^e.$$

We may calculate marginal social cost (MSC) from (13):

$$(16) \quad \begin{aligned} MSC^e &= \frac{\partial TC^e}{\partial N} = \frac{2\delta N}{s} \\ &= 2 \frac{TC^e}{N} = 2ATC^e. \end{aligned}$$

Note that marginal social cost is independent of departure time. The reason is that other drivers adjust their departure times in such a way that the departure-rate function (including the additional traveler) is independent of the additional traveler's departure time.

Several observations are in order:

1. In the no-toll equilibrium, total schedule delay cost equals total travel time cost. The model highlights the costs of traveling at inconvenient times, which are hidden in the standard model.
2. Equation (15), after substitution of (13) and (14), characterizes price as a function of  $N$  and  $s$ . The resulting function is therefore the reduced-form supply function, to which a trip demand function may be added to determine the full equilibrium. The supply function captures the technology, the pricing regime considered, and commuters' time-of-use decisions. The demand function captures commuters' frequency-of-use decisions.
3. Downs's law (Anthony Downs, 1962), which is based on observation, states that an expansion of road capacity will reduce the length of the peak period more than the level of congestion at the peak. In the bottleneck model with inelastic demand, a doubling of capacity halves the length of the rush hour and the maximum travel time but leaves the maximum queue length unaltered.
4. Queuing time in the model is pure deadweight loss. If the departure rate were set at  $s$  between  $t_q$  and  $t_q'$ , queuing would be eliminated. Furthermore, since the time pattern of arrivals would be the same as in the no-toll equilibrium, total schedule delay cost would be unchanged. Thus, the social saving from the change would equal total travel time cost in the no-toll equilibrium, which equals one-half of total travel cost. This example provides a stark illustration of the potential social saving that can be achieved through altering the time pattern of departures, even when the total number of trips and the distribution of arrival times are held constant.
5. Private cost and marginal social cost are determined by the pattern of congestion over the entire period of use, and not, as is assumed in the standard analysis, by the number of users over some portion of the period.
6. Marginal social cost should be computed *mutatis mutandis*, not *ceteris paribus*; that is, marginal social cost is computed incorrectly if one adds a commuter and computes the increase in social cost from doing so, without allowing other drivers to adjust their departure times. The reason why computing marginal social cost *ceteris paribus* is incorrect is that the envelope theorem does not hold. Because there is no toll, prices are distorted. Consequently, the adjustments that commuters make in their departure times in response to the added driver alter the deadweight loss from unpriced congestion.
7. We shall see later that a change in the form of pricing alters the private and marginal social cost functions. Thus, contrary to the standard model, the private and marginal social cost functions are not completely technologically determined, but depend as well on the form of pricing. This is because the cost functions capture consumers' time-of-use decisions, which depend on the form of pricing, even with  $N$  and  $s$  fixed.

## II. Elastic Demand, Capacity Arbitrary

In the previous section, we characterized the reduced-form supply function in the ab-

sence of a toll. We now incorporate variable trip demand.

The demand for trips is assumed to be

$$(17) \quad N = N(p) \quad \frac{dN}{dp} < 0.$$

The consumers' surplus from travel with price  $p$  is

$$(18) \quad CS(p) = \int_p^\infty N(p') dp'$$

while the social surplus (gross of capacity costs) from travel, with price  $p$  and average toll  $\tau$ , is the sum of consumers' surplus and toll revenues:

$$(19) \quad SS(p, \tau) = CS(p) + N\tau.$$

#### A. The No-Toll Equilibrium

From (13) and (15),

$$(20) \quad p^e = \frac{\delta N^e}{s}$$

which is the supply function for trips in the absence of a toll. Solving this simultaneously with the demand function (17) yields the equilibrium trip price and number of trips,  $\hat{p}^e(s)$  and  $\hat{N}^e(s)$ , where a “^” over a variable indicates that it is a function of only  $s$  (plus exogenous parameters). We also have

$$(21) \quad \widehat{CS}^e(s) = \widehat{SS}^e(s) = \int_{\hat{p}^e(s)}^\infty N(p') dp'.$$

We now derive the reduced-form supply functions for the three toll regimes, as was done in the previous section for the case of no toll. Subsequently, we shall characterize equilibrium for the various toll regimes and then compare the four equilibria.

#### B. Alternative Tolling Regimes

We shall consider three different tolling regimes: a uniform toll, a fine toll, and a

coarse toll. The *uniform toll* is constant throughout the day. It does not alter the pattern of congestion over the rush hour, but with elastic trip demand it causes a reduction in the number of trips. The *fine toll* is a completely flexible time-dependent toll. The *coarse toll* is intermediate between the uniform and fine tolls. It combines a lower off-peak toll with a higher peak toll and is characterized by four parameters: the magnitudes of the peak and off-peak tolls, as well as the times at which the peak toll is applied and later removed. The average level of the coarse toll discourages travel, and the toll differential between the peak and off-peak periods shifts traffic from the peak to the off-peak period.

For each toll, we assume throughout the paper that the parameters of the toll are set optimally. The derivations of the optimal tolls are presented in Arnott et al. (1990); here we record the results and provide some intuition for them.

*Uniform Toll.*—Since the uniform toll adds a constant fee to each trip, for fixed  $N$  it does not alter the departure pattern [recall the derivations of (9) and (10)]. Thus, total travel cost is related to  $N$  and  $s$  in the same way as in the no-toll equilibrium; that is,

$$(22) \quad TC^u = \delta \left( \frac{N^2}{s} \right)$$

where superscript  $u$  denotes the uniform toll. Thus,

$$(23) \quad MSC^u = 2ATC^u.$$

Since price and the toll are the same for all commuters, so too is private cost. Also, average travel cost equals (average) private cost. Thus,

$$(24) \quad p^u = C^u + \tau^u = ATC^u + \tau^u.$$

For efficiency, the price of a trip equals marginal social cost. Hence, using (23) and

(24), one obtains

$$(25) \quad \text{ATC}^u + \tau^u = p^u = \text{MSC}^u = 2\text{ATC}^u$$

$$\tau^u = \text{ATC}^u.$$

Thus, the optimal toll equals average travel cost.

*Fine Toll.*—The social optimum can be decentralized with a time-varying toll. The social optimum entails no queue and therefore zero total travel time cost. Also, for a given number of travelers, the beginning and end of the rush hour are the same as in the no-toll equilibrium. Thus,  $\text{SDC}^o(N, s) = \text{SDC}^e(N, s)$  while  $\text{TTC}^o(N, s) = 0$ , so that, from (13),

$$(26) \quad \text{TC}^o = \text{SDC}^o = \frac{\delta}{2} \left( \frac{N^2}{s} \right)$$

where a superscript o denotes the social optimum. Efficiency requires that each traveler pays marginal social cost which equals twice the average travel cost. In the absence of a queue the person who arrives on time faces zero travel cost and for efficiency must therefore pay a toll equal to twice the average travel cost. Furthermore, the persons who depart first and last incur twice the average schedule delay and hence twice the average travel cost and should therefore pay no toll. Finally, because of the linearity of (5), the fine toll increases linearly from  $t_q$  to  $t^*$  and then decreases linearly from  $t^*$  to  $t_q$ . The average toll equals average travel cost.

*Coarse Toll.*—For this toll, too, the average toll paid equals average travel cost. This toll is more efficient than a uniform toll since it alters the queuing pattern over the rush hour, but since it does not completely eliminate queuing, it is not as efficient as the fine toll. It is shown in Arnott et al. (1990) that, with application of the coarse toll,

$$(27) \quad \text{TC}^c = \frac{\delta}{4} \left[ 3 - \frac{(\gamma - \alpha)\beta}{(\beta + \gamma)(\alpha + \beta)} \right] \frac{N^2}{s}$$

where superscript c denotes the coarse toll.<sup>6</sup> We collect the above results in the following proposition.

**PROPOSITION 1:** *Let  $j = e, u, c, o$  index the pricing regime. Then,*

$$(28a) \quad \text{TC}^j(N, s) = \frac{\Gamma^j \delta N^2}{s}$$

$$(28b) \quad \text{ATC}^j(N, s) = \frac{\Gamma^j \delta N}{s}$$

$$(28c) \quad \text{MSC}^j(N, s) = 2\text{ATC}^j(N, s)$$

where

$$(29a) \quad \Gamma^e = \Gamma^u = 1$$

$$(29b) \quad \Gamma^o = \frac{1}{2}$$

$$(29c) \quad \Gamma^c = \frac{1}{4} \left[ 3 - \frac{(\gamma - \alpha)\beta}{(\beta + \gamma)(\alpha + \gamma)} \right].$$

Also,

$$(30a) \quad p^e(N, s) = \text{ATC}^e(N, s).$$

Furthermore, for  $j = u, c, o$ , where  $\tau^j$  denotes the corresponding average toll,

$$(30b) \quad p^j(N, s) = \text{ATC}^j(N, s) + \tau^j(N, s)$$

and

$$(31) \quad p^j(N, s) = \text{MSC}^j(N, s) = 2\text{ATC}^j(N, s)$$

<sup>6</sup>The (optimal) coarse toll has the following characteristics. It is applied at the front of the queue. The peak toll is turned on before  $t^*$  and off after  $t^*$ . In the early morning, well before the peak toll is applied, a queue builds up. Then, for a period of time immediately before the peak toll is applied, there are no departures, and the queue length decreases to zero at the moment the peak toll takes effect. The queue then rises until  $t^*$  and falls after  $t^*$  until it reaches zero just before the peak toll is lifted. Right after the peak toll is lifted, there is a mass of departures and no departures subsequently.

which together imply

$$(32) \quad \tau^j(N, s) = ATC^j(N, s) \\ = \frac{MSC^j(N, s)}{2} = \frac{p^j(N, s)}{2}.$$

With the tolls set at their optimal levels, commuters pay the marginal social cost of each trip. Thus, the marginal-social-cost functions for the three toll regimes given in (28) and (29) are the corresponding reduced-form supply functions.

C. Equilibrium for the Toll Regimes

From (28) and (31), for  $j = u, c, o$ , the trip supply function for the various toll regimes is

$$(33) \quad p^j(N, s) = \frac{2\Gamma^j \delta N}{s}$$

and equations (33) and (17) together implicitly give  $\hat{N}^j(s)$  and  $\hat{p}^j(s)$ , while (32), (33), and (17) give  $\hat{\tau}^j(s)$ . Equilibrium for a particular toll regime is portrayed graphically in Figure 2. The diagram is standard. However, the demand and cost curves are defined over the entire rush hour, and it makes no sense to define them over an interval of the rush hour. Thus, the standard graphical treatment of traffic congestion is consistent with the analysis of this paper when the rush hour is treated as a single period, but not when the rush hour is divided into time intervals.

We also have, for  $j = u, c, o$ ,

$$(34) \quad \widehat{CS}^j(s) = \int_{\hat{p}^j(s)}^{\infty} N(p) dp$$

and

$$(35) \quad \widehat{SS}^j(s) = \widehat{CS}^j(s) + \hat{N}^j(s) \hat{\tau}^j(s).$$

D. Comparison of the Four Regimes

Together (28), (29), (31), and (17) imply

$$(36) \quad \hat{p}^u(s) > \hat{p}^c(s) > \hat{p}^o(s) = \hat{p}^e(s)$$

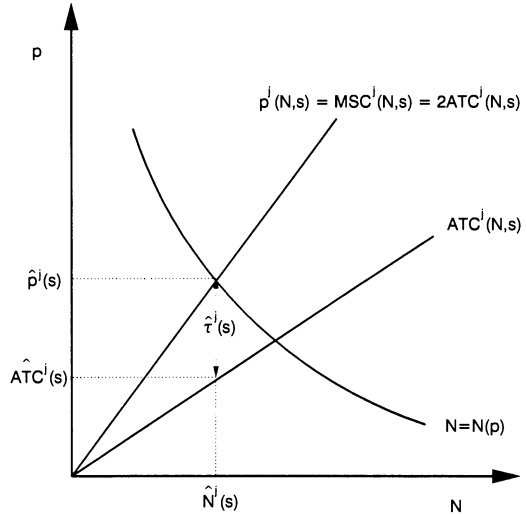


FIGURE 2. EQUILIBRIUM WITH THE OPTIMAL TOLL UNDER TOLL REGIME  $j$ :  $\hat{\tau}^j(s) = \widehat{ATC}^j(s)$

which, with variable demand, in turn implies

$$(37) \quad \hat{N}^u(s) < \hat{N}^c(s) < \hat{N}^o(s) = \hat{N}^e(s).$$

The explanation of these inequalities is straightforward. For a given number of travelers, the price of a trip is lowest for the no-toll equilibrium (since no toll is charged) and the fine-toll equilibrium (since travel on the road is efficient), intermediate for the coarse-toll equilibrium, and highest for the uniform-toll equilibrium. The coincidence of the no-toll and fine-toll supply functions reflects the particular congestion technology assumed.<sup>7</sup>

One can calculate the efficiency loss, relative to the fine toll, of the other pricing regimes:

$$(38) \quad \widehat{EL}^j(s) = \widehat{SS}^o(s) - \widehat{SS}^j(s) \quad j = c, u, e.$$

<sup>7</sup>In particular,  $SDC^e = TTC^e = SDC^o$  and  $TTC^o = 0$ , which imply that  $ATC^e = 2ATC^o$ , so that  $p^o = 2ATC^o = ATC^e = p^e$ .

The efficiency loss in the no-toll equilibrium equals the toll revenue collected at the fine-toll equilibrium. The efficiency loss in the uniform-toll and coarse-toll equilibria equals the corresponding loss in consumers' surplus plus the loss (possibly negative) of toll revenue. Since the no-toll situation is a special case of a uniform toll, which in turn a special case of a coarse toll, it is evident that

$$(39) \quad \widehat{EL}^e(s) > \widehat{EL}^u(s) > \widehat{EL}^c(s) > 0.$$

### III. Elastic Demand, Optimal Capacity

Optimal capacity is that capacity which maximizes gross social surplus less capacity costs; it occurs where the marginal benefit from capacity expansion (the gain in social surplus) equals the marginal cost. Let  $K(s)$  be the capacity-construction-cost function. The marginal capacity-expansion cost is

$$(40) \quad MC(s) = \frac{dK(s)}{ds}.$$

#### A. The No-Toll Regime

Since there is no toll, the marginal benefit is simply the gain in consumers' surplus. From (21),

$$(41) \quad \widehat{MB}^e(s) = \frac{d\widehat{CS}^e(s)}{ds} \\ = -\hat{N}^e(s) \frac{d\hat{p}^e(s)}{ds}.$$

#### B. The Toll Regimes ( $j = u, c, o$ )

From (35), the marginal benefit from capacity expansion equals the marginal consumers' surplus plus the marginal toll revenue; that is,

$$(42) \quad \widehat{MB}^j(s) = \frac{d\widehat{CS}^j(s)}{ds} + \frac{d[\hat{N}^j(s)\hat{\tau}^j(s)]}{ds}.$$

From (34),

$$\frac{d\widehat{CS}^j(s)}{ds} = -\hat{N}^j(s) \frac{d\hat{p}^j(s)}{ds}$$

and from (32),

$$\frac{d\hat{\tau}^j(s)}{ds} = \left(\frac{1}{2}\right) \frac{d\hat{p}^j(s)}{ds}.$$

Thus, from the above, and using (28b) and (32),

$$(43) \quad \widehat{MB}^j(s) = -\hat{N}^j(s) \frac{d\hat{p}^j(s)}{ds} \\ + \frac{d\hat{N}^j(s)}{ds} \hat{\tau}^j(s) + \hat{N}^j(s) \frac{d\hat{\tau}^j(s)}{ds} \\ = \frac{d\hat{N}^j(s)}{ds} \hat{\tau}^j(s) - \hat{N}^j(s) \frac{d\hat{\tau}^j(s)}{ds} \\ = \hat{N}^j(s) \hat{\tau}^j(s) \left[ \frac{1}{\hat{N}^j(s)} \frac{d\hat{N}^j(s)}{ds} \right. \\ \left. - \frac{1}{\hat{\tau}^j(s)} \frac{d\hat{\tau}^j(s)}{ds} \right] \\ = \frac{\hat{N}^j(s) \hat{\tau}^j(s)}{s} \\ = \frac{\widehat{TC}^j(s)}{s}$$

which has the following interpretation: since the tolls are set at the optimal level conditional on the tolling regime (with the result that commuters pay the marginal social cost of a trip), the analysis is first-best.<sup>8</sup> As a result, the envelope theorem holds, which in this context implies that the marginal benefit of an incremental capacity expansion is the same before and after individuals adjust

<sup>8</sup>There is a difficulty of terminology here. There are two qualitatively different sources of inefficiency possible: the *level* (or absence) and the *form* of the toll. We refer to a situation where travelers face the marginal social cost of a trip as first-best, even when the form of the toll is not optimal.

trip frequency in response to the capacity expansion. Since total travel cost with  $N$  fixed is proportional to the inverse of  $s$ , the marginal benefit of capacity expansion with trip frequency fixed is  $\widehat{TC}^j/s$ , which equals the marginal benefit of capacity expansion with trip frequency variable.

Recall from (36) that  $\hat{p}^u(s) > \hat{p}^c(s) > \hat{p}^o(s)$ . With elasticity of demand ( $\epsilon$ ) less than unity, it follows that  $\hat{N}^u(s)\hat{p}^u(s) > \hat{N}^c(s)\hat{p}^c(s) > \hat{N}^o(s)\hat{p}^o(s)$ , which in turn implies from (43) and (31) that

$$(44) \quad \widehat{MB}^u(s) > \widehat{MB}^c(s) > \widehat{MB}^o(s) \quad \text{for } \epsilon < 1.$$

We assume that for each regime the marginal-cost curve cuts the marginal-benefit curve from below. Then optimal capacity is determined by setting marginal benefit equal to marginal cost, and from (40) and (44)

$$(45) \quad s_*^u > s_*^c > s_*^o \quad \text{for } \epsilon < 1$$

where “\*” subscripts denote “at optimal capacity.” Thus, when  $\epsilon < 1$ , optimal capacity is larger the less sensitive the tolling regime. The less sensitive the tolling regime, the higher is average travel cost given capacity. Furthermore, since price equals twice average travel cost for each toll regime, with  $\epsilon < 1$  total travel cost is higher the coarser is the tolling regime given capacity. Finally, the marginal benefit from capacity expansion, for each  $s$ , is directly proportional to total travel cost. Hence, for each level of capacity, the marginal benefit from capacity expansion is higher, the less sensitive the tolling regime. The same line of argument implies that when  $\epsilon > 1$ , optimal capacity is smaller, the coarser the toll regime.

From (40) and (43) and using  $p_*^j = 2ATC_*^j = 2TC_*^j/N_*^j$ ,

$$N_*^j p_*^j = 2s_*^j \left. \frac{dK(s)}{ds} \right|_{s_*^j}$$

Thus, as long as the elasticity of marginal construction costs with respect to capacity is

greater than  $-1$ , a condition we assume to hold,

$$(46) \quad N_*^u p_*^u > N_*^c p_*^c > N_*^o p_*^o \quad \text{for } \epsilon < 1.$$

This implies that, for  $\epsilon < 1$ ,

$$(47) \quad p_*^u > p_*^c > p_*^o$$

and, with  $\epsilon > 0$ ,

$$N_*^u < N_*^c < N_*^o.$$

The cases of  $\epsilon = 1$  and  $\epsilon > 1$  can be found in Arnott et al. (1987).<sup>9</sup>

### C. Comparison of No-Toll and Toll Regimes

It remains to compare the no-toll equilibrium with optimal capacity to the various toll equilibria with optimal capacity. Unfortunately, the results are not as clean as for the comparison of the toll equilibria with optimal capacity since, with the price of a trip in the no-toll equilibrium below marginal social cost, the determination of optimal capacity is a second-best problem. Equation (43) does not apply in the no-toll equilibrium because the envelope theorem does not hold (the marginal benefit from capacity expansion in the no-toll equilibrium includes the change in the efficiency loss due to underpriced congestion caused by the induced increase in trip frequency). However, some results can be obtained, and these are presented in Arnott et al. (1987).

The empirical evidence strongly supports the hypothesis that rush-hour car travel, which is predominantly commuting, is price-inelastic (e.g., Daniel McFadden, 1974; Small, 1983; John Pucher and Jerome Rothenberg, 1976). Combining this stylized fact with (45), we have the prediction that optimal capacity falls as the “sensitivity of the tolling regime” increases. The model also predicts that, with constant elasticity of

<sup>9</sup>It is interesting to note that, with optimal capacity and constant unit construction cost,  $\widehat{MB}^j(s) = \widehat{TC}^j/s = k \equiv MC(s)$ , which implies  $TC_*^j = ks_*^j$ ; that is, capacity construction cost equals total travel cost. This result is due to the form of congestion.

demand less than 1, optimal capacity in the no-toll regime is higher than under all the tolling regimes (see Arnott et al., 1987).

It is worth restating that there are two separate issues involved in comparing optimal capacity between regimes. The first is whether "second-best capacity" (optimal capacity conditional on congestion being underpriced) is less than or greater than first-best capacity (where congestion is efficiently priced). This is a classic problem that has been well researched in the context of the standard model of traffic congestion (e.g., William C. Wheaton, 1978; John D. Wilson, 1983; Edmond L. d'Ouille and John F. McDonald, 1990). Broadly speaking, our results are consistent with these papers, though not fully comparable because their treatment of congestion is static. The second issue entails the comparison of first-best toll regimes that differ in their sensitivity. Here our analysis is new, and the results are clean.

For future reference, we note that with optimal capacity the efficiency losses associated with the coarse-, uniform-, and no-toll equilibria compared to the optimal fine-toll equilibrium are

$$(48) \quad EL_*^j = \begin{cases} \int_{p_*^o}^{p_*^c} N(p) dp + \frac{N_*^o p_*^o}{2} - K(s_*^o) + K(s_*^e) & j = e \\ \int_{p_*^o}^{p_*^j} N(p) dp + \frac{N_*^o p_*^o}{2} - \frac{N_*^j p_*^j}{2} - K(s_*^o) + K(s_*^j) & j = u, c. \end{cases}$$

The results of the previous two sections are brought together in the following proposition.

**PROPOSITION 2:**

- (i)  $\hat{p}^u(s) > \hat{p}^c(s) > \hat{p}^o(s) = \hat{p}^e(s)$ ;
- (ii)  $\widehat{EL}^e(s) > \widehat{EL}^u(s) > \widehat{EL}^c(s) > 0$ ,  $EL_*^e > EL_*^u > EL_*^c$ ;
- (iii) with price-sensitive trip demand,  $\hat{N}^u(s) < \hat{N}^c(s) < \hat{N}^o(s) = \hat{N}^e(s)$ ;
- (iv) with  $\epsilon < 1$ ,  $s_*^u > s_*^c > s_*^o$ ;

(v) with  $\epsilon < 1$  and  $[d^2K(s)/ds^2]s + dK(s)/ds > 0$ ,  $p_*^u > p_*^c > p_*^o$ , and (if  $\epsilon > 0$ )  $N_*^u < N_*^c < N_*^o$ .

**IV. The Self-Financing of Capacity**

Mohring and Harwitz (1962) and Strotz (1965) investigated the extent to which a network of roads should be self-financing. In the single-period version of their analysis, the private cost ( $C$ ) is a function of capacity ( $s$ ) and the number of trips ( $N$ ). The planner chooses price and capacity so as to maximize social surplus [consumers' surplus plus toll revenue,  $R = (p - C)N$ ] minus capacity costs,  $K(s)$ ; that is,

$$(49) \quad \max_{p,s} \int_p^\infty N(p') dp' + [p - C(N(p), s)]N(p) - K(s).$$

The corresponding first-order conditions for  $p$  and  $s$  are:

$$(50a) \quad -N + \left(1 - \frac{\partial C}{\partial N} \frac{dN}{dp}\right)N + (p - C) \frac{dN}{dp} = 0$$

and

$$(50b) \quad -\frac{\partial C}{\partial s}N - \frac{dK}{ds} = 0$$

respectively. Equation (50a) reduces to

$$(50a') \quad p = C + \frac{\partial C}{\partial N}N.$$

Equation (50a') states that price should be set equal to marginal social cost, the sum of the private cost ( $C$ ) and the marginal congestion externality ( $[\partial C/\partial N]N$ ), and (50b) states that capacity should be constructed up to the point where marginal cost ( $dK/ds$ ) equals marginal benefit ( $-[\partial C/\partial s]N$ ). Assume to simplify that  $C(\cdot)$  is homogeneous of degree  $h^c$  in  $N$  and  $s$  and that  $K(\cdot)$  is homogeneous of degree  $h^K$  in  $s$ . Then, using Euler's Theorem, (50b)

may be rewritten as

$$(51) \quad N \left( \frac{\partial C}{\partial N} N \right) = h^K K + N h^C C.$$

The optimal toll,  $\tau$ , should be set equal to the marginal congestion externality,  $(\partial C / \partial N)N$ . Thus,

$$(52) \quad R \equiv N\tau = h^K K + N h^C C$$

where  $R$  is toll revenue. Equation (52) implies that, if there are constant costs to capacity expansion so that  $h^K = 1$  and if a doubling of capacity and the number of trips leaves the cost of a trip unchanged so that  $h^C = 0$ , then  $R = K$ ; that is, optimal toll revenue exactly covers the cost of constructing optimal capacity. One can extend the analysis to compute the proportion of the cost of capacity construction that is financed from toll revenue when either the toll or the level of capacity is nonoptimal.

The issue of interest here is the form of the self-financing results in the bottleneck model. Obviously, in the no-toll regime, no toll revenue is collected. We therefore restrict our attention to the uniform, coarse, and fine tolls. The analog to (49) is

$$(53) \quad \max_{p^j, s} \int_{p^j}^{\infty} N(p') dp' + [p^j - ATC^j(N(p^j), s)] N(p^j) - K(s) \quad j = u, c, o.$$

It is evident that (53) is simply a particularization of (49) and, hence, that the MHS self-financing results hold for our model. Since  $ATC(\cdot)$  is homogeneous of degree zero in  $N$  and  $s$  for each tolling regime (recall Proposition 1), equation (52) reduces to

$$(54) \quad R^j_* = h^K K(s^j_*) \quad j = u, c, o.$$

Independent of the tolling regime, the ratio of the revenue collected from the optimal toll to the construction cost of optimal capacity equals the elasticity of construction

cost with respect to capacity. This is remarkable since it indicates that *the optimal degree of self-financing of a road is independent of the form of the pricing system employed*; for example, if a road system should be self-financing when a sophisticated tolling system is employed, it should also be self-financing when only a flat parking fee is applied.

The major result of this section is summarized in the following proposition.

**PROPOSITION 3:** *The ratio of the revenue collected from the optimal toll to the costs of constructing optimal capacity equals the elasticity of construction cost with respect to capacity, whatever the tolling regime. With constant costs of capacity, the road should be self-financing; with increasing costs, it should generate a surplus; and with decreasing costs, it should operate at a loss.*

### V. A Numerical Example

We start by considering the situation in which capacity is fixed. Subsequently, we extend the example to treat optimal capacity.

#### A. Capacity Fixed

We normalize so that  $N^e = 1$ , and then set  $s = 0.4$  cars/hr so that the length of the rush hour in the no-toll equilibrium,  $t_q - t_q = N^e / s = 2.5$  hours. Thus,  $p^e = \delta(N^e / s)$  [from (20)] = \$2.5( $\delta$ ). Consistent with column 1 of table 2 in Small (1982), we take  $\alpha = \$5.00/\text{hr}$ ,  $\beta = \$3.05/\text{hr}$ , and  $\gamma = \$11.88/\text{hr}$ .<sup>10</sup> Hence,  $\delta \equiv \beta\gamma / (\beta + \gamma) = \$2.425/\text{hr}$ ,  $p^e = \$6.063$ , and  $\Gamma^c = 0.7292$

<sup>10</sup>Small (1982) estimated the ratios  $\beta/\alpha$  and  $\gamma/\alpha$ . To construct an estimate of  $\alpha$  we draw on Small (1991 pp. 2-54), who concludes that "...a reasonable average value of time for journey to work is 50 percent of the gross wage rate..." Bureau of Labor Statistics (1991) reports average 1990 hourly earnings of \$10.03. Half this is \$5.00/hr in round figures, which we take as our value of  $\alpha$ . Since the Bureau of Labor Statistics figure excludes supervisory and government workers, it understates the average wage for the whole population and thus leads to a conservative value for  $\alpha$  as constructed.

TABLE 1—NUMBER OF COMMUTERS, COST OF TRIP, AND EFFICIENCY LOSS WITH VARIOUS TOLLING REGIMES WHEN ROAD CAPACITY IS FIXED ( $N^c = 1$ ,  $s = 0.4$  vehicles/hr,  $\alpha = \$5.00$ /hr,  $\beta = \$3.05$ /hr,  $\gamma = \$11.88$ /hr)

Variable	$\epsilon$	Tolling regime			
		e	u	c	o
$\hat{N}^j$	0	1	1	1	1
	0.2	1	0.8909	0.9390	1
	1	1	0.7071	0.8280	1
$\hat{p}^j$	0	6.063	12.125	8.842	6.063
	0.2	6.063	10.802	8.304	6.063
	1	6.063	8.574	7.322	6.063
$\widehat{EL}^j$	0	3.031	3.031	1.390	0
	0.2	3.031	2.671	1.302	0
	1	3.031	2.101	1.144	0
$\left(\frac{\widehat{EL}^j}{\widehat{EL}^c}\right)$	0	1	1	0.458	0
	0.2	1	0.881	0.429	0
	1	1	0.693	0.377	0

[from (29c)]. Finally, we assume a constant-elasticity demand function,  $N = np^{-\epsilon}$ . Then  $n = N^c(p^c)^\epsilon = 1$  for  $\epsilon = 0$ , 1.434 for  $\epsilon = 0.2$ , and 6.063 for  $\epsilon = 1$ . Throughout the example, we treat three demand elasticities: (i)  $\epsilon = 0$ , the extreme of completely inelastic trip demand; (ii)  $\epsilon = 1$ , the highest possible reasonable value; and (iii)  $\epsilon = 0.2$ , our best guess on the basis of McFadden (1974), Pucher and Rothenberg (1976), and Small (1983).<sup>11</sup>

From formulas in Arnott et al. (1987), we obtain the results given in Table 1. Recall that the average toll equals one-half the price in the three tolling regimes. Several points are worthy of note. First, observe that the uniform toll tends to be more efficient (i.e.,  $\widehat{EL}^u / \widehat{EL}^c$  is lower) the higher the

elasticity of demand. This is as expected. There are two sources of efficiency loss from imposing no toll: first, because travel is underpriced, too many trips are taken; second, given the number of trips taken, cars do not distribute themselves efficiently over the rush hour. The uniform toll affects only the first source of efficiency loss, which is relatively more important the more elastic is trip demand. Second, with our best-guess demand elasticity of 0.2, more than half the efficiency gain from applying the fine toll can be achieved by applying the coarse toll. Third, the efficiency gain from applying congestion tolls can be substantial: relative to the no-toll situation, \$3.03 per trip can be saved by applying the optimal fine toll. As argued in Arnott et al. (1990), this suggests that the potential gain from applying urban auto tolls may have been considerably underestimated. Fourth, with our best-guess demand elasticity, the efficiency gain from applying the optimal coarse toll is almost five times as high as that from applying the optimal uniform toll. Thus, the efficiency gain from the rescheduling of trips induced by an urban auto toll is likely to be larger than the efficiency gain from the reduction in trips caused by the toll. This in turn strengthens the argument for more serious investigation of technologically advanced tolling schemes.

### B. Capacity Optimal

We assume that there are constant costs to capacity expansion [ $K(s) = ks$ ] and that construction costs are such that the length of the rush hour is 2.5 hours when road width is chosen optimally for the no-toll regime. Using formulas in Arnott et al. (1987), we find  $k = 15.157$  for  $\epsilon = 0$ ,  $k = 12.630$  for  $\epsilon = 0.2$ , and  $k = 7.578$  for  $\epsilon = 1$ , and we obtain the results given in Table 2. They are qualitatively much the same as those presented in Table 1 and are consistent with the results given in Proposition 2. It is worth remarking, however, that with our best-guess demand elasticity, optimal capacity with the fine toll is only about 75 percent of that with no toll.

<sup>11</sup>Our model is sufficiently stylized that it is not clear what elasticity should be employed, though it should be long-run. Note also that since we set fixed travel time to zero,  $\epsilon$  is the elasticity of trips with respect to the variable component of trip price. The empirical literature, on the other hand, measures the elasticity of trips with respect to the total trip price.

TABLE 2—NUMBER OF COMMUTERS, PRICE OF TRIP, AND EFFICIENCY LOSS WITH VARIOUS TOLLING REGIMES WHEN ROAD CAPACITY IS SET OPTIMALLY ( $N_*^c = 1, s_*^c = 0.4$  vehicles/hr,  $\alpha = \$5.00$ /hr,  $\beta = \$3.05$ /hr,  $\gamma = \$11.88$ /hr)

Variable	$\epsilon$	Tolling regime			
		e	u	c	o
$N_*^j$	0	1	1	1	1
	0.2	1	0.8866	0.9150	0.9502
	1	1	0.7071	0.8281	1
$p_*^j$	0	6.063	12.125	10.355	8.574
	0.2	6.063	11.069	9.452	7.827
	1	6.063	8.574	7.321	6.063
$EL_*^j$	0	3.551	3.551	1.780	0
	0.2	3.334	2.970	1.515	0
	1	3.031	2.101	1.144	0
$\left(\frac{EL_*^j}{EL_*^e}\right)$	0	1	1	0.501	0
	0.2	1	0.891	0.454	0
	1	1	0.693	0.377	0
$s_*^j$	0	0.4000	0.4000	0.3416	0.2828
	0.2	0.4000	0.3885	0.3424	0.2944
	1	0.4000	0.4000	0.4000	0.4000

VI. Extensions

A. Heterogeneous Commuters

The extension to treat commuter heterogeneity is conceptually (though not algebraically) straightforward; indeed, the extension has already been done for the optimal fine-toll and no-toll equilibria with inelastic demand by Gordon F. Newell (1987), Cohen (1987), and Arnott et al. (1989), who consider heterogeneity in  $\alpha, \beta, \gamma,$  and  $t^*$ . The essential insight is that in equilibrium user types follow a specific departure order; for example, “assembly-line workers” (those with a high  $\gamma$ ) depart so as to arrive early. Consequently, one cannot simply apply the analysis of the paper to an average user.

With heterogeneous commuters, the results are not as neat. The simple reduced-form diagram does not extend; the relationship between aggregates is no longer as simple; and the characteristics of the various toll equilibria, as well as the generaliz-

ability of the MHS self-financing results, may depend on whether the congestion toll is anonymous.

While user heterogeneity undermines the simplicity of the results presented in the paper, it reinforces the basic message. In order to model user heterogeneity correctly, it is necessary to model different groups’ departure-time decisions and, hence, to adopt a structural approach.

B. Other Congestible Facilities

We cast our analysis in terms of a very specific congestible facility: a bottleneck on a point-input, point-output road in the morning rush hour. An obvious question is to what extent our results generalize to other congestible facilities. The short answer is that there are advantages to modeling all congestible facilities structurally and explicitly treating users’ decisions and the congestion technology, but the details of the analysis will differ from one congestible facility to the next. We have seen, for a specific case, that the standard model is poorly specified. Structural modeling imposes the discipline to ensure proper specification. The standard procedure of dividing the period of use into intervals and solving separately for equilibrium in each interval may, however, provide a good approximation for some congestible facilities.

The nature of users’ decisions will differ according to the congestible facility. For all congestible facilities, time of use is an important margin of user choice. Other margins may be important too: with flow congestion in transportation, the cost functions depend on the vehicle headway drivers choose (Julio J. Rotemberg, 1985); a realistic model of parking would treat the decision concerning length of time parked (Amihai Glazer and Esko Niskanen, 1990); in telephone traffic, the user decides on call duration; in the use of public facilities, such as a swimming pool or zoo, visit length is important; in an art gallery, the extent of congestion depends on the distance viewers stand from pictures; and in wilderness areas, the cost functions depend on hikers’

choice of path and speed of travel (Mordechai Shechter and Robert C. Lucas, 1978).

The nature of the congestion technology also depends on the facility, and for most facilities there are several congestion-prone elements of capacity. For example, in going to a baseball game, a spectator will cruise for a parking spot, walk from the parking spot to the entrance, queue for a ticket, experience crowding while watching the game, and then after the game join a car queue to exit the parking lot. The extent to which congestion is dynamic (i.e., to which cumulative usage effects are important) also varies across facility types.

It remains to be seen how the details of the analysis differ according to the type of congestible facility. An analysis similar to that of this paper has been undertaken in de Palma and Arnott (1990), but for telephone congestion on a single line. In that study, modeling of telephone congestion is the same as that for the road bottleneck, except that (i) the queue discipline is such that customers in the queue are served on a random basis, while on the road they are served on a first-come first-served basis, and (ii) telephone capacity is a stock (the number of users that can be talking on a line *at a point in time*) while road capacity is a flow. The reduced-form formulas obtained were very similar to those in the present study. This demonstrates that the approach taken in the paper can be adapted to other congestible facilities. However, the extension to other congestible facilities for which there are multiple elements of capacity or for which congestion takes a form other than queueing will not be trivial.

We now comment on the extent to which the results recorded in the propositions generalize. Proposition 1 recorded the congestion cost function in our model under various pricing regimes. The results are specific and do not generalize. Proposition 2 compared the various tolling equilibria. The result that the efficiency loss is larger the less sensitive the tolling regime is obviously general. The other major result concerned optimal capacity: with  $\varepsilon < 1$  ( $> 1$ ) optimal capacity was larger (smaller) the less sensitive

the tolling regime. These results generalize as follows. Since the tolls are set optimally conditional on the tolling regime, with identical individuals, each user faces a price equal to marginal social cost, and the envelope theorem holds. Consequently, the marginal benefit from capacity expansion is  $-\partial TC^j / \partial s$ , and as long as marginal cost cuts marginal benefit from below, optimal capacity is larger under regime  $j'$  than under regime  $j$  if  $-\partial TC^{j'} / \partial s > -\partial TC^j / \partial s$  for all  $s$ . With  $\varepsilon < 1$ ,  $-\partial TC^j / \partial s$  should normally be larger the less sensitive the tolling regime, but the precise conditions will depend on the facility. Proposition 3 related the self-financing results. Equation (52) generalizes, provided each user faces a trip price equal to marginal social cost.

Our analysis suggests that previous studies have significantly underestimated the gains from the sophisticated tolling of urban roads. Those studies were based on models that ignored both cross-price effects and the dynamic nature of congestion and, hence, failed to capture the change in the distribution of departure times induced by tolls. Studies of other congestible facilities have generally treated cross-price effects but not the dynamic nature of congestion. Thus, it is unclear whether structural models of those facilities would uncover significant previously unperceived gains from more sophisticated pricing policies.<sup>12</sup>

<sup>12</sup>Compared to roads, the pricing of most other congestible facilities (e.g., telephone, computer, electrical, and water networks) is relatively sophisticated. Time-dependent pricing is the rule for long-distance telephone traffic and is not uncommon for electrical power. Priority pricing (whereby jobs with a higher priority move up the queue more quickly) is the rule on mainframe computers. Usage-dependent pricing is employed in some contexts; the University of Cambridge employs it for computer usage, and Hydro Québec employs it in some of its electrical pricing. Other forms of peak-load pricing include priority-service pricing (Robert Wilson, 1989) and demand-layer pricing (Robert S. Main, 1973). In all contexts, however, more sophisticated pricing is possible. For example, long-distance telephone rates could presumably be made very sensitive to the level of usage: a customer would dial a long-distance number, the efficient tariff based on current system usage would be displayed, and she would then decide whether to proceed with the call.

## VII. Conclusion

In the Introduction we argued that the standard model of peak-period traffic congestion is poorly specified because it fails to model commuters' departure-time decisions and the congestion technology. To our knowledge, the first paper in the literature that addressed these deficiencies was Vickrey (1969). He constructed a model of the morning commute on a road with a single bottleneck in which a fixed number of identical individuals wish to arrive downtown at the same time and choose when to depart, trading off schedule delay against travel time. The queue length adjusts until, in equilibrium, trip price is uniform over the peak period. The solution to the model provides a function relating trip price and marginal social cost to capacity, the number of users over the entire rush hour, and the form of pricing. This yields the correctly specified reduced-form supply and cost functions. The trip demand functions, too, should be specified over the entire rush hour. Thus, with identical commuters, the version of the standard model which treats the period of use as a single interval is sound. However, the procedure employed in much of the literature, of dividing up the period of use into intervals and solving separately for equilibrium in each interval, is logically flawed.

In this paper, we examined some of the economic implications of the Vickrey model and also extended it to treat elastic demand and optimal capacity under a variety of pricing regimes. Two particular findings merit emphasis. First, with identical individuals, the MHS results on the degree of self-financing of congestible facilities with optimal capacity and optimal tolls were shown to apply regardless of the tolling regime. Hence, if a road of optimal capacity should be self-financing with an optimal time-varying toll, it should also be self-financing when only a gasoline tax (set at the optimal level) can be employed. Second, we computed the gains from efficient pricing to be considerably greater than those given in the empirical literature on urban auto congestion. The reason is that previous empirical

estimates are based on a model that ignores the efficiency gain that results from a toll's ability to redistribute travelers over the rush hour. This suggests that more serious consideration should be given to sophisticated tolling schemes on urban roads.

While the analysis focused on rush-hour auto congestion, we discussed the generalization of our approach to other congestible facilities. We argued that there are always advantages to providing a *structural* model of a congestible facility, which explicitly treats users' decisions and the facility's congestion technology.

The appropriate directions for future research are evident. Vickrey's (1969) paper, supplemented by this one, provides the *method* for determining structural models of facilities subject to peak-load congestion. However, the modeling of the congestion technology is primitive. What we need now are more realistic models of congestion technologies and of consumers' behavioral decisions, along with empirical estimation of them. In the context of rush-hour traffic congestion, for example, models should be developed which derive hypercongestion (traffic-jam situations) from driving behavior, solve for equilibrium on a congested network, and account for heterogeneity among users in addition to accident and road damage costs. As well, it is clearly desirable to apply Vickrey's (1969) approach to other types of congestible facilities, notably computers, public utilities, airports, telecommunications, and recreational facilities.

## REFERENCES

- Aigner, Dennis, J. and Hirschberg, Joseph G., "Commercial/Industrial Customer Response to Time-of-Use Electricity Prices: Some Experimental Results," *Rand Journal of Economics*, Autumn 1985, 16, 341-55.
- Arnott, Richard, de Palma, André and Lindsey, Robin, "Bottleneck Congestion with Elastic Demand," Discussion Paper 690, Institute for Economic Research, Queen's University, 1987.

- \_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_, "Schedule Delay and Departure Time Decisions with Heterogeneous Commuters," *Transportation Research Record*, 1989, 1197, 56-67.
- \_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_, "Economics of a Bottleneck," *Journal of Urban Economics*, January 1990, 27, 111-30.
- Arnott, Richard and Kraus, Marvin**, "The Ramsey Problem for Congestible Facilities," National Bureau of Economic Research (Cambridge, MA), Technical Working Paper No. 84, 1990.
- Ben-Akiva, Moshe, de Palma, André and Kanaroglou, Pavlos**, "Dynamic Model of Peak Period Traffic Congestion with Elastic Arrival Rates," *Transportation Science*, May 1986, 20, 164-81.
- Boiteux, Marcel**, "La Tarification des Demandes en Pointe: Application de la Théorie de la Vente au Coût Marginal," *Revue Générale de l'Électricité*, August 1949, 58, 321-40; reprinted in English translation, "Peak-Load Pricing," *Journal of Business*, April 1960, 33, 157-79.
- Borins, Sanford F.**, "The Political Economy of Road Pricing: The Case of Hong Kong," in *Proceedings of the World Conference on Transport Research, Vancouver, BC*, Vol. 2, Vancouver: Center for Transportation Studies, University of British Columbia, 1986, pp. 1367-78.
- Braeutigam, Ronald R.**, "Optimal Policies for Natural Monopolies," in Richard L. Schmalensee and Robert D. Willig, eds., *Handbook of Industrial Organization*, Amsterdam: North-Holland, 1989, pp. 1289-1346.
- Braid, Ralph M.**, "Uniform versus Peak-Load Pricing of a Bottleneck with Elastic Demand," *Journal of Urban Economics*, November 1989, 26, 320-7.
- Cohen, Yuval**, "Commuter Welfare under Peak-Period Congestion Tolls: Who Gains and Who Loses?" *International Journal of Transport Economics*, October 1987, 14, 238-66.
- Crew, Michael A. and Kleindorfer, Paul R.**, *The Economics of Public Utility Regulation*, Cambridge, MA: MIT Press, 1986.
- de Palma, André and Arnott, Richard**, "Usage-Dependent Peak-Load Pricing," *Economics Letters*, 1986, 20 (2), 101-5.
- \_\_\_\_\_, \_\_\_\_\_ and \_\_\_\_\_, "The Temporal Use of a Telephone Line," *Information Economics and Policy*, 1990, 4 (2), 155-74.
- d'Ouille, Edmond L. and McDonald, John F.**, "Optimal Road Capacity with a Suboptimal Congestion Toll," *Journal of Urban Economics*, July 1990, 28, 34-49.
- Downs, Anthony**, "The Law of Peak-Hour Expressway Congestion," *Traffic Quarterly*, July 1962, 16, 393-409.
- Dupuit, Jules**, "De l'Influence des Péages sur l'Utilité des Voies de Communication," *Annales des Ponts et Chaussées: Mémoires* 1, 1849, 817, 170-248; translated by Elizabeth Henderson, "On Tolls and Transport Charges," *International Economic Papers*, 1962, 11, 7-31.
- Glazer, Amihai and Niskanen, Esko**, "Parking Fees and Congestion," *Regional Science and Urban Economics*, March 1992, 22, 123-32.
- Hendrickson, Chris and Kocur, George**, "Schedule Delay and Departure Time Decisions in a Deterministic Model," *Transportation Science*, February 1981, 15, 62-77.
- Keeler, Theodore E. and Small, Kenneth A.**, "Optimal Peak-Load Pricing, Investment, and Service Level on Urban Expressways," *Journal of Political Economy*, February 1977, 85, 1-25.
- Knight, Frank**, "Some Fallacies in the Interpretation of Social Costs," *Quarterly Journal of Economics*, August 1924, 38, 582-606.
- Kraus, Marvin**, "The Welfare Gains from Pricing Road Congestion Using Automatic Vehicle Identification and On-Vehicle Meters," *Journal of Urban Economics*, May 1989, 25, 261-81.
- \_\_\_\_\_, **Mohring, Herbert and Pinfeld, Thomas**, "The Welfare Costs of Nonoptimum Pricing and Investment Policies for Freeway Transportation," *American Economic Review*, September 1976, 66, 532-47.
- Main, Robert S.**, "Periodic vs. Demand-Layer Pricing for Utility Loads," Ph.D. dissertation, University of California at Los Angeles, 1973.
- McFadden, Daniel**, "The Measurement of Urban Travel Demand," *Journal of Public Economics*, November 1974, 3, 303-28.
- Mohring, Herbert and Harwitz, Mitchell**, *High-*

- way *Benefits*, Evanston, IL: Northwestern University Press, 1962.
- Newell, Gordon F.**, "The Morning Commute for Non-identical Travellers," *Transportation Science*, May 1987, 21, 74-88.
- Pigou, Arthur C.**, *The Economics of Welfare*, London: Macmillan, 1920.
- Pucher, John and Rothenberg, Jerome**, "Pricing in Urban Transportation: A Survey of Empirical Evidence on the Elasticity of Travel Demand," mimeo, Department of Economics, Massachusetts Institute of Technology, 1976.
- Rotemberg, Julio J.**, "The Efficiency of Equilibrium Traffic Flows," *Journal of Public Economics*, March 1985, 26, 191-206.
- Schechter, Mordechai and Lucas, Robert C.**, *Simulation of Recreational Use for Park and Wilderness Management*, Baltimore: Johns Hopkins University Press, 1978.
- Small, Kenneth A.**, "The Scheduling of Consumer Activities: Work Trips," *American Economic Review*, June 1982, 72, 467-79.
- \_\_\_\_\_, "The Incidence of Congestion Tolls on Urban Highways," *Journal of Urban Economics*, January 1983, 13, 90-111.
- \_\_\_\_\_, *Urban Transportation Economics*, Chur, Switzerland: Harwood, 1991.
- Steiner, Peter O.**, "Peak Loads and Efficient Pricing," *Quarterly Journal of Economics*, November 1957, 71, 585-610.
- Strotz, Robert H.**, "Urban Transportation Parables," in Julius Margolis, ed., *The Public Economy of Urban Communities*, Washington, DC: Resources for the Future, 1965, pp. 127-69.
- Vickrey, William**, "Pricing in Urban and Suburban Transport," *American Economic Review*, May 1963 (*Papers and Proceedings*), 53, 452-65.
- \_\_\_\_\_, "Congestion Theory and Transport Investment," *American Economic Review*, May 1969 (*Papers and Proceedings*), 59, 251-61.
- \_\_\_\_\_, "Responsive Pricing of Public Utility Services," *Bell Journal of Economics*, Spring 1971, 2, 337-46.
- Walters, Alan A.**, "The Theory and Measurement of Private and Social Cost of Highway Congestion," *Econometrica*, October 1961, 29, 676-99.
- Wheaton, William C.**, "Price-Induced Distortions in Urban Highway Investment," *Bell Journal of Economics*, Autumn 1978, 9, 622-32.
- Williamson, Oliver E.**, "Peak Load Pricing and Optimal Capacity under Indivisibility Constraints," *American Economic Review*, September 1966, 56, 810-27.
- Wilson, John D.**, "Optimal Road Capacity in the Presence of Unpriced Congestion," *Journal of Urban Economics*, May 1983, 13, 337-57.
- Wilson, Robert**, "Efficient and Competitive Rationing," *Econometrica*, January 1989, 57, 1-40.
- Bureau of Labor Statistics**, *Employment and Earnings*, U.S. Department of Labor, Washington, DC: U.S. Government Printing Office, 1991.