JOURNAL OF URBAN ECONOMICS 27, 111–130 (1990)

Economics of a Bottleneck*

RICHARD ARNOTT,[†, ||] ANDRÉ DE PALMA,[‡] AND ROBIN LINDSEY[§]

[†]*Queen's University, Kingston, Ontario, Canada K7L 3N6;* [‡]*Northwestern University, Evanston, Illinois 60201; and* [§]*University of Alberta, Edmonton, Alberta, Canada T6G 2E1*

Received October 31, 1986; revised September 9, 1987

Drawing on Vickrey (*Amer. Econ. Rev.* **59**, 251–261 (1969)) and subsequent literature, this paper provides a thorough economic analysis of the simplest bottleneck model of road congestion with peak-load demand: In the morning rush hour, a fixed number of identical individuals, one per car, must travel from home to work, between which is a bottleneck of given capacity. The costs of travel include queuing time and schedule delay (time early or late for work). The frequency distribution of departure times adjusts so that in equilibrium all individuals have the same travel costs. The paper extends the previous literature by examining a coarse toll and solving for optimal capacities. The application of congestion tolls can generate efficiency gains by altering the frequency distribution of departure times. Previous estimates of the efficiency gains from congestion tolling are likely to be substantially downward-biased because they have ignored this effect.

Most urban economists are familiar with rush-hour traffic models which treat flow congestion (Strotz [35], Vickrey [37, 38], Mohring [29] *inter alia*, and reviewed in Arnott [3]). The simplest such model assumes that a fixed number, N, of identical individuals, one per car, must travel between home and work in the morning rush hour on a single road. An individual's travel time, T, is given by the congestion function $T = T(N, w)$ (where w is road width and $\partial T/\partial N > 0$, $\partial T/\partial w < 0$) which may be interpreted as implicitly assuming that the length of the rush hour is exogenous and that traffic flow is uniform over the rush hour. We term this the 'naive' flow congestion model to contrast it with the more sophisticated models of flow congestion employed by traffic engineers, which explicitly treat the physics of traffic flow. The model has been extended to allow for elastic trip demand (e.g., Wilson [40]) and multiple modes (e.g., Mohring [30]) and has been incorporated into the monocentric city model (e.g., Arnott [2], Kraus [24]). While

*We thank SSHRCC for financial support; seminar participants at Tel-Aviv, Concordia, Guelph, Ben-Gurion, U.B.C., and Western Ontario, as well as Ralph Braid and Dan Klein, for useful comments; the referee for a very helpful report; and Marvin Kraus for providing us with a list of references on the efficiency gains from congestion tolls. An earlier version of the paper was presented at the World Conference on Transportation Research, Vancouver, British Columbia, 1986.

[||] Present address: Boston College, Chestnut Hill, MA, 02167.

0094-1190/90 $3.00

widely used, this treatment of traffic congestion suffers from a severe conceptual weakness: *It fails to treat the individual's departure time decision —when to leave home*. In making this decision, the individual must face a tradeoff. The simplest way to model this tradeoff is in terms of travel time *vs* schedule delay (early or late arrival at work); an individual can choose to depart in the tails of the rush hour when travel time is low and pay the penalty of arriving at work early or late, or at the peak when travel time is high but schedule delay costs are low.

In the past few years, a new model of traffic congestion has been developed which makes the departure time decision endogenous. To the best of our knowledge, the original contribution is Vickrey [37, 38]. He considered a situation where a fixed number of identical commuters have to travel from a single entrance (home) to a single exit (work) along a single road during the morning rush hour. There is a single bottleneck on the road with a fixed capacity or service rate, and if the arrival rate at the bottleneck exceeds this capacity a queue develops. The distribution of arrival times is such that it is physically impossible for all commuters to arrive at work exactly on time and to experience no queue; consequently, they face the tradeoff described in the previous paragraph. Each commuter chooses his departure time to minimize his trip costs, which are linear in queuing or travel time and schedule delay, and include the toll charge where applicable. Equilibrium obtains when no individual has an incentive to alter his departure time. In addition to characterizing the no-toll equilibrium, Vickrey [38] determined the social optimum and solved for the toll which decentralizes it. He also provided an illuminating discussion of the benefits of capacity expansion on a single route and for routes in parallel, with and without the toll.

Vickrey's model was independently formulated by Hendrickson and Kocur [18] and Fargier [15]. Fargier considered both the morning and afternoon (characterized by a desired *departure* time) rush hours. Subsequent work has extended these papers. For example, Smith [34] proved the existence of equilibrium under general assumptions on the function relating trip costs to travel time and schedule delay, Daganzo [9] provided a proof of uniqueness of equilibrium, and Hendrickson and Kocur [18] provided an empirical application. Our paper continues this line of development, investigating the characteristics of equilibrium with an optimal step function or coarse toll and solving for optimal capacity under various toll régimes.

While an exhaustive literature review would be out of place here, we should mention other lines of development in which the departure rate function has been endogenous. Henderson [16, 17] considered a model of rush-hour congestion similar to Vickrey's except that a form of flow congestion, rather than queuing congestion, was treated; specifically, he assumed that the travel time of a commuter is an increasing function of the

departure rate *at his departure time* and is therefore independent of downstream traffic conditions, a strong assumption.[1] Hurdle [22], meanwhile, provided essentially the same treatment of congestion as Vickrey, but assumed that the departure rate at time t is a function only of t, and of travel time with departure at t, and is therefore independent of travel time for other departure times—a zero cross-price elasticity assumption. Alpha and Minh [1] and de Palma *et al.* [11] considered probabilistic demand resulting in a stochastic equilibrium. A dynamic extension of this approach to a day-to-day adjustment is considered in Ben-Akiva *et al.* [6].

Since rush-hour traffic flow models with endogenous departure times have, with the exception of Vickrey and Henderson, been developed by engineers, little attention has been paid to their economic implications. The principal aim of this paper is to examine some of these implications. Perhaps the most interesting finding is that the efficiency gains from applying congestion tolls in our model may be substantially larger than the gains computed using the naive flow congestion model. The reason is that a major benefit from congestion tolls derives from the change they induce in the departure rate function, an effect which is excluded by assumption in the conventional naive flow congestion analysis. If the benefits from applying congestion tolls are as large as an example in the paper suggests, more serious consideration should be given to the implementation of urban traffic tolling schemes of the sort advocated by Vickrey [37].

While this paper examines urban automobile traffic congestion, it should be possible to adapt and extend it to treat other congestible facilities with peak-load demand such as telecommunications and air travel networks, electricity grids, computers, and public facilities such as swimming pools.

Section I provides a formal description of the model. Section II characterizes the no-toll equilibrium. Section III obtains the social optimum and describes the toll which decentralizes it. Section IV examines equilibrium with a coarse toll. Section V contains an example. Section VI treats optimal capacity under various toll régimes. Extensions are discussed in Section VII.

I. THE MODEL

Every morning during the rush hour, a fixed number, N, of individuals travel between home (A) and work (B). Each individual travels by his own car along the single road joining (A) and (B). Travel is uncongested except at a single bottleneck through which at most s cars can pass per unit of

[1] Henderson has to impose the restriction that those departing later must arrive later. His treatment of congestion, while physically unrealistic, has some nice properties. For example, in contrast to Vickrey's queuing model, he obtains the result that the no-toll equilibrium rush hour is shorter than optimal, which accords with intuition.

time; if the arrival rate at the bottleneck exceeds s, a queue develops. We term s either the capacity or the service rate of the bottleneck.

Travel time from A to B is

$$T(t) = T^f + T^v(t). \tag{1}$$

where T^f is the fixed component of travel time from A to B. $T^v(t)$ is the variable component, waiting time at the bottleneck, and t is *departure time from home*. Without restriction, we set $T^f = 0$; thus, an individual arrives at the bottleneck as soon as he leaves home, and arrives at work immediately upon leaving the bottleneck. Let $D(t)$ be the length of the queue. Then

$$T^v(t) = \frac{D(t)}{s}; \tag{2}$$

an individual's queuing time equals queue length at the time he joins the queue divided by the service rate of the bottleneck. Where $\hat{t}$ is the most recent time at which there was no queue, and $r(t)$ is the departure rate function,

$$D(t) = \int_{\hat{t}}^{t} r(u)\, du - s(t - \hat{t}), \tag{3a}$$

which implies that

$$\dot{D}(t) = r(t) - s \qquad \text{for } D(t) > 0, \tag{3b}$$

where $\cdot$ denotes a time derivative.

Besides travel time costs, individuals incur schedule delay costs—the costs of arriving at work early or late. To simplify the analysis,[2] we assume that all individuals want to arrive at work at t^*. Let $\tilde{t}$ be the departure time for which an individual arrives at work on time. Then

$$\tilde{t} = t^* - T^v(\tilde{t}). \tag{4}$$

If an individual departs at $t < \tilde{t}$, he is early by an amount $t^* - t - T^v(t)$, while if he departs at $t > \tilde{t}$, he is late by an amount $t + T^v(t) - t^*$.

[2] In an earlier version of the paper, Arnott, de Palma, and Lindsey [4], we allowed for an "on-time" window, a period of on-time arrival time. We eliminated this in the current draft to simplify the algebra.

As in Vickrey [38], it is assumed that the cost of a trip, C, is linear in travel time and schedule delay;[3] specifically

$$\begin{aligned} C &= \text{travel time costs} + \text{time early costs} + \text{time late costs} + \text{toll} \\ &= \alpha(\text{travel time}) + \beta(\text{time early}) + \gamma(\text{time late}) + \text{toll}, \end{aligned} \tag{5}$$

where α is the shadow value of travel time, and β and γ are the shadow values of time early and late.

II. NO-TOLL EQUILIBRIUM

Equilibrium obtains when no individual has an incentive to change his departure time. Since individuals are identical, this implies that trip cost must be the same at all times at which departures occur.

With $\alpha > \beta$,[4] all vehicles except the first and last experience congestion, and the departure rate is piecewise constant and given by

$$r(t) = \begin{cases} s + \dfrac{\beta s}{\alpha - \beta} & \text{for } t \in [t_q, \tilde{t}), \\ s - \dfrac{\gamma s}{\alpha + \gamma} & \text{for } t \in (\tilde{t}, t_{q'}], \end{cases} \tag{6}$$

where t_q and $t_{q'}$ are the times at which the rush hour queue begins and ends, respectively. The intuition underlying this result is that the departure rate function must be such that the marginal benefit from postponing departure by a unit of time equals the marginal cost. In the case of departure prior to $\tilde{t}$, the marginal benefit from postponing departure is the reduction in time early costs, $\beta(1 + \dot{D}/s)$, and the marginal cost the increase in travel time costs, $\alpha\dot{D}/s$. Application of (3b) then gives $r(t)$ for $t \in [t_q, \tilde{t})$. The reasoning for departure after $\tilde{t}$ is analogous. The arrival rate at work, meanwhile, is constant at s over the rush hour. Thus, a queue builds up linearly from t_q to $\tilde{t}$ and then dissipates linearly until it disappears at $t_{q'}$.

[3] This function was introduced by Vickrey [38], was employed in Hendrickson and Kocur [18], and has been estimated by Small [32]. Assuming it to be linear considerably simplifies the algebra and improves the clarity of the paper.

[4] This accords with experimental data (e.g., McFadden, Talvitie, *et al.* [28] and Small [32]). The case $\alpha < \beta$ is discussed briefly in Appendix 1 of Arnott, de Palma, and Lindsey [4].

It remains to compute t_q, $t_{q'}$, and $\tilde{t}$. For this purpose, we write the following three equations:

$$(\tilde{t} - t_q)\left(s + \frac{\beta s}{\alpha - \beta}\right) + (t_{q'} - \tilde{t})\left(s - \frac{\gamma s}{\alpha + \gamma}\right) = N \tag{7a}$$

$$(\tilde{t} - t_q)\frac{\beta s}{\alpha - \beta} = (t_{q'} - \tilde{t})\frac{\gamma s}{\alpha + \gamma} \tag{7b}$$

$$\tilde{t} + \frac{\beta}{\alpha - \beta}(\tilde{t} - t_q) = t^*. \tag{7c}$$

The first equation states that the total number of vehicles which depart is N, the second specifies that congestion disappears at $t_{q'}$, and the last follows from the definition of $\tilde{t}$. Solving these equations yields

$$t_q = t^* - \left(\frac{\gamma}{\beta + \gamma}\right)\left(\frac{N}{s}\right) \tag{8a}$$

$$t_{q'} = t^* + \left(\frac{\beta}{\beta + \gamma}\right)\left(\frac{N}{s}\right) \tag{8b}$$

$$\tilde{t} = t^* - \left(\frac{\beta\gamma}{\alpha(\beta + \gamma)}\right)\left(\frac{N}{s}\right). \tag{8c}$$

Naturally, $t_q < \tilde{t} < t_{q'}$. Note furthermore that N and s enter the equations only as N/s. The comparative static properties of (8) are intuitive and left to the reader.

Cumulative departures *from home* and arrivals *at work* are plotted as a function of time in Fig. 1. The slopes of the two curves give the departure and arrival rates; the vertical distance between the two curves is queue length, and the horizontal distance indicates travel time.

The trip cost of the individual departing at t_q, who experiences only schedule delay, is

$$C(t_q) = \beta(t^* - t_q). \tag{9}$$

Substituting (8a) into (9) gives

$$C(t_q) = \left(\frac{\beta\gamma}{\beta + \gamma}\right)\left(\frac{N}{s}\right). \tag{10}$$

Since everyone has the same trip cost, the total travel cost (TC) for all commuters is

$$\mathrm{TC}^e = \left(\frac{\beta\gamma}{\beta + \gamma}\right)\left(\frac{N^2}{s}\right), \tag{11}$$

where superscript e denotes the no-toll equilibrium.

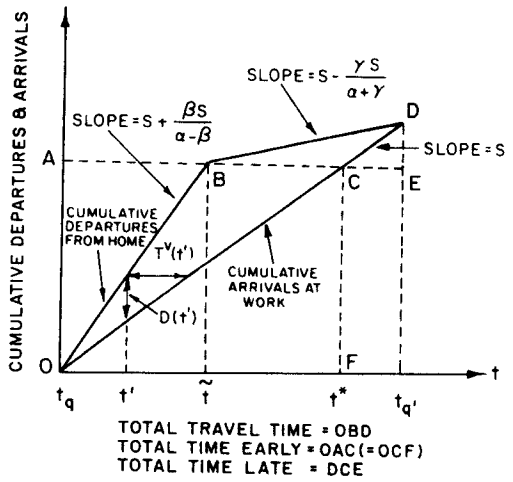

FIG. 1. Queue length, travel time, cumulative departures and arrivals, total travel time, total time early, and total time late: the no-toll equilibrium.

Total travel time may be computed as $\int_{t_q}^{t_{q'}} D(t)\,dt$—the area between the cumulative departures and arrivals functions in Fig. 1. Total time early is $\int_{t_q}^{t^*} s(t^* - t)\,dt$—the area under the cumulative arrivals schedule from t_q to t^*; and total time late, $\int_{t^*}^{t_{q'}} s(t - t^*)\,dt$. From these areas, using (8), one can compute total travel time costs (TTC) and total schedule delay costs (SDC),[5]

$$\mathrm{TTC}^e = \mathrm{SDC}^e = \frac{\beta\gamma}{2(\beta + \gamma)}\left(\frac{N^2}{s}\right). \tag{12}$$

It is interesting to observe that TC^e, SDC^e, and TTC^e are all independent of α. An explanation of this result is as follows: The length of the rush hour is independent of α. Since the first and last individuals to depart experience only schedule delay and no travel time, and since their trip costs are the same in equilibrium, their schedule delay costs must be the same. It follows that the start and end of the rush hour are independent of α (see (8a) and (8b)). Since the trip cost of the first person to depart is independent of α and equals everyone else's trip cost, total travel costs are independent of α. With knowledge of the start and end of the rush hour, total schedule delay costs can be calculated, independent of α. And since total travel time costs equal total travel costs minus schedule delay costs,

[5]The equality between SDC^e and TTC^e is essentially a geometric result, and stems from the linearity of trip costs in schedule delay and travel time.

they too are independent of α. As the shadow value of travel time increases, total travel time decreases by the same proportional amount, so that total travel time costs remain constant.

III. COMPARISON OF THE SOCIAL OPTIMUM AND NO-TOLL EQUILIBRIUM

We provide a heuristic derivation of the social optimum, in which total travel costs are minimized. First, waiting time is pure deadweight loss, so that at the social optimum there is no queue and hence total travel time costs are zero. Second, travel will occur over a continuous time interval; otherwise, total schedule delay costs are unnecessarily large. These two results together imply that the departure and arrival rates are s throughout the rush hour. Third, the schedule delay cost of the first and last travelers must be the same; otherwise, total schedule delay costs could be reduced by transferring individuals from one end of the rush hour to the other. Since the bottleneck is fully utilized over the rush hour in both the no-toll equilibrium and the social optimum, and since the number of commuters is the same in both cases, the length of the rush hour is the same in the no-toll equilibrium and the social optimum.[6] Furthermore, in both the no-toll equilibrium and the social optimum, the schedule delay costs must be the same for the first and last travelers. In the former case, this follows from the equal trip cost condition for equilibrium while in the latter this is an efficiency condition. These results together imply that *the start and end of the rush hour and the cumulative arrivals function are the same in the no-toll equilibrium and the social optimum.*

The social optimum can be decentralized by employing the following time-dependent toll, which we sometimes refer to as the optimal fine toll:[7]

$$\tau^{o}(t) = \begin{cases} 0 & t < t_q \\ a - (t^* - t)\beta & t \in [t_q, t^*) \\ a - (t - t^*)\gamma & t \in [t^*, t_{q'}] \\ 0 & t > t_{q'}. \end{cases} \tag{13}$$

Here $a \leq (\beta\gamma/(\beta + \gamma))(N/s)$ and superscript o denotes the social optimum. From the argument of the previous paragraph, it follows that, relative to the no-toll equilibrium, *application of the optimal fine toll does not change*

[6] This result obtains because of the form of congestion assumed; with flow congestion and an endogenous departure rate function, the rush hour is longer in the social optimum than in the no-toll equilibrium (Henderson [17]).

[7] Throughout the paper, we express the toll as a function of *arrival* time. In the context of auto congestion, this is natural since toll gates are at the front of queues.

total schedule delay costs but eliminates travel time, resulting in a social saving of TTC^e (see (12)).

The time-dependent toll discussed above could be practically difficult to administer. Not only does it require knowledge of the parameters β, γ, N/s, and t^*, but also toll collection costs could be quite high. De Palma and Arnott [10] have proposed a usage-dependent toll, whereby individuals pay a toll proportional to the length of the queue they join (or, equivalently in the context of the model, proportional to travel time). If the toll is set arbitrarily high, no individual will join a queue. All the deadweight loss associated with the queue is thereby eliminated and the social optimum attained.

IV. THE OPTIMAL COARSE TOLL

One does not observe tolls as fine as the one described above. Instead, most tolls are uniform over the day or are step functions. Since travel demand is inelastic in the model, a uniform toll has no effect. In this section, the simplest case of the latter class of tolls is examined, a toll with a single step over the rush hour.

A coarse toll is defined to be a fee ρ^c paid at the front of the queue over a time interval $[t^+, t^-]$, where superscript c denotes the optimal coarse toll equilibrium and the superscripts + and − indicate the turning on and off respectively of the toll. The object is to find the optimal fee and time interval. A rigorous derivation of the coarse toll is tedious. Here we simply present the results; calculations and proofs are available from the authors on request.

The qualitative properties of equilibrium depend on whether $\gamma \gtreqless \alpha$. Empirical estimates (Small [32]) suggest that the shadow value of a minute late is significantly larger than the shadow value of travel time, and hence that $\gamma > \alpha$. We treat only this case here.

The qualitative effects of applying the optimal coarse toll are shown in Fig. 2. First, the toll is applied at $t^+ \in (t_q, t^*)$ and lifted at $t^- \in (t^*, t_{q'})$. Second, t^+, t^-, and ρ^c are chosen so that the length of the queue is zero at the moment the toll is applied and also immediately before it is lifted, but at no other times in the interior of the rush hour. Third, whether or not the coarse toll is optimal, there are no departures for a period ρ^c/α before the toll is applied, while a mass of individuals of size $2s\rho^c/(\alpha + \gamma)$ departs immediately after the toll is lifted. The explanation of the former result is as follows: Since the bottleneck is full throughout the rush hour, the first person to arrive (at work) after the toll is applied arrives just after the last person to arrive before the toll is applied. Both therefore have approximately the same schedule delay. For both to have the same trip cost, the latter must incur travel time costs that are higher by the amount of the toll and consequently must depart ρ^c/α earlier. The explanation for the mass of

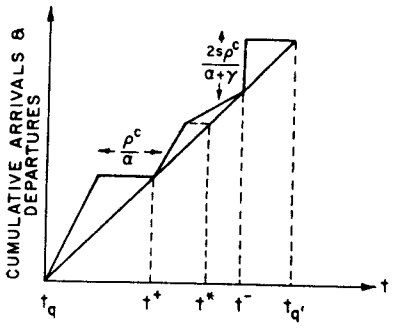

FIG. 2. Traffic flow over the rush hour with an optimal coarse toll and $\gamma > \alpha$.

departures immediately after the toll is lifted is similar. The last person to arrive before the toll is lifted must have the same trip cost as the "first" person to arrive after the toll is lifted. The latter must therefore incur travel time plus schedule delay costs that are ρ^c higher than the former. This is impossible unless there is a mass of departures just after the toll is lifted. We assume that position in the mass is random. On this assumption, it is easily demonstrated that the size of the mass is $2s\rho^c/(\alpha + \gamma)$. A final characteristic of equilibrium is that no one departs after the mass.

Solving for the optimal toll gives

$$\rho^c = \frac{\beta\gamma}{2(\beta + \gamma)}\left(\frac{N}{s}\right) \tag{14a}$$

$$t_q^c = t^* - \frac{\gamma}{\beta + \gamma}\left(\frac{N}{s}\right) + \frac{(\gamma - \alpha)\rho^c}{(\beta + \gamma)(\alpha + \gamma)} \tag{14b}$$

$$t^+ = t_q^c + \frac{\rho^c}{\beta} \tag{14c}$$

$$t^- = t_q^c + \frac{N}{s} - \frac{2\rho^c}{\alpha + \gamma}. \tag{14d}$$

To determine optimal capacity in Section VI, it will be necessary to know the values of various aggregates. These can be derived as follows: First, schedule delay costs can be computed (per Fig. 2) as

$$\mathrm{SDC}^c = \frac{\beta s}{2}\left(t^* - t_q^c\right)^2 + \frac{\gamma s}{2}\left(t_q^c + \frac{N}{s} - t^*\right)^2, \tag{15}$$

TABLE 1
Some Aggregates

Case	SDC	TTC	TC
No toll	ϕ	ϕ	2ϕ
Social optimum	ϕ	0	ϕ
Optimal coarse toll $\gamma > \alpha$	$\chi\phi$	$(\psi - \chi)\phi$	$\psi\phi$

Note.

$$\phi \equiv \frac{\beta\gamma}{2(\beta+\gamma)}\left(\frac{N^2}{s}\right), \qquad \chi \equiv 1 + \frac{\gamma\beta(\gamma-\alpha)^2}{4(\beta+\gamma)^2(\alpha+\gamma)^2},$$

$$\psi \equiv \frac{3}{2} - \frac{(\gamma-\alpha)\beta}{2(\beta+\gamma)(\alpha+\gamma)}.$$

where t_q^c is given by (14b). Second, total toll revenues can be computed as

$$R^c = \rho^c s(t^- - t^+), \tag{16}$$

from (14). Third, total travel costs may be computed using the relation that

$$-\frac{R^c}{N} + \beta(t^* - t_q^c) = \frac{TC^c}{N}, \tag{17}$$

which states that, when toll revenues are equally redistributed, the trip cost of the first person to depart (LHS of (17)) equals average trip cost (RHS of (17)), which equals total travel costs divided by the population, since toll revenues net out. Finally, TTC = TC − SDC. The results of the computations are given in Table 1. Of course, $TC^e > TC^c > TC^o$. Also, $SDC^c > SDC^e = SDC^o$. This pair of inequalities, along with $TTC^o = 0$, imply $TTC^e > TTC^c > TTC^o$. All these inequalities accord with intuition.[8] On the basis of the results in Table 1, one can compute the ratio of total travel

[8]The inequality $SDC^c > SDC^e$ is not so obvious. Total schedule delay costs are minimized when the schedule delay cost of the first and last individuals to arrive at work are equal. In the no-toll equilibrium, this condition is satisfied since the first and last individuals to arrive incur no travel time, and to have equal trip costs must have the same schedule delay costs.

With the optimal coarse toll and $\gamma > \alpha$, no one departs at $t_{q'}^c$, implying that $C(t_q^c) < C(t_{q'}^c)$. Since $C(t_q^c)$ equals the schedule delay cost of the first person to arrive and $C(t_{q'}^c)$ the schedule delay cost of the last person to arrive, the schedule delay cost of the first person to arrive is less than that of the last. Hence, $SDC^c > SDC^e$. This argument also explains why application of the optimal coarse toll postpones both the beginning and the end of the rush hour (see Fig. 3).

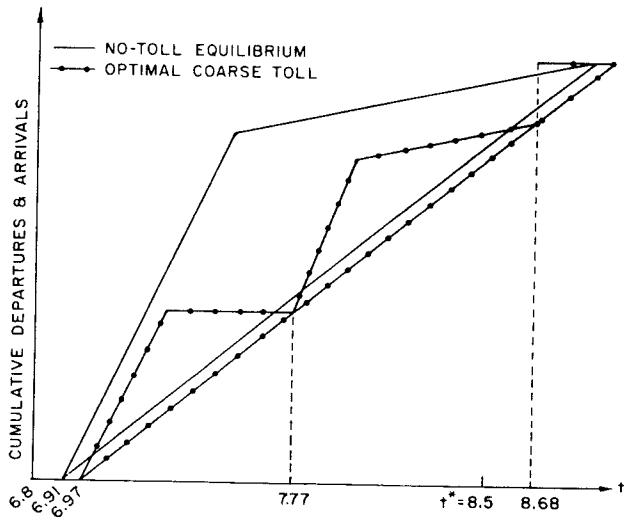

FIG. 3. Cumulative departures and arrivals. Example: $N/s = 2$, $\alpha = 6.4$, $\beta = 3.9$, $\gamma = 15.21$.

time costs to aggregate schedule delay costs, the social savings from applications of the toll, the proportionate savings, and so on.

V. AN EXAMPLE

To get some idea of magnitudes, we consider an example. The rush hour lasts for 2 hours and is centered around 8.5 AM (i.e., $t^* = 8.5$ with time indicated in decimal form). The shadow values of travel time, early arrival time, and late arrival time (α, β, and γ) are \$6.40/hour, \$3.90/hour, and \$15.21/hour, respectively; these shadow values are chosen on the basis of Small [32]. Since $\gamma > \alpha$, the case treated in the text for the coarse toll applies; thus, from (14), $\rho^c = 3.10$, $t_q = 6.97$, $t^+ = 7.77$, and $t^- = 8.68$. The cumulative departures and arrivals for the no-toll equilibrium and for the optimal coarse toll are shown in Fig. 3. Application of the coarse toll causes the rush hour to start slightly later, which is explained in footnote 8.

The results for the aggregates in per capita terms are given in Table 2. Imposition of the optimal coarse toll from 7.77 to 8.68 AM causes travel time costs to fall by an average of \$1.70, while average schedule delay costs rise by \$.02. Thus, imposition of the toll causes travel costs to fall by \$1.68 per person. Application of the optimal fine toll, meanwhile, causes travel costs to fall by \$3.10 per person. Thus, the optimal coarse toll in this example is 54.2% efficient, where the efficiency of a coarse toll is defined

TABLE 2
Example: Per-capita Aggregates, Capacity Fixed

Case	SDC/N	TTC/N	TC/N
No-toll equilibrium	\$3.10	\$3.10	\$6.21
Social optimum	\$3.10	0	\$3.10
Optimum coarse toll	\$3.12	\$1.40	\$4.53

naturally as

$$\text{eff}^c = \frac{TC^e - TC^c}{TC^e - TC^o}. \tag{18}$$

It is of interest to compare the social savings from congestion tolling using our model with the figures obtained from other papers. To our knowledge, all previous papers, except several by Small [33, e.g.],[9] which have estimated the efficiency gains from congestion tolling have employed the naive flow model of congestion. They differ significantly from one another in other respects, however. Limitation of space precludes a detailed description of each paper. We can, however, outline their main features.

The papers are usefully distinguished according to which margins congestion tolls affect. First, Arnott and MacKinnon [5] and Segal and Steinmeier [31] examined the effects of congestion tolling in a simple monocentric city model with a fixed CBD and fixed trip frequency. The congestion toll affected only lot size choice. The papers obtained efficiency gains of 0.07 and 0.08% of average household income, respectively. (Since some papers report efficiency gains as a proportion of transport costs, and others as a proportion of income, we shall make the approximation that transport costs are 10% of income and subsequently present efficiency gains as a percentage of transport costs.) Second, Sullivan [36] employed a more sophisticated monocentric city model with a CBD whose size is determined endogenously. As a result of this extra margin, he obtained efficiency gains equal to 8%. Third, Kraus *et al.* [25] estimated the efficiency gains on a freeway of

[9]Small [33] examines congestion on a transport corridor served by both bus and car. The number of trips is fixed and car and bus travel are congested by a bottleneck. The departure rate is assumed fixed over the rush hour, as a result of which the queue builds up linearly. Imposition of an optimal toll on car travel causes just enough modal switching to completely eliminate the bottleneck. Small obtains social savings of the same magnitude as we do, but these are due to modal switching rather than a change in the time pattern of departures.

optimal capacity from imposing optimal differential peak and off-peak tolls compared to the situation where an optimal uniform toll is applied. Depending on parameter values (including price elasticities) they obtained a range of estimates for the efficiency gains, varying from 0.1 to 1%. The margin of adjustment was the number of trips in the two periods. Fourth, Mohring [30] examined the effects of alternative policies on the modal split between bus and car, when overall travel demand is fixed, and found (his Table 6.3) that with volume-to-capacity ratios and without reserved bus lanes, the peak-hour efficiency gain from marginal social cost pricing both modes is almost 50% of travel costs in the equilibrium in which no car toll is charged and the bus fare is set at zero. Fifth, Kraus [24], in the richest model of all those discussed, which made endogenous both the frequency of non-work trips and modal choice in a monocentric city, obtained maximum welfare gains of about 8%. Combining the results of Small [33] and the above papers suggests that the lion's share of the efficiency gains from congestion tolls derive from their effect on modal choice.

In our bottleneck model, the congestion toll affects only the pattern of departure times, a margin that is ignored in all the above papers. With the optimal fine toll, efficiency gains are 50% of total variable transport costs. The monetary implications of such proportional savings are remarkable. Suppose, in the example above, that 200,000 individuals travel in the morning rush hour for 200 days a year, and that the discount rate is 5%. The present value of the (gross) social saving from applying the optimal coarse toll during the morning rush hour would be $1,344,000,000 ($6720 per car). Even after the costs of toll collection (administration, equipment, and congestion caused by the toll collection) are subtracted, the net social saving might still be impressive.

Our model is extremely simple. Not only does it ignore all the other margins of adjustment noted above, but it also assumes that everyone is identical and has the same desired arrival time. In fact, one expects firms to respond to congestion by choosing a distribution of work start times and introducing flextime, and individuals to spread themselves out over the rush hour so that those with low shadow values of travel time journey at the peak of the rush hour while those with high values travel at the tails. These effects would normally reduce the efficiency loss due to congestion and hence the potential efficiency gains from imposing congestion tolls. Contrarily, our model ignores another beneficial effect of tolls, the lengthening of the rush hour they induce, which Henderson [17] discusses. On balance, the above figure for the gross efficiency gain from imposing an optimal coarse toll seems too high, though not out of line with the figures mentioned in Vickrey [37]. Nevertheless, the model does suggest that technologically advanced schemes for urban traffic toll implementation deserve more serious attention than they have received to date.

TABLE 3
Optimal Capacity

	No toll	Social optimum	Optimal coarse toll
$\hat{s}$	$\sqrt{2}\,\S$	$\S$	$\sqrt{\psi}\,\S$

Note.

$$\S \equiv N\left(\frac{\beta\gamma}{2(\beta+\gamma)k}\right)^{1/2} \qquad \psi \equiv \frac{3}{2} - \frac{(\gamma-\alpha)\beta}{2(\beta+\gamma)(\alpha+\gamma)}.$$

VI. COST-BENEFIT ANALYSIS AND OPTIMAL CAPACITY

Since our model assumes a fixed demand for trips, the optimal capacity, $\hat{s}$, minimizes the sum of total travel costs and amortized construction costs, $K(s)$, i.e.,

$$\hat{s} = \operatorname{argmin}(\mathrm{TC}(s) + K(s)). \tag{19}$$

To simplify the analysis, we assume that there are constant costs to capacity expansion, i.e., $K(s) = ks$. The expressions for optimal capacity under the three toll régimes are presented in Table 3.

The following inequalities hold:

$$\hat{s}^{\mathrm{e}} > \hat{s}^{\mathrm{c}} > \hat{s}^{\mathrm{o}}. \tag{20}$$

The finer the toll, the less the congestion, the lower the marginal benefit from road expansion,[10] and the smaller optimal capacity.

The example of the previous section is now extended to treat capacity. It is assumed that, with actual capacity chosen to minimize total travel costs plus amortized construction costs in the no-toll equilibrium, the rush hour is 2 hours long. From Table 3, this implies that $k = 12.42$. With this value of k and the parameter values assumed previously, the results shown in

[10] In the analysis of the optimal capacity of a point-input, point-output road of constant width and subject to naive flow congestion, there are two offsetting effects determining whether second-best (i.e., subject to the toll being zero or suboptimal) road width is larger or smaller than the first-best (see Wheaton [39] and Wilson [40]). On the one hand, pricing travel below marginal social cost causes too many individuals to travel on the road. To counteract this, the road should be narrower than in the first best. On the other hand, widening the road may speed up the traffic so much that even though the number of cars may increase, the excess burden associated with unpriced congestion may fall. Which effect dominates depends on the trip demand function and the congestion technology. The effects here are analogous, and since auto travel demand is assumed to be completely inelastic, only the latter effect is operative.

TABLE 4

Example: Per Capita Aggregates, Capacity Optimal

Case	$\frac{\hat{s}}{N}$	$\frac{TC}{N}\Big\vert_{\hat{s}}$	$\frac{k\hat{s}}{N}$	$\frac{TC + k\hat{s}}{N}$
No-toll equilibrium	0.500	\$6.21	\$6.21	\$12.42
Social optimum	0.354	\$4.39	\$4.39	\$8.78
Optimum coarse toll	0.427	\$5.30	\$5.30	\$10.60

Table 4 are obtained. We note three features of interest: First, since the optimal length of the rush hour is the reciprocal of optimal capacity per individual, the rush hour is longest at the social optimum and shortest in the no-toll equilibrium. Second, whatever the toll régime, since TC is homogeneous of degree minus one in s (see Table 1) the first-order condition for optimal capacity, $\mathrm{TC}'(\hat{s}) + k = 0$, implies $-\mathrm{TC}/\hat{s} + k = 0$ or $\mathrm{TC} = k\hat{s}$—with optimal capacity, capacity construction costs equal total travel costs. Third, the social saving from application of the coarse toll is again significant, about \$1.82 per trip, and the optimal coarse toll is about 49.7% efficient, where the efficiency of the coarse toll here is

$$\mathrm{eff}^{\mathrm{c}} = \frac{(\mathrm{TC} + k\hat{s})^{\mathrm{e}} - (\mathrm{TC} - k\hat{s})^{\mathrm{c}}}{(\mathrm{TC} + k\hat{s})^{\mathrm{e}} - (\mathrm{TC} - k\hat{s})^{\mathrm{o}}}. \tag{21}$$

VII. CONCLUDING COMMENTS

This paper examined the economics of congestion at a single bottleneck. It extended previous work by examining a coarse toll and optimal capacity under various toll régimes. One interesting result was that the efficiency gains from application of an optimal toll, arising from the change in the frequency distribution of departure times it causes, can be substantially greater than the efficiency gains that have been computed employing the more familiar naive flow model of congestion. Furthermore, in our model a significant fraction of the gains from a fully optimal toll can be achieved by applying a single-step function toll—the optimal coarse toll—which should be implementable at significantly lower cost that the more complex time-dependent toll. It therefore appears that technologically sophisticated tolling systems, of the sort that have been proposed by Vickrey, may be cost-efficient and merit serious study.

Another interesting result is that total schedule delay costs are of the same order of magnitude as total variable travel time costs. Thus, current

cost-benefit practice, which typically ignores schedule delay costs, may yield seriously erroneous conclusions.

While the model of bottleneck congestion described in the paper provides a simplified description of both travel demand and the congestion technology, we believe that, appropriately extended, it can provide a rich description of urban rush-hour traffic flow. Desirable extensions include:

(a) Heterogeneity of Drivers

Drivers order themselves in a systematic way over the rush hour (de Palma *et al.* [11]). "Factory workers"—those with a relatively low shadow value of travel time and high schedule delay costs (low α/β and low α/γ) —tend to travel at the peak of the rush hour, while "professionals"—those with a relatively high shadow value of time and more flexible working hours (high α/β and high α/γ)—tend to travel in the tails. In cost-benefit practice, some average shadow value of time is employed in measuring the benefits of a capacity expansion. However, since drivers tend to order themselves over the rush hour so that those with a lower shadow value of time have greater travel time, the use of an average shadow value of time may introduce a bias in the direction of overinvestment in roads. It is also necessary to treat driver heterogeneity in determining the distributional effects of transport policies.

(b) Stochastic Capacity and Demand

Variability in the number of cars travelling over the rush hour can be considerable, as can be the variability in road capacity, at least in countries such as Canada with harsh winters. How are the equilibria in the various toll régimes, as well as optimal capacities, affected by such variability? The answer clearly depends on the knowledge commuters have concerning the state of the system, when deciding whether and when to leave for work. De Vany and Saving [14] have investigated this issue for the case of naive flow congestion, stochastic demand, and uninformed commuters (viz., commuters know only the probability density of demand). In cost-benefit practice, the stochastic nature of capacity and demand is treated crudely by choosing capacity on the basis of travel time saved under ideal conditions for, say, the 60th busiest hour of the year (Highway Research Board [21]).

(c) A Road Network

Extension to treat a road network with multiple origins and destinations and multiple routes between any origin and destination is necessary before the theory can be empirically implemented. Any change in the transport system will affect the pattern of rush-hour travel over the entire network. Preliminary results have been derived by Mahmassani and Herman [27] and de Palma *et al.* [12, 13]. It will be particularly interesting to see to what

extent the benefits from tolls are reduced when tolls can be applied on some roads but not on others.

(d) Hypercongestion

When flow is plotted against velocity at a given point on a road for different times, they tend to be negatively correlated in less busy periods, but positively correlated in at least the peak of the rush hour in major cities. The latter phenomenon is termed hypercongestion,[11] and corresponds to traffic jams, gridlock, and stop-start travel. We guess that in major cities most of the efficiency loss due to congestion occurs when traffic is hypercongested. For this reason, it is very important to provide a persuasive treatment of hypercongestion. While our model does not treat hypercongestion, it can be modified to do so.

(e) Variability in Working Hours

In our model, all firms have the same desired arrival time. In fact, however, different firms choose different working hours, and some firms give their workers a choice as to working hours (flextime). In deciding on its employees' work schedules, a firm should realize that workers are willing to pay (in the form of lower wages) not to travel in peak periods. Firms will weigh this saving against the loss that derives from their employees working at different times and at times that may be inconvenient to suppliers and customers. Hendrickson and Kocur [18] and Henderson [17] have examined how traffic flow over the rush hour is altered when employees have different working hours.

(f) Disequilibrium

We solved for the equilibrium distribution of departure times on the assumption that drivers are fully informed. In fact, however, because drivers have imperfect knowledge concerning traffic conditions which vary in a complex and largely unpredictable way, they can be expected to experiment with alternative departure time strategies and to learn only gradually which strategy is best. De Palma *et al.* [12] and Ben-Akiva *et al.* [6, 7, 8] have constructed computer models which simulate this adaptive behavior, and are continuing their work.

The model also needs to be extended to treat modal choice and elastic travel demand.

While the model developed was for morning-rush-hour road congestion, it can be modified to treat any congested facility subject to peak-load

[11] For an introduction to traffic flow models which permit hypercongestion, such as the kinetic theory of traffic flow or car-following theory, see, e.g., Institute of Traffic Engineers [23], Lindsey [25], and Herman [20].

demand. One potential application is telephone usage. Demands for different periods are interdependent, and this interdependence can be captured by modeling the telephone user as facing a tradeoff between the analog to travel time—the time spent waiting for a free line—and the analog to schedule delay—the cost of placing a call at an inconvenient time.[12]

REFERENCES

1. A. S. Alpha and D. L. Minh, A stochastic model for the temporal distribution of traffic demand—The peak hour problem, *Transp. Sci.* **13**, 315–324 (1979).
2. R. Arnott, Unpriced transport congestion, *J. Econ. Theory* **21**, 294–316 (1979).
3. R. Arnott, Some issues related to the economics of non-stationary state traffic flow, *Revue Econ.* **36**, 11–43 (1985) [in French].
4. R. Arnott, A. de Palma, and R. Lindsey, "Economics of a Bottleneck," Queen's University, Institute for Economic Research, discussion paper #636 (1985).
5. R. Arnott and J. MacKinnon, Market and shadow land rents with congestion, *Amer. Econ. Rev.* **68**, 588–600 (1978).
6. M. Ben-Akiva, A. de Palma, and P. Kanaroglou, Capacity constraints in traffic models with elastic demand, "Proceedings of the 10th planning and research colloquium," The Netherlands (1983).
7. M. Ben-Akiva, M. Cyna, and A. de Palma, Dynamic model of peak period congestion, *Transp. Res.* **18B**, 339–355 (1984).
8. M. Ben-Akiva, A. de Palma, and P. Kanaroglou, Dynamic model of peak period traffic congestion with elastic arrival rates, *Transp. Sci.* **20**, 164–181 (1986).
9. C. Daganzo, The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck, *Transp. Sci.* **19**, 29–37 (1985).
10. A. de Palma and R. Arnott, Usage-dependent peak-load pricing, *Econ. Lett.* **20**, 101–105 (1986).
11. A. de Palma, M. Ben-Akiva, C. Lefèvre, and N. Litinas, Stochastic equilibrium model of peak period traffic congestion, *Transp. Sci.* **17**, 430–453 (1983).
12. A. de Palma, C. Lefèvre, and M. Ben-Akiva, A dynamic model of traffic congestion in a corridor, *Int. J. Comput. Math. Appl.* **14**, 201–223 (1987).
13. A. de Palma, P. Hansen, and M. Labbé, "Commuters' paths with penalties for early or late arrival time," The Center for Mathematical Studies in Economics and Management Science, Northwestern University, discussion paper 727 (1987).
14. A. De Vany and T. Saving, "Competition and highway pricing for stochastic traffic," *J. Bus.* **53**, 45–60 (1980).
15. P. H. Fargier, "Influence du mécanisme de choix de l'heure de départ sur la congestion du traffic routier," Institut de Recherche des Transports, Arcueil, France (1981).
16. J. V. Henderson, "Economic Theory and the Cities," Chap. 8, Academic Press, New York (1977).
17. J. V. Henderson, The Economics of Staggered Work Hours, *J. Urban Econ.* **9**, 349–364 (1981).
18. C. Hendrickson and G. Kocur, Schedule delay and departure time decisions in a deterministic model, *Transp. Sci.* **15**, 62–77 (1981).

[12] The telephone problem differs from the traffic problem in that (i) in the former, all those in the queue (trying to place a call) have an equal probability of being served, while in the latter it is "first-come, first-served"; and (ii) the traffic bottleneck provides a flow constraint, while a telephone line a stock constraint.

19. C. Hendrickson and E. Planck, The flexibility of departure times for work trips, *Transp. Res. A* **18**, 25–36 (1984).
20. A. Herman, Remarks on traffic flow theories and the characterization of traffic in cities, *in* "Self-Organization and Dissipative Structures" (C. Schieve and P. M. Allen, Eds.), Univ. of Texas Press, Austin (1982).
21. Highway Research Board, "Highway Capacity Manual," Special Report 87 Highway Research Board, Washington, D.C. (1965).
22. V. F. Hurdle, Equilibrium flows on urban freeways, *Transp. Sci.* **15**, 255–293 (1981).
23. Institute of Traffic Engineers, "Traffic Engineering Handbook," Prentice–Hall, Englewood Cliffs, NJ (1976).
24. M. Kraus, The welfare gains from pricing road congestion using automatic vehicle identification and on-vehicle meters, *J. Urban Econ.* **25**, 261–281 (1989).
25. M. Kraus, H. Mohring, and T. Pinfold, The welfare costs of nonoptimum pricing and investment policies for freeway transportation, *Amer. Econ. Rev.* **66**, 532–547 (1976).
26. R. Lindsey, "Non-steady-state Traffic Flow," mimeo (1980).
27. H. Mahmassani and R. Herman, Dynamic user equilibrium departure time and route choice on idealized traffic arterials, *Transp. Sci.* **18**, 362–384 (1984).
28. D. McFadden, A. Talvitie, *et al.*, "Demand Model Estimation and Validation," The Urban Travel Demand Forecasting Project, Phase I Final Report Series, Vol. V, mimeo (no date).
29. H. Mohring, "Transportation Economics," Ballinger, Cambridge, MA (1976).
30. H. Mohring, The benefits of reserved bus lanes, mass transit subsidies, and marginal cost pricing in alleviating traffic congestion, *in* "Current Issues in Urban Economics" (P. Mieszkowski and M. Straszheim, Eds.), Johns Hopkins Press, Baltimore (1979).
31. D. Segal and T. Steinmeier, The incidence of congestion and congestion tolls, *J. Urban Econ.* **7**, 42–62 (1980).
32. K. Small, The scheduling of consumer activities: Work trips, *Amer. Econ. Rev.* **72**, 467–479 (1982).
33. K. A. Small, The incidence of congestion tolls on urban highways, *J. Urban Econ.* **13**, 90–111 (1983).
34. M. J. Smith, The existence and calculation of traffic equilibria, *Transport. Res.* **17B**, 291–303 (1983).
35. R. H. Strotz, Urban transportation parables, *in* "The Public Economy of Urban Communities" (J. Margolis, Ed.), Resources for the Future, Washington, D.C. (1965).
36. A. M. Sullivan, Second-best policies for congestion externalities, *J. Urban Econ.* **14**, 105–123 (1983).
37. W. S. Vickrey, Pricing in urban and suburban transport, *Amer. Econ. Rev.* **53**, 452–465 (1963).
38. W. S. Vickrey, Congestion theory and transport investment, *Amer. Econ. Rev.* **59**, 251–261 (1969).
39. W. Wheaton, Price-induced distortions in urban highway investment, *Bell J. Econ.* **9**, 622–632 (1978).
40. J. Wilson, Optimal road capacity in the presence of unpriced congestion, *J. Urban Econ.* **13**, 337–357 (1983).