

RECENT DEVELOPMENTS IN
THE BOTTLENECK MODEL¹

September 1995

Richard Arnott*

André de Palma**

and

Robin Lindsey***

Preliminary: Please do not cite or quote without the permission of one of the authors.

*Richard Arnott
Department of
Economics
Boston College
Chestnut Hill
MA 02169, USA

**André de Palma
HEC
Université de Genève
102 Bd. Carl Vogt
CH-12111 Genève 4
Switzerland

***Robin Lindsey
Department of
Economics
University of Alberta
Edmonton
Alberta T6G 2H4
Canada

¹We would like to thank Fabrice Marchal for technical assistance

Recent Developments in the Bottleneck Model

1. Introduction

The basic economic model of urban traffic congestion treats just one margin of individual choice – how many trips to take. In reality the array of decisions facing an individual is far more complex. It is useful to conceptualize the individual as deriving utility from activities (dining in an Italian restaurant, going skiing, meeting Joe over a beer, taking the kids to school). Given a choice of activities, the individual must decide which activities she will perform and how best to schedule them. The individual's transportation demand is derived from the solution to this choice and scheduling problem. The modern economic theory of urban traffic congestion falls far short of this level of conceptual sophistication, but this is the general direction in which the theory is being developed. Thus far, significant progress has been made in modeling only the individual's participation, departure time and route choice decisions for the morning commute. The seminal paper was by Vickrey (1969). Since Vickrey assumed that traffic congestion takes the form of queueing behind a bottleneck, the Vickrey model has come to be termed 'the bottleneck model' – hence the title of this paper.

The principal innovation of Vickrey's bottleneck model was to endogenize individuals' departure times. In the simple version of the model, identical morning commuters travel from home to work along a single road which has a bottleneck of fixed flow capacity. All commuters wish to arrive at work at the same time. But because the bottleneck capacity is finite this is physically impossible – some must arrive early and/or others late. The costs of early and late arrival are termed schedule delay costs. Each individual decides when to

2. The Analytics of the Simple Bottleneck Model¹

Every morning a fixed number, N , of identical individuals travel from home (O-origin) to work (D-destination). Drivers are treated as a continuum of measure N . Each individual travels by her own car along the single road joining O and D. Travel is uncongested except at a single bottleneck with a deterministic capacity of s cars per unit time. If the arrival rate at the bottleneck exceeds s , a queue develops. Travel time from O to D is

$$T(t) = \bar{T} + T^w(t),$$

where $\bar{T}$ is the fixed component of travel, $T^w(t)$ is waiting time (queueing time) at the bottleneck, and t is departure time from home. Let $Q(t)$ be queue length (number of cars in the queue) at time t . Then

$$T^w(t) = \frac{Q(t)}{s}; \quad (1)$$

an individual's queueing time equals queue length at the time she joins the queue divided by the bottleneck capacity (henceforth simply 'capacity'). To simplify algebra and terminology, and without loss of generality, we set $\bar{T} = 0$ so that $T(t) = \frac{Q(t)}{s}$; thus an individual arrives at the bottleneck as soon as she leaves home and arrives at work immediately upon getting through the bottleneck. With $r(t)$ denoting the departure rate, queue length evolves according to

$$\frac{dQ(t)}{dt} = \begin{cases} 0 & \text{for } Q(t) = 0 \text{ and } r(t) \leq s \\ r(t) - s & \text{otherwise} \end{cases} \quad (2)$$

¹ This section draws heavily on Arnott, de Palma and Lindsey (ADL hereafter) (1990a), and to a lesser extent on ADL (1993a).

over the period of departures, $\mathfrak{T}$, in such a way that no commuter can strictly reduce her trip price by changing her departure time. With identical individuals this equilibrium condition reduces to the pair of requirements that over $\mathfrak{T}$ trip price is the same for every commuter, whereas outside $\mathfrak{T}$ trip price is at least as high:

$$p(t) \begin{cases} = p & \text{for } t \in \mathfrak{T} \\ \geq p & \text{for } t \notin \mathfrak{T} \end{cases},$$

where p is the equilibrium trip price. The stage is now set to solve for the no-toll equilibrium. Note first that a queue must exist in the interior of the departure interval since otherwise a person departing when there was no queue would face a lower schedule delay cost than either the first or last person to depart, and the same (zero) travel time cost. The bottleneck therefore operates at capacity throughout $\mathfrak{T}$. Letting t_0 and t_e denote respectively the beginning and end of the departure interval, this implies

$$t_e - t_0 = \frac{N}{s}; \quad (4)$$

the length of the departure interval equals the number of commuters divided by capacity. The first person departs early (i.e. departs at such a time that she arrives early) and faces no queue; hence

$$p(t_0) = C(t_0) = \beta(t^* - t_0). \quad (5a)$$

The last person, who departs late, also faces no queue, since joining a queue would increase her travel time without reducing her time late. Hence

$$p(t_e) = C(t_e) = \gamma(t_e - t^*). \quad (5b)$$

Combining (4), (5a), (5b), the definition of t_n and the equal-trip-price condition

$$p(t_0) = p(t_e) = p \quad (5c)$$

and the queue dissipates linearly.³ Figure 1 depicts the no-toll equilibrium. Cumulative departures from home, ABC, and cumulative arrivals at work, AC, are plotted as functions

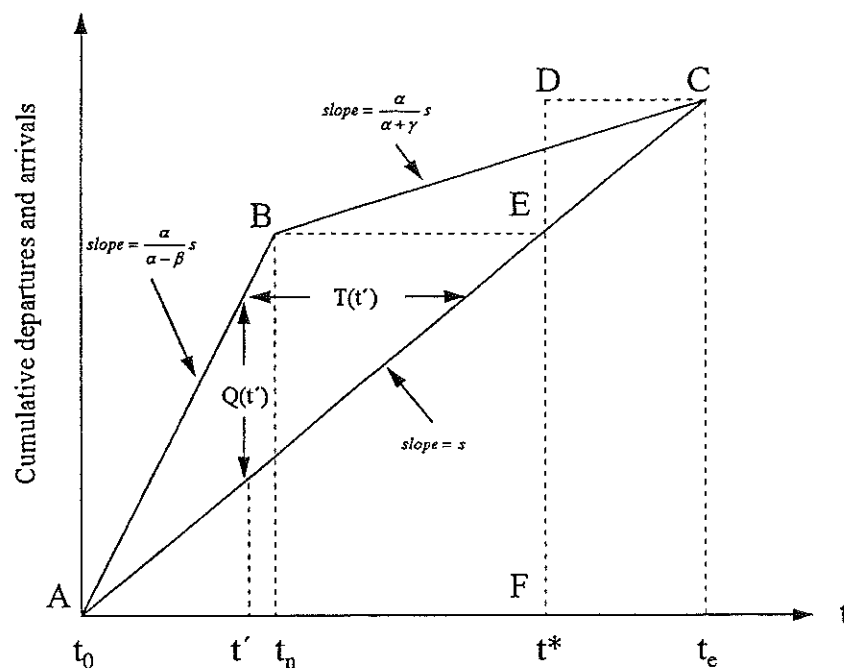


Figure 1: The no-toll equilibrium

of time. The slopes of the two curves are the departure rate from home, $r(t)$, and the arrival rate at work, s , respectively. The vertical distance between the two curves at t' equals the number of commuters who have left home by t' minus the number who have arrived at work, and hence equals queue length at t' , $Q(t')$. A point on the ordinate indexes a particular individual. Thus the horizontal distance between the cumulative departures and arrivals curves at t' measures an individual's travel time, $T(t')$.

Total travel time (the sum of travel times for all commuters) may be calculated by summing over time increments the number of commuters in the queue in each increment:

³ If, contrary to assumption, $\alpha < \beta$, all early arrivals wish to depart at the same time. A unique equilibrium exists that entails a mass of individuals departing at the same time (ADL 1985).

the added driver departs. Because of the linearity of the trip cost function, total schedule delay cost and total travel time cost are both half of total travel cost:

$$TTC = SDC = \frac{\delta}{2} \left(\frac{N^2}{s} \right). \quad (11a,b)$$

Social optimum

The social optimum is most easily established by intuitive reasoning. First the bottleneck should be used to capacity throughout the rush hour, since otherwise schedule delay costs are unnecessarily high. Thus $t_e - t_0 = \frac{N}{s}$. Second, as noted earlier the departure rate should never exceed capacity so that no queue develops; otherwise unnecessary travel time costs are incurred. Third, the schedule delay costs of the first and last persons should be the same; otherwise transferring commuters from one end of the rush hour to the other would reduce total costs. These conditions imply that the rush hour departure interval is the same for the social optimum as for the no-toll equilibrium; so too is the arrival interval and hence total schedule delay costs. Since total travel costs are double total schedule delay costs in the no-toll equilibrium, but equal to total schedule delay costs at the social optimum, total travel costs at the social optimum are only half those in the no-toll equilibrium:

$$TC^o = \frac{\delta}{2} \left(\frac{N^2}{s} \right), \quad (12)$$

where o denotes the social optimum.

Several features of the model and its solution are worthy of note.

- There exist a unique no-toll equilibrium and a unique social optimum.

- Travel time costs are pure deadweight loss. This points to the potentially very substantial travel cost savings (see eq. (10a) and eq. (10b)) that can be achieved simply by altering the time pattern of departures, which we show in Section 3 can be achieved by tolling.

3. Tolling in the basic bottleneck model⁵

In this section we show that the social optimum can be decentralized by an optimal time-varying toll. We then briefly consider the relative efficiency of step function tolls.

Decentralization of the social optimum via a toll

The time-varying toll which decentralizes the social optimum can be established by intuitive reasoning: impose a toll that at each point in time equals the travel time cost of the person who arrives at work at that time in the no-toll equilibrium. This induces a new equilibrium in which the equal trip-price condition holds without queueing. The toll simply replaces queueing cost as the rationing mechanism for arrival time slots. Furthermore the toll does not alter the rush-hour interval. It is easy to show that the optimal toll is⁶

$$\tau(t) = \begin{cases} \delta\left(\frac{N}{s}\right) - \beta(t^* - t) & \text{for } t \in [t_o, t^*] \\ \delta\left(\frac{N}{s}\right) - \gamma(t - t^*) & \text{for } t \in [t^*, t_e] \end{cases},$$

Bernstein (1994) has shown that when the discreteness of individuals is taken into account, a standard Nash equilibrium does not exist. We suspect that this is a 'technical problem' in the sense that, with the introduction of noise that reflects uncertainty about other individuals' departure times, equilibrium would exist.

⁵ This section draws primarily on ADL (1990a).

⁶ The magnitude of the toll outside $[t_o, t_e]$ must be set so that trip price exceed the equilibrium trip price; otherwise it is arbitrary.

where r denotes the pricing regime and Γ^r is a coefficient that does not depend on N or s .

For the no-toll equilibrium, $\Gamma = 1$ per (10a). For the social optimum, $\Gamma = 1/2$ per (12). And for the optimal step toll, $\Gamma \in (1/2, 1)$ is a decreasing function of the number of steps.

The bottleneck model demonstrates very effectively the magnitude of the potential efficiency gains that congestion tolls can achieve by changing the time pattern of departures, even without any reduction in the number of trips. As a back-of-the-envelope calculation consider a city of 1 million people of whom 40% work. Suppose, roughly updating Small's (1982, Table 3, col. 1) estimates, that $\gamma = \$20/\text{hr.}$, $\alpha = \$10/\text{hr.}$, and $\beta = \$5/\text{hr.}$, and that the length of the rush hour is one hour ($N/s = 1$). According to (11a) the daily efficiency gains from application of the optimal time-varying toll would be

$$\frac{(5)(20)}{25} \left(\frac{(4 \times 10^5)^2}{4 \times 10^5} \right) = 1.6 \times 10^6. \text{ With 200 commuting days a year and an annual discount}$$

rate of 0.05, this translates into discounted present value efficiency gains of $\$6.4 \times 10^9$, or \$16,000 per commuter! Comparable gains would presumably be achieved in the afternoon rush hour.⁸

It should be noted that optimal capacity varies according to the tolling regime. Since the marginal benefit from capacity is directly proportional to total travel costs⁹, which are higher the 'coarser' the tolling regime, optimal capacity is higher the coarser the tolling

⁸ There are conceptual difficulties in modelling the afternoon rush hour. Suppose, for example, that one were to assume that everyone left work eight hours after arriving. Then the departure rate from work would be s and there would be no queueing in the afternoon rush hour. The best-known formulation, which ignores the dependence between a driver's departure time from work and her arrival time, is Fargier (1983).

⁹ Given (13) the marginal benefit of capacity is $-\frac{\partial TC^r}{\partial s} = \Gamma^r \delta \frac{N^2}{s^2} = \frac{TC^r}{s}$.

There is an important qualitative difference between the cases of inelastic and elastic demand in the choice of optimal capacity because with elastic demand latent demand must be considered. An expansion of capacity lowers trip price, which induces increased traffic. Whether this affects the net marginal benefit from capacity expansion depends on whether trip price equals marginal cost. Let $B(N)$ be gross benefit from N trips and $W(N,s)$ net benefits, defined as $W(N,s) = B(N) - TC(N,s)$. Then

$$\begin{aligned} \frac{dW}{ds} &= \frac{\partial W}{\partial s} + \frac{\partial W}{\partial N} \frac{dN}{ds} = \frac{\partial W}{\partial s} + \frac{\partial W}{\partial N} \frac{dN}{dp} \frac{dp}{ds} \\ &= -\frac{\partial TC}{\partial s} + \left[\left(\frac{\partial B}{\partial N} - \frac{\partial TC}{\partial N} \right) \frac{dN}{dp} \frac{dp}{ds} \right]. \end{aligned} \quad (14)$$

The first term on the right-hand side of (14) captures the direct benefit of the capacity expansion – the benefit that occurs with no change in travel behaviour. The second term, in square brackets, captures the indirect benefit which comes about because of behavioral changes induced by the capacity expansion – here an increase in the number of trips due to the reduction in trip price caused by the capacity expansion. With no toll the marginal social cost of the induced increase in traffic exceeds the marginal social benefit. Thus the indirect benefit is negative. With marginal-cost pricing, however, the indirect benefit is zero because the marginal social benefit of the induced increase in traffic equals the marginal social cost. This is an example of a general principle in economics which surfaces repeatedly in urban transportation. If pricing is efficient (i.e. with marginal-cost pricing) only the direct effects of a small change need be considered; the indirect effects can be ignored. This is called the envelope theorem. But if pricing is inefficient, the envelope theorem does not apply and indirect effects have to be considered.

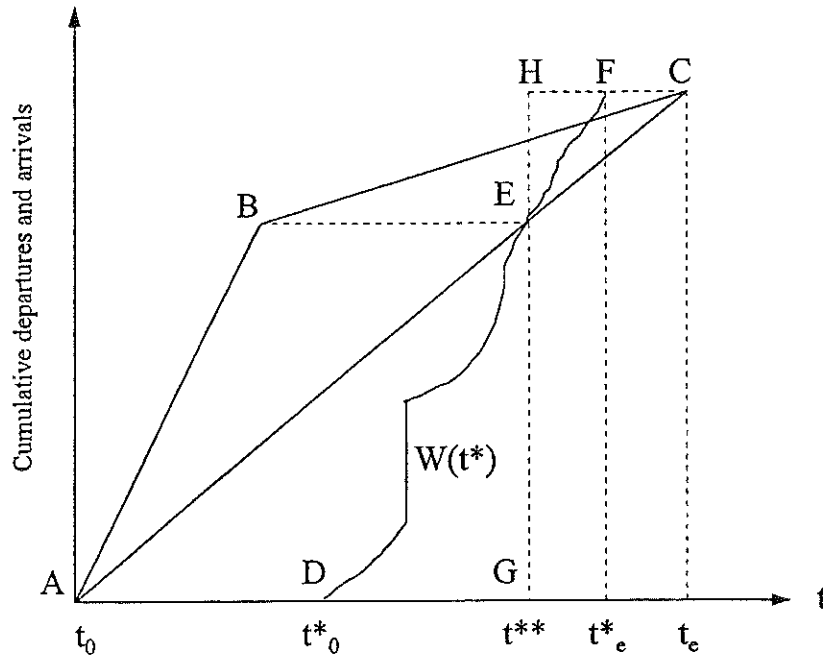


Figure 2: No-toll equilibrium with a distribution of t^*

individuals, $\frac{\gamma}{\beta + \gamma} N$ individuals arrive early, and $\frac{\beta}{\beta + \gamma} N$ arrive late. The departure rate, queueing pattern and total travel time costs are also the same as in the basic model where all commuters wish to arrive at t^{**} .¹³ But total time early is smaller by area DEGD, and total time late is smaller by area EHFE. A notable feature of the equilibrium is that, provided $s(t_e^* - t_0^*) < N$, total travel time is not reduced by spreading the distribution of desired arrival times. Staggering work hours is then ineffective for reducing congestion, although it does reduce total schedule delay costs.

Suppose now that, still with $s(t_e^* - t_0^*) < N$, $W(t^*)$ crosses the cumulative arrivals curve three or more times (the number must be odd). The no-toll equilibrium for this case is characterized in

¹³ The crucial property is that $W(t^*)$ cross the cumulative arrivals curve once. As illustrated in Figure 2, $W(t^*)$ can be flatter than s over part of the range $[t_0^*, t_e^*]$, and can also have discontinuities.

congestion is underpriced competitive equilibrium entails an excessively concentrated distribution of work start times. In the absence of congestion pricing, therefore, there is a role for government in encouraging greater dispersion of work start times.

Next suppose individuals differ in terms of α , β and γ , but have the same γ/β and t^* . This means that commuters differ in their absolute shadow costs of travel time and schedule delay, but not the ratio of their shadow cost of time early to time late. In the no-toll equilibrium commuters order themselves according to their relative cost of travel time to schedule delay, α/β . Those with the lowest α/β (the highest relative cost of schedule delay) travel at the peak. In contrast, with the optimal time-varying toll, which eliminates queueing, commuters order themselves according to the absolute cost of schedule delay, those with the highest cost traveling at the peak.

Roughly speaking, imposing tolls entails paying with money rather than with travel time. Thus, tolls tend to benefit disproportionately those with a high shadow cost of travel time, who are typically the rich. This may be one of the principal reasons why tolls are so unpopular politically. To counteract this adverse equity effect, several authors have recommended a package approach whereby toll revenue would be allocated to various uses (road construction and maintenance, park and ride facilities, improved transit, general revenues and so on), in such proportions that all major groups end up better off.¹⁵

Heterogeneity also has implications for optimal capacity. If one defines optimal capacity to be that capacity which minimizes total cost, then cost-benefit practitioners should employ shadow costs of time which are a weighted average of the shadow costs of

¹⁵ See for example Goodwin (1989), Bayliss (1992), Jones (1992) and Small (1993).

Mahmassani and Herman (1984) using Greenshields' flow congestion model in which travel speed falls linearly with density. The same network configuration was later studied by ADL (1990b) using the bottleneck model.

Regardless of the 'performance model' (the technology of congestion, e.g. flow or bottleneck congestion) or the number of routes, equilibrium can be characterized in two steps. The first step involves solving for the equilibrium departure rate on each route taking the number of drivers on the route as given. The second step involves calculating the route split by applying Wardrop's principle that travel cost be equal on all routes that are used, and equal or higher on routes that are not used. This implies that all drivers departing at the same time must experience the same travel time, whatever route they choose. If routes differ in free-flow travel costs, the first drivers to depart take the shortest route. Longer routes start to be used only when travel times on shorter routes have been driven up sufficiently by congestion.

In the case of bottleneck congestion and two routes in parallel that both get used, ADL (1990b) show that equilibrium travel cost is

$$C = \alpha \frac{s_1 \bar{T}_1 + s_2 \bar{T}_2}{s_1 + s_2} + \delta \left(\frac{N}{s_1 + s_2} \right),$$

where $\bar{T}_i$ is the fixed travel cost on route i , $i=1,2$. Comparing this with (7) it is apparent that in terms of costs the network is equivalent to a single route with capacity $s_1 + s_2$ and a fixed travel cost equal to the capacity-weighted average fixed cost on the two routes. This formula generalizes straightforwardly to any number of routes in parallel.

queueing time downstream and schedule delay costs of both groups that comes about through induced changes in departure rates. In this situation a reduction in capacity would

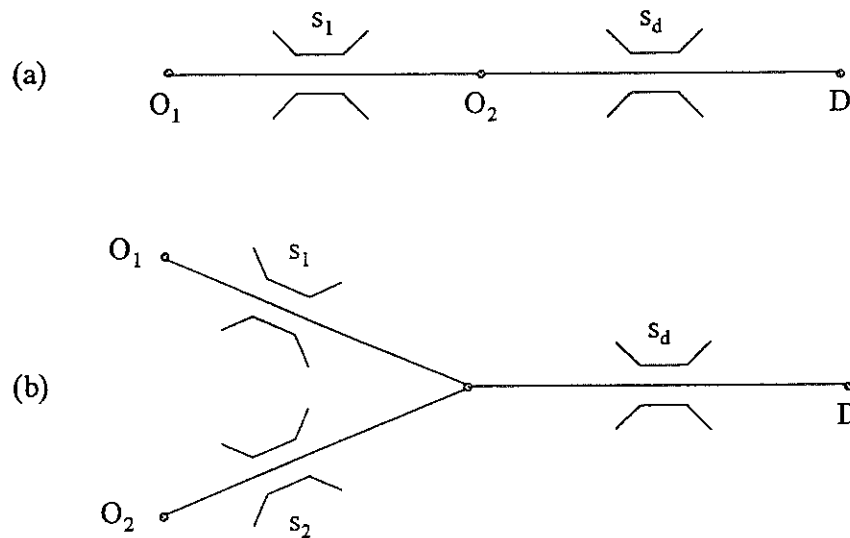


Figure 3: Networks with bottlenecks in series

be welfare-improving. In lieu of tearing up traffic lanes, this might be accomplished by metering access to the bottleneck in a manner analogous to the practice of metering freeway on-ramps.

This dynamic network equilibrium paradox is reminiscent of the celebrated Braess (1968) paradox for static user equilibrium whereby adding a link to a network results in an increase in total travel costs. In response to the link addition, some drivers alter routes to lower their private costs, but since congestion is unpriced this may entail their switching from routes with lower social costs to routes with higher social costs. The dynamic network equilibrium paradox results not from route choice adjustments but from departure time adjustments. (Note that drivers have no choice of route in Figure 3(b).)

regimes. We focus here on the case where optimal time-varying tolls are applied on the whole road network, thereby supporting the system optimum — at least on simple networks.²⁰

Some results for the system optimum with one O/D pair and two routes in parallel were derived by ADL (1990b). As with the basic model, the departure rate on each route is maintained at the capacity of the respective bottleneck to prevent queueing. This departure rate can be supported as an equilibrium by applying a time-varying toll on each route analogous to the toll for the single bottleneck model. If the two routes differ in fixed travel costs, then a shift in traffic from the shorter to the longer route is also required because in the no-toll equilibrium the shorter route has a longer queue and thus experiences greater congestion. The shift in traffic can be realized by adding to the time-varying toll on the shorter route a positive time-invariant component. Unless the routes differ appreciably in fixed travel costs, the efficiency gains derived from inducing the optimal route split are a small fraction of the gains achieved by eliminating queueing. This suggests that the substantial efficiency gains from time-varying tolls that obtain with the single bottleneck model carry over to networks.

For the networks in Figure 3 too it can be shown that the system optima entail no queueing and can be supported by time-varying tolls. One might be tempted to conclude from this that queueing can never be socially optimal on a network characterized by bottleneck congestion. Yet de Palma and Jehiel (1994) have shown that queueing can be socially optimal. One of their examples is depicted in Figure 4. A single O/D pair is

²⁰ Dafermos and Sparrow (1969) have shown that the system optimum can be decentralized using link-based tolls if and only if the number of paths does not exceed the number of links plus the number of OD pairs. Where this condition is not satisfied route-based tolls are required for decentralization.

The maximum number of drivers that can be accommodated without queueing is thus $(\overline{T}_B - \overline{T}_A)s_B + 2\Delta s$. Yet even more can be accommodated by feeding them onto route B at rate s for the period $\overline{T}_B - \overline{T}_A$. While they depart, a queue develops on route B at a rate $s - s_B$. But given (15) all drivers taking route B still arrive within the time window.

In sum: if the number of drivers falls in the range $((\overline{T}_B - \overline{T}_A)s_B + 2\Delta s, (\overline{T}_B - \overline{T}_A + 2\Delta)s]$, then all drivers can be accommodated within the time window, but only by allowing queueing. Given the assumption that arrival outside the window is much more costly than queueing time, queueing is socially optimal. The assumption of prohibitively costly schedule delay costs outside the arrival time window can obviously be relaxed somewhat and queueing will still occur in the social optimum.

de Palma and Jehiel (1994) provide another example, featuring the cost function (3) and two groups of drivers with different values of α , β and t^* , in which queueing is again optimal. These two examples suggest that queueing may be a common feature of the social optimum on networks with bottleneck congestion.

A final word is in order regarding cost-benefit analysis. In contrast to the no-toll equilibrium, the indirect effects of a small policy change on the system optimum can be ignored. Hence the cost-benefit analysis of a small, localized, policy change can ignore how it affects traffic flows on the rest of the network.

c) Tolling on part of the network

Full congestion pricing on an entire road network has never been achieved and is not a realistic prospect in the near future. Impediments to comprehensive pricing include the infrastructure and administration costs, equity concerns and (at least in the US) respect for

compromise between reducing total travel, which requires a relatively high toll, and achieving an appropriate division of traffic which (depending on the relative congestibility of the two routes) may dictate a toll that is negative. As expected, the relative efficiency of second-best pricing to first-best pricing declines as the elasticity of demand rises and conflict between the two goals becomes increasingly severe. Contrarily, second-best pricing is more efficient the less congestible and shorter is route 2 relative to route 1, because route 2 then carries a larger share of total traffic, leaving less unpriced congestion on route 1. This suggests that priority should be given to tolling short, high-capacity, roads.

Tabuchi (1993) has studied another variant of the second-best pricing problem in which a road subject to bottleneck congestion runs in parallel with a railway. Operating costs of the rail include a fixed cost and a constant marginal cost per rider. The railway is thus congestion-free, but uneconomic to run at low levels of demand. On the assumption of inelastic demand, and average-cost pricing of rail trips, Tabuchi shows that the road share of travel is highest with an optimal fine (i.e. time-varying) road toll and successively smaller with a step toll, a uniform toll, and no toll. This accords with Braid's analysis of two bottlenecks in parallel. But having an uncongested railway as the alternative mode, rather than a congestible highway, alters the relative efficiency gains of the various road tolls. At sufficiently high levels of demand, a uniform toll yields more than half the efficiency gains from the fine toll. These gains accrue from diverting people from the congested road to the uncongested rail. (By comparison, with inelastic demand and either a single road in isolation or two roads, a uniform toll is worthless.) Additionally, an

disablings and so on. Fluctuations also occur in travel demand. It is frequently asserted²⁵ that more than 50% of time loss due to congestion is 'non-recurrent', stemming from fluctuations in capacity and demand.

Traffic accidents and other disturbances that occur during a given travel period have been incorporated into dynamic network simulations, but have not been treated analytically. This section focuses on analytical work with the bottleneck model that has been done on day-to-day fluctuations in capacity and demand. The effects of such stochasticity depend on the extent to which realizations of capacity and demand are anticipated. At one extreme, if they are perfectly anticipated equilibrium is the same as that which obtains in the deterministic bottleneck model with the particular realization of capacity and demand. At the other extreme individuals know only the bivariate probability distribution of capacity and demand. More generally, they will obtain some but less than perfect information about driving conditions from radio and other information sources. A working assumption is that individuals make travel decisions to minimize expected trip price, with expectations conditioned on the information available to them.

Fluctuations in capacity and demand create fluctuations in travel times and costs. On days when capacity is low or demand is high, travel times will be high, whereas if capacity is high and/or demand is unusually low, there may be no congestion. Except in the extreme case of perfect anticipation, uncertainty about trip price tends to increase over the course of the rush hour, and will have consequences for the equilibrium distribution of trip timing.²⁶

²⁵ See for example Lindley (1987). This proposition has recently been challenged by Hall (1993).

²⁶ A characterization of the no-toll equilibrium under uncertainty using the bottleneck model is given in ADL (1995b). There it is shown that the provision of better, but less than perfect, information about

jams;²⁷ intersection congestion is predominant in downtown areas, leading in extreme cases to gridlock; and pedestrian and parking congestion are also important.

It remains an open question how urban auto congestion is best treated in simulation models aimed at practical implementation. At least five approaches have been taken: a) assume that traffic behaves as if there were bottleneck congestion on each link (Ghali and Smith 1995); b) assume flow congestion on each link (Mahmassani and Herman 1984); c) assume that capacity on a link depends on queue length (Yang and Huang 1995); d) assume flow congestion upstream on each link and queueing congestion downstream (Jayakrishnan et al 1994); and e) assume that travel speed on a link depends on either the entry rate onto that link (Henderson 1974) or the exit rate from it (Chu 1993).

There are at least three major considerations in deciding on the performance model: conceptual soundness, forecasting accuracy, and computational speed. Just as economists often insist that a model have solid microeconomic foundations, so might it be reasonable to require that performance models have a sound foundation in physics. The bottleneck model satisfies this criterion. Unfortunately, however, non-stationary flow congestion models, based on either fluid flow or car following theory, are analytically intractable (Newell 1988) and computationally demanding, even on massively parallel computers.

²⁷ One difficult issue is how to deal with traffic-jam type situations – referred to in the economics literature as hypercongestion – in which flow is positively related to velocity. The flow congestion models treat traffic jams by assuming that velocity, v , is negatively related to density, k , which is consistent with velocity being positively related to flow, f , at high densities ($f = kv$ and $k = k(v)$, implying $f'(v) = k'(v)v + k$). Hypercongestion can be incorporated into the bottleneck model by assuming that bottleneck capacity is negatively related to queue length. The current view, based especially on field studies by Hall and various coauthors (see Hall et al 1992), is that both these treatments are oversimplified, and that hypercongestion is a transient phenomenon triggered by some stochastic irregularity in flow, which generates a shock wave that travels backwards through the traffic.

5. METROPOLIS

METROPOLIS provides a simulation tool to compute a dynamic equilibrium in a general network. This simulator solves the following problem: Given a directed connected graph, dynamic congestion laws, travel cost functions and an O/D matrix, compute the departure time and route choice equilibrium. The links of the graph correspond to actual streets or to artificial links which connect the centroids (origins and/or destinations) to the actual street network. We use the bottleneck model to describe the congestion technology. The travel cost function is $\tilde{C}(t) = C(t) + \mu \varepsilon$ where ε is a normalized double exponential random variable, μ is a scale parameter which measures the degree of commuter heterogeneity, and $C(t)$ is given by eq. (3). The O/D matrix specifies the number of commuters making a home-to-work trip between each origin and destination. At equilibrium, no individual can decrease her travel cost by modifying either her departure time or route. This model reduces to Vickrey's model (extended to a network) in the limit as μ tends to zero (see Ben-Akiva et al. 1986).

We will now compare the results of the simulator with the no-toll equilibrium of the single bottleneck model. For this purpose we use data for the city of Geneva and its suburbs — an area of approximately 200 square miles. There are 79,300 commuters and 30,000 O/D pairs³⁰. Values of the schedule delay cost parameters are assumed to be $t^* = 8:06$ AM, $\alpha = \$6/\text{hr}$, $\beta = \$3/\text{hr}$ and $\gamma = \$12/\text{hr}$; hence $\delta = \$2.4/\text{hr}$. (The ratios $\frac{\beta}{\alpha}$ and $\frac{\gamma}{\alpha}$ are similar to Small's (1982) estimates.) The scale parameter is set at $\mu = 0.5$ and the time

³⁰ The network has 1,200 nodes, 300 centroids and 3,100 links. Each link is characterized by a fixed travel time and a capacity (the bottleneck is located at the end of each link).

Given $\frac{\hat{N}}{\hat{s}} = 110.2$ mins. this implies $\hat{s} = 677$ vehicles/min.³² Finally, from (9) $r_E = 1,354$

vehicles/min. and $r_L = 226$ vehicles/min.

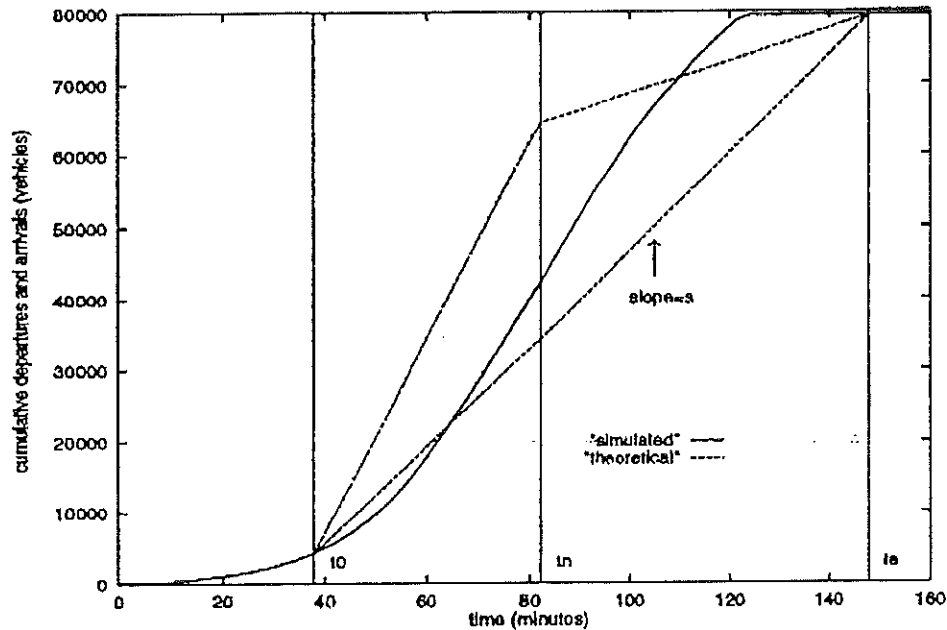


Figure 5: The simulated and simple bottleneck departure schedules

The equilibrium cumulative departures schedule for the single bottleneck model is indicated in Figure 5 by the dashed lines. The departure rate in the simulated equilibrium reaches a similar maximum value, and for a comparable duration of time. But there is no appreciable decrease in the departure rate until near the end of the rush hour, whereas in the bottleneck equilibrium the departure rate drops abruptly below capacity when individuals start to arrive late.

To investigate possible reasons for this contrast, METROPOLIS was re-run using successively smaller values of μ . This made no appreciable qualitative difference to the equilibrium departure schedule. We suspect that the geographical dispersion in origins and destinations on the Geneva network is primarily responsible. Individuals making

³² Note that about 6% of commuters do not experience congestion in the simulated equilibrium.

done to improve understanding of the complex interaction between parking and traffic congestion.

b) Mass transit

Mass transit has yet to be incorporated fully into either analytical bottleneck models³³ or dynamic network equilibrium models. To do so appears, conceptually at least, relatively straightforward.

c) Intersections

As noted earlier, probably the most important form of congestion in downtown areas is intersection congestion. While intersection congestion has been built into 'microsimulation models' which track individual cars but do not solve for equilibrium (e.g. NETSIM) it has yet to be treated in the bottleneck model.

d) Demand

The treatment of demand in the basic bottleneck model is very simple. The number of trips is a function of trip price, which includes travel time costs, schedule delay costs and toll costs, where the cost of travel time and the cost of time early or late are taken as exogenous. Several authors (e.g. Vickrey (1969, 1973), Henderson (1974, 1981)) have developed somewhat more sophisticated demand submodels in which individuals choose their allocation of time to maximize utility. But these still fall a long way short of the conceptualization described in the introduction, which is to solve for each individual's trip demand over time and space, derived as part of the solution to the individual's activity choice and scheduling problem. A satisfactory treatment of demand is actually even more

³³ Tabuchi (1993) has made a good start.

commuting trips, but to do this in a way that is both tractable and operational will require a conceptual breakthrough.

g) Location

The current generation of dynamic network equilibrium models is designed for short-term forecasting and policy simulation. Before these models are applied to longer-term forecasting, as would be necessary in the cost-benefit analysis of a transport improvement, it will be necessary to forecast how the locational pattern of trip demand will change. The trends toward subcentering and increased decentralization will probably continue. A further consideration is that any major change in transport policy or technology will affect urban spatial structure and the spatial structure of travel demand.

7. Conclusions

The Vickrey bottleneck model is over twenty-five years old. At first glance the model appears almost trivially simple. But its innovation, to model the evolution of traffic flow over the rush hour as an equilibrium phenomenon, incorporating commuters' trip-timing decisions has given rise to a new generation of dynamic urban traffic congestion models that are a quantum level more sophisticated and satisfactory than the previous generation of static rush-hour traffic flow models.

This paper has described the state of the art with respect to the bottleneck model. Economists have contributed primarily to analytical development of the model and to exploration of its economic implications. Transportation engineers/scientists have focused on its practical implementation via the development of dynamic network equilibrium

BIBLIOGRAPHY

- Arnott, R., A. de Palma and R. Lindsey (1985), 'Economics of a Bottleneck', Queen's University, Institute for Economic Research, discussion paper #636.
- Arnott, R., A. de Palma and R. Lindsey (1988), 'Schedule Delay and Departure Time Decisions with Heterogeneous Commuters', Transportation Research Record, 1197, 56-67.
- Arnott, R., A. de Palma and R. Lindsey (1990a), 'Economics of a Bottleneck', Journal of Urban Economics, 27, 111-130.
- Arnott, R., A. de Palma and R. Lindsey (1990b), 'Departure Time and Route Choice for Routes in Parallel', Transportation Research B, 24B(3), 209-228.
- Arnott, R., A. de Palma and R. Lindsey (1991a), 'Does Providing Information to Drivers Reduce Traffic Congestion?', Transportation Research A, 25A(5), 309-318.
- Arnott, R., A. de Palma and R. Lindsey (1991b), 'A Temporal and Spatial Equilibrium Analysis of Commuter Parking', Journal of Public Economics, 45(3), 301-335.
- Arnott, R., A. de Palma and R. Lindsey (1992), 'Route Choice with Heterogeneous Drivers and Group-Specific Congestion Costs', Regional Science and Urban Economics, 22, 71-102.
- Arnott, R., A. de Palma and R. Lindsey (1993a), 'A Structural Model of Peak- period Congestion: A Traffic Bottleneck with Elastic Demand', American Economic Review, 83(1), 161-179.
- Arnott, R., A. de Palma and R. Lindsey (1993b), 'A Dynamic Traffic Equilibrium Paradox', Transportation Science, 27(2), 148-160.
- Arnott, R., A. de Palma and R. Lindsey (1994), 'The Welfare Effects of Congestion Tolls with Heterogeneous Commuters', Journal of Transport Economics and Policy, 28(2), 139-161.
- Arnott, R., A. de Palma and R. Lindsey (1995a), 'Information and Usage of Free-Access Congestible Facilities', International Economic Review, forthcoming.
- Arnott, R., A. de Palma and R. Lindsey (1995b), 'Information and Time-of-Usage Decisions in the Bottleneck Model with Stochastic Capacity and Demand', manuscript.
- Arnott, R., A. de Palma and R. Lindsey (1995c), Congestion: A Dynamic Approach, Cambridge, Mass.: MIT Press, in progress.

- de Palma, A. F. Marchal and Y. Nesterov (1995), 'METROPOLIS: A Modular Architecture for Dynamic Traffic Simulation', manuscript.
- Douglas, R.W. (1975), 'A Parking Model - The Effect of Supply on Demand', American Economist, 19(1), 85-86.
- Fargier, P.H. (1983), 'Effects of the Choice of Departure Time on Road Traffic Congestion: Theoretical Approach', Proceedings of the 8th International Symposium on Transportation and Traffic Theory, Toronto, Canada, 223-263.
- Ghali, M. and M. Smith (1995), 'A Model for the Dynamic System Optimum Traffic Assignment Problem', Transportation Research B, 29B(3), 155-170.
- Glazer, A. and E. Niskanen (1992), 'Parking Fees and Congestion', Regional Science and Urban Economics, 22, 123-132.
- Goodwin, P.B. (1989), 'The "Rule of Three": A Possible Solution to the Political Problem of Competing Objectives for Road Pricing', Traffic Engineering & Control, 30(10), 495-497.
- Hall, F.L., V.F. Hurdle and J.H. Banks (1992), 'A Synthesis of Recent Work on the Nature of Speed-Flow and Flow-Occupancy (or Density) Relationships on Freeways', paper presented at the 71st annual meeting of the Transportation Research Board.
- Hall, W.R. (1993), 'Non-Recurrent Congestion: How Big Is the Problem? Are Traveler Information Systems the Solution?', Transportation Research C, 1C(1), 89-103.
- Henderson, J.V. (1974), 'Road Congestion: A Reconsideration of Pricing Theory', Journal of Urban Economics, 1, 346-355.
- Henderson, J.V. (1981), 'The Economics of Staggered Work Hours', Journal of Urban Economics, 9, 349-364.
- Hendrickson, C. and G. Kocur (1981), 'Schedule Delay and Departure Time Decisions in a Deterministic Model', Transportation Science, 15(1), 62-77.
- Jayakrishnan, R., H. Mahmassani and T-Y Hu (1994), 'An Evaluation Tool for Advanced Traffic Information and Management Systems in Urban Networks', Transportation Research C, 2(3), 129-147.
- Jones, P.M. (1992), 'Gaining Public Support for Road Pricing through a Package Approach', paper presented at the Sixth World Conference on Transportation Research, Lyon, France.

- Tabuchi, T. (1993), 'Bottleneck Congestion and Modal Split', Journal of Urban Economics 34, 414-431.
- Verhoef, E.T., R.H.M. Emmerink, P. Nijkamp and P. Rietveld (1994), 'Information Provision, Flat- and Fine Congestion Tolling and the Efficiency of Road Usage', Tinbergen Institute, Discussion paper TI 94-157.
- Vickrey, W.S. (1969), 'Congestion Theory and Transport Investment', American Economic Review (Papers and Proceedings), 59, 251-260.
- Vickrey, W.S., (1973), 'Pricing, Metering, and Efficiently Using Urban Transportation Facilities', Highway Research Record, 476, 36-48.
- Yang, H. and H-J. Huang (1995), 'Analysis of Time-Varying Pricing of a Bottleneck with Elastic Demand Using Optimal Control Theory', manuscript.
- Young, W., M.A.P. Taylor, R.G. Thompson, I. Ker and J. Foster (1991), 'CENCIMM, A Software Package for the Evaluation of Parking Systems in Central City Areas', Traffic Engineering & Control, 32(4), 186-193.

