The Bottleneck Model: An Assessment and Interpretation

Kenneth A. Small[a]

January 6, 2015

[a]Department of Economics, University of California, Irvine, CA 92697-5100, USA. Email: ksmall@uci.edu. Telephone: +1-949-824-5658

JEL Codes: R41, R48

Keywords: Congestion; bottleneck; scheduling; congestion pricing; parking; reliability

Forthcoming, *Economics of Transportation*

Abstract

The bottleneck model of congestion with endogenous scheduling has become a standard tool of transportation economics. It provides surprising insights about the time pattern of congestion, optimal pricing, and many distinct inefficiencies of unpriced equilibria including wrong departure order with heterogeneous preferences, wrong allocation of users across links of a network, and wrong order in which parking spaces are occupied. It illuminates the roles of travel-time reliability, traffic information, and extreme congestion ("hypercongestion"). It has been developed for use in practical network planning. Future use will probably emphasize greater realism, leading to more practical applications.

The Bottleneck Model: An Assessment and Interpretation

Kenneth A. Small

The so-called "bottleneck model", as formulated by Vickrey (1969) and elaborated especially in papers by Arnott, de Palma, and Lindsey (hereafter ADL),[1] is arguably the most fundamental advance in congestion analysis since the static congestion model of Walters (1961). It has provided significant new insights and computational tools for understanding many features of congestion. These insights include the nature of time-of-day shifts (e.g. the "shifting peak" phenomenon), various inefficiencies in unpriced equilibria, the temporal pattern of optimal pricing, and some surprising effects of pricing on travel patterns and travel costs. The model sheds new light on such diverse matters as residential location, parking, metering to improve traffic flow, and agglomeration. It suggests fruitful ways to analyze travel-time reliability, and to understand a form of extreme congestion known as "hypercongestion", in which traffic flow and speed covary positively. Furthermore, the model has a reduced form that is a special case of the Walters static model, making it possible to apply the many insights into congestion that have arisen from that more widely known approach.

In this brief review, I comment selectively on the nature and implications of this pioneering model, as well as its likely further use in research. Along the way, I consider how the model has shaped literature in economics and engineering, and how it is likely to do so in the future.

As part of a special issue honoring Richard Arnott, I cannot resist a personal note about how this model was developed. In the early 1980s, I was visited by Arnott, who enthusiastically described ambitious plans for collaboration with a visiting colleague and a former student (de Palma and Lindsey, respectively). He proceeded to outline a ten-year research program that would systematize the Vickrey model, create a transparent notation for it, provide an elegant derivation of key properties, and work out a number of generalizations. I could not imagine projecting a research agenda that far ahead, much less naming the collaborators; thus I tried to encourage him in the overall project while lowering his expectations to ones that seemed more

[1] The model was developed in a number of papers by ADL and others. Two of the most definitive early statements are ADL (1990, 1993a).

realistic. But his vision proved uncannily accurate: the resulting papers demonstrate these authors' success in bringing Arnott's (and perhaps Vickrey's) initial ideas to fruition, as well as in developing numerous and sophisticated additional directions. And as we shall see, this progress has engaged many other talented researchers as well.

1. The bottleneck model in essence

The model is a combination of two features, only one of which is indicated in its name. Congestion takes the form of queuing behind a simple deterministic bottleneck, usually interpreted as the entrance to a central business district (CBD). Demand results from a particular form of scheduling preferences: scheduling costs, which are piecewise linear in the discrepancy between desired and actual arrival time, are traded off against travel-time costs.

The supply side (i.e. congestion formation) accounts for the model's name; but it is the demand side that is more central to its importance. This is because endogenous scheduling relaxes the fundamental limitation of static models and opens an entire realm of behavior (endogenous scheduling shifts) to new understandings.

It is also the demand side that is most manifestly unrealistic, at least in the model's usual formulation. Demand consists of "$\alpha - \beta - \gamma$" preferences, in which travelers trade off travel time, valued at α per unit, against scheduling inconvenience. For the latter, there is a single predetermined preferred time t^* for arrival at the end of the bottleneck; deviations from arrival at t^* result in scheduling costs equal to β per unit of arrival time if early (i.e. arrival prior to t^*) and to γ per unit if late. Sometimes t^* is replaced by an interval of indifference (as in Ben-Akiva et al. 1984), with relatively minor effects on results.

What the "bottleneck" technology contributes is a practical way to close the model, thereby enabling equilibrium results to be computed, evaluated, and compared across different situations. This description of congestion has proven to be simpler and more amenable to analytical results than the flow-congestion approach pioneered by Henderson (1974) and updated by Chu (1995); flow congestion, while more flexible, creates a model that is exceedingly difficult to solve without making significant approximations such as that the speed for an entire trip depends on conditions at just one point in time.

Numerical simulations of the bottleneck model typically rely on any of the numerous estimates of "value of time" for α, and on one of the few empirical estimates of scheduling parameters, typically that of Small (1982). Small's results are often characterized in approximation as supporting ratios β/α=0.5 and γ/α=2. These estimates satisfy the condition $\beta<\alpha$, which is important for existence of equilibria and thus is often assumed.

The assumption of a single universal preferred exit time from the bottleneck is curious. A moment's thought suffices to realize that even if everyone wanted to be at work at the same time (itself a gross simplification), the diversity of destinations would prevent them from wanting to exit the bottleneck at the same time. Interestingly, this homogeneity assumption was not made in the seminal paper by Vickrey (1969), nor in an important generalization of it by Newell (1987).[2] Rather, Vickrey assumed a uniform distribution of t^*, which does not greatly complicate the analysis. Why, then, did ADL and nearly all subsequent elaborations of the model choose to assume homogeneity in preferred arrival time?[3] Probably because it facilitates easier and more transparent generalizations, for example to two bottlenecks in a network or to random capacity; and because it greatly simplifies welfare analysis, as it implies that everyone achieves the same utility in equilibrium. Thus the homogeneity simplification has a significant advantage for developing theory. Nevertheless I believe further progress will require its removal, especially for empirical application. It is encouraging that generalizations to more realistic distributions of preferred arrival times appear to be tractable, at least for the simplest versions of the model.

2. Basic Insights

The model calls attention to several features of equilibria with traffic congestion, some of which are surprising and many of which survive, in modified form, when assumptions are relaxed.

[2] See Small and Verhoef (2007), Section 4.1.2. Other early derivations of certain properties include those in Hendrickson and Kocur (1981), Fargier (1983), Ben-Akiva et al. (1984), Daganzo (1985), and Braid (1989).

[3] A few authors have assumed stochastic rather than deterministic demand, which implies a different kind of heterogeneity in desired arrival time. These include Ben-Akiva et al. (1984), Ben-Akiva et al. (1986), and the developers of the METROPOLIS model discussed later in this paper. Stochastic demand greatly facilitates finding the unique equilibrium via an adjustment process.

Time pattern of congestion. Perhaps the most fundamental feature is that the time pattern of congestion has a shape determined mainly by scheduling preferences. Bottleneck capacity affects the duration and severity of the congested period, but not the rate at which queuing time rises or falls. This result depends on an equilibrium condition. If users are competitive, in the sense of each taking the travel environment as given, then each user will choose a schedule that equates the marginal temporal variation in scheduling cost to that in travel-time cost. That is, departing a little earlier must produce changes in scheduling and travel-time costs that balance each other.[4] For example, with $\alpha - \beta - \gamma$ preferences, a traveler arriving before t^* will choose a particular departure time (i.e. time entering the queue) such that the marginal scheduling cost β of traveling still earlier is balanced by an identical marginal travel-time cost saving. Applying this condition at each point determines the shape of the function plotting travel delay against departure time: namely, it must rise with slope $\beta/(\alpha-\beta)$ and then fall with slope $-\gamma/(\alpha+\gamma)$.

Working backward from this and a consistency condition that everyone is accommodated yields the function describing the arrival rate over time. Note that the nature of congestion (the supply side) first enters the calculation in this consistency condition. The equilibrium is unique, as proved by Daganzo (1985) with a more general distribution of desired arrival times.

Costs of congestion. Even more surprising, in unpriced equilibrium the aggregate costs due to congestion—namely travel-time (queuing) and scheduling costs—are each completely independent of value of time α, so long as α is positive. If the value of time rises, departures become less clustered as travelers try harder to avoid congestion; but arrival times, which are constrained by bottleneck capacity, are unaffected and so are aggregate scheduling and travel-time costs.

Furthermore, exactly half these aggregate costs are travel-time costs, the rest being scheduling costs—although this ratio is different for different individuals. So not only does the value of time have no effect on aggregate user cost of congestion, half of those costs are scheduling and thus not even measured directly by observing travel time. This is a drastic

[4] Here I adopt the terminology that is most common in the economics literature on this model, in which "departure" means departure from home, and "arrival" means arrival at work. Given the usual simplification of ignoring travel time to or from the bottleneck, this "departure time" is thus the time of arrival at the back of the queue, and "arrival time" is the time of departure from the bottleneck. Therefore, authors occasionally interchange the meanings of "departure" and "arrival" relative to that here and in all of ADL's papers. One solution, adopted by Small and Verhoef (2007), is to call them "queue entry" and "queue exit," respectively.

revision of intuition of and normal rhetoric regarding congestion, both of which focus on the cost of time wasted while driving slowly. Of course, the exact 50-50 split applies only in the simplest version of the model, but the fundamental point remains: observing travel time captures only one of two major sources of congestion cost.

Effects of pricing. Another insight is that optimal time-varying pricing completely eliminates travel-time costs, while having no effect on scheduling costs. It accomplishes this by using the toll to mimic the pattern of travel-time costs that would occur in unpriced equilibrium. The toll thereby maintains the equilibrium arrival pattern (which cannot be improved upon due to limited bottleneck capacity) with a price incentive instead of a travel-time incentive. This was the main point stressed by Vickrey (1969). Given the result stated in the previous paragraph, it implies that pricing reduces aggregate travel costs (travel time plus scheduling) by exactly half.

Note that with a pricing regime in place, a casual observer might falsely think that because there is no longer any congestion, pricing is no longer appropriate. This is yet another paradoxical implication of the bottleneck model, at variance with most static models.[5]

Static model as reduced form. Finally, consider the reduced form of the model in which an optimal time-varying congestion toll is applied; it has an arbitrary constant which can be chosen at any convenient value without affecting results, presuming the total number of travelers N is fixed. A natural choice sets the toll at zero outside the period of congestion, ensuring that it is continuous. The time-varying toll combines with time-varying scheduling cost to produce a constant generalized price (toll plus travel cost). This generalized price turns out to equal the marginal aggregate cost of adding another traveler to the system, given that the toll pattern is adjusted accordingly and that everyone then readjusts. This marginal aggregate cost is rising in total quantity demanded N: specifically, it is proportional to N. Thus, looked at in aggregate, the problem looks exactly like a static Walters model in which congestion technology has travel time rising proportionally to travel volume. Furthermore, the average level of the time-varying price in the bottleneck model is exactly the marginal-cost price called for in the static model. This

[5] There is one other such case, namely a static model based on deterministic bottleneck congestion. Then optimal pricing rations traffic flow to exactly the bottleneck capacity, so that no congestion occurs and yet marginal cost exceeds average cost because equilibrium occurs at a point where the marginal cost curve is vertical.

insight, due to ADL (1993a), links the dynamic and static models while generalizing the dynamic model to price-sensitive total demand.[6]

This finding implies that the self-financing theorem of Mohring and Harwitz (1962), derived from a static model, also applies here. Optimal pricing generates just enough revenue to cover long-run investment cost provided there are no economies or diseconomies of scale in providing capacity.[7]

Remarkably, the results of the two previous paragraphs apply even to a flat toll or an otherwise constrained time-varying toll, so long as the toll *level* can be chosen optimally. The reason is that the generalized price is constant in time, no matter what the pricing regime; so the price level can be set equal to the (constant) social marginal cost of adding a traveler to the system.

It is worth noting that with the recognition of realistic degrees of randomness in conditions or in decision-making criteria, neither the static nor dynamic equilibria described here will ever be realized precisely. As noted by de Palma et al. (1997), they are rather benchmarks which are useful so long as conditions change only slowly.

3. Forms of inefficiency

One of the primary paths of model development has been to explore situations that produce additional margins of behavior for which private decisions may be inefficient. I discuss several of these situations in this section; see also ADL (1998) and de Palma and Fosgerau (2011).

[6] The generalization to elastic total demand was also done earlier by Ben-Akiva et al. (1986).

[7] More generally, with optimal pricing the ratio of revenue to capacity cost is equal to the elasticity of capital cost with respect to capacity. This result relies on the fact that bottleneck queuing meets a condition that may be viewed as constant returns to scale in producing congestion: namely, that equi-proportional changes in number of travelers and capacity do not affect travel time. See Small and Verhoef (2007), Sects. 5.1.1-5.1.2.

3.1. Heterogeneity in value of time and scheduling preferences

First, suppose there are two types of travelers, differing not in their desired arrival times (as discussed earlier) but in their scheduling cost and travel-time cost parameters (ADL 1988, 1994; Small and Verhoef 2007, pp. 133-134). Specifically, suppose they differ in the value of time α and in the parameters β and γ that depict how utility varies with arrival time in the cases of early and late arrival, respectively. The discussion in the previous section implies that the equilibrium departure pattern will now have two different slopes where it rises, and two different slopes where it falls; the two groups of travelers will allocate themselves accordingly. Will the shape of this pattern be second-best optimal? (That is, will it maximize welfare given that queuing is not eliminated via pricing?)

Generally not. Travelers will sort themselves by their ratios β/α and γ/α, those with the lower ratios traveling toward the outside of the rush hour. But an omniscient planner would instead sort them by the absolute values of β and γ. If those two orderings differ, the departure order of the two groups in unpriced equilibrium is inefficient. As a result, optimal pricing not only eliminates all travel-time costs, it also reduces aggregate scheduling costs by inducing those with lower costs of schedule delay to choose arrival times farther from the (common) desired arrival time.

Welfare results are derived by ADL (1994) for several specific forms of parameter heterogeneity—e.g. in ratio γ/β or in desired arrival time t^*—with an emphasis on the distribution of benefits from a toll or a capacity expansion. Heterogeneity in preference parameters is extended to continuous distributions by Newell (1987), by van den Berg and Verhoef (2011), and in the METROPOLIS model described in a subsequent section.

Remarkably, user heterogeneity by itself does not destroy the self-financing result for optimal tolls.[8] However, self-financing no longer applies when heterogeneity is combined with a constraint on the time-varying toll, such as that it be a flat toll (Arnott and Kraus 1995, 1998). The reason is that a flat toll, combined with heterogeneous users, allows some queuing. The external cost of such queuing may vary by user type, and this destroys the self-financing result.

[8] This was earlier shown by Mohring (1970) in a static model, for the case of separate tolls for two time periods.

3.2. Networks

Instead of a single bottleneck, one might have two or more bottlenecks in series or parallel. For example, if travelers differ in which bottleneck(s) they pass on their journey to work, the equilibrium can be such that an upstream bottleneck is actually helpful in alleviating the consequences of non-optimal departure times of those using a downstream bottleneck. Two different configurations with this property are studied by Kuwahara (1990) and ADL (1993b). In both, aggregate travel costs can sometimes be reduced by purposefully restricting the upstream flow or the downstream queue priority. This situation is an example of a Braess paradox, in which building or expanding a link in the network can raise aggregate congestion costs. It also illustrates a potential value of entry-ramp metering—completely aside from the usual motivation of stabilizing flow through the bottleneck.

The situation is even more interesting if there is also heterogeneity in user-cost parameters. ADL (1992) consider two user groups and two bottleneck in parallel. They show that equilibrium may be any of three types: the two user groups can travel together (sharing both roads), they can be partially separated (one group using one road, the other dividing among the two roads), or they can be fully separated. Similarly, the optimum may be any of these same three types; but the configuration of an equilibrium need not match that of the corresponding optimum. Thus, yet another welfare gain from pricing or some other intervention may be to better allocate diverse users across the two parallel roads.

Parallel bottlenecks provide an opportunity to study tolled express lanes, the primary form of congestion pricing in the United States. Braid (1996) obtains the second-best optimal time-varying toll on one bottleneck when a parallel bottleneck is unpriced. Bernstein and Muller (1993) focus on the policy tradeoff between extracting revenue from the tolled bottleneck (e.g. to facilitate privatization) versus operating it efficiently—the latter usually calls for a toll much smaller than the revenue-maximizing one. Van den Berg and Verhoef (2011) consider continuously parameterized heterogeneity and focus on the distribution of benefits from such second-best pricing. They find that users are no longer indifferent between the equilibrium and priced situations (when revenues are not returned to them)—in fact, a majority benefit from

pricing in numerical simulations. Generally, these authors find that second-best pricing produces a pattern of benefits quite different from that obtained in static models.

Yang and Meng (1998) analyze a more general network of bottlenecks, with travelers distributed among various origins and destinations. Those in each origin-destination pair have $\alpha - \beta - \gamma$ scheduling preferences with parameters unique to that destination. The authors find the system optimum for such a situation, along with required tolls, by applying network tools to a "space-time expanded network." It seems that this approach could identify many practical situations where optimal pricing takes interesting and unexpected forms.

This literature generally assumes a "vertical queue," i.e. it ignores any consequences of the spatial extent of a queue—in particular, it does not allow for "spillback" in which a queue at one intersection interferes with traffic through another. An exception is the METROPOLIS model reviewed in Section 4. Another approach to dealing with spillbacks, which can be especially virulent forms of congestion, is to model them aggregately, a topic considered in Section 3.6.

3.3. Residential location

In the basic bottleneck model, an optimal time-varying price produces no change in utility or behavior, prior to redistribution of revenues. Therefore, any changes in residential location or other land-use incentives would come only from such redistribution. Arnott (1998) formalizes this result in the context of two locations, a suburb and a downtown area.

But when the spatial extent of trips is more varied, more can be said. Fosgerau and de Palma (2012) extend the bottleneck model to incorporate a non-zero free-flow time to the bottleneck, the free-flow time being heterogeneous and interpreted as an indicator of residential distance from a central bottleneck. Importantly, they also use a more general model of preferences, in which utility is a general concave function of departure time and arrival time. They show that under reasonable conditions, travelers sort themselves such that those living farthest away arrive latest. The optimal price removes the queue, and it preserves this sorting. However, it does not preserve the actual clock times of arrivals. Nor does it leave all travelers indifferent: on the contrary, those living farthest away benefit and those living closer are hurt, prior to redistribution of revenues. This occurs because the arrival period shifts earlier, which

disadvantages those who previously arrived early in order to beat the queue (they live close by), but helps those who previously arrived at the end of the rush hour (they live far away). This result could affect the politics of introducing time-varying pricing at a point, or on a cordon, that is crossed by travelers from varied distances.

What about the other direction of causation: can scheduling behavior influence land-use decisions? Gubins and Verhoef (2014) consider this question within a version of the model in which scheduling preferences arise because spending time at different locations creates different rates of utility acquisition.[9] Specifically, they use a particular form of utility that produces $\alpha-\beta-\gamma$ preferences: namely, spending time at home produces utility at a constant rate α, but spending time at work produces utility at rates that jump from (α–β) to (α+γ); the jump occurs at the desired arrival time. They then generalize these preferences so that the utility of spending time at home depends positively on the size of one's residential lot, which is determined endogenously using the monocentric city structure that is standard in urban modeling. They then embed these preferences (summarized by X, the total subutility achieved from time spent at home and work) within an overall Cobb-Douglas utility function that also depends on numéraire consumption Q: specifically, $U=X^bQ^{(1-b)}$. These modifications alter the usual first-order conditions relating travel time T or toll τ (in an equilibrium or optimum, respectively) to arrival time t_a. In the standard model, those conditions are, in the case of early arrivals:

$$\frac{\partial T}{\partial t_a}=\frac{\beta}{\alpha}; \qquad \frac{\partial \tau}{\partial t_a}=\beta\cdot$$

But in the extended model, the parameters on the right-hand sides of these expressions depend on lot size L. Furthermore, the derivative in the second expression acquires an additional factor accounting for the difference between subutility X and overall utility U. The result is:

$$\frac{\partial T}{\partial t_a}=\frac{\beta(L)}{\alpha(L)}; \qquad \frac{\partial \tau}{\partial t_a}=\beta(L)\cdot\frac{bQ}{(1-b)X}$$

[9] This insightful interpretation and generalization of preferences is postulated by Vickrey (1973) and developed thoroughly by Tseng and Verhoef (2008). Among other implications, this formulation causes value of time to vary by time of day in all but very special cases such as when it reduces to $\alpha-\beta-\gamma$ preferences.

where quantities $\beta(L)$ and $\alpha(L)$ express the implications of lot-size dependence for the disutilities of early arrival and of time spent traveling, respectively. Since X and Q are solutions to the consumer's choice problem, and thus also vary, this formulation introduces two changes to the equilibrium conditions; both cause the marginal effects of arrival time to no longer be constant, implying that travel time and optimal toll are no longer piecewise-linear functions of clock time.

Another result from the Gubins-Verhoef model is that introducing congestion pricing causes lot size L to expand, because pricing permits more time to be spent at home and this enhances the incentive for a larger lot. This means that introducing pricing will cause the city to expand, becoming less dense—the exact opposite of the result when the usual static model of congestion is applied to a monocentric city. This also occurs using a more general formulation of utility (Fosgerau and Kim 2014).

One important feature of the static model is suppressed in the usual bottleneck formulation of dynamic congestion: namely, some congestion may also occur on the road connecting further-out locations to the bottleneck. In that case, an optimal toll pattern would penalize travel from such locations in order to reduce outlying congestion, presumably causing the city to become more compact just as in the static monocentric model. On the other hand, work locations are partially decentralized in reality, and pricing may make them spread out more, taking residences with them. Thus it remains to be seen what will be the balance of centralizing and decentralizing tendencies from pricing once these and other realistic features of urban structure are considered.

3.4 Parking

Parking has been studied by introducing another distance into the model. ADL (1991a) use a setup in which travel through a bottleneck is followed by a choice of where to park, whether close to or far from the (common) destination. They show that the unpriced equilibrium is inefficient in the order in which parking spaces are occupied. In equilibrium, under reasonable conditions, the first travelers to pass through the bottleneck choose the spaces closest to the CBD. This causes later arrivals to have to allow extra time for a longer walk from parking space to CBD. This in turn shifts the entire travel pattern earlier than it needs to be. An optimal

allocation instead has the earliest travelers park farthest from the CBD, so that parking spaces fill from the outside in.[10]

This optimum cannot be achieved with time-varying pricing of the bottleneck alone. Instead, it also requires a parking-price schedule that makes close-in spaces more expensive. This schedule is quite simple: it is piecewise linear with a steep slope for early arrivals, a less steep one for late arrivals.

By contrast, a parking-price schedule alone, even when it does not vary over time, can induce substantial queue reduction while also inducing the correct parking order. In fact, it can eliminate the queue entirely, although doing so is not second-best optimal because the steep parking-price schedule would induce people to arrive too early in order to get cheaper parking spaces. The second-best optimal parking-price schedule (conditional on no road pricing) is intractable to derive, but an intuitive schedule is presented by ADL (1991a) which, in numerical simulations, performs comparably to and often better than a policy of road pricing alone. It vastly outperforms parking fees set competitively by private operators. Furthermore, as the authors note, if only one of the two kinds of pricing is to be adopted, a spatially varying parking price has considerable administrative advantages, is less likely to be regressive in impact, and is a more familiar policy relative to a congestion toll.

This model is remarkable in that the analytic framework of the bottleneck model carries over quite intuitively, albeit with some additional complexity, to the simultaneous analysis of congestion and parking. Other work on parking suggests that this will not be the case when other aspects of parking, such as cruising to find an open parking space, are taken into account.

One may also ask whether a time-varying parking fee, unrelated to distance, might substitute for an optimal toll. Fosgerau and de Palma (2012) show that such a fee, when constrained to be non-decreasing in arrival time, can reduce queuing and shorten its duration but cannot eliminate queuing entirely. In the case of $\alpha - \beta - \gamma$ preferences, the fee achieves a welfare gain that is a fraction $\beta/(\beta+\gamma)$ of the maximum welfare gain from an optimal toll. Queuing can be completely eliminated, however, if an evening commute is considered, with parking fee also depending on evening departure time and with sufficient dependence (e.g. via fixed work-day

[10] Recall that schedule delay costs for early arrivers are linear in arrival time. If they were a convex function instead, this conclusion would be modified.

duration) between the morning and evening commutes. This is one of the very few results in the literature to consider the morning and evening commutes simultaneously, and suggests that such consideration may turn up interesting new possibilities for welfare-enhancing policies.

There are many such other aspects of parking: duration, search, information, reserved permits, and imperfect competition among private providers, to name a few. Search for parking, known as cruising, interacts especially strongly with congestion (Arnott and Inci 2006) and is often credited with responsibility for a large portion of downtown travel times. A thorough review of the economics of parking is carried out by Inci (2014). In general, results in parking models seem sensitive to the specific assumptions made about these features as well as about congestion technology. While the bottleneck model is one useful way to view parking, it is far from the only one and it is not yet clear whether any one approach will prove more generally useful than others.

3.5. Reliability and information

When capacity or other factors affecting speed are uncertain, travelers are subjected to unreliability: that is, they cannot predict precisely how long their trip will take. In a model with endogenous scheduling, this greatly complicates the decision process because travelers must now consider not just the most likely travel cost but an expectation of travel cost over varying conditions. Indeed, the most prominent theoretical model of how consumers value reliability applies the same $\alpha - \beta - \gamma$ preferences as those used in the bottleneck model.[11]

Since reliability involves lack of information about how a stochastic process is realized, its analysis naturally invites considering the effects of information provision. Travelers may obtain partial or full information about travel conditions before or during a trip, with different effects on their choices. This creates a rich environment for considering all kinds of information technologies. It also invites consideration of the adjustment process, as travelers respond to new conditions and/or information through trial, error, and learning; and of how the fraction of

[11] Noland and Small (1995). Fosgerau and Karlström (2010) show how to calculate results for a quite general distributions of the uncertain travel time. Fosgerau and Engelson (2011) extend the theory to scheduling preferences arising from different rates of utility acquisition at different locations. For reviews of reliability, see Li et al. (2010) and Small (2012).

travelers who receive information affects the adjustment process and the ultimate equilibrium. For example, day-to-day adjustment has been studied both theoretically and in driver simulation laboratories (de Palma et al. 1997; Mahmassani and Jayakrishnan 1991). Needless to say, studies of these phenomena have major implications for the economics of advanced traveler information systems.

Given $\alpha - \beta - \gamma$ preferences as a tool to derive aversion to unreliability, adding bottleneck queuing is one natural way to complete a model of how reliability develops in equilibrium. Doing so can be quite enlightening. Fosgerau (2010) uses the bottleneck to explain rigorously what has been observed empirically by Small et al. (2005): namely, that unreliability does not track congestion exactly over the course of a rush hour, but rather peaks later and lasts well beyond the time when congestion has dissipated. This turns out to be important for empirically disentangling the disutilities of congestion and of unreliability.

ADL (1991b) are able to obtain strong and apparently robust results regarding information provision.[12] They use a setup where drivers can choose between two parallel bottlenecks on their way to work, each with stochastic capacity. Expected travel costs are found to rise with the variance of capacities, as intuition suggests. However, costs vary in complex ways with the amount of information provided. Full information is beneficial, but partial information may be harmful by causing drivers to switch to a time or route with a smaller expected cost but a greater social marginal cost. For example, as shown explicitly by Mahmassani and Jayakrishnan (1991), drivers may over-react by switching in large numbers to a less congested route, causing it to become more congested than the route they are switching from.[13]

3.6. Hypercongestion

It is possible for types of traffic equilibria to occur in which flow rates vary inversely with incoming traffic volume, a situation known as “hypercongestion.” The inefficiency of

[12] Similar results are obtained ADL (1999), who also derive implications of both reliability and information for optimal capacity provision.

[13] In addition drivers may become oversaturated with information and make bad decisions, even for themselves, as a result: see Ben-Akiva et al. (1991).

congestion is then especially great because the capacity of the system is effectively reduced just when it is most needed. Small and Chu (2003) argue that hypercongestion is inherently dynamic because it is unstable, existing only as a temporary situation due to a surge in demand. Thus, it is natural to analyze it using a dynamic model.

A few papers have used the bottleneck model for this purpose, by postulating that bottleneck capacity varies inversely with the length of the queue.[14] Indeed, such a model is essentially equivalent to the kind of "macroscopic fundamental diagram" analyzed by Carlos Daganzo and several coauthors, in which a functional relationship between speed and density applies to areawide averages within a downtown area.[15] Details depend on specific assumptions made about queuing; but all analyses agree that an optimal price would eliminate the queue and thus maintain capacity at its highest possible level, thereby eliminating multiple sources of inefficiency.

The treatment of a variable-capacity bottleneck by Fosgerau and Small (2013) is closest to the original bottleneck model, in that it assumes a first-in-first-out (FIFO) queuing discipline. While FIFO is usually a realistic assumption, it is not necessarily so when "capacity" refers to a macroscopic property of an area containing a dense street network. However, it is made more realistic in this model because the length of the queue affects bottleneck capacity, and therefore later entrants to a queue can affect the delay experienced by earlier entrants—just as in a real street network. In this respect, the FIFO assumption is perhaps less restrictive than assumptions made by other treatments of hypercongestion, which typically rely on flow congestion and so, as noted earlier, require that a given vehicle's speed depends solely on the density of vehicles at a single point in time.

In equilibrium, the cumulative departure and arrival patterns generated by the Fosgerau-Small model of variable bottleneck capacity resemble those from applications of the standard bottleneck model with heterogeneity. The welfare gains, however, can be dramatically greater because pricing can eliminate the periods of limited capacity. Furthermore, the model naturally accommodates an analysis of flow metering as an alternative policy. Effectively, metering

[14] Yang and Huang (1997), Geroliminis and Levinson (2009), Fosgerau and Small (2013).

[15] See especially Daganzo (2007) and Geroliminis and Daganzo (2008). Small and Chu (2003) also define a macroscopic fundamental diagram, although without using that terminology, and Ardekani and Herman (1987) estimate one empirically.

consists of moving part of the queue outside the area where it affects capacity. This draws attention to the tradeoff, in determining metering policies, between the advantages of maintaining full bottleneck capacity and the costs of maintaining a separated queue, which may require some very expensive storage space.

3.7. Other topics

Insights from the bottleneck model extend to many other areas of inquiry. I will just mention three here.

Airport runways. Airplanes are subject to queuing for takeoff and landing slots, due to limited runway capacity. A prominent line of analysis postulates bottleneck congestion in order to analyze such situations (Daniel 1995). One recent paper even addresses the effects of various policies using a model very similar to the "bottleneck model" of roadway congestion (Silva et al. 2014).

Work hours. Travel patterns during morning and afternoon commuting hours can be interrelated through factors that influence the duration of a work day (Zhang et al. 2005). This implies a two-way causality between labor-market choices and scheduling choices.

Agglomeration. The advantages of having many people engaged simultaneously in activities is well recognized as an important part of urbanization and innovation. Fosgerau and Small (2014) make a case that this applies to both leisure and productive activities, and show that such agglomeration can endogenously produce the Vickrey-like scheduling preferences that are usually taken as exogenous. Furthermore, making preferences endogenous in this way results in the bottleneck model yielding quite different predictions.

All these topics can be analyzed with other models, but the bottleneck model has proven insightful for generating valuable and often unexpected insights.

4. Applications to real networks: The METROPOLIS model

André de Palma and several colleagues have developed a model, known as METROPOLIS, whose applications to date have used the usual components of the bottleneck model, but applied separately to the various links in the network and to different origin-

destination pairs. Thus there is heterogeneity in cost parameters, including desired arrival time. METROPOLIS also includes additive stochastic heterogeneity in the form of a continuous logit choice among departure times. De Palma et al. (1997) and De Palma and Marchal (2002) provide readable descriptions, including details on how it can be calibrated in order to analyze a real city.

The model is solved using microsimulation on the demand side, but like the bottleneck model itself is "mesoscopic" in its description of congestion: specifically, an individual vehicle's speed depends on total number of vehicles ahead of it on the same link. The model does not require that this dependence be that of a point bottleneck, although the model's developers have found that to be the form of congestion that works best. METROPOLIS also does not impose an analytical equilibrium as its solution; rather, it follows an iterative process based on a heuristic model of day-to-day behavioral adjustments, the latter informed by laboratory experiments.

METROPOLIS allows for a number of add-on functions such as response to uncertainty in and information about traffic. It has been calibrated for a few different cities or regions, including Geneva and Paris (De Palma et al. 1997, Saifuzzaman et al. 2012).

The model has also been applied to small artificial networks to study particular questions. Two studies by de Palma et al. (2005a, 2005b) illustrate this nicely. Both papers use the network structure to analyze toll cordons (where a toll is imposed when crossing a ring surrounding a CBD) and area charging (where a toll is imposed for any travel within the ring). They consider optimal tolls and two types of second-best tolls: a flat toll (just one level) and a step toll (two or more levels at different times). The first of these papers also considers a "third-best" policy, closely resembling policies actually used extensively in the United States, in which the toll is adjusted in order to just prevent queuing. The third-best policy is found to be very inefficient when implemented with flat tolls, in line with results by Small and Yan (1991) and Verhoef and Small (1994) using a static model; but it often performs quite well if implemented with dynamic tolls.

METROPOLIS has been implemented both with a "vertical" queue for each link, meaning the model does not account for the physical length of a queue, and with a "horizontal" queue in which case the queue can affect other links. This opens the possibility of using it to study hypercongestion.

5. What's next?

The model reviewed here is a mature one, elaborated in many ways and widely known among specialists in transportation economics. Along with the static model of Walters (1961), it has become a standard tool for understanding congestion and predicting the results of policies or other factors related to it. What is the future of this line of analysis? What should it be?

The situation seems ripe for something with greater realism to take center stage in research: perhaps some alternative approach to hypercongestion. But the current bottleneck model is too well embedded and too important to intuition to simply fade away. Rather, we are more likely to see a process comparable to the displacement but not the replacement of the static model: the older model will remain an important base to which new results will be compared, and newer models will be more successful if they have the older one as a special case or as a reduced form. In this way, the rich set of insights the model has generated will continue to influence the thinking of transportation analysts, and will help them understand the underlying nature of new results.

At the same time, the current bottleneck model is likely to be developed so as to apply more accurately to real situations. As an example, Zhang et al. (2010) work out the results of a bottleneck whose capacity varies exogenously over time, in discrete steps. (Their paper does not address hypercongestion because capacity does not depend on flow rate or queue length.) This developmental process is analogous to the one by which various practical constraints have been added to the basic static model in order to analyze real-world pricing policies, generally within a "second-best" framework (Small and Verhoef 2007, Sect. 4.2). Thus, such development with the bottleneck model can be expected to further enrich the ability of transportation professionals to describe the results of real-world pricing proposals, especially their dynamic aspects.

By incorporating more realism, the bottleneck model will also continue to provide new theoretical insights, although these will be somewhat more specialized than up to now. It will gradually become more suitable for empirical work in economics and for planning applications. This will further increase its use in engineering, especially in practical design applications, and in analytical urban planning. Thus it is very likely that insights from the bottleneck model will influence urban design and transportation policy, as well as research in economics and transportation, for many years to come.

References

Ardekani, S., Herman, R., 1987. Urban network-wide traffic variables and their relations. Transportation Science 21, 1-16.

Arnott, R., 1998. Congestion Tolling and the Urban Spatial Structure. Journal of Regional Science 38, 495-504.

Arnott, R., de Palma, A., Lindsey, R., 1988. Schedule delay and departure time decisions with heterogeneous commuters. Transportation Research Record 1197, 56-67.

Arnott, R., de Palma, A., Lindsey, R., 1990. Economics of a bottleneck. Journal of Urban Economics 27, 111-130.

Arnott, R., de Palma, A., Lindsey, R., 1991a. A temporal and spatial equilibrium analysis of commuter parking. Journal of Public Economics 45, 301-335.

Arnott, R., de Palma, A., Lindsey, R., 1991b. Does providing information to drivers reduce traffic congestion? Transportation Research Part A, 25, 309-318.

Arnott, R., de Palma, A., Lindsey, R., 1992. Route choice with heterogeneous drivers and group-specific congestion costs. Regional Science and Urban Economics 22, 71-102.

Arnott, R., de Palma, A., Lindsey, R., 1993a. A structural model of peak-period congestion: A traffic bottleneck with elastic demand. American Economic Review 83, 161-179.

Arnott, R., de Palma, A., Lindsey, R., 1993b. Properties of dynamic traffic equilibrium involving bottlenecks, including a paradox and metering. Transportation Science, 27, 148-160.

Arnott, R., de Palma, A., Lindsey, R., 1994. The welfare effects of congestion tolls with heterogeneous commuters. Journal of Transport Economics and Policy 28, 139-161.

Arnott, R., de Palma, A., Lindsey, R., 1998. Recent developments in the bottleneck model, in Button, K.J., Verhoef, E.T. (Eds.), Road Pricing, Traffic Congestion and the Environment: Issues of Efficiency and Social Feasibility. Edward Elgar, Cheltenham, UK, pp. 79-110.

Arnott, R., de Palma, A., Lindsey, R., 1999. Information and time-of-usage decisions in the bottleneck model with stochastic capacity and demand. European Economic Review 43, 525–548.

Arnott, R. Inci, E., 2006. An integrated model of downtown parking and traffic congestion. Journal of Urban Economics 60, 418-442.

Arnott, R. Kraus, M., 1995. Financing capacity in the bottleneck model. Journal of Urban Economics 38, 272-290.

Arnott, R. Kraus, M., 1998. When are anonymous congestion charges consistent with marginal cost pricing? Journal of Public Economics 67, 45-64.

Ben-Akiva, M., Cyna, M., de Palma, A., 1984. Dynamic model of peak period congestion. Transportation Research Part B 18, 339-355.

Ben-Akiva, M., de Palma, A., Kanaroglou, P., 1986. Dynamic model of peak period traffic congestion with elastic arrival rates. Transportation Science 20, 164-181.

Ben-Akiva, M., de Palma, A., Kaysi, I., 1991. Dynamic network models and driver information systems. Transportation Research Part A 25, 251-266.

Bernstein, D., Muller, J., 1993. Understanding the competing short-run objectives of peak period road pricing. *Transportation Research Record* 1395, 122-128.

Braid, R.M., 1989. Uniform versus peak-load pricing of a bottleneck with elastic demand. *Journal of Urban Economics* 26, 320-327.

Braid, R.M., 1996. Peak-load pricing of a transportation route with an unpriced substitute. Journal of Urban Economics 40, 179-197.

Chu, X., 1995. Endogenous trip scheduling: The Henderson approach reformulated and compared with the Vickrey approach. Journal of Urban Economics 37, 324-343.

Daganzo, C.F., 1985. The uniqueness of a time-dependent equilibrium distribution of arrivals at a single bottleneck. Transportation Science 19, 29-37.

Daganzo, C.F., 2007. Urban gridlock: macroscopic modeling and mitigation approaches. Transportation Research Part B 41, 49–62.

Daniel, J.I., 1995. Congestion Pricing and Capacity of Large Hub Airports: A Bottleneck Model with Stochastic Queues. Econometrica 63, 327-370.

de Palma, A., Fosgerau, M., 2011. Dynamic traffic modeling, in: de Palma, A., Lindsey, R., Quinet, E., Vickerman, R. (Eds.), Handbook in Transport Economics. Edward Elgar, Cheltenham, UK, pp. 188–212.

de Palma, A., Kilani, M., Lindsey, R., 2005a. Comparison of second-best and third-best tolling schemes on a road network. Transportation Research Record: Journal of the Transportation Research Board 1932, 89-96.

de Palma, A., Kilani, M., Lindsey, R., 2005b. Congestion pricing on a road network: A study using the dynamic equilibrium simulator METROPOLIS. Transportation Research Part A 39, 588-611.

de Palma, A., Marchal, F., 2002. Real cases applications of the fully dynamic METROPOLIS tool-box: An advocacy for large-scale mesoscopic transportation systems. Networks and Spatial Economics 2, 347-369.

de Palma, A., Marchal, F., Nesterov, Y., 1997. METROPOLIS: Modular system for dynamic traffic simulation. Transportation research Record 1607, 178-184.

Fargier, P.-H., 1983. Effects of the choice of departure time on road traffic congestion, in: Hurdle, V.F., Hauer, E., Steuart, G.N. (Eds.), Proceedings of the Eighth International Symposium on Transportation and Traffic Theory. University of Toronto Press, Toronto, pp. 223-262.

Fosgerau, M., 2010. On the relation between the mean and variance of delay in dynamic queues with random capacity and demand. Journal of Economic Dynamics and Control 34, 598-603.

Fosgerau, M., de Palma, A., 2012. Congestion in a city with a central bottleneck. Journal of Urban Economics 71, 269-277.

Fosgerau, M., Engelson, L., 2011. The value of travel time variance. Transportation Research Part B 45, 1–8.

Fosgerau, M., Karlström, A., 2010. The value of reliability. Transportation Research Part B 44, 38-49.

Fosgerau, M., Kim, J., 2014. Vickrey Meets Alonso: Commute Scheduling and Congestion in a Monocentric City. Working paper, Danish Technical University. http://home.sogang.ac.kr/sites/econdept/seminars/Lists/b6/Attachments/178/Vickrey%20Meets%20Alonso%20Commute%20Scheduling%20and%20Congestion%20in%20a%20Monocentric%20City.pdf

Fosgerau, M., Small, K.A., 2013. Hypercongestion in Downtown Metropolis. Journal of Urban Economics 76, 122–134.

Fosgerau, M., Small, K.A., 2014. Endogenous Scheduling Preferences and Congestion. UC Irvine Economics Working Paper 13-14-03. http://www.socsci.uci.edu/~ksmall/EndogenSchedPrefs.pdf

Geroliminis, N., Daganzo, C.F., 2008. Existence of urban-scale macroscopic fundamental diagrams: some experimental findings. Transportation Research Part B 42, 759–770.

Geroliminis, N., Levinson, D.M., 2009. Cordon pricing consistent with the physics of overcrowding, in: Lam, W.H.K., Wong, S.C. Lo, H.K. (Eds.), Transportation and Traffic Theory, Springer.

Gubins, S., Verhoef, E.T., 2014. Dynamic bottleneck congestion and residential land use in the monocentric city. Journal of Urban Economics 80, 51-61.

Henderson, J.V., 1974. Road congestion: A reconsideration of pricing theory. Journal of Urban Economics 1, 346-365.

Hendrickson, C. Kocur, G., 1981. Schedule delay and departure time decisions in a deterministic model. Transportation Science 15, 62-77.

Inci, E., 2014. A review of the economics of parking, working paper, Sabanci University.

Kuwahara, M., 1990. Equilibrium queueing patterns at a two-tandem bottleneck during the morning peak. Transportation Science 24, 217-229.

Li, Z., Hensher, D.A., Rose, J.M., 2010. Willingness to pay for travel time reliability in passenger transport: A review and some new empirical evidence. Transportation Research Part E 46, 384–403.

Mahmassani, H.S., Jayakrishnan, R., 1991. System Performance and User Response Under Real-Time Information in a Congested Traffic corridor. Transportation Research Part A 25, 293-307.

Mohring, H., 1970. The peak load problem with increasing returns and pricing constraints. American Economic Review 60, 693-705.

Mohring, H. Harwitz, M., 1962. Highway benefits: An analytical framework. Northwestern University Press, Evanston, Illinois.

Newell, G.F., 1987. The morning commute for nonidentical travelers. Transportation Science 21, 74-88.

Noland, R.B., Small, K.A., 1995. Travel-time uncertainty, departure time choice, and the cost of morning commutes. Transportation Research Record 1493, 150-158.

Saifuzzaman, M., de Palma, A., Motamedi, K., 2012. Calibration of METROPOLIS for Ile-de-France. SustainCity Working Paper 7.2, CES, ENS-Cachan, France. http://www.sustaincity.org/publications/WP_7.2_Paris_Case_Study_METROPOLIS.pdf

Silva, H.E., Verhoef, E.T., van den Berg, V.A.C., 2014. Airlines' strategic interactions and airport pricing in a dynamic bottleneck model of congestion. Journal of Urban Economics 80, 13-27.

Small, K.A., 1982. The scheduling of consumer activities: Work trips. American Economic Review 72, 467-479.

Small, K.A., 2012. Valuation of Travel Time. Economics of Transportation 1, 2-14.

Small, K.A., Chu, X., 2003. Hypercongestion. Journal of Transport Economics and Policy 37, 319–352.

Small, K.A., Verhoef, E.T., 2007. The Economics of Urban Transportation. Routledge (Taylor & Francis), London.

Small, K.A., Yan, J., 2001. The Value of 'value pricing' of roads: Second-best pricing and product differentiation. Journal of Urban Economics 49, 310-336.

Small, K. A., Winston, C., Yan, J., 2005. Uncovering the distribution of motorists' preferences for travel time and reliability. Econometrica 73, 1367-1382.

Tseng, Y. Y. and Verhoef, E. T., 2008. Value of time by time of day: A stated-preference study Transportation Research Part B 42, 607–618.

van den Berg, V., Verhoef, E.T., 2011. Winning or losing from dynamic bottleneck congestion pricing? The distributional effects of road pricing with heterogeneity in values of time and schedule delay. Journal of Public Economics 95, 983-992.

Verhoef, E.T., Small, K.A., 2004. Product Differentiation on roads: Constrained congestion pricing with heterogeneous users. Journal of Transport Economics and Policy 38, 127-156.

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.

Walters, A.A., 1961. The theory and measurement of private and social cost of highway congestion. Econometrica 29, 676-699.

Yang, H., Huang, H. J., 1997. Analysis of the time-varying pricing of a bottleneck with elastic demand using optimal control theory. Transportation Research Part B 31, 425–440.

Yang, H., Meng, Q., 1998. Departure time, route choice and congestion toll in a queuing network with elastic demand. Transportation Research Part B 32, 247-260.

Zhang, X., Yang, H., Huang, H.-J., Zhang, H.M., 2005. Integrated scheduling of daily work activities and morning–evening commutes with bottleneck congestion. Transportation Research Part A 39, 41-60.

Zhang, X., Zhang, H.M., Li, L., 2010. Analysis of user equilibrium traffic patterns on bottlenecks with time-varying capacities and their applications. International Journal of Sustainable Transportation 4, 56-74.