Economics of Transportation 4 (2015) 110–117

Contents lists available at ScienceDirect

Economics of Transportation

journal homepage: www.elsevier.com/locate/ecotra

The bottleneck model: An assessment and interpretation

Kenneth A. Small *

Department of Economics, University of California, Irvine, CA 92697-5100, USA

ARTICLE INFO

Article history:
Received 18 July 2014
Received in revised form
11 October 2014
Accepted 4 January 2015

Keywords:
Congestion
Bottleneck
Scheduling
Congestion pricing
Parking
Reliability

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.

© 2015 Elsevier Ltd. All rights reserved.

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 the 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 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

* Tel.: +1 949 824 5658.
E-mail address: ksmall@uci.edu

[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).

http://dx.doi.org/10.1016/j.ecotra.2015.01.001
2212-0122/© 2015 Elsevier Ltd. All rights reserved.

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 $\beta/\alpha=0.5$ and $\gamma/\alpha=2$. 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.

2.1. 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 under this consistency condition. The equilibrium is unique, as proved by Daganzo (1985) with a more general distribution of desired arrival times.

2.2. 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 revision of intuition 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.

2.3. 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

[2] See Small and Verhoef (2007), Sect. 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.

[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.

Download English Version:

https://daneshyari.com/en/article/5062939

Download Persian Version:

https://daneshyari.com/article/5062939

Daneshyari.com